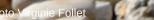
FRCCS 2023

French Regional Conference on Complex Systems

Le Havre, France 31 May - 02 June

Book of Abstracts



Contents

Foreword	10
Committee	11
Advisory Board CSS France	11
General Chairs	11
Program Chairs	11
Poster Chairs	11
Publication Chairs	11
Finance Chair	11
Registration Chair	11
Award Chairs	11
Student Grant Chairs	12
Web Chair	12
Sponsor Chair	12
Local Chair	12
Local Committee	12
Invited Speakers	13
Luca Maria Aiello	
ITU Copenhagen Denmark	14
Keynote: Coloring Social Relationships	14
Ginestra Bianconi	
Queen Mary University UK \ldots \ldots \ldots \ldots \ldots	15
Keynote: The dynamics of higher-order networks: the effect of	
topology and triadic interactions	15
Víctor M. Eguíluz	
University of the Balearic Islands Spain	16
Keynote: Complex systems perspective on the ocean ecosystem	16
Adriana Iamnitchi	. –
Maastricht University Netherlands	17
Keynote: Taming the Wild West of Social Media: The Digital	
Services Act and its Effects on Computational Social	
Science	17
Rosario N. Mantegna	10
Palermo University Italy	18
Keynote: Social anatomy of a financial bubble	18
Céline Rozenblat	20
University of Lausanne Switzerland	20
Keynote: Resilience and transitions from local Digital Twins to	00
global complex urban systems	20

Network Analysis Diagnosing network attacks by a machine-learning approach	22
Davide Coppes and Paolo Cermelli	23
to better understand the context of humanitarian operations Aurélie Charles, Guillaume Bouleux and Giacomo Kahn . Filtering Real World Networks: A Correlation Analysis of Statistical	28
Backbone Techniques Ali Yassin, Hocine Cherifi, Hamida Seba and Olivier Togni	32
Interaction Network and Graphlets to Identify Sport Teams' Signa- ture	02
Quentin Bourgeais, Guillaume Hacques, Rodolphe Char- rier, Eric Sanlaville and Ludovic Seifert	36
Measuring Movie Script Similarity using Characters, Keywords, Lo- cations, and Interactions	
Majda Lafhel, Mohammed El Hassouni, Benjamin Renoust and Hocine Cherifi	42
NetBone: A Python Package for Extracting Backbones of Weighted Networks	
Ali Yassin, Abbas Haidar, Hocine Cherifi, Hamida Seba and Olivier Togni	46
Social Complexity	50
Visualizing Mobile Phone Communication Data in Criminal Inves- tigations: the Case of Media Multiplexity	
Martina Reif, Bruno Pinaud, Thomas Souvignet, Guy Melançon and Quentin Rossy	51
Bounded confidence models generate one more cluster when the number of agents is slowly growing	
Yerali Gandica and Guillaume Deffuant	56
Antoine Houssard, Federico Pilati, Maria Tartari, Pierluigi	
Sacco and Riccardo Gallotti	60
Amina Azaiez and Robin Salot $\ldots \ldots \ldots \ldots \ldots \ldots$	64
Opinion dynamics model revealing yet undetected cognitive biases Guillaume Deffuant	89
Dynamics & Self-Organization	108
A toy model for approaching volcanic plumbing systems as complex systems	100
Remy Cazabet, Catherine Annen, Jean-Françcois Moyen and Roberto Weinberg	109
	100

Reconstruction of variables of interest in nonlinear complex systems:	
application to a C. elegans biological neural network	110
Nathalie Verdière, Sébastien Orange and Loïs Naudin	112
POSTER: Agent-based modelling to simulate realistic self-	
organizing development of the mammalian cerebral cortex	110
Umar Abubacar and Roman Bauer	116
How to Grasp the Complexity of Self-Organised Robot Swarms?	
Jérémy Rivière, Aymeric Henard, Etienne Peillard,	110
Sébastien Kubicki and Gilles Coppin	119
Analysis of a Network of Hodgkin-Huxley Excitatory and Inhibitory	
Neurons	
B. Ambrosio, M.A. Aziz-Alaoui, M. Maama and S.M.	101
$Mintchev \dots \dots$	131
Hopf Bifurcation in Oncolytic Therapeutic Model with Viral Lytic	
Cycle	100
Fatiha Najm, Radouane Yafia and M.A. Aziz Alaoui	136
Poster presentation: Preterm birth indicates higher neural rich club	
organisation than term counterparts	1.40
Katherine Birch, Dafnis Batalle and Roman Bauer	140
Diffusion & Epidemics	143
Supply, demand and spreading of news during COVID-19 and	
assessment of questionable sources production	
Pietro Gravino, Emanuele Brugnoli, Giulio Prevedello,	
Martina Galletti and Vittorio Loreto	144
A local Agent-Based Model of COVID-19 Spreading and Interven-	
tions	
Perrette Benjamin, Cruz Christophe and Cherifi Hocine	146
Towards a Generic Agent Based Vector-Host Model	
Cyrine Chenaoui, Nicolas Marilleau and Slimane Ben Miled	151
Exploring and optimising infectious disease policies with a stylised	
agent-based model	
Jeonghwa Kang and Juste Raimbault	179
Contact networks in daily-life pedestrian crowds and risks of viral	
transmission	
Alexandre Nicolas and Simon Mendez	197
Minimizing epidemic spread through mitigation of anti-vaccine opin-	
ion propagation	
Sarah Alahmadi, Markus Brede and Rebecca Hoyle	201
	205
Linguistics & Multilayer	205
Imbalanced Multi-label Classification for Businessrelated Text with	
Moderately Large Label Spaces	200
Muhammad Arslan and Christophe Cruz	206

SINr: a python package to train interpretable word and graph em-	
beddings Thile and Brandson, Nicelas Durmá, Simon Cavillat and Ar	
Thibault Prouteau, Nicolas Dugué, Simon Guillot and An- thony Perez	215
Topics evolution through multilayer networks	210
Andrea Russo, Antonio Picone and Vincenzo Miracula	219
Towards efficient multilayer network data management	215
Georgios Panayiotou, Matteo Magnani and Bruno Pinaud	223
Knowledge Graph for NLG in the context of conversational agents	220
Hussam Ghanem, Massinissa Atmani and Christophe Cruz	227
Urban	239
How to Reduce Streets-Network Sprawl?	
Supharoek Chattanachott, Frédéric Guinand and Kittichai	
Lavangnananda	240
A hybrid network: Sea-land connectivity in the global system of	
cities	044
Cesar Ducruet, Barbara Polo and Bruno Marnot	244
Agent-based modelling of urban expansion and land cover change: a prototype for the analysis of commuting patterns in Geneva,	
Switzerland.	
Flann Chambers, Christophe Cruz and Giovanna Di Marzo	
Serugendo	249
Radial analysis and scaling law of housing prices in French urban	- 10
areas using DVF data	
Gaëtan Laziou, Rémi Lemoy and Marion Le Texier	256
Towards a geographical theory different from that of the natural	
sciences: foundations for a relational complexity model	
Olivier Bonin	260
	0.05
Structure & Dynamics	265
On the impact of introducing random modifications to the neigh-	
borhood of the abelian sandpile Paulin Héleine, Juan Luis Jiménez Laredo, Frédéric	
Guinand and Damien Olivier	266
Asymptotic Dynamic Graph Order Evolution Analysis	200
Vincent Bridonneau, Frédéric Guinand and Yoann Pigné	274
Who to Watch When? Strategic Observation in the Inverse Ising	- 1 - 1
Problem	
Zhongqi Cai, Enrico Gerding and Markus Brede	285
The architecture of multifunctional ecological networks	
Mar Cuevas-Blanco, Sandra Hervías-Parejo, Victor	
Martínez Eguiluz, Lucas Lacasa, Isabel Donoso and Anna	
Traveset	289
Temporal Betweenness Centrality on Shortest Paths	
Mehdi Naima, Matthieu Latapy and Clémence Magnien .	293

Mobility	297
Delineation of city districts based on intraday commute patterns	
Yuri Bogomolov, Alexander Belyi, Ondrej Mikes and	
Stanislav Sobolevsky	298
Analysis of the German Commuter Network	
Christian Wolff, Markus Schaffert, Christophe Cruz and	
Hocine Cherifi	302
Academic Mobility as a Driver of Productivity: A Gender-centric	
Approach	
Mariana Macedo, Ana Maria Jaramillo and Ronaldo	
Menezes	306
Mobility networks as a predictor of socioeconomic status in urban	
systems	
Devashish Khulbe, Stanislav Sobolevsky, Alexander Belyi	
and Ondrej Mikes	309
Is Paris a good example for a X-minute city? Modeling city compo-	
sition on POI data and X-minute statistics in Paris	
Sarah J Berkemer and Paola Tubaro	313
Impact of pedestrian flocking tactics on urban networks	
Guillaume Moinard and Matthieu Latapy	318
From CONSumers to PROSumers: spatially explicit agent-based	
model on achieving Positive Energy Districts	
Erkinai Derkenbaeva, Gert Jan Hofstede, Eveline van	
Leeuwen and Solmaria Halleck Vega	322
Communities	205
Communities	325
Filtering the noise in consensual community detection	
Antoine Huchet, Jean-Loup Guillaume and Yacine Ghamri- Doudane	326
Doudane Deep Learning Attention Model For Supervised and Unsupervised	520
Network Community Detection	330
Stanislav Sobolevsky	330
• •	334
Fabrice Lecuyer	504
Framework	
Stephany Rajeh and Hocine Cherifi	338
Backbone Extraction of Weighted Modular Complex Networks based	000
on their Component Structure	
on their Component Structure	340
on their Component Structure Sanaa Hmaida, Hocine Cherifi and Mohammed El Hassouni	342
Sanaa Hmaida, Hocine Cherifi and Mohammed El Hassouni	342 346
Sanaa Hmaida, Hocine Cherifi and Mohammed El Hassouni	

A pattern of diffusion of artificial intelligence in science: the devel-	
opment of an AI scientific specialty in neuroscience	
Sylvain Fontaine, Floriana Gargiulo, Michel Dubois and	
Paola Tubaro	354
Junk science bubbles and the abnormal growth of giants	
Floriana Gargiulo, Tommaso Venturini and Antoine Hous-	
sard	357
Unpacking popularity: volume, longevity, connectivity and globality	
Mariana Macedo, Melanie Oyarzun and Cesar A. Hidalgo	360
Socio-Technical Systems	363
Integrated bi-objective model for berth scheduling and quay crane assignment with transshipment operations	
Marwa Samrout, Adnan Yassine and Abdelkader Sbihi	364
Network structures of a centralized and a decentralized market. A	004
direct comparison.	
Sylvain Mignot and Annick Vignes	369
Alignment of Multinational Firms along Global Value Chains: A	005
network-based perspective	
Charlie Joyez	373
Is Conservation Agriculture the Future of Farming in France?	515
Damien Calais	403
Let's Tweet about Soccer? A Gender-centric Question	400
Akrati Saxena and Mariana Macedo	409
From geographic data to spatial knowledge in agent-based modeling	405
applied to land use simulation	
Severin Vianey Kakeu Tuekam, Eric Fotsing and Marcellin	
Julius Antonio Nkenlifack	412
	112
Blockchain	428
Deep Reinforcement Learning for Selfish Nodes Detection in a	
Blockchain	
Md Muhidul Islam Khan	429
Blockchain for the maritime: A modular proposition	
Rim Abdallah, Cyrille Bertelle, Jerome Besancenot,	
Claude Duvallet and Frederic Gilletta	438
Automation and fluidity of logistics transactions through blockchain	
technologies	
Maxence Lambard, Cyrille Bertelle and Claude Duvallet	442
Privileging permissioned blockchain deployment for the maritime	
sector	
Rim Abdallah, Cyrille Bertelle, Jérôme Besancenot,	
Claude Duvallet and Frederic Gilletta	446
Secure access control to data in off-chain storage on blockchain-based	
consent systems by cryptography	
Mongetro Goint, Cyrille Bertelle and Claude Duvallet	450

Learning on Graphs	454
Are Networks Really Useful? — Interplay Between Structures and	
Vector Representations	
Tsuyoshi Murata	455
Quantile regression: an approach based on GEV distribution and	
machine learning	
Lucien M. Vidagbandji, Laurent Amanton, Alexandre	
Berred and Cyrille Bertelle	459
DyHANE: Dynamic Heterogeneous Attributed Network Embedding	
Liliana Martirano, Roberto Interdonato, Dino Ienco and	
Andrea Tagarelli	463
Index by Author	468

Foreword

Dear colleagues and participants,

It is our great pleasure to welcome you to the French Regional Conference on Complex Systems (FRCCS 2023). This year's conference promises to be an exciting event, bringing together distinguished researchers and practitioners from around the world to share their latest findings and insights on complex systems.

We are particularly pleased to introduce our six invited speakers, who are among the most renowned experts in the field. Luca Maria Aiello from ITU Copenhagen Denmark, Ginestra Bianconi from Queen Mary University UK, Víctor M. Eguíluz from University of the Balearic Islands Spain, Adriana Iamnitchi from Maastricht University Netherlands, Rosario N. Mantegna from Palermo University Italy, and Céline Rozenblat from University of Lausanne Switzerland, will each give a keynote address on their latest research and insights.

The year the conference is held in the beautiful city of Le Havre, France, a UNESCO World Heritage site known for its stunning architecture, rich history, and cultural significance. Le Havre is an ideal location for our conference, offering a unique blend of history and modernity, which perfectly reflects the theme of the conference.

We would like to extend our sincere thanks to the University of Le Havre for hosting this event, and to our sponsors, Springer Nature and Frontiers, for their support. Their contributions have helped make this conference possible.

Finally, We would like to express our appreciation to all of our participants and presenters. We are confident that this conference will provide a valuable opportunity to exchange ideas, learn from each other, and develop new collaborations and partnerships.

I hope that you will find this book of abstracts to be a useful resource and an excellent preview of the exciting research presented at the conference.

Sincerely,

Cyrille Bertelle and Roberto Interdonato General Chairs, FRCCS 2023

Committes

Advisory Board CSS France

- Cyrille Bertelle, Université Le Havre Normandie
- Chantal Cherifi, DISP, Lyon
- Hocine Cherifi, ICB, Dijon
- Hamamache Kheddouci, LIRIS, Lyon
- Benjamin Renoust, Median Technologies, Sophia Antipolis

General Chairs

- Cyrille Bertelle, Université Le Havre Normandie
- Roberto Interdonato, CIRAD, Montpellier

Program Chairs

- Moulay A. Aziz-Alaoui, Université Le Havre Normandie
- Hocine Cherifi, Université de Bourgogne

Poster Chairs

- Nicolas Dugué, Université Le Mans
- Bruno Pinaud, Université de Bordeaux

Publication Chairs

- Christophe Cruz, Université de Bourgogne
- Yoann Pigné, Université Le Havre Normandie

Finance Chair

• Benjamin Renoust, Median Technologies, Sophia Antipolis

Registration Chair

• Valentina Lanza, Université Le Havre Normandie

Award Chairs

- Konstantin Avrachenkov, INRIA Sophia Antipolis
- Avner Bar-Hen, CNAM Paris
- Annick Lesne, Université Pierre et Marie Curie
- Denise Pumain, Université Paris 1 Panthéon-Sorbonne

Student Grant Chairs

- Chantal Cherifi, Université Lyon 2
- Maxime Lenormand, INRAE UMR TETIS Montpellier
- Pascal Poncelet, LIRMM Université de Montpellier

Web Chair

• Nathalie Corson, Université Le Havre Normandie

Sponsor Chair

• Pierre Parrend, Université de Strasbourg

Local Chair

• Claude Duvallet, Université Le Havre Normandie

Local Committee

- Rim Abdallah, Université Le Havre Normandie
- Laurent Amanton, Université Le Havre Normandie
- Benjamin Ambrosio, Université Le Havre Normandie
- Alexandre Berred, Université Le Havre Normandie
- Rodolphe Charrier, Université Le Havre Normandie
- Christophe Duhamel, Université Le Havre Normandie
- Mongetro Goint, Université Le Havre Normandie
- Nathalie Verdière, Université Le Havre Normandie
- Lucien Vidagbandji, Université Le Havre Normandie

Invited Speakers



Luca Maria Aiello ITU Copenhagen Denmark 1	.4
Ginestra Bianconi Queen Mary University UK	.5
Víctor M. Eguíluz University of the Balearic Islands Spain 1	.6
Adriana Iamnitchi Maastricht University Netherlands	.7
Rosario N. Mantegna Palermo University Italy	.8
Céline Rozenblat University of Lausanne Switzerland 2	20

Luca Maria Aiello ITU Copenhagen Denmark



Luca Aiello holds a PhD in Computer Science from the university of Turin, Italy. He is currently an Associate Professor at the IT University of Copenhagen. Previously, he worked for 10 years as a Research Scientist in the industry: at Yahoo Labs in Barcelona, and at Bell Labs in Cambridge (UK). He conducts research in Computational Social Science, an interdisciplinary field of studies that uses Social Science theories to guide the solution to Data Science problems. He is currently working on text analysis techniques that, when applied to conversations, can help understand people's social behavior and psychological well-being. His work has been covered by hundreds of news articles published by news outlets worldwide including Wired, WSJ, and BBC.

Keynote: Coloring Social Relationships

Social relationships are the key determinant of crucial societal outcomes, including diffusion of innovation, productivity, happiness, and life expectancy. To better attain such outcomes at scale, it is therefore paramount to have technologies that can effectively capture the type of social relationships from digital data. NLP researchers have tried to do so from conversational text but mostly focusing on sentiment or topic mining, techniques that fall short on either conciseness or exhaustiveness. We propose a theoretical model of 10 dimensions (colors) of social relationships that is backed by decades of research in social sciences and that captures most of the common relationship types. We trained a deep-learning model to accurately classify text along these ten dimensions. By applying this tool on large-scale conversational data, we show that the combination of the predicted dimensions suggests both the types of relationships people entertain and the types of real-world communities they shape. We believe that the ability of capturing interpretable social dimensions from language using AI will help closing the gap between the oversimplified social constructs that existing social network analysis methods can measure and the multifaceted understanding of social dynamics that has been developed by decades of theoretical research.

Ginestra Bianconi Queen Mary University UK



I work in statistical mechanics and network theory and I enjoy combining statistical mechanics with graph theory, topology, and other mathematical subjects to study the network complexity.

The field has a rich interdisciplinary character since complex networks describe the interactions of a large variety of complex systems, from the Internet to the brain, the climate and social networks. My main focus in on the theory of networks, but am also very interested in applications ranging from biological networks social networks.

In am expert in network modelling and for this I use both equilibrium and nonequilibrium statistical mechanics approaches. Currently my research focuses on generalized network structures including multilayer and higher-order networks (simplicial complexes).

Keynote: The dynamics of higher-order networks: the effect of topology and triadic interactions

Higher-order networks capture the interactions among two or more nodes in complex systems ranging from the brain, to chemical reaction networks. Here we show that higher-order interactions are responsible for new dynamical processes that cannot be observed in pairwise networks.

We will cover how topology is key to define synchronization of topological signals, i.e. dynamical signals defined not only on nodes but also on links, triangles and higher-dimensional simplices in simplicial complexes. Interesting topological synchronization dictated by the Dirac operator can lead to the spontaneous emergence of a rhythmic phase where the synchronization order parameter displays low frequency oscillations which might shed light on possible topological mechanisms for the emergence of brain rhythms.

We will also reveal how triadic interactions can turn percolation into a fullyfledged dynamical process in which nodes can turn on and off intermittently in a periodic fashion or even chaotically leading to period doubling and a route to chaos of the percolation order parameter.

Víctor M. Eguíluz University of the Balearic Islands Spain



PhD in Physics, University of the Balearic Islands (1999). Postdoctoral researcher at DELTA (ENS-CNRS-EHESS) (France, 2000) and Niels Bohr Insitute (Denmark, 2001). Ramon y Cajal Fellow (2003-2007). Staff member of IFISC (UIB-CSIC) since 2007. Main research areas: complex systems, complex networks, social dynamics, biological systems.

Keynote: Complex systems perspective on the ocean ecosystem

The ocean plays a central role in the Earth system. It regulates geophysical processes and is thus crucial to understand the pace of global change. It is also the largest ecosystem and thus host of biodiversity. Understanding the interactions among physical, natural, and human processes, how they evolve in time and how they participate in the functioning of ecosystems is thus of crucial importance. The development of novel techniques under the umbrella of Big Data analytics and Data science offers new opportunities for complex systems to analyze the Ocean system. We will present opportunities and challenges for complex systems in this context and in particular recent advances in the analysis of the largest multi-taxa marine megafauna tracking dataset recently assembled.

Adriana Iamnitchi Maastricht University Netherlands



Adriana Iamnitchi is Professor, Chair of Computational Social Sciences at Maastricht University. Her research spans different aspects of data and computer science, with a particular focus on social media forensics, network science, and distributed systems. Until recently she has been professor of computer science in the United States, where her work was funded by the National Science Foundation, Office for Naval Research, and DARPA. She holds a PhD in Computer Science from The University of Chicago and is an ACM Distinguished Member, IEEE Senior Member, and recipient of the National Science Foundation Early CAREER award.

Keynote: Taming the Wild West of Social Media: The Digital Services Act and its Effects on Computational Social Science

The surge in social media use over the last decade brought a host of unintended and unanticipated complications, such as disinformation, polarization, and undisclosed content monetization. Some of these problems can be traced back to the lack of regulation governing the digital interactions, algorithmic decisions, monetization, and business strategies that shape the social media landscape. In an effort to address these issues, the European Union has recently approved the Digital Services Act (DSA), aiming to better regulate online spaces, including social media platforms. This talk will delve into challenges and opportunities that the implementation and enforcement of the DSA may bring for computational social scientists.

Rosario N. Mantegna Palermo University Italy



He is one of the leading pioneers in the field of econophysics. He started to work in the area of the analysis and modeling of social and economic systems with tools and concepts of statistical physics as early as in 1990. He published the first econophysics paper in a physics journal in 1991. He co-authored the first econophysics paper in Nature, in 1995. In 1999 he published the first book on econophysics. Just after Mantegna earned his tenured position in 1999, he founded the Observatory of Complex Systems (http://ocs.unipa.it), a research group of the Dipartimento di Fisica of Palermo University. Mantegna has participated in several international research projects contributing to the management and coordination of them. Examples are the COST P10 action "Physics of Risk" and the GIACS (General Integration of the Applications of Complexity in Science) coordination action of the "Jerusalem Declaration on Data Access, Use and Dissemination for Scientific Research".

Keynote: Social anatomy of a financial bubble

The study of financial bubbles is a highly controversial topics in economics and finance. Despite a large number of economic analyses, anecdotal evidence and increased theoretical attention, the quantitative monitoring and modeling of financial bubbles still miss standards broadly accepted by scholars. Our study focuses on the famous dotcom bubble that inflated financial markets during the period 1995-2000. Specifically, we investigate Nokia share ownership during the onset of the bubble and during its aftermath up to 2010 by investigating a unique database that tracks the financial ownership of all Finnish legal entities. We document a persistent flow of investment from foreign investors in the Nokia company during the inflation period of the bubble. This is a typical anecdotical scenario observed in the setting of financial bubbles. A second fundamental observation concerns the number of Finnish investors having an open investment position in Nokia at a given day. This number increased more than exponentially during the 1998-2000, reflecting a dramatic raise of attention at a country-wise level during bubble inflation. We exploit the unique combination of studying a multinational company that was among worldwide protagonists during dotcom bubble and a complete coverage of daily financial ownerships for all Finnish investors. The distribution of investment gains and losses was strongly inhomogeneous across different categories of investors. Financial professionals were better equipped to obtain gains during bubble inflation and limit losses when the bubble bursts. On the contrary, investors with limited financial expertise gained during bubble inflation but incurred in significant losses — or struggled to limit them — after the bubble burst. Joint

work with Federico Musciotto (University of Palermo, Italy) and Jyrki Piilo (University of Turku, Finland)

This keynote is sponsored by Applied Network Science, Springer Nature

Céline Rozenblat University of Lausanne Switzerland



Céline Rozenblat is professor of Urban Geography at the Institute of Geography and Sustainability of the University of Lausanne, Switzerland (former Director of the Institute 2018-2022), vice-president of the International Geographical Union (IGU) and member of the Complex Systems Society. She studies systems of cities at European and world scales, multinational firm networks, inter-urban dynamics, comparative urban data, mapping and visualization of networks in geography, and spatial analysis. For several years she has worked on the relations between the evolution of multi-level urban processes and dynamics in city-system networks. To study these topics comparatively, she has built many databases on European and worldwide cities and the networks they form underlying cities' properties and evolution in a multi-dimensional and long temporal approach. Diachronic and dynamic studies supply materials to develop spatial and dynamic models and visualizations. Former member of the commission of Urban Health & Well-being of the International Science Council, she recently developed in parallel some studies on the planetary urban health and is coordinating an international MOOC on Urban Health Systems.

Keynote: Resilience and transitions from local Digital Twins to global complex urban systems

In the context of the climate change and energetic and political crises, new tendencies in the long/medium term evolution of urban systems, together with new data and methods, require that existing theoretical assumptions and conceptualizations be challenged as global urban hierarchies are reconfigured and citizens' aspirations in their urban environments are in strong transformations. The connection between urban systems at different levels of organization becomes more and more relevant for understanding urban systems and their transitions. But the current inter-urban perspective is not sufficient to encompass these dynamics. The evolution of power distributions inside and between cities reshapes the world organization of central/peripheral cities and the complexity of the global urban system. Actors as multinational firms, or high-level innovation centers, participate actively in these reconfigurations that concentrate wealth, control, innovation, and attractiveness in a few cities. At the local level, citizens wish more health and well-being implying better policy coordination between local and national/international scales. In the complexity of this multi-level system, how is regionalization of the world reshaping in a multipolar urban world? How does the multi-level perspective highlight some resilience properties if one integrates environmental care in the urban system? The theories and methodologies derived from complex systems sciences bring new perspectives for urban transitions towards more sustainability and resilience.

Network Analysis



Diagnosing network attacks by a machine-learning approach Davide Coppes and Paolo Cermelli	23
Characterisation of the robustness of weighted networks, a first step to better understand the context of humanitarian op- erations Aurélie Charles, Guillaume Bouleux and Giacomo Kahn	28
 Filtering Real World Networks: A Correlation Analysis of Statistical Backbone Techniques Ali Yassin, Hocine Cherifi, Hamida Seba and Olivier Togni	32
Interaction Network and Graphlets to Identify Sport Teams' Sig- nature Quentin Bourgeais, Guillaume Hacques, Rodolphe Charrier, Eric Sanlaville and Ludovic Seifert	36
 Measuring Movie Script Similarity using Characters, Keywords, Locations, and Interactions Majda Lafhel, Mohammed El Hassouni, Benjamin Renoust and Hocine Cherifi 	42
 NetBone: A Python Package for Extracting Backbones of Weighted Networks Ali Yassin, Abbas Haidar, Hocine Cherifi, Hamida Seba and Olivier Togni	46



Diagnosing network attacks by a machine-learning approach

Davide Coppes · Paolo Cermelli

Keywords Network science · Network attacks · Machine learning

1 Introduction

In this note we explore a simple machine-learning procedure to establish whether and how a given network has been attacked, without requiring the knowledge of the structure of the network before the attack.

We characterize a graph by a list of four normalized metrics: the ratio between the average and the maximum degree, the global clustering coefficient, the ratio between the average path length and the diameter, and the degree assortativity. Focusing on three basic random graphs, Erdős-Rényi (ER), Barabasi-Albert (BA) and Watts-Strogatz (WS), we train two popular classification algorithms, k-Nearest-Neighbor and Random Forest, to recognize whether a given network has been attacked as well as the type of attack. We test our procedure on both artificial and real networks, first performing either targeted attacks or random failures, and then applying our classification scheme to the resulting network.

Even though the training set used in this paper is quite limited, our procedure is surprisingly successful in identifying the network type and distinguishing between random failures and targeted attacks, and could therefore provide a basis for more sophisticated approaches to the diagnosis and detection of damaged networks.

Our approach is different from those usually employed in the huge literature on network attacks, where the network damage is measured in terms of the

D. Coppes

P. Cermelli

Department of Physics, University of Torino, Italys E-mail: davide.coppes@edu.unito.it

Department of Mathematics, University of Torino, Italy E-mail: paolo.cermelli@unito.it

size of the giant component, as, for instance, in the seminal works [1,2], and, in general, the knowledge of the intact network is required (cf. e.g., the reviews [3–5]). A notable exception is the recent work [6], in which a machine learning approach similar to ours has been employed to predict the robustness of various real networks.

2 The classification procedure

In order to compare graphs with different numbers of nodes, we use here normalized versions of the maximum degree and average path length, defined as

$$\delta = d_{\rm ave}/d_{\rm max}, \qquad \lambda = \ell/{\rm diam}(G),$$

where d_{ave} and d_{max} are the average and maximum degree, respectively, ℓ is the average path length and diam(G) is the diameter of the network. Henceforth, each graph in this paper will be identified by the list

$$(\delta, C, \lambda, r), \tag{1}$$

i.e., the normalized reciprocal maximum degree δ , the global clustering coefficient C, the normalized average path length λ , and the degree assortativity r (Fig. 1). We chose these four metrics for their ability to condense many network characteristics; adding other metrics would make the model more complex and probably more performing, even if further studies need to be done.

We perform three types of attacks, consisting in the removal of a given fraction of nodes (here 1, 5 and 10%) according to the following criteria: random failure, in which the nodes to be removed are drawn from a uniform distribution, and targeted attacks, in which the deleted nodes are those with decreasing degree or betweenness. Using NetworkX [7], we first generate intact random graphs with n = 500, 800 and 1000 nodes. Then, the assigned fraction of nodes is removed, so that the attacked graphs have 495, 475, 450 or 792, 760, 720 or 990, 950, 900 nodes. If the attack disconnects the network, this is discarded, so that only connected graphs are retained. Subsequently, intact ER, BA and WS random graphs are generated with these new numbers of nodes. These sets of intact and attacked networks are those that serve as benchmarks, as explained below.

In order to determine whether and how a given network has been attacked, we use two popular supervised learning algorithms, namely the k-Nearest Neighbors and the Random Forest classifiers [8]. The first was chosen for the simplicity of both use and interpretation of the results; the second one, although it is a more complex model and therefore requires more resources, is flexible and provides excellent generalization performances.

In our work the data are lists of metrics (1) of a specific network, and the labels used for the classification contain the following information:

 $-L_1$ characterizes the type of graph, i.e. ER, BA or WS.

- $-L_2$ characterizes the type of attack and can assume four possible values: intact (i), random failure (r), maximum degree (d), or maximum betweenness (b).
- $-L_3$ corresponds to the attack intensity, i.e. the fraction of removed nodes (1, 5 or 10%).

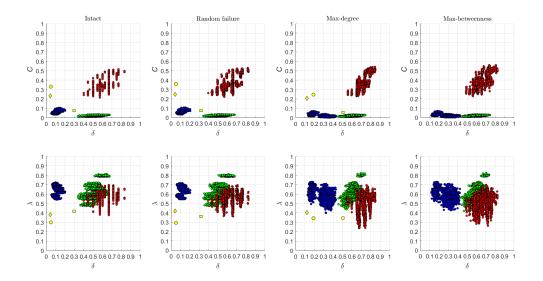


Fig. 1 Data points used to train the k-Nearest Neighbors classifier. Each data point corresponds to a network, colored in green for ER, in blue for BA and in red for WS. Data points are represented in the (δ, C) and (δ, λ) planes because the assortativity is very low in all cases. Notice that, since targeted attacks consist in the removal of nodes with maximal connectivity, which affect the maxima of the degree and path-length distributions more severely than the corresponding averages, the quantities δ and λ are consistently larger and smaller (respectively) in damaged networks relative to intact ones. The yellow circle, square and diamond are the data points corresponding to the intact and damaged fb-pages-food, power-bcspwr09 and web-polblogs networks, respectively (cf. Section 3.1).

3 Results

Since the k-Nearest Neighbors algorithm yields a lower accuracy than Random Forest, we only discuss here the performance of the latter, as evaluated both by the accuracy of the results, i.e. the fraction of correct predictions in the test set, and by the confusion matrix, which is the matrix whose entry C_{ij} is the number of networks that we know to belong to class i and which the model has classified as belonging to j; the number of data entries is 8000 for each class.

In this work, we have applied the classification procedure at the finest scale, but we group below some classes together in order to make the results more understandable and generalisable.

- The fine classification procedure allows to distinguish the type of graph, the type of attack and its intensity, so that each network is classified according to the values of $L_1 = ER, BA, WS, L_2 = i, r, d, b$ and $L_3 = 1\%, 5\%, 10\%$. At this level of classification, the accuracy of the algorithm is 0.6193.
- By grouping together intact and randomly-attacked graphs in a single class, and targeted attacks (degree and betweenness) in another, we obtain a coarser classification of the data. Here the attack intensity is neglected, so that the only label is L_{12} with values ER/(i,r), ER/(d,b), BA/(i,r), BA/(d,b), WS/(i,r), WS/(d,b). For this classification the accuracy rises to 0.9713, which yields the confusion matrix in Table 1.

	BA/i,r	ER/i,r	WS/i,r	BA/d,b	ER/d,b	WS/d,b
BA/i,r	0.9993	0.0	0.0	0.0007	0.0	0.0
$\mathbf{ER/i,r}$	0.0	0.9507	0.0	0.0014	0.0479	0.0
WS/i,r	0.0	0.0	0.9593	0.0	0.0	0.0407
BA/d,b	0.0	0.0007	0.0	0.9993	0.0	0.0
ER/d,b	0.0	0.0368	0.0	0.0007	0.9625	0.0
WS/d,b	0.0	0.0	0.0343	0.0	0.0	0.9657

Table 1 The confusion matrix relative to the coarsest label L_{12} . All values have been normalized so that the sum over each row is 1, so that the elements C_{ii} indicate the accuracy in classifying the specific class i and the elements C_{ij} , with $i \neq j$, the errors in classifying that class.

The remarkable increase in accuracy demonstrates that the main error of the algorithm is in distinguishing intact networks from those that have undergone random failures and between the two targeted attacks. Instead, the errors related to the recognition of the type of graph or the distinction between targeted and intact-random attacks are minimal.

3.1 Real networks

Although the classification algorithm was trained on networks generated by theoretical models only, we have applied it to some real networks as well, based on the (hazardous) assumption that at least one of the three models behave similarly to the real network.

We chose three networks from contexts as dissimilar as possible but where a random failure and a targeted attack made sense. The first, fb-pages-food, is a social network made up of Facebook; the second, power-bcspwr09, is a power network, while the third, web-polblogs, is a web graph [9].

We have applied to these networks both random and maximum-degree attacks deleting 10% of the nodes; in terms of the coarsest classification, with label L_{12} , the two networks obtained from the attack of the social network have been classified as WS/i, r and WS/d, b. Similarly, the networks obtained from the power network were classified as ER/i, r and ER/d, b and, finally, those deriving from the web graph as WS/i, r and WS/d, b. Even though the procedure always associates the same graph label L_1 to each pair of attacked networks, in these cases the interesting outcome does not lie in the type of graph that has been predicted, but in the correct assessment of the type of attack that the network has undergone.

4 Discussion

The main result of our work is that the Random Forest classifier gives highly satisfactory results for the classical random graphs, and is able to identify correctly whether a graph has been attacked by a targeted attack, while it is unable to distinguish between intact and randomly-attacked networks. This could be due to the fact that the random graphs used for the training have special, artificially enhanced features, and therefore it could be comparatively easy to assess whether an attack has significantly affected these features.

However, applying the classification algorithm to three real networks, both subject to random failures and targeted attacks, shows that, somewhat surprisingly, this is able to understand correctly whether the graph has been attacked.

This result suggests that machine-learning approaches may indeed be successful in performing the complex task of diagnosing whether a network has been attacked.

References

- 1. R. Albert, H. Jeong, A.L. Barabasi, Nature 406, 378 (2000)
- 2. R. Cohen, K. Erez, D. ben Avraham, S. Havlin, Physical Review Letters 85, 4626 (2000)
- 3. S. Iyer, T. Killingback, B. Sundaram, Z. Wang, PLOS ONE 8 (2013)
- M. Bellingeri, D. Cassi, S. Vincenzi, Physica A: Statistical Mechanics and its Applications 414, 174 (2014)
- 5. S. Wandelt, X. Sun, D. Feng, M. Zanin, S. Havlin, Scientific Reports 8, 13513 (2018)
- N.K.K. Nguyen, Q. Nguyen, H.H. Pham, T.T. Le, T.M. Nguyen, D. Cassi, F. Scotognella, R. Alfieri, M. Bellingeri, Complexity **2022**, 3616163 (2022)
- A. Hagberg, P. Swart, D. Schult, Technical Report, Los Alamos National Lab.(LANL) (2008)
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Journal of Machine Learning Research 12, 2825 (2011)
- 9. R.A. Rossi, N.K. Ahmed, in AAAI (2015). URL https://networkrepository.com



Characterisation of the robustness of weighted networks, a first step to better understand the context of humanitarian operations

Aurelie Charles · Guillaume Bouleux · Giacomo Kahn

Abstract In situations where crises are succeeding one to another, it is important to understand and measure the strengths and weaknesses of one's logistics network. This observation applies to many companies. It is even more relevant for humanitarian organizations, who are confronted to increased demand for humanitarian aid without having a sufficient budget to cover all present and future needs. Our proposal allows to visualize these strategic points, using complex networks. We measure the robustness of local infrastructures (health and logistics) by simulating its response in the event of a crisis. Non binary attacks, where nodes and/or links are damaged but not removed entirely are used in order to remain as close as possible to the real phenomenon, where the damages suffered by infrastructures may hinder their capacity but not always totally destroy it. We also use weighted networks. This work is carried out in close collaboration with Handicap International, so as to validate the relevance of the approach and its applicability through real applications.

Keywords Complex Systems \cdot Robustness \cdot Percolation \cdot Humanitarian Logistics

1 Non Binary Percolation

According to Bellingeri et al., assessing the robustness of real-world systems with a percolation threshold is relevant only if at least two assumptions are

A. Charles

Université Lumière Lyon2 et INSA Lyon DISP-UR4570 Tel.: +33 6 64 52 58 06 E-mail: a.charles@univ-lyon2.fr

G. Bouleux Université Jean Monnet et INSA Lyon DISP-UR4570

G. Kahn Université Lumière Lyon2 et INSA Lyon DISP-UR4570 verified. Firstly, during the percolation process, the binary (i.e. unweighted) modeling of the network should not over simplify the system. Secondly, the binarity of the theoretical attack (spared or deleted) should match with the actual pressure [2]. However, some complex real-world systems do not satisfy these assumptions. Networks are not only specified by their topology but also by the dynamics of information or traffic flow taking place on the structure [1]. In particular, the intensity of connections may be very important in the understanding of some systems. For example, in a transportation network, roads have different traffic capacity, a highway will be much busier than a trunk road. Here assumption (i) would not be verified. Plus, in real-world context, the articles [7, 3, 2] warn that a system can collapse long before the connections between the elements are broken. In other worlds, accidents or traffic jams could seriously slow down traffic while the roads remain open.

This summary proposes an approach that we call the non binary percolation (NBP) to evaluate the robustness of networks. This theory is in line with the percolation but it allows to take into account the weighting of networks and the resistance of elements under pressure. We consider that the weights represent a flow, a service or an exchange capacity. When a system is under pressure, its elements struggle to remain functional but do not systematically collapse. This observation is modeled by a non-binary attack. Elements can be spared or deleted, but also damaged, and therefore less effective.

2 Robustness Indicators

Xing Pan and Huixiong Wang highlight that the conclusions about robustness depend to a large extent on the chosen indicators [8]. Bellingeri et al. establishe that deleting a very small fraction of selected edges does not necessarily affect the largest connected component (LCC) but it can produce a rapid collapse of weighted condition measures [2]. Consequently, it is of primary importance to choose the right health indicator.

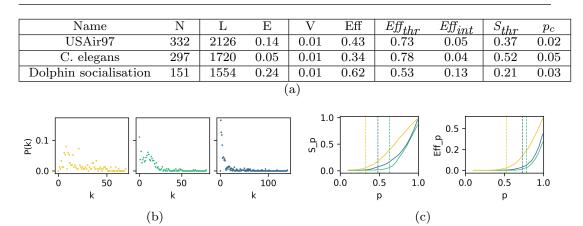
We consider four potential robustness indicators :

S. Usual percolation uses S (the fraction of nodes belonging to the LCC) to evaluate the state of the network and defines the robustness by the percolation threshold p_c . It is commonly used, but it evaluates only the topological connectedness of the network, and neglects the weights [2, 8].

Eff. Efficiency allows a precise analysis of the information flow on weighted networks and informs on the average distance between nodes [5].

 Eff_{thr} . Similarly to percolation threshold, we define Eff_{thr} as the smallest p such that Eff is not zero. When the attack is sufficiently destructive, the simulation shows a phase transition at point Eff_{thr} exactly as in percolation.

 Eff_{int} . As proposed in [6] in a classical percolation context, we suggest to use the area under the curve of the Eff during the NBP process. The larger the area, the greater the Eff during the process. This measure allows to hierarchize the robustness of the networks regardless of the Eff_{thr} value and reflects the overall behavior of the network during all of the attacks.



French Regional Conference on Complex Systems – Le Havre, May 31 June 2, 2023

Fig. 1 (a) Statistics on reals networks (b) Degree distributions. (c) NBP process. Legend : Dolphin socialisation (yellow), C. elegans (green) and USAir97 (blue).

3 Simulation on Real networks

In this section, we apply our approach to real networks : **USAir97**, which describes air routes between American airports, **C. elegans**, which describes the worm brain network [4], and **Dolphin socialisation**, which is the result of 124 days of dolphin watching in the Cedar Key area, Florida.

The usual percolation gives the following theoretical thresholds 0.02, 0.03 and 0.05 for respectively USAir97, Dolphin socialisation and finally C. elegans. NBP delivers more contrasting and different scores. The ranking is not the same either, as illustrated in figure 1(a). We observe a phase transition of Eff_{int} . Figure 1(c) reveals that Dolphin socialisation are much more robust with $Eff_{int} = 0.13$, then USAir97 with 0.05, and C. elegans with 0.04. We can see, especially for the dolphin socialisation network, that considering the condition of a network by its topology leads to an overestimation of its health.

With those robustness indicators, we intend to measure the robustness of huanitarian response networks, such as the one presented by Figure 2.

4 Conclusion

We propose to perform a non binary percolation to measure the robustness of existing networks if a large-scale natural crisis occurs. Different indicators of robustness are possible here. We can cite the classic indicator, S, which is calculated by determining the fraction of the remaining nodes located in the giant component. Other indicators related to the efficiency of the system, its ability to transmit product flows even when damaged are being studied. This summary focuses on various possibilities to measure the robustness of the networks. It is a preliminary study, before application to real humanitarian logistics networks. The possibilities offered by network theory are numerous and make it possible to better exploit the available datasets in order to improve the crisis response process.

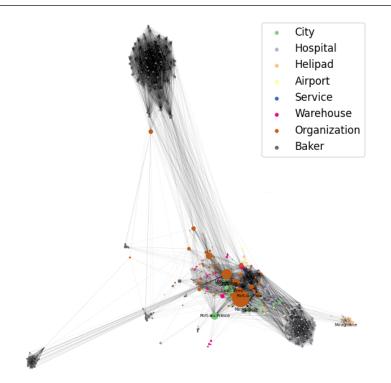


Fig. 2 Force Atlas representation of the network during relief operation in Haiti after both the earthquake and the cholera outbreak, end of 2010

References

- Alain Barrat et al. "The architecture of complex weighted networks". In: Proceedings of the national academy of sciences 101.11 (2004), pp. 3747– 3752.
- [2] M Bellingeri et al. "A comparative analysis of link removal strategies in real complex weighted networks". In: Scientific reports 10.1 (2020), pp. 1– 15.
- [3] M Bellingeri et al. "The heterogeneity in link weights may decrease the robustness of real-world complex weighted networks". In: Scientific reports 9.1 (2019), pp. 1–13.
- [4] Mohamad Khajezade, Sama Goliaei, and Hadi Veisi. "A game-theoretical network formation model for C. elegans neural network". In: *Frontiers in Computational Neuroscience* 13 (2019), p. 45.
- [5] Vito Latora and Massimo Marchiori. "Efficient behavior of small-world networks". In: *Physical review letters* 87.19 (2001), p. 198701.
- [6] Wenguo Li et al. "Maximizing network resilience against malicious attacks". In: Scientific reports 9.1 (2019), pp. 1–9.
- [7] Giannis Moutsinas and Weisi Guo. "Node-level resilience loss in dynamic complex networks". In: *Scientific reports* 10.1 (2020), pp. 1–12.
- [8] Xing Pan and Huixiong Wang. "Resilience of and recovery strategies for weighted networks". In: *PloS one* 13.9 (2018), e0203894.



Filtering Real World Networks: A Correlation Analysis of Statistical Backbone Techniques

Ali Yassin · Hocine Cherifi · Hamida Seba · Olivier Togni

Abstract Networks are an invaluable tool for representing and understanding complex systems. They offer a wide range of applications, including identifying crucial nodes, uncovering communities, and exploring network formation. However, when dealing with large networks, the computational challenge can be overwhelming. Fortunately, researchers have developed several techniques to address this issue by reducing network size while preserving its fundamental properties [1–9]. To achieve this goal, two main approaches have emerged: structural and statistical methods. Structural methods aim to keep a set of topological features of the network while reducing its size. In contrast, statistical methods eliminate noise by filtering out nodes or links that could obscure the network's structure, utilizing advanced statistical models.

In a previous work [10] we compared a set of seven statistical backbone filtering techniques in the World Air Transportation network. Results show that the Marginal Likelihood Filter, Disparity Filter, and LANS Filter give more importance to high-weight edges. The other techniques emphasize both small and high-weighted edges.

This study extends the previous research on seven statistical filtering techniques, namely Disparity, Polya Urn, Noise Corrected, Marginal Likelihood, LANS, ECM, and GloSS filters, through the analysis of 39 real-world networks

A. Yassin

H. Cherifi

ICB UMR 6303 CNRS - Univ. Bourgogne - Franche-Comté, Dijon, France

H. Seba

Laboratoire d'Informatique de Bourgogne, University of Burgundy, Dijon, France

This material is based upon work supported by the Agence Nationale de Recherche under grant ANR-20-CE23-0002.

Laboratoire d'Informatique de Bourgogne, University of Burgundy, Dijon, France E-mail: ali_yassin@etu.u-bourgogne.fr

Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France O. Togni

of diverse origins. These networks range in size from 18 to 13,000 nodes and include character, web, biological, economic, infrastructural, and offline/online social networks. In the first experiment, we aim to evaluate and compare the similarities between the seven statistical filtering techniques. Each method assigns a probability value, called a p-value, to each edge. To compare the methods, we use these p-values to conduct correlation analysis. Specifically, we compute the Pearson correlation between each pair of techniques' p-value edges. However, it is important to note that Pearson correlation examines linear relationships, whereas Jaccard similarity compares the similarities of two sets. Therefore, we use Jaccard similarity to compare the fraction of shared edges in each backbone. In a second experiment, we investigate the relationship between edge significance and edge properties. To do this, we compute the Pearson correlation between the p-values and edge properties, including weight, edge degree, and edge betweenness. Fig 1 illustrates these results.

The heatmaps present the mean and standard deviation of Pearson correlation between filtering technique pairs across all networks. The couples (LANS, Disparity filter) and (Noise Corrected, ECM) are well correlated (0.8). Conversely, the Polya Urn filter does not exhibit a noticeable correlation with any other filtering method. The standard deviation heat map shows a low standard deviation validating these findings.

The middle graphs illustrate the typical behaviors of how the mean Jaccard score changes as a function of the top fraction of edges sorted by various backbone filtering techniques. The top left panel shows a low Jaccard score between the Polya Urn filter and the Noise Corrected filter. The other techniques also have a low Jaccard score between the Polya Urn filter. The top right panel shows that the GloSS filter shares at least 20% of its edges with the Marginal Likelihood filter. The other techniques have the same behavior as the Marginal Likelihood filter with the GloSS filter except for the Polya Urn filter. The bottom right panel shows that the set of edges obtained by the Disparity filter shares on average at least 50% of its edges with the LANS filter. The ECM Filter and Marginal Likelihood Filter (ECM-MLF) and ECM Filter and Noise Corrected Filter (ECM-NC) behave similarly. Finally, in the bottom left panel the set of edges obtained by the Marginal Likelihood filter shares on average at least 70% of its edges with the Noise Corrected filter. On the other hand, the couples DF-NC, DF-ECM, DF-MLF, LANS-ECM, LANS-NC, and LANS-MLF behave the same, sharing at least around 30% of the edges. However, they have a high standard deviation.

The boxplots illustrate the Pearson correlation coefficient between edge pvalues and edge weights, degrees, and betweenness across all networks. Results indicate a greater demonstration of the distinct behavior of the Polya Urn filter. The edge p-values were found to be uncorrelated with edge weights, degree, and betweenness, with a very low standard deviation. In contrast, the top panel shows that the edge p-values obtained through the Disparity filter and Marginal Likelihood filter were correlated with weights, with average correlation higher than 0.6. This indicates that these techniques prioritize edges with high weights. In the middle panel, the Noise Corrected filter and ECM filter have an average correlation higher than 0.6. This means that these methods give importance to edges that connect hubs, as these edges have a high edge degree, which is used indirectly by these methods to determine edge significance. Finally, the bottom panel shows that the edge p-values from all techniques have no correlation with edge betweenness, indicating that none of the methods prioritize edges that play a significant role in communication between nodes through the shortest paths.

In conclusion, correlation analysis is crucial in highlighting similarities and differences between backbone edge filtering techniques, identifying areas for improvement, and advancing knowledge in this field. This study can help to identify areas where improvements can be made in the development of new techniques or in the refinement of existing ones.

Keywords Complex Networks · Backbone Filtering Techniques · Network Compression · Graph Summarization · Sparsification

References

 C.H. Gomes Ferreira, F. Murai, A.P. Silva, M. Trevisan, L. Vassio, I. Drago, M. Mellia, J.M. Almeida, Plos one 17(9), e0274218 (2022)

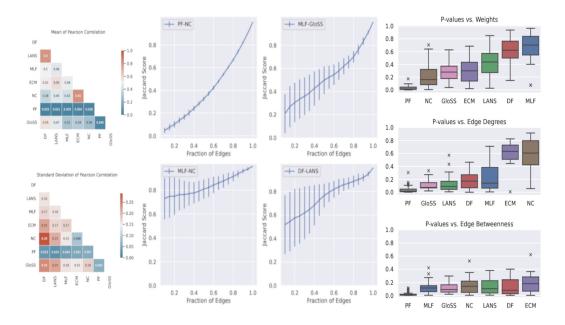


Fig. 1 At the left, the mean and standard deviation heatmap of Pearson Correlation between pairs of filtering techniques p-values across all networks. In the middle the typical behaviors of the mean Jaccard score between the set of edges of pairs of different filtering techniques as a function of fraction of edges preserved across all networks. In the right, the boxplots of Pearson correlation coefficient between edge p-values and edge weights, degrees, and betweenness across all networks. The MLF is the Marginal Likelihood Filter, DF is the Disparity Filter, LANS is the Local Adaptive Network Sparsification, PF is the Polya Urn Filter, NC is the Noise Corrected Filter, and GloSS is Global Statistical Significance Filter. Note that we took the absolute value of the Pearson correlation.

- 2. Z. Ghalmane, C. Cherifi, H. Cherifi, M. El Hassouni, Scientific Reports 10(1), 1 (2020)
- 3. Z. Ghalmane, C. Cherifi, H. Cherifi, M. El Hassouni, in 9th International Conference on Complex Networks and Their Applications (2020), pp. p-3
- 4. V. Gemmetto, A. Cardillo, D. Garlaschelli, arXiv preprint arXiv:1706.00230 (2017)
- Z. Ghalmane, C. Cherifi, H. Cherifi, M. El Hassouni, Information Sciences 576, 454 (2021)
- S. Rajeh, M. Savonnet, E. Leclercq, H. Cherifi, in Network Science: 7th International Winter Conference, NetSci-X 2022, Porto, Portugal, February 8-11, 2022, Proceedings (Springer International Publishing Cham, 2022), pp. 67-79
- A. Yassin, H. Cherifi, H. Seba, O. Togni, in 2022 IEEE Workshop on Complexity in Engineering (COMPENG) (IEEE, 2022), pp. 1–8
- A. Yassin, H. Cherifi, H. Seba, O. Togni, in Complex Networks and Their Applications XI: Proceedings of The Eleventh International Conference on Complex Networks and their Applications: COMPLEX NETWORKS 2022—Volume 2 (Springer International Publishing Cham, 2023), pp. 551–564
- 9. L. Dai, B. Derudder, X. Liu, Journal of Transport Geography 69, 271 (2018)
- A. Yassin, H. Cherifi, H. Seba, O. Togni. Air transport network: A comparison of statistical backbone filtering techniques (2023). DOI 10.1007/978-3-031-21131-743



Interaction Network and Graphlets to Identify Sport Teams' Signature

Quentin Bourgeais · Guillaume Hacques · Rodolphe Charrier · Eric Sanlaville · Ludovic Seifert

Abstract The traditional way of analyzing team sport performance has limitations in terms of understanding the interactions between players and the performance context. Therefore, it has been proposed to analyze teams through the lens of complexity science paradigm and to use graph theory to analyze interaction networks between players in order to assess the collective behavior. In this study, we aim at (1) investigating how a defensive imbalance constraint the emergence of interactions patterns between players and (2) identifying "team's signature" defined as their preferences in this emergence. 24 rugby teams and 18 basketball teams of 3 young elite players played a small-sided game in 2 situations characterized by different levels of defensive imbalance (high/low). We established a list of all possible network structures ("graphlets") and associated each possession with a graphlet to design a "profile" as the frequency of each graphlet. We evaluated the effect of the manipulated constraint on the collective behavior by comparing the mean profile of both situations, and we detected teams' signature by clustering teams' profiles. Results suggest that the defensive imbalance constraints more basketball teams than rugby teams, whereas team preferences seem more significant in rugby. By mobilizing complexity science paradigm and graph theory to assess collective behavior, we are able to explore the effect of a given constraint on interaction between players and to identify each team's preferred patterns of interaction. It provides a more performance-contextualized and interaction-driven analytical framework which could easily be extended afterwards.

Keywords team sport \cdot interaction \cdot network

R. Charrier \cdot E. Sanlaville

Q. Bourgeais \cdot G. Hacques \cdot L. Seifert

CETAPS UR3832, Université de Rouen Normandie, France

Normandie Univ, UNIHAVRE, UNIROUEN, INSA Rouen, LITIS, Le Havre, France

1 Introduction

Team sport performance has been usually analyzed through notation systems registering actions of players and/or critical events of the game over time, but it faces some limitations such as the little reference to the performance context and to the ongoing interactions between performers: that is why it has been proposed to analyze teams through the lens of complexity science paradigm [1]. Indeed, team sports players are coordinating themselves to achieve a common goal: a team defined as a complex system means that its parts (i.e. players) are interacting and that its collective behavior emerges from these interactions. Such systems have several ways of organizing and re-organizing in response to a change within it or within the surrounding constraints: as each team is doing it in its own way we can refer to the existence of a "team signature" [2]. Graphs are a useful tool to analyze complex systems by allowing a focus on the network of interactions: it's typically used in team sport analysis through the network of aggregated passes over at least 1 game (usually called "passing network") to provide a direct visual inspection of a team's strategy and to calculate some network metrics from Social Network Analysis (e.g. [3]). Although aggregated passing networks show some differences between team's organizations, the temporal dimension needs to be considered to obtain a more detailed profile of a team and so identify its signature [4]. In a graph, we can try to identify the existence of "motifs", defined as small subgraphs that are over-represented in comparison to a randomized graph [5]. One can also define motifs as specific patterns of edges between vertices that appear to be statistically significant in the network: they are considered as structural signatures of the function of a network, and a motif profile can be used to discriminate networks [6]. Inspired by motifs, some authors proposed to analyze "flow motif' in passing networks that are statistically significant pass sequence pattern, in a way to identify football team's playing style by making a profile of flow motif by team [7]. Because "flow motifs" consider the sequential order of the passes, they contain temporal information about the network structure and are more representative of the actual interactions between players. If motifs are a special type of subgraphs (i.e. those that are over-represented in a network compared to a given model) some authors proposed the concept of "graphlets" (defined as connected networks with a small number of nodes) to describe all possible structures with a given number of nodes [8]. This concept has been extended to various types of networks such as directed networks [9] or temporal networks [10]. However, all of these works focused on subgraphs and their frequency to obtain what we call here a "profile" of the network. In line with it, we investigated profiles of basketball and rugby teams to describe team's behavior. In this work, we set up a controlled Small-Sided Game (or SSG, which is a game-like training situation played on a smaller pitch and with fewer players than in the real performance context) in which we manipulated a specific constraint (i.e. the defensive imbalance) and we characterize the collective behavior through a profile of players interactions patterns that have occurred. By comparing profiles, our objective is twofold: (1) to investigate how the constraint we manipulate during the SSG shapes the emergence of interactions patterns between players and (2) to identify team's signatures. A team's signature is more precisely defined here as the preferences that appear in the organization (when facing a given constraint) and re-organization (when the constraints changes) of a team when players repeatedly interact with each other. It suits common objectives in team sport analysis: to understand which constraints lead to which reorganizations of inter-personal coordination [11] and to identify some regularities in the way a team plays [12].

2 Methods

Data consists in 24 rugby and 18 basketball teams of 3 young elite players who repeated 24 times a SSG played in 2 situations (12 times each) characterized by 2 different levels of defensive imbalance (high and low). The situation is controlled 3 vs 3 situation replicated each time identically because we fixed the starting position of all the attackers and all the defenders. We induced a defensive imbalance by changing the positioning of one of the defenders: by positioning him further away from the zone he has to defend, we create a defensive delay (or imbalance) that can be exploited by the attacking team. Thus, it is expected that a higher defensive imbalance facilitates the attack, whereas a lower defensive imbalance tends to help the defensive team. This is confirmed by the success rate of the possessions, in both sport: from 57% of success in the balanced situation to 64% of success in the imbalanced one in rugby, and from 81% of balanced possessions ending with a shot (with 35%of shooting efficiency) to 85% of imbalanced ones (with 50% of efficiency). Data represents a total of 1008 possessions (576 in rugby, 432 in basketball) from which we extract the passing network performed by the team that we classify according to their structure (Figure 1). Because in this SSG time was constrained (i.e. 8 seconds maximum to play a possession) we are able to associate a unique (directed) graphlet to each possession. Although we have more possible graphlets than the original graphlets' classification because we consider directed edges, all directed graphlets are not possible with our specific data: the interaction being the transmission of a ball, it means you must receive it before passing it. Interestingly, our classification closely aligns with weighted subgraphs detectable by a random walker [13], but we choose to simply talk about graphlet for language ease.



Fig. 1 List of possible graphlets with 0 to 3 passes between 3 players ('other' includes structures with 4 passes or more).

We design a profile as the relative proportion (or frequency) of each possible graphlet performed among all possible ones during the 12 trials in a given situation. First, we compare the average profiles (i.e. including all teams) performed to achieve both situations in order to evaluate the effect of the defensive imbalance on the collective behavior. Then, using JASP software (JASP Team, 2023) we cluster profiles (within each sport separately) to identify teams' signature with a neighborhood-based algorithm: occurrences of each graphlet constitutes our 9 features (algorithmic settings: Hartigan-Wong algorithm, using means as center type and optimizing the BIC value with a maximum of 10 clusters). A team behavior is described as the average profile of their assigned cluster in a given condition (i.e. balanced, imbalanced) and the effect of the defensive imbalance on their behavior is evaluated by considering the couple of clusters they are assigned to in both situations. Doing so, we characterize preferences in teams' organization and re-organization, what we defined as their team's signature.

3 Results

For consistency, we have to exclude possessions in which the team fails: losing the ball before reaching the target ends the passing network too early and biases the comparison, so we focus on possessions in which rugby team scores a try (excluding 39% of the trials) and in which basketball team takes a shot (excluding 18%). Firstly, we focused on the average profile by situation. Results shows that players do not interact in the same way to achieve the 2 situations: the χ^2 independence test indicates that the profile is related to the imbalance level in basketball ($\chi^2 = 21.403$, df = 8, p = .006) as in rugby ($\chi^2 = 28.366$, df = 8, p; .001). We observe the reinforcement of the preferred interaction pattern's prevalence while increasing the defensive imbalance: pattern '12' in basketball and pattern '1223' in rugby. Secondly, we focused on each profile separately (i.e. for a single team in a single condition). We observe that several profiles exist, as we can differentiate clusters of profiles in basketball (Clusters = 3, N $= 36, R^2 = 0.294, BIC = 319.020, Silhouette = 0.150$ and in rugby (Clusters $= 5, N = 48, R^2 = 0.455, BIC = 404.610, Silhouette = 0.180$). Clusters are sorted from C1 to C5 according to the number of teams constituting the cluster (Figure 2).

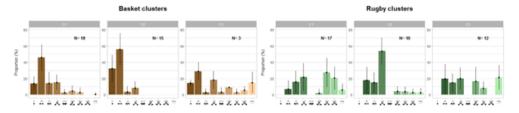


Fig. 2 Average team's profile by clusters, within each sport (with 'N' the number of teams within each cluster, and the standard deviation displayed for each pattern within each cluster). On the left, the 3 for basketball teams. On the right, 3 of the 5 for rugby teams (we exclude C4 and C5 representing respectively 2 teams and 1 team's).

The rugby clustering solution is more complex but explains a higher proportion of the variance than the basketball one (45.5% compared to 30%). Even if the clustering solution has some structure, there may be room for improvement according to the silhouette scores (respectively 0.150 and 0.180 in basketball and rugby). Finally, we focused on "cluster couples" (i.e. the cluster assignment of each team in both situations).

4 Discussion & Conclusion

In basketball, we are able to identify a prevalent graphlet ('12') regardless of the defensive situation, with an increase of its prevalence when there is more defensive imbalance because the second most recurrent ('1') decreased in favor of more complex forms (e.g. '1221', '1223', '122331'). The basketball teams' clustering mostly matches to the manipulation of the constraint (i.e. defensive imbalance), plus a third smaller category of behaviors consisting of more recurrent use of more complex networks. In rugby, we also identify a prevalent graphlet ('1223') with an increase of its prevalence when the defensive imbalance increases. The most balanced situation makes more varied collective behaviors emerge, and especially more complex ones (e.g. '122331', '122332', 'others'). The rugby teams' clustering is more complex and shows more diversity of behavior between teams, with 5 clusters and multiple couples: 2 of the 3 main clusters have profiles that are relatively balanced between simple and complex structures, while the 3rd one is characterized by a very strong prevalence of the pattern '1223' (the last two clusters contain negligeable data and therefore are not relevant to be discussed). In this work, we mobilized the complexity science paradigm and graph theory to analyze sports teams. More precisely, we aim at designing a profile of the interaction network within a team, modeled by the graphlets' frequency, in a way to describe the team behavior. We argue that this framework allows us to (1) investigate the effect of a given constraint on the emergence of team's behavior (what we particularly see in basketball) and (2) explore team's signature, defined as the behavioral preferences of a team that emerge in relation to the constraints (more observed in rugby). This approach offers a more performance-contextualized and interaction-driven analytical framework, which is essential for team sport analysis. Obviously, this framework requires further development in particular by going beyond the limits of current data (e.g. not so many observations because of the difficulties of implementing an interventional protocol with elite players, small passing networks because possessions last for a maximum of 8 seconds, a limited number of possible graphlets because we have only 3 players per team). For example, future work could perform such analysis using larger dataset of games played in a real performance context. There is also a need to consider more effectively spatial and temporal dimensions, and even try to go beyond the pass to focus on other interactions.

References

- B. Travassos, K. Davids, D. Araújo, T.P. Esteves, International Journal of Performance Analysis in Sport 13(1), 83 (2013)
- 2. M. Hughes, I. Franks, Journal of sports sciences 23(5), 509 (2005)
- 3. J.L. Pena, H. Touchette, arXiv preprint arXiv:1206.6904 (2012)
- R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Science 298(5594), 824 (2002)
- 5. J. Buldú, J. Busquets, I. Echegoyen, F. Seirul. lo, Scientific reports 9(1), 13602 (2019)
- F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.G. Young, G. Petri, Physics Reports 874, 1 (2020)
- 7. L. Gyarmati, H. Kwak, P. Rodriguez, arXiv preprint arXiv:1409.0308 (2014)
- 8. N. Pržulj, D.G. Corneil, I. Jurisica, Bioinformatics 20(18), 3508 (2004)
- 9. D. Aparício, P. Ribeiro, F. Silva, arXiv preprint arXiv:1511.01964 (2015)
- 10. Y. Hulovatyy, H. Chen, T. Milenković, Bioinformatics **31**(12), i171 (2015)
- N. Balague, C. Torrents, R. Hristovski, K. Davids, D. Araújo, Journal of Systems Science and Complexity 26, 4 (2013)
- 12. J. Gudmundsson, M. Horton, arXiv preprint arXiv:1602.06994 (2016)
- F. Picciolo, F. Ruzzenenti, P. Holme, R. Mastrandrea, New Journal of Physics 24(5), 053056 (2022)



Measuring Movie Script Similarity using Characters, Keywords, Locations, and Interactions

Majda Lafhel · Mohammed El Hassouni · Benjamin Renoust · Hocine Cherifi

Abstract Measuring similarity between multilayer networks is difficult, as it involves various layers and relationships that are challenging to capture using distance measures. Existing techniques have focused on comparing layers with the same number of nodes and ignoring inter-relationships. In this research, we propose a new approach for measuring the similarity between multilayer networks while considering inter-relationships and networks of various sizes. We apply this approach to multilayer movie networks composed of layers of different entities (character, keyword, and location) and inter-relationships between them. The proposed method captures intra-layer and inter-layer relationships, providing a comprehensive overview of the multilayer network. It can be used in various applications, including analyzing movie story structures and social network analysis.

Keywords Movie Script Multilayer Network · Inter layer relationships · Multilayer Graph Distance measure, Comparing movies

1 Introduction

In recent years, multilayer network analysis has achieved widespread use in various fields, including social networks, transportation networks, biological networks, and communication networks. Measuring the similarity between multilayer networks is a complex task. That is because the multilayer network consists of different entities of layers and relationships, making it challenging to define a distance measure that captures the overall multilayer network structures. Brodka et al. (2018) [1]

Mohammed El Hassouni

E-mail: mohamed.elhassouni@gmail.com

Benjamin Renoust E-mail: renoust@gmail.com

Hocine Cherifi E-mail: hocine.cherifi@gmail.com

Majda Lafhel E-mail: majdalafhel1@gmail.com

have proposed a property matrix that represents a multiplex network. The property matrix maps layers and nodes into structures. Brodka et al. have used three methods to compare multiplex networks: aggregations(min, max, entropy), layer distributions(Jensen-Shannon Divergence), and similarity functions(Jaccard, cosine, correlation). Giordano et al. (2019) [2] used factorial methods for quantifying multiplex networks visually. Ghawi et al. (2022) [3] have used community detection to quantify the similarity between multilayer networks.

In previous works [4][5], we investigated Network Portrait Divergence [6] and Laplacian Spectra Descriptor [7] to compare the similarity between movie stories. We extracted for each movie a multilayer network [8] composed of three layer entities (character, keyword, and location). We ignored inter-relationships and compared monolayers of the same entities.

This research aims to quantify the similarity between movies. There have been multiple approaches to quantifying visual content [9–12,11,13–16]. Here, we consider multilayer network movies with inter-relationships between layers. To the best of our knowledge, there is currently no approach for measuring the similarity between multilayer networks considering inter-relationships. Analyzing the structure of interlayer relationships provides extra information and a comprehensive overview of the multilayer network. Moreover, previous studies have focused on comparing layers with the same number of nodes. Multilayer movie networks consist of layers of different sizes, making finding an appropriate measure challenging. We propose an approach that captures intralayer and intralayer relationships in the multilayer network, considering networks of various sizes.

2 Methodology

A graph \mathcal{G} is a set of nodes \mathcal{N} connected by edges \mathcal{E} . Based on this property, we consider nodes of the same entity linked by intra-relationships as graphs \mathcal{G}_{intra} and nodes of different entities connected by inter-relationships as graphs \mathcal{G}_{inter} . We work on multilayer network movie scripts with three entities (character, keyword, and location). So, the multilayer network includes six types of networks: $\mathcal{G}_{intra_{CC}}$ is the character graph, $\mathcal{G}_{intra_{KK}}$ is the keyword graph, $\mathcal{G}_{intra_{LL}}$ is the location graph, $\mathcal{G}_{inter_{CK}}$ consists of inter-relationships connecting character and keywords, $\mathcal{G}_{inter_{KL}}$ consists of inter-relationships connecting keyword and location nodes, and $\mathcal{G}_{inter_{KL}}$ consists of inter-relationships connecting keyword and location nodes.

The proposed algorithm (Algorithm 1) maps \mathcal{G}_{intra} , \mathcal{G}_{inter} , and network features into one matrix \mathcal{P} as follows.

- (i) Six rows, where the first three rows represent the three intralayers ($\mathcal{G}_{intra_{CC}}$, $\mathcal{G}_{intra_{KK}}$, and $\mathcal{G}_{intra_{LL}}$), and the last three rows represent the three interlayers ($\mathcal{G}_{inter_{CK}}$, $\mathcal{G}_{inter_{KL}}$, and $\mathcal{G}_{inter_{CL}}$).
- (ii) Six columns, each one represents a network features: max degree, max centrality, density, adjacency, Laplacian, and network portrait.
- (iii) Each cell c_{ij} encodes a network feature j of the network type i.

French Regional Conference on Complex Systems - Le Havre, May 31 June 2, 2023

Algorithm 1 Matrix property extraction **input:** $\mathcal{G}_{intra_{CC}}$, $\mathcal{G}_{intra_{KK}}$, $\mathcal{G}_{intra_{LL}}$, $\mathcal{G}_{inter_{CK}}$, $\mathcal{G}_{inter_{KL}}$, $\mathcal{G}_{inter_{CL}}$ **output:** matrix property \mathcal{P} 1: for i in $\mathcal{G}_{intra_{CC}}$, $\mathcal{G}_{intra_{KK}}$, $\mathcal{G}_{intra_{LL}}$, $\mathcal{G}_{inter_{CK}}$, $\mathcal{G}_{inter_{KL}}$, $\mathcal{G}_{inter_{CL}}$ do 2: $\mathcal{D} \leftarrow \max((\deg(1), \deg(2), ..., \deg(\mathcal{N}))) //return the max node defined as a statement of the statement$ //return the max node degree of i. 3: $\mathcal{BC} \leftarrow \max(\sum_{s \neq t \in V} \frac{\sigma_{st}(\mathcal{N})}{\sigma_{st}})$ //return the max node betweeness centrality of i $//\sigma_{st}$: total shortest paths passing from a node s to a node t $//\sigma_{st}(\mathcal{N})$: total number of σ_{st} that passing through a node n //return density of i 4: $Dens \leftarrow \mathcal{E}/(\mathcal{N} * (\mathcal{N} - 1))$ 5: $\mathcal{A} \leftarrow Extract_Adjacency_matrix(i)$ 6: $\mathcal{L} \leftarrow Extract_Laplacian_matrix(i)$ 7: $\mathcal{B} \leftarrow Extract_NetworkPortrait_matrix(i)$ 8: $s_A \leftarrow sum(eigenvalues(\mathcal{A}))$ 9: $s_L \leftarrow sum(eigenvalues(\mathcal{L}))$ 10: $s_B \leftarrow sum(\mathcal{B})$ 11: $v_i \leftarrow [\mathcal{D}, \mathcal{BC}, Dens, s_{\mathcal{A}}, s_{\mathcal{L}}, s_{\mathcal{B}}]$ 12: end for $13: \ \mathcal{P} \leftarrow [v_{\mathcal{G}_{intra}_{CC}}, v_{\mathcal{G}_{intra}_{KK}}, v_{\mathcal{G}_{intra}_{LL}}, v_{\mathcal{G}_{inter}_{CK}}, v_{\mathcal{G}_{inter}_{KL}}, v_{\mathcal{G}_{inter}_{CL}}]$

Consider a pair of multilayer network movies M and M'. In the first step, we extract property matrices \mathcal{P} from M and \mathcal{P}' from M'. Second, we flatten matrices \mathcal{P} and \mathcal{P}' to vectors \hat{v} and $\hat{v'}$. Then, we compute the distance between \hat{v} and $\hat{v'}$ using the Euclidean Distance.

$$\hat{D} = \sqrt{\sum_{i=\mathcal{G}_{intra_{CC}}}^{\mathcal{G}_{inter_{CL}}} (\lambda_{\hat{v}_i} - \lambda'_{\hat{v'}_i})^2} \tag{1}$$

3 Experimental Results

We performed experiments using movie scripts from various categories. In our previous work, it appears that the romance films were the more challenging. Therefore we concentrate on these movies. We compare: Titanic (1997), episode I of Twilight (2008), and episode II of Twilight (2009). For each movie, we extracted three layers (character, keyword, location), intra-relationships and inter-relationships. We collected ground-truth data by inviting a group of individuals to rank the similarity between romance movies. Based on the evaluation, we obtained the following ranking: Episodes I and II of Twilight are in the first rank, Titanic and I of Twilight in the second, and episodes II and three also in the second.

To illustrate the efficiency of the proposed method in quantifying the similarity between romance movies (Titanic, episodes I and II of Twilight) we compare the obtained results (Table 2) to those of previous studies (Table 1).

In a previous investigation (Table 1), the Network Portrait Divergence revealed the similarity between character layers, and the Network Laplacian Spectra detected the similarity between location layers. But, no measure indicated the similarity between keyword layers.

French Regional Conference on Complex Systems – Le Havre, May 31 June 2, 2023

Table 1 Checklist table for similarity between romance movies				
Measures/Layers	Character	Keyword	Location	
NetLSD	X	X	X	
NetMF	X	X	X	
D-measure	X	X	X	
Network Portrait Divergence	\checkmark	×	×	
Laplacien Spectra	X	×		

According to Table 2, the distance between episodes I and II of Twilight (272.93) is the smallest. That means episodes I and II are the most similar. Indeed, the ground-truth data shows episodes I and II of Twilight are in the first rank. On the other hand, Titanic is closer to episode II of Twilight (403.52) than Episode I (450.24). However, according to the ground-truth data, both pairs of movies are second, which reveals how Titanic is far from episodes I and II of Twilight.

Table 2 Distance between romance movies using	the proposed method
Romance movies	Distance
episode I of Twilight & episode II of Twilight	272.93

episode I of Twilight & episode II of Twilight	272.93
Titanic & episode I of Twilight	450.24
Titanic & episode II of Twilight	403.52

In brief, the proposed method revealed the high similarity between episodes I and II of Twilight and the distance between Titanic compared to both movies. In contrast, in the previous research, no measure revealed the similarity between keyword layers. Furthermore, the time complexity of the proposed technique is much smaller than the prior one. That is because we compare the overall multilayers at one time.

References

- 1. P. Bródka, A. Chmiel, M. Magnani, G. Ragozini, Royal Society open science 5(8), 171747 (2018)
- G. Giordano, G. Ragozini, M.P. Vitale, Social Networks 59, 154 (2019)
- 3. R. Ghawi, J. Pfeffer, Social Networks 68, 1 (2022)
- 4. M. Lafhel, H. Cherifi, B. Renoust, M. El Hassouni, Y. Mourchid, in International Conference on Complex Networks and Their Applications (Springer, 2020), pp. 284–295
- 5. M. Lafhel, L. Abrouk, H. Cherifi, M. El Hassouni, in 2022 IEEE Workshop on Complexity in Engineering (COMPENG) (IEEE, 2022), pp. 1–5
- 6. J.P. Bagrow, E.M. Bollt, Applied Network Science 4(1), 1 (2019)
- 7. A. Tsitsulin, D. Mottin, P. Karras, A. Bronstein, E. Müller, in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2018), pp. 2347-2356
- 8. Y. Mourchid, B. Renoust, H. Cherifi, M. El Hassouni, in International Conference on Complex Networks and their Applications (Springer, 2018), pp. 782-796
- 9. S. Rital, H. Cherifi, S. Miguet, in International Conference on Pattern Recognition and Image Analysis (Springer, Berlin, Heidelberg, 2005), pp. 522–531
- 10. C. Demirkesen, H. Cherifi, in International conference on advanced concepts for intelligent vision systems (Springer, Berlin, Heidelberg, 2008), pp. 752-763
- 11. S. Rital, A. Bretto, H. Cherifi, D. Aboutajdine, in International Symposium on VIProm-Com Video/Image Processing and Multimedia Communications (IEEE, 2002), pp. 351-355
- 12. A. Lasfar, S. Mouline, D. Aboutajdine, H. Cherifi, in Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, vol. 1 (IEEE, 2000), vol. 1, pp. 1031–1034
- 13. M. Hassouni, H. Cherifi, D. Aboutajdine, IEEE Transactions on Image Processing 15(3), 572 (2006)
- 14. R.R. Pastrana-Vidal, J.C. Gicquel, C. Colomes, H. Cherifi, Proc. 5th Int. WIAMIS (2004)
- 15. R.R. Pastrana-Vidal, J.C. Gicquel, J.L. Blin, H. Cherifi, in Human Vision and Electronic Imaging XI, vol. 6057 (SPIE, 2006), vol. 6057, pp. 276–286
- 16. M. Messadi, H. Cherifi, A. Bessaid, arXiv preprint arXiv:2106.04372 (2021)



NetBone: A Python Package for Extracting Backbones of Weighted Networks

Ali Yassin · Abbas Haidar · Hocine Cherifi · Hamida Seba · Olivier Togni

Abstract NetBone is a new open-source Python package designed to simplify analyzing complex networks. With a wide range of techniques available, Net-Bone allows researchers to extract the backbone of a network while preserving its essential structure. The package includes nine structural methods and five statistical techniques, offering users a comprehensive solution to network analysis. It is user-friendly and straightforward to use, with easy installation. The package accepts different types of inputs, including data frames or Networkx graphs, and provides evaluation measures for comparative purposes. Additionally, NetBone offers an option to generate plots. Its versatility makes it a valuable tool for data scientists and social scientists, significantly enhancing their research and data analysis capabilities.

Keywords Complex Networks · Backbone Filtering Techniques · Network Compression · Graph Summarization · Sparsification

A. Yassin

A. Haidar

Computer Science Department, Lebanese University, Beirut, Lebanon

H. Cherifi

ICB UMR 6303 CNRS - Univ. Bourgogne - Franche-Comté, Dijon, France

H. Seba Univ Lvc

This material is based upon work supported by the Agence Nationale de Recherche under grant ANR-20-CE23-0002.

Laboratoire d'Informatique de Bourgogne, University of Burgundy, Dijon, France E-mail: ali_yassin@etu.u-bourgogne.fr

Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France O. Togni

Laboratoire d'Informatique de Bourgogne, University of Burgundy, Dijon, France

1 Introduction

Networks are increasingly important in scientific research across various fields, providing a valuable means of understanding relationships between different entities from different domains, including biological, infrastructural, and social sciences. Network analysis has been applied to diverse research questions, from studying the spread of diseases to understanding social hierarchies and beyond. However, as networks become more complex, analyzing them can be challenging, especially for those with a large number of nodes and edges. To address this issue, many approaches have been developed for reducing the size of networks while preserving their essential structure.

One of the most common methods for analyzing complex networks is to extract their backbone, which involves identifying the most significant edges and nodes while discarding extraneous information [1–9]. There are two primary categories of backbone extraction techniques: structural and statistical. Structural techniques focus on the network's topological properties and extract a backbone with specific topological features. Statistical techniques, on the other hand, assess the importance of edges and nodes based on hypothesis testing or empirical distribution, frequently used to eliminate noise in the network.

Despite the variety of approaches available for extracting the network backbone, there is currently no common framework in the field of network analysis that simplifies their application. To overcome this issue, we have introduced a new Python package called "NetBone" that offers a range of filtering techniques for extracting the network backbone.

2 Netbone presentation

The package provides nine structural techniques, such as the maximum spanning tree, global threshold, doubly stochastic, metric backbone, ultra-metric backbone, high salience skeleton, h-backbone, modularity backbone, and overlapping nodes and hubs backbone. Additionally, it offers five statistical techniques, including the disparity filter, marginal likelihood filter, noise corrected filter, locally adaptive network sparsification filter, and enhanced configuration mode filter.

The installation and usage of the NetBone package for network analysis are user-friendly and straightforward. With just a few simple steps, users can install the package via pip or directly from the repository ¹. Once installed, the user can simply import the package and call the desired method, giving it the appropriate input. NetBone accepts different types of inputs, including data frames or Networkx graphs. Then the user can then choose from a variety of filters, such as the boolean filter, fraction filter, or threshold filter. For instance, the boolean filter is suitable for methods that extract one subgraph that cannot be modified, such as the metric backbone or maximum spanning tree filter.

¹ https://gitlab.liris.cnrs.fr/coregraphie/netbone/

On the other hand, the threshold or fraction filter is useful for methods like the high salience skeleton or disparity filter that assign scores or p-values to the edges, allowing the user to set a threshold or keep only a specified fraction of edges or edges within the desired threshold. Listing 1 shows a sample code snippet on how to install and use the NetBone python package for network analysis.

In addition, NetBone not only offers various filtering techniques for extracting the backbone of a network, but it also includes evaluation measures utilized by Serrano [10] for comparing different methods. These measures encompass the fraction of edges, nodes, weights, and the number of components preserved in the backbone, as well as the cumulative weight and degree distribution. Moreover, NetBone provides an option for users to visualize the results by generating plots through the use of matplotlib and seaborn python libraries.

3 Conclusion

NetBone is a versatile and easy-to-use tool for network analysis that provides a comprehensive solution. Its open-source nature allows users to adapt and customize it to their specific needs. With its wide range of filtering techniques and evaluation measures, NetBone is suitable for researchers, data scientists, and social scientists alike. By providing an efficient way to analyze complex networks, NetBone has the potential to enhance your research and data analysis greatly.

```
# install NetBone
1
 !pip install netbone
2
 3
4
 # import NetBone and used libraries
5
6 import NetBone as nb
 import networkx as nx
7
 import pandas as pd
8
9
 10
11
12 # read the weighted edge list using networkx
 G = nx.read_weighted_edgelist(filename)
13
14
 # read the weighted edge list using pandas
15
 G = pd.read_csv(filename)
16
17
 ******
18
19
 # apply the Metric Backbone filter
20
 m_backbone = nb.metric_backbone(G)
21
22
 # apply the High Salience Skeleton filter
23
24 hss_backbone = nb.high_salience_skeleton(G)
25
```

```
27
28 # apply a boolean filter to extract the backbone
29 m_backbone = nb.boolean_filter(m_backbone)
30
31 # apply a threshold filter on the scores to keep all edges with
        scores higher than 0.5
32 hss_backbone = nb.threshold_filter(hss_backbone, threshold=0.5)
33
34 # apply a fraction filter on the scores to keep 30% of the network
35 hss_backbone = nb.fraction_filter(hss_backbone, fraction=0.3)
```

Listing 1 Here is a sample code snippet showing how to install and use the NetBone python package for network analysis

References

- C.H. Gomes Ferreira, F. Murai, A.P. Silva, M. Trevisan, L. Vassio, I. Drago, M. Mellia, J.M. Almeida, Plos one **17**(9), e0274218 (2022)
- 2. Z. Ghalmane, C. Cherifi, H. Cherifi, M. El Hassouni, Scientific Reports 10(1), 1 (2020)
- Z. Ghalmane, C. Cherifi, H. Cherifi, M. El Hassouni, in 9th International Conference on Complex Networks and Their Applications (2020), pp. p–3
- 4. V. Gemmetto, A. Cardillo, D. Garlaschelli, arXiv preprint arXiv:1706.00230 (2017)
- Z. Ghalmane, C. Cherifi, H. Cherifi, M. El Hassouni, Information Sciences 576, 454 (2021)
- S. Rajeh, M. Savonnet, E. Leclercq, H. Cherifi, in Network Science: 7th International Winter Conference, NetSci-X 2022, Porto, Portugal, February 8-11, 2022, Proceedings (Springer International Publishing Cham, 2022), pp. 67-79
- A. Yassin, H. Cherifi, H. Seba, O. Togni, in 2022 IEEE Workshop on Complexity in Engineering (COMPENG) (IEEE, 2022), pp. 1–8
- A. Yassin, H. Cherifi, H. Seba, O. Togni, in Complex Networks and Their Applications XI: Proceedings of The Eleventh International Conference on Complex Networks and their Applications: COMPLEX NETWORKS 2022—Volume 2 (Springer International Publishing Cham, 2023), pp. 551–564
- 9. L. Dai, B. Derudder, X. Liu, Journal of Transport Geography 69, 271 (2018)
- M.A. Serrano, M. Boguna, A. Vespignani, Proceedings of the National Academy of Sciences 106, 6483 (2009). DOI 10.1073/pnas.0808904106

Social Complexity



Visualizing Mobile Phone Communication Data in Criminal In- vestigations: the Case of Media Multiplexity	
Martina Reif, Bruno Pinaud, Thomas Souvignet, Guy Melançon and Quentin Rossy	51
Bounded confidence models generate one more cluster when the number of agents is slowly growing Yerali Gandica and Guillaume Deffuant	
 Monetization in online streaming platforms: an exploration of inequalities in twitch.tv Antoine Houssard, Federico Pilati, Maria Tartari, Pierluigi Sacco and Riccardo Gallotti 	60
Political Participation and Voluntary Associations : A Hyper- graph Case Study Amina Azaiez and Robin Salot	64
Opinion dynamics model revealing yet undetected cognitive bi- ases Guillaume Deffuant	80
	00



Visualizing Mobile Phone Communication Data in Criminal Investigations: the Case of Media Multiplexity

Martina Reif · Bruno Pinaud · Thomas Souvignet · Guy Melançon · Quentin Rossy

Abstract Given the multiple channels offered by social media apps, the analysis of communication data in criminal investigations has become a challenging task. A multivariate graph, gathering information of different types, can be inferred from communication events (calls, group discussions, etc.) and contact information (e.g. phone directory or app "friends"). Astute transformations are however required to properly associate virtual entities used by a single physical person. This paper proposes a visual analytics approach to support this task relying on graph transformations and proper visual encodings.

Keywords Mobile phone data \cdot Communication data \cdot Multivariate graphs \cdot Network Visual Analytics \cdot Crime Analysis

1 Introduction

Mobile phone communication data is frequently used in criminal investigations to reconstruct activities of interest, identify key actors and significant locations, or study the relationships between relevant entities of a case [1,2]. Visualization is an essential step in the analysis process as it facilitates information interpretation. Traditional communication data admitted a straightforward model since a phone could only be used for voice conversation, SMS or MMS, and this typically using a single phone number. As a consequence, communications could be modeled as a one-mode graph where two phone numbers (nodes) were linked whenever a communication took place between them (edges, Fig. 1, the graph on the left).

However, nowadays, users have access to a plethora of applications, such as WhatsApp, Telegram, Instagram or Facebook Messenger. The simultaneous

B. Pinaud and G. Melançon

M. Reif, T. Souvignet, and Q. Rossy

ESC, University of Lausanne E-mail: firstname.lastname@unil.ch

Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800 E-mail: firstname.lastname@u-bordeaux.fr

use of such applications, a phenomenon called "Media Multiplexity" [3], turns the analysis and visualization of communication data into a subtle and complex task for several reasons. One of them is the multiplication of identifiers: as each application uses its own referencing system, traditional graph representations of combined communication data generate multiple one-mode graphs, rather than a single connected graph (Fig. 1). However, in order to reconstruct a user's activity based on phone data, a main objective in criminal investigations, all traces generated by any means of communication should be combined and visualized in a coherent manner. A simple way to link the identifiers between them is to use contact information.

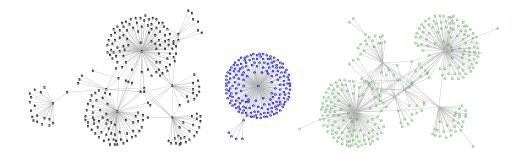


Fig. 1 "Media Multiplexity": visualizing communication data from different applications (from left to right: phone, Facebook and WhatsApp) leads to multiple one-mode networks.

This paper brings forward contributions to help solve this challenge. Communication data as well as contact information is gathered into a unique multivariate graph. Simultaneous links between nodes of different types allow to compute a quotient graph grouping nodes into "metanodes". This multilevel model thus reduces the size of the considered network making it more legible, and favors the emergence of hypotheses during the investigation. Multilevel representations of graphs are displayed as nested graph layouts. The capability to examine established associations by looking into metanodes assists analysts in assessing the relevance of the graph model.

Our approach was developed in a multidisciplinary context using data collected by a Swiss Law Enforcement Agency, anonymized and made freely available to the community [4].

2 Related Work

Visualization and analysis of phone data in criminal investigations has been studied on multiple occasions [5,6]. However, most studies focus on call detail records and data from wiretaps. As far as we know, there is no literature available on the process of visualizing multi-source communications data obtained from phone extractions. This could be due to limited access to actual case data for researchers. Additionally, the concept of media multiplexity in this process has yet to be explored. Regarding graph representation and visualization, our work builds on past work focusing on multilevel network visualization [7,8] making use of features implemented in the TULIP graph visualization framework [9].

3 Use Case

Our starting point is data extracted from several drug trafficking offenders' seized phones [4]. Communication data such as call and messaging logs allow linking communication application identifiers $(id_1 \text{ for } app_1, id_2 \text{ for } app_2, ...)$ used on the seized phones with correspondents' application identifiers. Communications are represented as directed and weighed edges between two nodes, where each node represents an application identifier $(id_i, id_j, ...)$ and the edge weight represents the number of communications. We call this type of data "communication data".

A phone directory gathers contacts of the phone owner. For each contact, it lists the associated application identifiers of this contact (id_u, id_v, \ldots) . We link them based on their co-occurrence under the same contact entry in the phone directory¹. We call this type of data "association data": data that allow the establishment of relationships between nodes, other than communications.

Integrating and visualizing the two resulting data sets (see Fig. 2) considerably reduces the risk of linkage blindness, i.e., the inability to detect links between relevant entities [10]. A recurring question in criminal investigations is the identification of mutual correspondents of two or more known suspects. When dealing with media multiplexity, having association data is required. For example, in cases where suspect A communicates with person C via WhatsApp, and suspect B with the same person C via Facebook, it is crucial to establish a link between the two identifiers (id_1 of C as a WhatsApp user and id_2 of C as a Facebook user).

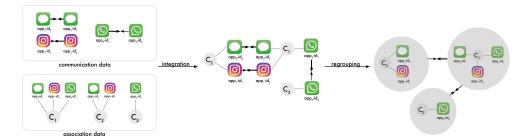


Fig. 2 Communication and association data are integrated into a combined view, then clustered with metanodes to improves its readability.

The resulting graph representation becomes multivariate. Nodes represent phone numbers or application identifiers. Edges can represent communica-

¹ In reality, the situations to be taken into account are numerous. We limit ourselves to this description for the sake of simplicity and because of space constraints.

tions (directed) or associations (not directed). The integration of associations allow to visually link different communications attributed to a same user and thus better understand the activity of a user in its entirety. However, such multivariate graph quickly become dense, cluttered, and difficult to read and analyze.

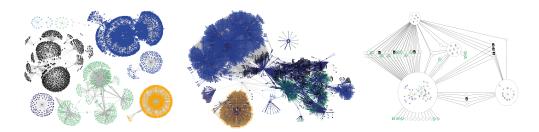


Fig. 3 From left to right: (1) graphs resulting from visualization of communication data, (2) graph integrating communication and association data, (3) simplified graph with metanodes

To improve graph legibility, we propose clustering associated entities, without merging them, using metanodes from the TULIP visual analytics platform [9]. Nodes being linked via association data thus become one single metanode, and communications carried out by identifiers used by the same person are regrouped on the metanode level. This reduces the number of visible nodes on the graph, without losing the underlying information as illustrated in Fig. 3. This has two major advantages for crime analysts. On the one hand, metanodes give analysts the possibility to verify established associations, and modify them if needed. Most criminal analysis visualization tools require users to work with static data structures, without much possibility to dynamically adapt the latter. In order to regroup identifiers used by a single person, a user is therefore compelled to merge the entities in question. On the contrary, TULIP allows for temporary regrouping of entities, without actually merging them. Analysts can thus properly evaluate established associations between identifiers and easily adapt the structure if necessary. On the other hand, original data are kept untouched. A chain of custody is a requirement for anyone working with any kind of trace in law enforcement context. TULIP is not merging entities when creating metanodes. Creating metanodes implies the creation of a graph hierarchy allowing to easily find how each node of the quotient graph was created.

4 Conclusion

An approach to visualize communication and association data for criminal investigations more effectively is promoted through the use of graph transformations and metanodes. Integrating communication with association data, such as contact information extracted from mobile phones, allow for data interpretation on a person-level rather than an application-level. Grouping associated entities improves legibility, helps rendering the visualization more efficient and supports the establishment of hypotheses. The use of metanodes allows analysts to verify established associations and, if necessary, adjust them in light of new clues.

References

- O. Ribaux, O. Delémont, C. Roux, F. Crispino, La science forensique au service de l'action de sécurité (2019), pp. 114–125
- S. Keating, Q. Rossy, P. Esseiva, Criminologie 53(2), 171 (2020). DOI 10.7202/1074192ar
- K.V. Cleemput, Bulletin of Science, Technology & Society 30(2), 75 (2010). DOI 10.1177/0270467610363143
- M. Reif, B. Pinaud, Q. Rossy, G. Melançon. Multivariate phone communication network [data set] (2023). DOI 10.5281/zenodo.7648495
- E. Ferrara, P. De Meo, S. Catanese, G. Fiumara, Expert Systems with Applications 41(13), 5733 (2014)
- D. McAndrew, in *The Social Psychology of Crime*, ed. by L. Alison, D. Canter (Routledge, London, 2000), p. 44
- D. Archambault, T. Munzner, D. Auber, IEEE Transactions on Visualization and Computer Graphics 13(2), 305 (2007)
- C. Rozenblat, G. Melançon, Methods for Multilevel Analysis and Visualisation of Geographical Networks, Methodos Series, vol. 11 (2013)
- D. Auber, D. Archambault, R. Bourqui, M. Delest, J. Dubois, A. Lambert, P. Mary, M. Mathiaut, G. Melançon, B. Pinaud, B. Renoust, J. Vallet, in *Encyclopedia of Social Network Analysis and Mining*, ed. by R. Alhajj, J. Rokne (Springer, New York, NY, 2017), pp. 1–28. DOI 10.1007/978-1-4614-7163-9_315-1
- 10. S.A. Egger, Journal of Police Science & Administration (1984)



Bounded confidence models generate additional clusters when the number of agents is growing

Yérali Gandica · Guillaume Deffuant

1 Introduction

Opinion dynamics models express mathematically some hypotheses about social interactions and provide means to investigate their effect in large populations. For instance, bounded confidence models [1,2] assume that when an agent's opinion is too far from the one of its interlocutor, it has no influence. This hypothesis can explain the emergence of macro-behaviours such as consensus, polarization or plurality of opinion clusters. Many papers are devoted to studying these models and their variants. For instance, several studies focus on including so-called extremists agents, whose opinion is at the border of the opinion interval [3–5], others consider agents with confidences drawn in a given interval[6], with different types of networks of interactions. Introducing noise in these models also significantly modifies their qualitative behaviour [7–9]. For a recent review of these models and related ones, see [10].

While in most of the models, the population of interacting agents is fixed, in this contribution, we consider a growing population of agents. Like in the model studied in [11], new agents are progressively added to the population. This model can be related to online communities of agents that are created, and then may grow more or less rapidly. Recent models inspired by the physics of gels address, more specifically, the dynamics of aggregation and desegregation of online groups [12]. In our model, new agents are added to the population over time while the agents are interacting, which is not the case in [11]. Our intention is to study this model in different network types, and particularly on scale-free networks. However, starting from fully mixed agents seems a necessary first step in order to understand the model and compare its results to the fixed population versions, with or without noise.

Y. Gandica

G. Deffuant

Valencian International University, Spain. E-mail: ygandica@gmail.com

Université Clermont-Auvergne, INRAE, UR LISC, France.

2 The model

We consider a population of a growing number N(t) of agents, with $N(0) = N_0$. An agent $i \in \{1, ..., N(t)\}$ characterised by an opinion $a_i(t) \in [0, 1]$. All the agents share the same confidence bound ϵ . At each time step, we perform the classical bounded confidence interaction as in [1]. Two agents i and j are chosen at random and:

If
$$|a_i(t) - a_j(t)| < \epsilon$$
 then
$$\begin{cases} a_i(t+1) = a_i(t) + \mu(a_j(t) - a_i(t)), \\ a_j(t+1) = a_j(t) + \mu(a_i(t) - a_j(t)), \end{cases}$$
 (1)

where μ is a parameter of the model, fixed to 0.5 in our simulations.

Moreover, with a probability $\omega \frac{N_0}{N(t)}$, ω being a strictly positive parameter, at each time step, a new agent is added to the population, with an opinion chosen at random uniformly in the opinion interval. The idea is that, on average, the same number ωN_0 agents are added to the population every N(t) interactions.

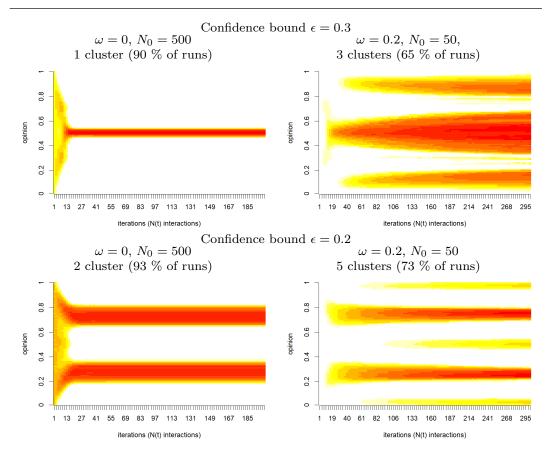
3 Preliminary simulation results

Figure 1 is obtained by running 1000 times the models with the same parameters, for 300 iterations (an interaction being N(t) binary encounters). In each case, we compute the number of opinion clusters. We selected the number of clusters that appears in the majority of the runs, and we superposed the opinions for all the simulations that yielded that number of opinion clusters. Finally, we take the logarithm of the number of opinions located in intervals of 0.01. The left panels of the figure show the results with an initial population of size N0 = 500 and a growth parameter $\omega = 0$, thus corresponding to the standard BC model without population growth. The right panels show the results with N0 = 50 and $\omega = 0.2$, hence an average growth of 10 agents at each round of N(t) pair interactions. In both cases, we observe that, on the left panels, the number of clusters is the one expected from the standard model with a large fixed population [1]. On the right panels, when the population is growing, the model generates additional secondary clusters located on average at a distance a bit higher than epsilon from the primary clusters. Indeed, we observe two secondary clusters for $\epsilon = 0.3$ and three for $\epsilon = 0.2$, taking shape much more slowly than with $\epsilon = 0.3$.

These are preliminary results, and we need to study the transition between the two behaviours of the model in more detail, in the space $(N0, \omega)$.

4 Discussion

At this stage of our study, we can only discuss first hypotheses explaining the observed transition of the cluster number.



French Regional Conference on Complex Systems – Le Havre, May 31 June 2, 2023

Fig. 1 Log density obtained with 1000 replicas of simulations for 300 iterations (N(t)) interactions). On the left, $N_0 = 500$ initial agents the population is fixed. On the right, $N_0 = 50$ initial agents and growing parameter $\omega = 0.2$ (on average 10 more agents at each round of N(t) interactions. The density is the average of all the runs leading to the configuration of clusters which is the most frequent one.

When starting with a relatively large population (like 500 agents as in our examples), for confidence bounds like 0.2 or 0.3, all the agents are connected in a chain of agents which are closer than the bound to each other. Therefore, the clusters tend to attract all the agents located at a distance lower than the bound ϵ from the cluster centre. As a result, the clusters are distant of approximated 2ϵ from each other. Moreover, it is well-known that there is a low probability that some opinions remain between two clusters if they are distant from more than twice the confidence bound. These are the so-called minor clusters, which appear clearly in the version of the model that considers a continuous distribution of opinions instead of discrete values [13].

When there are constantly new agents that appear on the opinion axis, these agents can appear in the regions that correspond to the standard minor clusters, and form new clusters progressively, once the major clusters have appeared. However, these clusters can maintain themselves only if they are much smaller than the primary clusters. Indeed, the regular appearance of new opinions between clusters of similar sizes that are distant of a bit more than the bound ϵ brings them closer and closer and they ultimately merge. If one of the clusters is much more smaller than the other, an opinion between them has a very high chance to be attracted only by the bigger cluster, except when it is out of its reach. Therefore, both clusters can maintain themselves, and their difference in size keeps increasing, because of a process that resembles the "preferential attachment" in networks.

These are only first hypotheses that we should evaluate through more systematic experiments and hopefully improve with a more accurate theoretical understanding.

References

- G. Deffuant, D. Neau, F. Amblard, G. Weisbuch, Advances in Complex Systems 3, 87 (2000)
- R. Hegselmann, U. Krause, Journal of Artificial Societies and Social Simulation 5(3) (2002)
- G. Deffuant, F.A.G. Weisbuch, T. Faure, Journal of Artificial Societies and Social Simulation 5(4) (2002)
- 4. G. Deffuant, Journal of Artificial Societies and Social Simulation 9(6) (2006)
- J.D. Mathias, S. Huet, G. Deffuant, Journal of Artificial Societies and Social Simulation 19(1) (2016)
- 6. H. Schawe, S. Fontaine, L. Hernandez, Physical Review Research 3(2) (2021)
- 7. M. Pineda, R. Toral, E. Hernandez-Garcia, Jounral of Statistical Mechanics: Theory and Experiment (2009)
- M. Pineda, R. Toral, E. Hernadez-Garcia, The European Physical Journal D 62, 109 (2011)
- 9. A. Flache, M. Macy, The Journal of Mathematical Sociology 35(1-3), 146 (2011)
- A. Flache, M. Maes, T. Feliciani, E. Chattoe-Brown, G. Deffuant, S. Huet, J. Lorenz, Journal of Artificial Societies and Social Simulation 20(4) (2017)
- F. Gargiulo, Y. Gandica, Journal of Artificial Societies and Social Simulation 20(3) (2017)
- P.D. Manrique, M. Zheng, Z. Cao, E.M. Restrepo, N.F. Johnson, Physical Review Letters 121(4) (2018)
- E. Ben-Naim, P. Krapivsky, S. Redner, Physica D: Nonlinear Phenomena 183(3-4), 190 (2003)



Monetization in online streaming platforms: an exploration of inequalities in twitch.tv

A. Houssard^{1,2} · F. Pilati³ · M. Tartari³ · P.L. Sacco⁴, · R. Gallotti² ¹CIS, CNRS, 59-61 Rue Pouchet, 75017 Paris, France ²CHuB Lab, Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo, Italy ³IULM University, Via Carlo Bo, 1, 20143 Milan, Italy ⁴DiSFiPEQ, University of Chieti-Pescara, Viale Pindaro 42, 65127 Pescara, Italy

Abstract The live streaming platform Twitch underwent in recent years an impressive growth in terms of viewership and content diversity. The platform has been the object of several studies showcasing how streamers monetize their content via a peculiar system centered around para-sociality and community dynamics but the lack of data regarding streamers revenues left wandering about the effectiveness of the strategies described. Using data from a recent leak we were able to characterize the activity and popularity dynamic and link them to actual revenue. Employing methods from social physics and econometrics, we analyzed audience building and retention dynamics and linked them to observed inequalities. We found a high level of inequality across the platform as well as some . Our results demonstrate that, even if the platform design and affordances favor monetization for smaller creators its non-algorithmic design leaves room for classical choice biases and allows a few streamers to emerge, retain and renew a massive audience.

Keywords Social media \cdot Twitch \cdot Monetization \cdot Inequalities

1 Introduction

Twitch is a live streaming platform allowing creators to broadcast, react and interact with their audience via a chat system. The platform was initially

A. Houssard

59-61 Rue Pouchet, 75017 Paris

E-mail: antoine.houssard@cnrs.fr

centered around the gaming subculture, but it consistently gained popularity (from 1.26 million average concurrent viewers in 2019 to 2.78 million in 2021 [4]) and largely extend the types of content offered. The platform has built itself around technical features allowing and encouraging great social interactivity features and connectivity affordances in both its consumption and monetization systems [7] [5]. Viewers find in Twitch streams a "third place" [7] where they can feel part of a community and engage with other members and streamers can leverage those feeling of attachment to nudge viewers toward the paying features (such as subscriptions) of their stream. Consequently, the platforms relies on follow, co-opting and classification rather than on heavy algorithmic recommendation systems. In fact such system would lead to incidental consumption and rapid switch (like in other platforms such as Youtube, Tiktok etc...) opposing those monetization systems. These dynamics have largely been characterized by authors such Johnson & Woodcock [12] [11] showing how streamers work relates mostly to parasociality and community formation ("capital communautaire [5]") but, despite their relevance, those work weren't able to gage the macro impact of such a system on inequalities.

2 Data and methods

The first set of data collected for our study derives from a leak of Twitch data which happened in 2021. Even if those leaks contained multiple types of data, in this paper we only consider the information relating to the streamers' revenue (total amount from 2019 to 2021). Those data have been enriched through multiple sources (posing significant challenges and leading to some discrepancies in the samples considered) characterizing the current visibility of streamers on the platforms [3][4], time series of their popularity [2] and activities [1]. In order to determine inequality levels of the variables of interest in the platform, we calculated the Gini coefficient for three fixed metrics (revenue, followers, subscribers) and fitted a power law and parabolic fractal to extend our curve. Moreover, in order to characterize the dynamics of audience pools we make use of classical methods used in the study of financial time series, such as those in [6] and [9]. To confirm our results, we modeled our data using the ranking dynamic method proposed by Iñiguez and colleagues [8]. Finally we compared the diversity in activity as a function of the ranking, using the number of games and variety of genres played.

3 Results

First, we observe that the revenue distribution in the Top 10000 streamers is highly unequal (Gini 0.57, Fig) and the prolongation of the curve using a parabolic fractal indicates that, if we consider the entirety of twitch users (having stream during the year), those inequalities would be even greater (Gini 0.94, 1A;B). Looking at the subscribers ranking for a sample of a 1000 streamers (February 2022 [4]) and its extension, we see comparable inequalities (Gini

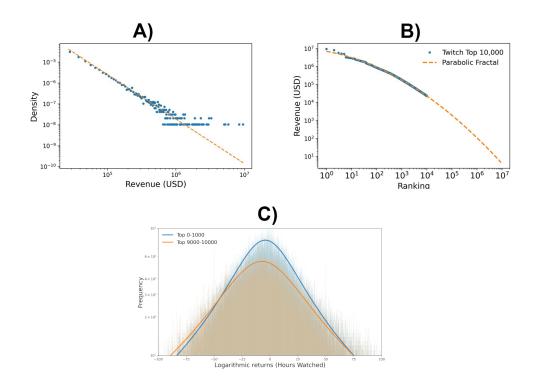


Fig. 1 A) The Revenue distribution August 2019 and October 2021 (in US\$). The distribution function appears long tailed and can be fit with a power law with exponent -2.13. The Gini index associated with this distribution is ≈ 0.57 . B) Rank plot of Streamers revenue (in US\$). The curve is fit with a parabolic fractal function (orange dashed line, $y = \log(x^{-b}) - c\log(x)^2 + norm$, with parameters, $b \approx 0.259$, $c \approx 0.0392$, $norm \approx 15.78$). Using the fit we could extend our estimate for the Gini index to all 9.2 million streamers, for which we find the considerably larger value of ≈ 0.93 . C) Histogram of the Logarithmic returns of the hours watched and the fitted Cauchy curve for the Top 0-1000 (Location ≈ 3.15 , Scale ≈ 29.4) and Top 9000-10000 streamers (Location ≈ -7.88 , Scale ≈ 39.61). These fits show that top streamers manage to keep a more stable viewership through time whereas the low end of the distribution has a more volatile visibility.

0.63 for the top 10,000) showing that the gap is in fact caused by the mobilization of the platform specific monetization systems and not only by differences in advertisement revenues. Moreover, inequalities are further increased by the differences in revenue sources (observed through the collection of URL present in streamers description). Secondly, following Wolff & Shen [10], we took interest in how well streamers capitalize on their audience and found that streamers having the biggest audiences [3] also have the worst conversion rate from followers to subscribers and drawn the least money from their current subscriber (confirmed by the sublinear relation between revenue and subscriber ; $\alpha \approx$ 0.8), which in principle should reduce inequality. This contradiction appears to find a resolution in the observed stability of top streamers. Using time series provided by StreamCharts, we compared the variation (using analogue methods to those used in the study of financial data [9]) in the volume of consumed content and viewers retention and, by fitting a Cauchy curve (1.C) and computing the auto-correlation for different lags, those methods showcased the ability of top streamers (top 500) to renew and retain their audience. This idea is confirmed by the fitted parameter found using Iñiguez and colleague model and, could be explained, by the greater diversity of content provided by top streamers (Spearman ≈ 0.76 between number of games played in the period and streamer rank).

4 Discussion

Despite some limitations, both ethical/legal and regarding the nature of the data, this study has allowed for a better understanding of the mechanism underlying the content monetization on Twitch.tv and, to a larger extent, question the relevance of algorithmic free platforms in reducing inequalities.

Acknowledgements This research project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870792. We thank Gerardo Iñiguez and the other authors of the Ranking Dynamics article47 for making their code available and for their guidance in using it. RG wants to thank Marta Arisi and Francisco Duque Lima for useful discussions.

References

- 1. Internet video game database. https://www.igdb.com/. Accessed: 2022-08-24.
- $2. Streamscharts. {\tt https://streamscharts.com/overview}. Accessed: 2022-08-24.$
- 3. Twitch tv api. https://dev.twitch.tv/docs/api/. Accessed: 2022-08-24.
- 4. Twitchtracker. https://twitchtracker.com/. Accessed: 2022-08-24.
- 5. Cocq, M. Constitution et exploitation du capital communautaire: Le travail des streamers sur la plateforme Twitch. *nrt*, 13 (July 2018).
- GOPIKRISHNAN, P., PLEROU, V., LIU, Y., AMARAL, L., GABAIX, X., AND STANLEY, H. Scaling and correlation in financial time series. *Physica A: Statistical Mechanics and its Applications 287*, 3-4 (Dec. 2000), 362–373.
- HAMILTON, W. A., GARRETSON, O., AND KERNE, A. Streaming on twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto Ontario Canada, Apr. 2014), ACM, pp. 1315–1324.
- IÑIGUEZ, G., PINEDA, C., GERSHENSON, C., AND BARABÁSI, A.-L. Dynamics of ranking. Nat Commun 13, 1 (Dec. 2022), 1646.
- STANLEY, M. H. R., AMARAL, L. A. N., BULDYREV, S. V., HAVLIN, S., LESCHHORN, H., MAASS, P., SALINGER, M. A., AND STANLEY, H. E. Scaling behaviour in the growth of companies. *Nature* 379, 6568 (Feb. 1996), 804–806.
- WOLFF, G. H., AND SHEN, C. Audience size, moderator activity, gender, and content diversity: Exploring user participation and financial commitment on twitch. tv. new media & society (2022), 14614448211069996.
- 11. WOODCOCK, J., AND JOHNSON, M. R. The affective labor and performance of live streaming on twitch. tv. *Television & New Media 20*, 8 (2019), 813–823.
- WOODCOCK, J., AND JOHNSON, M. R. Live Streamers on Twitch.tv as Social Media Influencers: Chances and Challenges for Strategic Communication. *International Journal* of Strategic Communication 13, 4 (Aug. 2019), 321–335.



Political Participation and Voluntary Associations A Hypergraph Case Study

Amina Azaiez · Robin Salot

Abstract Civic organizations, ranging from interest groups to voluntary associations, constantly influence policy formation in representative democracies. This work presents a local case study that examines the relationship between voluntary associations and local political institutions in a city with almost two thousand residents. Traditionally, sociologists approaches focus on individual characteristics such as age, gender, or socio-professional statues. Here, we model social interactions between members of organizations through a hypergraph and explain political involvement. Specifically, we model interactions as hyperedges that correspond to activities proposed by organizations and involve the individuals who participate in those activities. Our analysis reveals a community-based structure, in which members of similar type of organization tend to interact more frequently. To quantify 'political participation', we introduce an interactional-based measure that extends the degree centrality. We also introduce the 'diversity coefficient' as an extension of the degree centrality to capture individual's ability to participate in activities composed of members from different communities. Among other centrality measures, we find that the diversity coefficient is the most significant factor in explaining political participation among members of associations.

Keywords Political Participation \cdot Voluntary Association \cdot Social Networks \cdot Hypergraphs

1 Introduction

The interdependence between social participation and political involvement has long been studied in social and political sciences. One of the earliest and most

Robin Salot

Amina Azaiez

Université Paris 1 Panthéon-Sorbonne, Centre d'économie de la Sorbonne, Maison des sciences économiques, 106-112 Bd de l'Hôpital, 75647 Paris Cedex 13, France E-mail: amina.azaiez@univ-paris1.fr

Laboratoire d'Anthropologie Sociale, Ecole des Hautes Etudes en Sciences Sociales, Collège de France, 52 rue du Cardinal Lemoine, 75005 PARIS, France E-mail: robin.salot@ehess.fr

well-known works on the subject is Tocqueville's Democracy in America, which emphasizes the importance of social participation and intermediary association in democracy. Some organizations such as 'intermediary organizations' or 'interest groups' (e.g. : trade unions or business organization) are explicitly linked to politics since they take part in institutionalized decision-making procedures and their function is to influence those decisions in favour of the group they represent [1-3]. On the other hand, other kinds of voluntary association, such as educational associations or art and sports clubs, do not have an explicit political function. However, it has been argued that even these 'apolitical' associations spur political participation [4,5]. For example, they give their members the opportunity to develop skills and build networks that could benefit them in future political career. These associations can also influence political leaders at the city level, as they may be in frequent contact with city council members, for example when organizing social events or setting up social services such as childcare or assistance to the elderly. Moreover, politicians can view these associations as potential channels to contact and convince their members. Hence, they can be considered 'parapolitical'[4]. Historically, scientists have sought to explain the political commitment of citizens by appealing to personal characteristics such as age, gender, social status, membership in associations [6,7]. Scholars in socio-psychology argue that the social context and micro-interpersonal interactions can influence and change individuals' opinions [8]. In this context, data are usually gathered using name generator surveys, where individuals are asked with whom they talk 'important matter'. This method produces several egocentric networks sampled independently, and the social interactions between ego and alters are considered as potential channels for social influence of alter on ego [9–11]. However, this method has limitations. The small size of each egocentric network and the uniqueness of the network's shape (star network) prevent macro-scale network analysis. Networks theory provides tools for the analysis of both the overall structure of the network and individualistic characteristics of individuals, such as centrality measures [12, 13]. Some limitation of this method can be overcome by studying larger networks and allowing for complex social interaction as is the case with hypergraphs¹. Furthermore, this literature has to a large extent focused on the influential processes at play in political opinion formation. This constrains the study of "political participation" to voting patterns and participation rates in national elections. However, social participation and political participation can be seen as a two-way flow, since the origin of the word "politics" encompasses all activities that serve to organize social groups and is not limited to the act of voting [5, p.11] [14].

In this paper, we study social interactions in a village of approximately two thousand residents to reveal the interdependency between social and political participation. We focus our study on members of voluntary non-profitable associations and elected officials through the interactions occurring within and between these organizations. We model these interactions through a hypergraph where the hyperedges correspond to activities proposed by the organizations and involve the individuals who participate in those activities. At the macro-scale, these data can

¹ A Hypergraph $\mathcal{H} = (V, E)$ is a set V of vertices and a set E of subset of V called hyperedges. Each hyperedge can connect any number of vertices. In contrast, an edge of an ordinary graph connects exactly two vertices.

be seen as a snapshot of the underlying social fabric of a village. At the micro-scale, these interactions can serve to understand complex social behaviour that permits access to local politics.

Through our study, we find that the network is organized into distinct communities and that the structure of these communities overlies the structure of organizations within the village. To better understand the interplay between social and political participation, we introduce two extensions of the degree centrality: one measures the "political participation" of agents, the other, called "diversity coefficient", captures the ability to participate in heterogeneous activities gathering individuals from different communities. We compare the correlation between the political participation and various centralities: the diversifying coefficient, and others from the literature. Our findings suggest that, for members of associations, engaging in heterogeneous activities is the most effective way to gain access to politics.

2 Related work and Hypothesis

Although there has been long-standing theoretical support for linking micro-level interactions to macro-level patterns [15], and despite the significant development of network theory in recent years, relative to other domains of research, political sciences have shown little interest in studying networks to explain political phenomena [6, 7, 16, 17]. Most previous studies investigating the link between networks and political behaviour have focused on egocentric networks of survey respondents, examining how social interactions with "alters" may influence the political opinions of the "ego". Researchers have examined a variety of influential processes in this context, including frequency of political discussion, level of political knowledge of alters, homogeneity of political preferences of alters [9], social diversity of the network [11]. These surveys had the advantage of being addressed to a large number of respondents with diverse social capital (eg. socio-economic status, ethnicity, age etc.) and making them more representative and generalizable at the scale of the concerned country. However, this method has many theoretical drawbacks. First, the ego-networks typically consist of only a few friends or acquaintances, usually between three and six. Knoke [9] admits that 'investigating these generic processes requires both network data observed over extended time and measures collected directly from all network participants, not just ego's perceptions of alters' opinions and behaviours', but due to time limitations, surveys only ask respondents to report information about a few of their intimates and sometimes, even the ties between alters were not reported, without mentioning the absence of second degree neighbours, i.e., friends of friends. This results in a star-shaped network where ego is at the center and all other alters are directly connected to ego, but not to each other. This creates an implicit assumption of a one-directional causal relationship, where alters influence ego but not vice versa.

A second important limitation of this method is the restriction of ego-network to core ties when asking respondents with whom they talk about 'important matter'. Indeed, scholars have significantly argued for the importance of weak ties, [15, 18,19] especially for the diffusion of information and access to new knowledge, or job opportunities. Granovetter ([15]) theorized this concept by schematizing communities as social groups where members are connected by strong ties and with a high density of ties within the community. He argued that the only way of connecting these communities is through weak ties, since the probability of having a close friend (with whom I share a strong tie) that does not belong to my core community is very low. Therefore, only weak ties can be bridging ties and are able to connect disjoint communities. The importance of these bridging weak ties for dynamical diffusion processes within social groups follows strikingly. In this paper, Granovetter specify the location of weak ties in voluntary associations and workplaces and argued the importance of bridging ties between community organizations in mobilizing resources to face an external threat.

From then on, various empirical studies have addressed the importance of weak ties. For instance, Eveland and Kleinman [20] studied 25 formal student activity groups and compare their general and political discussion networks by asking each member their relations with all the other members of the group. Among other results, they found that 'the ratio of political to general discussion frequency is about five to one. Although we do not have comparable data for strong tie networks, it is unlikely that 1/5 of interactions among spouses or best friends are political, which suggests a non trivial amount of political interaction in these weak ties networks when viewed as a proportion of all interactions '. This encourages the choice of voluntary association as social foci for the study of political processes. However, this comparative analysis of 25 disconnected, and assumed independent networks, does not trigger 'bridging ties'. Other studies concerning political involvement and social networks confuse 'weak' and 'bridging' ties [11]. Granovetter argued that bridging ties are weak ties whilst the converse is not necessarily true. Indeed, to locate bridging ties, the whole structure of the network is needed, while the strength of a tie between two individuals is a very local property of the network.

Nevertheless, an interesting comparative study in six urban neighbourhoods and their voluntary associations conducted by Crenson [21] suggests that bridging ties and individuals involved at the end points of these ties are of special interest. He defines residents with 'close-knit' neighbourhood friendship networks as individuals whom "most of their friends lived in the neighbourhood and, second, that most of their friends knew one another" and 'loose-knit' residents, individuals whom "reported that most of their friends lived inside the neighbourhood but that most of their friends were not acquainted with one another". One may argue that 'loose-knit' residents are bridges that connect their friends, since the latter do not know each other. He found that not only loose-knit residents were more aware of community associations in their neighbourhood and had a high rate of participation in these associations, but also that voluntary associations located in the neighbourhood with dominant loose-knit residents were better able to fit the expectations of their members. Thus, without engaging heavy mathematical tools from networks theory, this result suggests that first the overall structure of the neighbourhood network may impact the functioning of voluntary associations, and second, that in a community shaped network, bridging individuals occupy a central place in the network.

Thus we have the intuition that in networks governed by 'structural holes' [18], the people who link distinct social groups play an important role, but the overall network must be community shaped. Indeed, this is pointed out by Siegel [22], who proposed a theoretical dynamical model for political participation and highlights the role of the structure of the network. In this model, agents in a network are assigned two sources of motivation to participate : an 'internal motivation' that

is independently distributed over individuals, and an 'external motivation' that depends on the rate of participation of agent's neighbours in the network at a given iteration. He compared the dynamics of this model according to 4 different networks (Small-World, village (clique), opinion leader, hierarchical) with different distribution of the internal motivation. He found that, in village networks composed of cliques tied together by weak ties, increasing the number of weak ties was of major importance to spur the mean political participation rate of all agents, while this is not necessarily the case for the other networks, where adding weak ties may even lower political participation depending on the distribution of internal motivations.

In this paper, we study social interactions between members of voluntary associations and political institutions in a village of two thousand residents by interviewing members of these organizations. We model social interactions through a hypergraph where each hyperedge corresponds to an activity proposed by an organization and includes the people who participate in that activity. As explained in section 3.1, this rural municipality is more likely to be considered as a village than an urban city. Hence, we expect that the network to be dense in a village where 'everybody knows everybody else'. However, Given that our data concerns specific organizations and that individuals likely interact more frequently with members of their own organization, we expect to observe a community structure within the hypergraph. Hence we propose the following hypotheses :

H1: The hypergraph has a community structure based on the structure of the organizations.

H2: People located at the interface of these communities have more access to local politics.

The organization of the paper goes as follows. Section 3 presents the sociological and political context of the village where our study was conducted, as well as the method of data collection and the hypergraph model used to analyse social interactions. In section 4.1 we describe the hypergraph in detail and analyse its overall structure. In section 4.2, we review and introduce various centrality measures. We define two extensions of the degree centrality: the *political participation*, which quantifies individuals' access to local politics, and the *diversity coefficient* which assesses individuals' ability to participate in heterogeneous activities. We review eigenvector centralities for hypergraphs. In section 7.3, we compare the correlations between the political participation and these centralities for members of associations and political institutions separately. In section 5 we combine all the results from the previous sections and discuss them qualitatively. In section 6, we conclude, review governance modes' of association better able to involve members in the political arena, argue for the use of complex structures such hypergraphs rather than graphs to model social interactions, and encourage analytical research in this direction. Further technical analysis of the comparison between the hypergraph and its projection is detailed in section 7.

3 Data and Method

3.1 Context

Since our study was conducted in a particular social and geographical context, we felt it is important to specify this context so that other studies with similar or contradictory results could draw conclusions about the generalizability of these results or the sources of incivilities.

The village of the survey is located in Seine et Marne, France. Its centre, built at the gateway to a valley, contrasts with the scattering of its hamlets established on the agricultural plateaus. The population of this small village doubled between 1968 and 2022, from 1000 to 2000 inhabitants. After the retirees, who represent 25.4% of the population, service sector employees (20.2%) and workers (15.4%) are largely represented. Caught up in the daily migration of tertiary work, which is characteristic of rural areas, these employees largely abandon associative places. As is often the case in rural areas, these associations, like the municipality, are mostly run by retired people. There is a first striking structural similarity between these two institutions: each association is made up of a general assembly of its members, which elect a board of directors, which in turn elects a bureau consisting of a president, a treasurer and a secretary. In the same way, in each municipality, a municipal council is elected by direct universal suffrage by the inhabitants of the municipality, which in turn elects a mayor and his deputies. In addition to this organizational resemblance, associations and the town hall are in frequently in contact : solicitation of associations by the town hall during national events (national day, heritage festival, etc.), request for funding from the town hall, invitation of elected officials to the general meetings of associations, or participation of elected officials in associative activities. The village under consideration is part of a community of municipalities², which is managed by a community council made up of the elected representatives of the municipalities. The inter-municipality gathers 31 communes and 27000 inhabitants, which shows the relatively important size of the considered village. This political institution is mainly concerned with economic development and land use planning, but it is in close contact with several associations, notably through the cultural, social, sports, environmental and early childhood commissions.

3.2 Data collection

In order to collect information about the interactions between citizens, we used mainly two methods of data collection: interviews and official documents. We selected 10 voluntary associations in the village, two from each of the following categories : art, educational, environmental protection, sport, social ³. We inter-

 $^{^2}$ A community of municipalities is an EPCI grouping together several municipalities that are all adjoining and without enclaves. Its purpose is to bring together municipalities within a solidarity area so as to develop a joint urban development project for the area (Insee).

 $^{^3}$ Two of the selected associations were not based in the village, but in surrounding villages. However, they often offer activities in the village and are in contact with the other associations

viewed 3 additional elected officials of the city hall, the mayor and two deputies⁴. We end up with 23 interviews. To avoid redundancy, from now on, we will use the term *association* to name voluntary associations and the word *organization* refers to voluntary associations and political institutions.

During the interviews, members of associations were asked three topics concerning the association(s) 5 to which they belong:

- The activities proposed by the association(s) he or she belongs to, during the last year and the members who attend these activities.
- The collaboration with other local associations.
- The interactions between the association and political institutions such as the city council and the community of municipalities.

Members of the city council were asked about the composition of the committees (urbanism, culture, human service...) and the frequency of meetings. They were also asked about their contact with the community of municipalities. With regard to political institutions, we widely use public reports of meetings where the attending persons were marked. Annual reports of associations provide clarity and were usually used as a memory basis for the interviews.

These interviews and documents provide 544 different activities engaging 618 individuals. As our focus is on associations and politics, we restrict the set of individuals to those whose membership to an association or a political institution is known. This reduces the number of individuals to 474 and the number of interactions that involve at least two individuals to 429. Table 1 resumes the composition of the sample. Association are categorized according to their principal activity. We distinguish 8 categories of associations with corresponding examples from our case study: Human services (e.g. :social work, childcare), sports clubs (e.g. : running, canoe kayaking), art (e.g. : Theatre, music, sewing), educational (e.g. : association for the conservation and enhancement of cultural heritage), environmental protection (e.g. : recency of species, litter pickups, fishing, vegetable gardening), recreational (e.g. : organizing festive events), Occupational (e.g. : inter-communal union for the management of schools, water management unions)⁶. Political members correspond to elected officials (city council, community of municipalities, the department...) or individuals from public institutions of an administrative nature. Lastly, we decide to keep in our study communal maintenance agents and the secretaries of the town hall because they participate very frequently, namely daily, to the organization tasks of the city, as well as meetings and decision-making procedures in city hall. The professional category refers to these individuals. This table shows the number of memberships per category. Note that, since one person can hold multiple memberships, the total number of memberships is larger than the total number of individuals.

 $^{^4\,}$ Among the twenty interviewed members of associations, two are also member of the city council. Hence, we interviewed five elected officials

 $^{^5}$ These respondents are members of 24 distinct associations and political institutions, in contrast with the 11 targeted ones, resulting in a total of 47 memberships. This means that some of them have more than one membership. In fact, the mean number of memberships is 2.1

 $^{^{6}}$ This last type of association is not the focus of our study, and for this reason, only 4 memberships are included, which result from collaborations with political institutions.

	Women(%)	$\operatorname{Men}(\%)$	Total Number
Art	57.4	42.6	54
Environmental-Protection	27.0	73.0	63
Human Service	68.2	31.8	66
Educational	56.1	43.9	66
Political	38.3	61.7	227
Professional	22.2	77.8	9
Sport	42.6	57.4	68
Occupational	50.0	50.0	4
Recreational	45.5	54.5	11
Total Number	255	313	568

French Regional Conference on Complex Systems – Le Havre, May 31 June 2, 2023

Table 1: Characteristics of memberships.

3.3 Creation of the Hypergraph

To model these interactions, a natural representation would be a hypergraph rather than a graph since the activities under consideration possibly involve more than two individuals. Not only is this choice made up for convenience, but also because hypergraphs have been gaining recent traction and a recent studies reveal the interest in preserving the higher order interactions in various structural analyses such as centrality measurement or community detection algorithms [16]. We will see an illustration of this assessment later (section 7) when comparing results obtained from the clique expansion of the hypergraph and the original one.

The hypergraph is built simply as follows: Each hyperedge corresponds to an activity, where the nodes are the individuals who participate in it. The frequency of the activity gives a weight to the hyperlink: (Daily: 200), (Weekly: 45), (Monthly: 12), (Quarterly: 4), (Annually: 1). We denote the resulting hypergraph $\mathcal{H} = (V, E, w)$ where V is the set of vertices E is the set hyperedges and w(e) the weight of the hyperedge e. From now, for simplicity, we will use the word edge instead of hyperedge. To clarify this point, here is an example of edge building: "Alice, Tom and Peter go to the canoe club every week" gives rise to the following edge: {Alice, Tom, Peter } with weight 45.

4 Data Analysis

4.1 Analysis of the Hypergraph Structure

In this section, we present some general features of the hypergraph \mathcal{H} to better understand the shape of its architecture. Our intuition is that \mathcal{H} has a community structure that matches the organizations at play. To assess this hypothesis we start by examining the distributions of some of its features in Table 2 and introduce their corresponding notations. The cumulative distribution functions of these features are displayed in figure 1 in log-log scale: edge cardinalities (1b), edge weights(1a), node degrees (1d) and node strengths (1c). There is no real analysis to be done concerning edge weights, the cumulative distribution function is shown mainly to have an idea about the frequency of the activities recorded in this study. We can still point the total weight of the edges 4391 which reveals the total number of

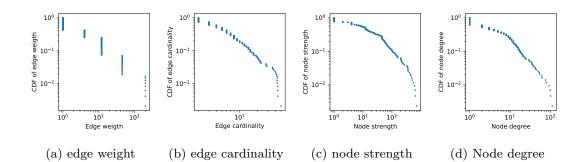


Fig. 1: Cumulative distribution functions of features of \mathcal{H} in log-log scale.

interactions stored for this study. The cardinality of an edge corresponds to its size. As we can see, the mean cardinality of 7.22 is significantly higher than the cardinality of an edge in a standard graph, which is 2. This observation already highlights the considerable number of relatively large group interactions and already justifies the use of hypergraphs.

In our case, the node degree corresponds to the variety of activities in which an individual is involved, while the node strength indicates the intensity of participation. First, the standard deviation of both features is significantly higher than the mean value, and the cumulative distribution functions suggest a positively skewed probability density with the existence of heavy tails. This shows that some individuals are very active while others appear to be peripheral. This unequal redistribution of strength among individuals can be explained in accordance with our hypothesis of community structure in two ways: there are communities that are much more active than others, or communities are made up of members with different positions, from the core to the periphery.

Since we are interested in revealing cohesive groups that are densely intra-connected

Feature	Notation	Mean	Std	Total
Number of nodes	n	-	-	429
Number of edges	g	-	-	474
Node degree	k_i	7.98	13.87	3423
Node strength	s_i	45.67	97.93	19592
Edge cardinality	d_e	7.22	7.98	3423
Edge weigth	w_e	9.26	27.01	4391

Table 2: General features of \mathcal{H}

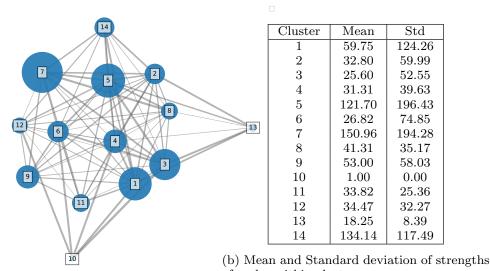
and loosely inter-connected, we perform a community detection algorithm based on the maximization of a quality function⁷. Barber ([24]) extended the modularity function initially proposed by Newman and Givran [25] to the case of bipartite graphs (see section ?? and ?? for a quick review on the modularity function and the clustering algorithm). Since the empirical literature lacks in evidence for the efficiency of multimodal networks, rather than their projection, we perform a comparative analysis between the partitions obtained using the hypergraph and those using its clique expansion in section 7.3. We find that hypergraph clustering

⁷ For a review on community detection algorithms, see [23]

provides a larger number of clusters and that these clusters are more balanced in size. We argue that preserving the complex structure of connections in hypergraph clustering better addresses the resolution problem associated with the modularity-based algorithm, but this assumption requires further analytical study.

The partition $\mathcal{C} = (C_1, \ldots, C_{14})$ obtained using the hypergraph clustering produces a good modularity equal to 0.790. Not only is this result a signature of a good partitioning of the clustering algorithm, but also reveals the community structure on the basis of the hypergraph. Note that the clustering algorithm produces mixed clusters containing both nodes and edges. However, for the purpose of our study, we focus on the partition of the nodes.

In figure 2a, each cluster C is represented by a node, and its volume $Vol(C) = \sum_{i \in C} s_i$ is proportional to its size. The intensity of communication between two clusters corresponds to the width of the edge connecting them. As we can see, clusters differ in sizes and in the way they interact. The mean volume is $\overline{Vol(C)} = 1399.42$ and the standard deviation of the volumes is std(Vol(C)) = 1104.18. This heterogeneity still holds inside clusters (see table 2b).



of nodes within clusters

(a) Community Graph

Fig. 2: (a) Size of clusters and the intensity of interactions between them. Each node represent a cluster. Node's size is proportional to cluster's volume. The width of an edges is proportional to intensity of interaction between clusters $\sum_{e:e\cap(C\cap C')\neq\emptyset} \frac{|e\cap(C\cup C')|}{d_e} w_e$. (b) Mean and Standard deviation of strengths within clusters.

To check whether the community structure is consistent with memberships of agents, we measure the cosine similarity for pairs of agents. For each individual i, we define a vector X^i such that $X^i_{\alpha} = 1$ if agent *i* is member of the organization α , 0 otherwise. The cosine similarity between nodes *i* and *j* is simply the normalized scalar product of X_i and X_j . Figure 3a shows the histograms of the *intra* and *inter* similarity. The intra-similarity is the similarity between a pair of agents in the same cluster and is represented in orange, whereas the inter-similarity, in blue, corresponds to the similarity between a pair of agents belonging to different clusters. The first observation is that most of the pairs of nodes in disjoint clusters have a vanishing similarity which indicates that they do not share any common membership. Still, there are a few pairs of agents in disjoint clusters with a non-zero similarity. This can be explained by the presence of agents with more than one membership. Those agents might have an unequal rate of participation in the organizations to which they are affiliated, thus they are probably put in clusters with members of the organization with the highest rate of participation, but they still hold common memberships with the others. Hence, this indicates that members of a given organization tend to belong to the same cluster.

On the other hand, looking at the intra-similarity density reveals a majority of pairs of nodes from the same cluster with a high similarity. Thus, individuals with common memberships are indeed in the same cluster. However, there is a non negligible proportion of pairs belonging to the same cluster without sharing any common membership. Hence, the community structure of \mathcal{H} seems to be explained by the organization memberships in a finer resolution, but these micro-communities merge to form bigger ones ⁸.

Now, one may wonder on what bases organizations tend to interact more or less intensively with others. Our guess is that organizations of the same category may share more common activities, as they have common interests. To answer this hypothesis we compute the inter and intra similarity, but this time accounting for categories of organizations. Namely, we define the vector Y^i for agent i with $Y^i_{\beta} = 1$ if agent i belongs to at least one organization of category β , 0 otherwise. Since this condition is less restrictive, we expect both density curves to reflect a higher probability for large similarity values, but this behaviour would not help us prove our hypothesis. Hence, we compute the same density curve of similarities with random assignment of associations' categories and compare them with the empirical ones. The random categorization of organizations is done while preserving the same number of organizations per category and memberships. The resulting curves are displayed in figure 3b. The histograms in blue correspond to the results of the random categorization with 200 runs, while the red ones represent the empirical case. We remark that in both cases, the red and the blue histograms have the same global shape: the inter-similarity is centred near 1 and the intra-similarity is centred near 0 and 1 with a gap a between the two extremes. This common form is explained by the fact that the composition of the organizations has been maintained during the random process 9 . We also remark that the red and the

 $^{^8}$ To compare the results obtained with the clique expansion clustering, Figure 6 shows the inter and intra-similarity in that case. In sec 7.4 we show that the hypergraph clustering permits a better recovery the organisations.

⁹ Let nodes *i* and *j* belonging to only one organization α and then have a similarity of 1. Changing the category of association α from β to β' do not change the similarity between *i* and *j* since their membership is unchanged.

blue histograms display differences. Concerning the inter-similarity, the red histogram is more peaked around 0 than the blue one. This indicates that a majority of individuals belonging to different groups not only do not share a common membership, but are not members of similar category organization either. Concerning the intra-similarity, the red histogram shows a higher frequency for a similarity of 1 than the blue one. This shows that individuals belonging to identical groups are, in majority, members of a similar category organization, which remains true even for people who are not members of the same association." This observation tends to affirm our hypothesis.

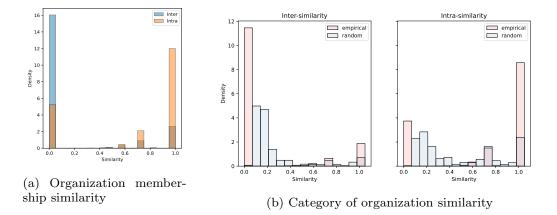


Fig. 3: **Distribution of inter and intra similarity**. Figure 3a Organization membership similarity: the intra-similarity is the similarity between a pair of agents in the same cluster and its distribution is represented in orange. The inter-similarity, in blue, corresponds to the similarity between a pair of agents belonging to different clusters. Figure 3b Category of organization similarity: the left panel correspond to the inter-similarity and the right panel correspond to the intra-similarity. Density histograms of similarities with random assignment of associations' categories are in blue, and the empirical ones are in red. The random categorization of organizations is done while preserving the number of organizations per category (200 runs).

Finally, to have a clear picture of the composition of the clusters, Figure 4 displays the cumulative categories of organizations to which agents belong, per clusters. Each bar represents the cumulative memberships of agents in a given cluster, grouped by category, with each category represented by a different color as indicated in the legend. The dominance of a particular color per cluster in most cases highlights the associativity of agents according to the categories of organizations they belong to.

We will not delve further into the analysis of the structure of the hypergraph. In summary, we have shown that the architecture of \mathcal{H} is based on a community structure where the communities consist of agents from the same organization at the micro-level, and from the same category of organizations at a larger scale. These communities are unequally distributed in sizes and this heterogeneity is

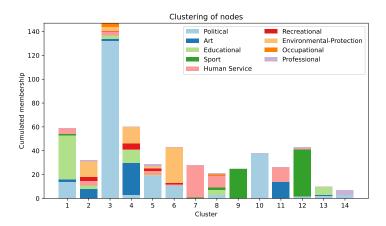


Fig. 4: Cluster Composition Each bar corresponds to each cluster indicated in the x-axis. Each unit rectangle in a bar represents one membership of an agent. Individuals with memberships in n different categories contribute n unit rectangles. Each unit area is coloured according to the category of the organization of the membership.

also present within the clusters. Thus, communities are composed of a core of very active individuals and others located on the periphery.

4.2 Centralities

Now that we have a general view of the interactions between members of associations and political institutions, in the following sections, we go further to answer the following question : what types of behaviours facilitate individuals' engagement with local politics? To this end, we introduce and recall multiple centrality measures. Note that one can normalize all the centrality vectors by setting their l_2 -norm to one. For our purpose, this is not necessary because we are interested in computing correlations in the next section.

4.2.1 Edge dependent degree centralities

Strength: First, we recall the strength centrality

$$s_i = \sum_{e:i \in e} w_e \tag{1}$$

This measure reveals the intensity of an individual's participation in all types of activities offered by associations and political institutions.

Political Participation: We also need to define a measure to quantify access to local politics. We propose the following one, which we call *political participation* and note p_i for agent *i*. First we define a set of agents that we call the *body politic* and note it \mathcal{P} . An agent belongs to \mathcal{P} if he or she is an elected official member of a political

institution. Let e an edge corresponding to an activity proposed by an organization, we make the following hypothesis: if the proportion of individuals from the body politics in e is large, then it is more probable that the activity represented by e is a 'political activity', i.e., contributes to making political decisions. On the other hand, if this proportion is very low, then it is more probable that the activity is 'apolitical'. This lead to the following expression :

$$p_i = \sum_{e:i \in e} \frac{|e \cap \mathcal{P}|}{d_e} w_e \tag{2}$$

where $|e \cap \mathcal{P}|$ is the number of agents in e that are from the political body. This measure can be seen as an extension of the usual degree centrality becoming an edge dependent degree centrality.

Diversity Coefficient: we introduce one further edge dependent degree centrality. This measure captures the ability of agents to participate in 'heterogeneous' activities involving individuals from different communities. The heterogeneity of an activity represented by an edge e is captured by the Shannon entropy :

$$h(e) = -\sum_{C_j \in \mathcal{C}} \frac{|e \cap C_j|}{d_e} \log\left(\frac{|e \cap C_j|}{d_e}\right)$$

 $\frac{|e \cap C_j|}{d_e}$ is the proportion of individuals in e that are in cluster C_j . A high entropy corresponds to an activity e that gathers individuals from various organizations. For a given agent i, the corresponding edge dependent degree centrality is denoted o_i and has the following expression

$$o_i = \sum_{e:i \in e} h(e)w_e \tag{3}$$

4.2.2 Eigenvector centralities

We recall centrality measures proposed by Tudisco and Higham ([26]). They define a general framework for computing nodes and edges eigenvector centralities for hypergraphs where the importance y_e of an edge e is a non-negative number proportional to a function of the importances of nodes in e, and similarly, the importance x_i of a node i is a non-negative number proportional to a function of the importances of the edges it participates in.

$$x_i \propto g\left(\sum_{e:i \in e} f(y_e)w_e\right) \quad y_e \propto \psi\left(\sum_{i \in e} \phi(x_i)\right)$$

where f, g, ψ, ϕ are four functions, possibly non linear, that must be specified by the user¹⁰.

 $^{^{10}}$ see the paper for the conditions on the properties of the functions that ensure the existence and uniqueness of the solutions

EV Linear: The choice of these four functions equal to the identity function $f = g = \psi = \phi(x) = id$ corresponds the EV linear centrality in table 3. As specified by the authors, this set of functions essentially corresponds to the standard eigenvector centrality applied to the graph and the line graph obtained by clique-expanding the input hypergraph and its dual respectively.

EV log-exp: The log-exp eigenvector centrality corresponds to the choices $f = id, \phi(x) = \ln(x), \psi(x) = \exp(x)$ and $g(x) = \sqrt{x}$.

 $EV \max$ The max eigenvector centrality corresponds to the choice $f = g = id, \phi = x^{\alpha}$ and $\psi(x) = x^{1/\alpha}$ with $\alpha = 10$.

4.2.3 Core to periphery

In section 4.1, we argue that the structure of the hypergraph is based on communities that are composed of an active core and a periphery. Let i be a node and C^* be the cluster to which i belongs. We define the core-to-periphery centrality as the relative strength of i in C^* :

$$ctp_i = \frac{s_i - \overline{s_{C^*}}}{\sigma_{C^*}}$$

where $\overline{s_{C^*}}$ is the mean strength of individuals present in C^* and σ_{C^*} is the standard deviations of the strengths of individuals in C^* . $\overline{s_{C^*}}$ and σ_{C_i} are in table 2b.

4.3 Correlation with Political Participation

Now that we have defined the political participation rate, and the other measures of centrality, in this section, we are able to answer the question asked in section 4.2 : what type of behaviour facilitates individuals' engagement with local politics? To this end, we compute the Pearson correlation between p_i and centrality measures. The results are shown in table 3. Betweenness¹¹ and closeness¹² centralities are calculated using the clique expansion of the hypergraph. The Pearson correlation coefficients are computed for two sets of nodes: the political body \mathcal{P} , and the set of individuals who are exclusively members of associations $V \setminus \mathcal{P}$.

First, when considering the set \mathcal{P} , we remark that the Pearson correlation is very high for almost all kinds of centralities except for betweenness and closeness. In this case, strength centrality captures well all the information. Note that, by definition, o_i is correlated to s_i , so to assess the relevance of the correlation between o_i and p_i , one needs to compare it to the correlation between s_i and p_i . Here, 0.990 > 0.800 indicates that for the set \mathcal{P} , agents with a high rate of political participation are the ones who participate in homogeneous activities rather than heterogeneous ones.

¹¹ For a given node i, betweenness centrality is the number of shortest paths that pass through i divided by the total number of shortest paths.

 $^{^{12}}$ For a given node *i* closeness centrality is the reciprocal of the sum of the length of the shortest paths between node *i* and all the other nodes

Centrality measure	\mathcal{P}	$V \setminus \mathcal{P}$
Strength	0.990	0.572
Diversity	0.800	0.807
EV linear	0.800	0.139
EV log expo	0.884	0.185
EV max	0.987	0.693
Core to Periphery	0.680	0.522
Betweenness	0.408	0.592
Closeness	0.091	0.134

French Regional Conference on Complex Systems – Le Havre, May 31 June 2, 2023

Table 3: Pearson Correlation between political participation and centrality measures

Concerning the set $V \setminus \mathcal{P}$, we remark that the diversity coefficient o_i obtains the best score (0.807) and significantly larger than s_i (0.572). Hence, for members of associations, participating in activities that involve individuals from different communities seems to be a favourable behaviour for political participation. We also remark that the strength and the core to periphery centrality are well correlated to political participation. This indicates that, people who are active, in general, and compared to others in their community, have more access to local politics.

We point out that the correlation obtained with betweenness centrality is also high. This measure reveals the importance of nodes in connecting other nodes, and in the case where the network is community based, is also reveals the importance of nodes in connecting communities. This notion is close to the one used to define the diversity coefficient. Indeed, heterogeneous activities gather individuals from different communities and form bridges between these communities. However, we note that the diversity coefficient and the political participation are extensions of the degree centrality, and therefore are, by definition, correlated to the strength centrality, while this not the case for betweenness centrality in its general definition.

We also notice that the EV max centrality correlates well with political participation. This is a surprising and interesting result since the choice of these sets of functions that build up the eigenvector centrality is not motivated by any empirical notion or concept, which is usually the case when defining centrality measures. Hence, there is no clear conclusion to be drawn concerning the behaviour of agents with high eigenvector centralities. On the other hand, the Pearson correlation is much lower for the EV linear centrality. This highlights the inability of standard graph analysis to capture the higher order interactions in a social group and advocates for pursuing analytical analysis of this work. Finally, we note that closeness centrality, which captures the ability of agent to be close to the others, does not not explain the access to local politics for members of associations. The same conclusion holds for the EV log exp, but same as for EV max, there is no clear interpretation of the behaviour that is captured by this measure.

5 Discussion

In this section, we combine all the results of the previous section and discuss them qualitatively.

First, we analyse the position of the political body in the network. In figure 4 we

see that agent in \mathcal{P} are mainly distributed in 5 clusters (1, 3, 5, 6 and 10). We note that the three larger clusters (3, 5 and 10) are almost exclusively composed of politicians. Not surprisingly, members of the political body have, on average, a much larger (8 times) political participation rate than members of associations. In addition, in Table 3, the fact the the strength coefficient correlates better than the diversity coefficient with the political participation indicates that participating in political activities and in homogenous activities are mutually reinforcing. All these observations show that the body politic forms rather exclusive communities where the activities, mainly political, are done in a political "entre-soi". We recall that we center our study around the associations and the city council present in a village. Many members of the political body are not elected officials of the town hall, but come, by snowball effect, from higher authorities such as the community of municipalities. This institution is composed of mayors and members of the municipal councils of the 31 municipalities that make up this organization. Thus, from the point of view of the members of the associations, these political communities seem rather closed. Getting in touch with them and participating in political activities is reserved for certain privileged members 13 .

Table 3 shows that members of associations whom participate in activities that gather individuals from diverse broads are those who have the most access to local politics. One might have thought that political activities are inherently heterogeneous, and that the opening coefficient is somehow an artifact, but when we calculate the Pearson correlation between the heterogeneity of an activity and its politicization, we find a positive correlation of 0.290^{-14} , but one that does not contain all the information. Furthermore, by comparing the results obtained for the strength centrality, we infer that an individual's social participation, in general, is less able to explain his or her political participation than when we focus more specifically on participation in heterogeneous activities. Thus participating in such diversifying activities reveals a structural behaviour in favour of accessing local politics. This idea is further support by the score of betweenness centrality for the clique expansion of the hypergraph. Indeed, these activities are all the more important since the hypergraph has a community structure based on that of organizations. Hence, they form bridges between the different organizations and allow the people who participate in them to have visibility in the public space, and thus, perhaps, legitimacy in the political space.

To date, this analysis has focused on interactional data to explain access to politics for association members. We now turn to the status of association members. Table 4 shows the point biserial correlation between being the president of an association and centrality measures.

Thus, the presidents of associations are mainly characterized by their proximity to local politics as well as their participation in heterogeneous activities. The difference in correlation observed between the core-periphery (0.314) and the strength (0.243) is consistent with the community structure based on those of the organizations. This indicates that presidents of associations are more active in their own organisation than outside them.

 $^{^{13}}$ This inequality in access to local politics for members of associations is reflected in the value of the Gini coefficient 0.79 close to 1.

¹⁴ p-value =0.000

Centrality	Point biserial	p-value
Political Participation	0.384	0.000
Strength	0.243	0.000
Diversity	0.397	0.000
Core to Periphery	0.314	0.000
Betweenness	0.363	0.000
Closeness	0.131	0.037

French Regional Conference on Complex Systems – Le Havre, May 31 June 2, 2023

Table 4: Point biserial correlation between being president of an association and centrality measures.

Putting everything together, the network has a community structure similar to that of organizations. The distribution of strengths within these communities is unevenly distributed. At the head of these communities are the presidents of the associations. In addition to being very active in their associations, the presidents are also the gateways to other organizations and participate in activities that bring together people from different communities. But, rather than their strong integration in their communities, it is their visibility in the public space and their ability to connect different organizations that are determining characteristics for their political legitimacy. These results highlight the inequalities that emanate from associations' modes of governance, which should be further examined and addressed.

6 Conclusion

In this paper, we have studied social interactions between members of associations and local political institutions in a city in France of nearly two thousand residents. We modelled group interactions using a hypergraph. We used a community detection algorithm designed for hypergraphs to reveal the architecture of the connexions that links individuals. We find that the hypergraph has a community structure based on the organizations that intervene in our study at a micro-level, and that these organizations gathers per category to form bigger communities. This confirms hypothesis 1 in section 2. These communities differ in sizes and in pattern of connexion with others. The fat tail of the distribution of strength among the vertices in the global network, is still observed inside the clusters.

We define an edge dependent degree centrality called *political participation* that quantifies the political involvement of individuals. To explain people disparity in political participation, we provide an other edge dependent degree centrality: the *diversity coefficient* that tend to measure the ability of an individual to participate in activities that involve people from different communities. We compare the correlations between the political participation and multiple centralities including the one that we define and others from the literature. We find that participating in heterogeneous activities is the key element to explain political participation for members of voluntary associations. This confirms hypothesis 2 in section 2, and in a more general context, is in line with the Granovetter theory of strength of weak ties [15] and the Burt theory of structural holes [18]. When we look at the status of association members, we realize that it is the association presidents who are the most likely to participate in this type of activities and who also have the most access to local politics. The purpose of this study is not to determine a causal relationship between people's interactional behaviour and their political participation. One may argue that the diplomatic trait of individuals is the one that allow them participating in activities that gather individuals from different backgrounds and is a key characteristic for becoming president of an association and gain some legitimacy in the political space. On the other hand, as being the president of an association, one should represent the association during formal meeting in political institutions and with other association partners. This role makes them the spokesperson of associations and forces them to participate in heterogeneous activities.

However, we can still point out that the associations' modes of governance involving a president assigned to the role of the representation of the association rather reinforce the inequality than balance the distribution of power between members. Indeed, even if the president is re-elected during the annual general assemble, in our case study, they accumulate a large number of mandates, sometimes they are at the creation of the organization. In studies supported by the Ministère de la Ville de la Jeunesse et des Sports conducted in 2011 and 2014 in France [27], they traced out a topology of governance modes of associations. The one that better describes our case study is the "tightened governance: in this type of association, governance is embodied in one or more omnipresent and charismatic persons: the president and/or the leader [...] the president does not leave much room for other internal stakeholders who tend to rely on his dynamism." They point out that a crucial problem that faces these organizations, is the renewal of the president, probably because other members do not feel as much integrated in the collective project and do not feel legitimate to run for the office. They also present the solutions proposed by some civic associations to tackle the problems associated with this type of governance. For instance, creating intermediate collegial bodies to feed back ideas to the decision-making body. This is intended to involve a variety of members and diversify the places of expression and decision. Other organizations fix the number of mandates such that the renewal of the leaders become mandatory.

Of course, the localisation of our study in a unique municipality in France, makes the generalization of our result questionable. For instance, the connexions between people would have a different pattern in an bigger urban cities, probably with a larger number of associations building up more clustered communities connected with fewer bridges. Also, in a cultural context were other modes of governance of associations are prominent should lead to different conclusions. But this limitation is almost inherent from the sociometric techniques since it would be costly to conduct this study at the scale of a country. We believe that the goal of such a study is to provide a more accurate understanding of the mechanisms that lead an unequal representativeness of members of associations in the political arena. This compensate the limitations of global national surveys where such precise interactional data are impossible to collect.

Finally, we would like to say a few words about the method used in this study, which probably makes up the originality of this work. We drew on methods often used in qualitative sociology and anthropology, by asking fairly open-ended questions during interviews rather than detailed, pre-formulated questionnaires. This method permits an adaptability according to the type of organization in question

(e.g. : size of the association, being part of a national federation or not, etc.). We mostly rely on documents provided by the respondents during the interviews (e.g. : annual report, agenda, photographs.). The collected data are then analysed quantitatively with macro-level network techniques. We choose a hypergraph modeling rather than a traditional graph because the structure of the data, in which people participate in activities, is better suited to hypergraph representation, and in doing so, rather than projecting the hypergraph, we preserve all the information. To advocate for the conservation of the higher order interactions, we compare the performance of the community detection algorithm for hypergraph and its clique expansion in section 7.3. We argue that preserving the complex structure of interactions in hypergraph clustering better addresses the resolution problem associated with the modularity-based algorithm and permits a better recovery of the organizations. This empirical observation should spruce an in-depth analytical analysis of this issue. Another argument for preserving the hypergraph structure is the opening of a wider range of centrality measures. Here, we find that the eigenvector centrality in linear mode (applied to the clique expansion of the hypergraph and the linegraph) do not correlate to political participation. On the other hand, introducing non-linearity, combined with the more complex structure of the hypergraph in the expression of eigenvector max, permits a better correlation with political participation. While this last centrality is not intended to mimic a social behaviour, we hope that this result encourages going further in this work and investigates other sets of functions and their corresponding social notions.

7 Appendix

In this section we explain the performed community detection algorithm and compare the results obtained on the one hand, with the hypergraph and on the other hand, with its clique expansion. The algorithm is a modularity based algorithm that tends to find the partitions corresponding to the maxima of the modularity function. 15

7.1 Modularity

Initially, this quality function was introduced by Newman and Girvan [25] and for a given partition of the vertices, it measures the extent to which links are formed within clusters rather than between them, comparatively to a null model. For simplicity, consider an unweighted networks with n vertices and m edges,

 $^{^{15}}$ Determining social groups in network science is a wild field of research. Freeman [28] provided a meta-analysis of 21 studies that attempts to specify social groups for the Southern Women data. Among the used methods are algebraic approaches, optimization algorithms, eigendecomposition and statical model. Some proposed overlapping social group, others don't. Some of them even proposed a measure for the positioning the individuals from core to periphery. For a review on community detection algorithms, see [23]

endowed with an adjacency matrix **A**, and a partition of vertices into q clusters $C = (C_1, \ldots, C_q)$, the corresponding modularity Q is :

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - P_{ij}) \delta(C_i, C_j)$$

where the P_{ij} are the probabilities in the null model that an edge exists between vertex *i* and *j*, and the $\delta(C_i, C_j) = 1$ if *i* and *j* are in the same cluster, 0 otherwise. The usual choice for the null model is a random graph where the expected degrees match the actual degrees in the original graph ¹⁶. In the case of graphs, fulfilling this condition implies $P_{ij} = \frac{k_i k_j}{2m}$, where $k_i = \sum_j A_{ij}$ is the degree of node *i*. In the case of hypergraphs, Barber ([24]) introduced a modularity function by

In the case of hypergraphs, Barber ([24]) introduced a modularity function by adapting the adjacency matrix and the null model. Consider an unweighted hypergraph with an incidence matrix **B** of size $n \times g$ (also called biadjacency matrix). He defines the incidence matrix **A**, and the corresponding probability matrix for the null model as follows :

$$\mathbf{A} = egin{pmatrix} \mathbf{O}_{n imes n} & \mathbf{B}_{n imes g} \ \mathbf{B}_{g imes n}^T & \mathbf{O}_{g imes g} \end{pmatrix} \ \ \mathbf{P} = egin{pmatrix} \mathbf{O}_{n imes n} & ilde{\mathbf{P}}_{n imes g} \ ilde{\mathbf{P}}_{g imes n}^T & \mathbf{O}_{g imes g} \end{pmatrix}$$

where the $\mathbf{O}_{i \times j}$ is the zero matrix of size $i \times j$. For the null model, he requested that the expected degrees of nodes are equal to the degrees in the actual graph, with the same constraint holding for the degrees of edges when considering the dual hypergraph. This leads to $\tilde{P}_{i,e} = \frac{k_i d_e}{m}$ where $k_i = \sum_e B_{ie}$ is the degree of node i, i.e., the number of edges e to which i belongs, $d_e = \sum_i B_{ie}$ is the cardinality the edge e, i.e., the number of nodes that are within it, and $m = \sum_e \sum_i B_{ie}$. In the case of weighted graphs and hypergraphs, the entries of the adjacency and

bi-adjacency matrices are no longer restricted to $\{0,1\}$ and the features k_i , d_e , and m are computed using the same formula as above.

Let us note that a major difference between the two methods, is that the second one provides a partitioning of both edges and nodes possibly in the same cluster. In our case, clusters should be composed of agents and events they participate in.

7.2 Algorithm

The modularity based-algorithms are designed to find the partition $\mathcal{C} = (C_1 \dots C_q)$ that maximizes the modularity. Of course, looking for all the possible clustering is not feasible since the number of ways in which is possible to partition the set vertices is huge. Among a variety of optimization algorithm (eg : genetic algorithms, extremal optimization, spectral optimization [23, p27-38]) we choose a greedy search method, known as the Louvain algorithm [29]. The algorithm starts with an initial clustering where nodes are put in different communities. Then, the algorithm iterate two successive phases. The first phase consists in shuffling nodes and, in turn, nodes are put in the communities that maximizes the modularity function. In the second phase, an aggregated graph is build where the nodes correspond to communities and the weight of edges are computed according to the edges present in the previous graph. These two steps are iterated until no increase in modularity is possible.

 $^{^{16}\,}$ For other definitions of the modularity function see [23] p 34 $\,$

7.3 Comparing clusters using the Hypergraph and its clique expansion

In this section we compare the performances of the clustering algorithms designed for hypergraph and for its clique expansion. A natural way of defining the clique expansion of the hypergraph is to put a link between each pair of nodes i and jpresent in a hyperedge e. The corresponding weight of the edge ij is then the aggregated weight. Formally, the clique expansion of the hypergraph $\mathcal{H} = (V, E, w)$ is the graph $\mathcal{G} = (V', E', w')$ where V' = V, $E' = \{(i, j) \in e | e \in E\}$ and $w'_{ij} = \sum_{e:(i,j) \in e} w_e$.

 $w'_{ij} = \sum_{e:(i,j) \in e} w_e$. First, the shuffling of nodes in the first phase of the algorithm introduces a stochastic effect. Hence, we run the algorithm 100 times for both cases, and find that the result is stable. We note the resulting partition of vertices resulting from the hypergraph $C_{\mathcal{H}}$ and the one resulting from its clique expansion $C_{\mathcal{G}}$

In order to measure to what extent the $C_{\mathcal{H}}$ and $C_{\mathcal{G}}$ are different, we compute the normalized mutual information I_{norm} ¹⁷ that takes its values between 0 and 1. We find $I_{norm}(C_{\mathcal{H}}, C_{\mathcal{G}}) = 0.766$ indicating that the partitions are similar but not that much.

Now the question is where does this difference lie in our case? One first answer for this question is the number of clusters and the distribution of clusters' size $Vol(C) = \sum_{i \in C} s_i$. Indeed, the clustering of the clique expansion graph leads to $q_{\mathcal{G}} = 10$ clusters with a standard deviation of cluster size $std_{\mathcal{G}} = 1659.65$ whereas the original hypergraph yields $q_{\mathcal{H}} = 14$ and $std_{\mathcal{H}} = 1104.18$. Hence, clustering nodes and edges at the same time provides a finer resolution and a more balanced partition. To go into further details, we introduce the so-called *resolution* parameter γ [30] in the modularity function :

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - \gamma P_{ij}) \delta(c_i, c_j)$$

Larger γ leads to more communities, while lower gamma leads to fewer communities. Now, the question is still there a difference in the resulting partitions when the number of clusters is fixed? To answer this, we run the algorithm for a range of values of γ . Then, we computed the normalized mutual information for partitions having the same number of clusters. Note that for a given γ , the number of clusters is higher in the case of hypergraph clustering than in the clique expansion case. Figure 5a displays the results. For values of q less than 13, the normalized mutual information (I_{norm}) is relatively low and unstable. However, for q values greater than or equal to 13, I_{norm} fluctuates around 0.85.

Figure 5b displays the standard deviation of clusters' size for a range of values of q, the number of clusters. The orange curve corresponds to the clustering obtained using the original hypergraph and the blue curve corresponds the clustering obtained using its clique expansion. In both cases, for q = 1, all the nodes are gathered in the same cluster, leading to a vanishing standard deviation. For low values of $q \leq 7$, the large values of the standard deviation indicate that both

¹⁷ Let X and Y be two partitions, $I_{norm}(X,Y) = \frac{2I(X,Y)}{H(X)+H(Y)}$ where I(X,Y) is the mutual information of X and Y and H(X) is the entropy of partition X. $I(X,Y) = \sum_{x,y} \mathbb{P}(X = x, Y = y) \log \frac{\mathbb{P}(X=x,Y=y)}{\mathbb{P}(X=x)\mathbb{P}(Y=y)}$ and $H(x) = -\sum_x \mathbb{P}(X = x) \log \mathbb{P}(X = x)$. $\mathbb{P}(X = x)$ if the probability that a vertex is assigned to community x in the scheme X.

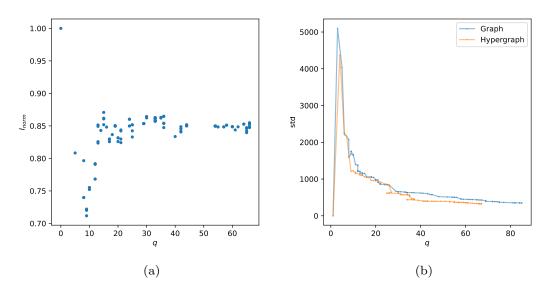


Fig. 5: (a) Normalized mutual information versus the number of clusters γ . (b) Standard deviation of clusters' size versus number of clusters. The orange curve corresponds to the clustering obtained using the original hypergraph and the blue curve corresponds the clustering obtained using its clique expansion.

algorithms provide a giant cluster that gathers most of the nodes. This tendency decreases more rapidly for hypergraph clustering. Moreover we can note that the orange curve is almost always under the blue curve, which indicates that the hypergraph clustering provides more balanced partitions in terms of sizes.

To conclude this part, we have shown that in our case of study, when using the original modularity function with the resolution parameter $\gamma = 1$, the hypergraph clustering provides a larger number of clusters and these clusters are more balanced in size. The partitions obtained with the clique expansion clustering are the expression of a well known problem in modularity-based algorithm, the so-called resolution limit [31]. One way of solving this problem is to introduce the resolution parameter γ in the expression of the modularity function. Indeed, when we tune the resolution parameter and compare the partitions that share the same number of clusters, partitions are relatively more similar, but the hypergraph clustering still provide more balanced partitions. Moreover, usually, we do not know the number of clusters behind the network's architecture and determining the resolution parameter that reveal the best this structure is a task on its on. Thus, in our case study, hypergraph clustering better faces resolution limit that the clique expansion clustering. This empirical observation deserves to be deepened. In addition, hypergraph clustering also provides edge partitions. We do not exploit this possibility in our study, but it can be useful for other cases of study.

7.4 Recovering the organization using the clustering algorithm

In comparing Figure 3a and 6, we can see that the orange histogram shows a higher proportion of individuals with an intra-similarity close to 1 and a lower proportion

of individuals with an intra-similarity close to 0 in hypergraph clustering compared to clique expansion clustering. This suggests that hypergraph clustering is more effective in recovering the organizations.

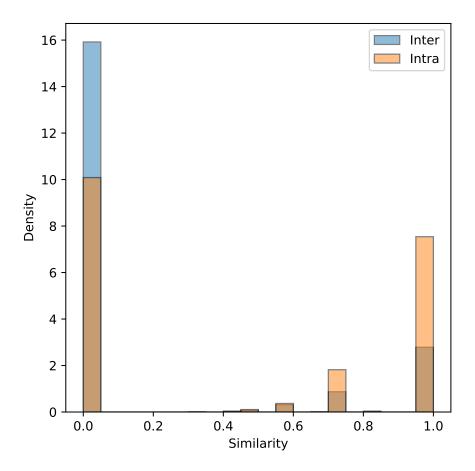


Fig. 6: Distribution of inter and intra similarity with the clusters obtained using the clique expansion of the graph. The similarity is computed using the vectors X corresponding to memberships of agents. The intra-similarity is the similarity between a pair of agents in the same cluster and its distribution is represented in orange. The inter-similarity, in blue, corresponds to the similarity between a pair of agents belonging to different clusters. The resolution parameter in the expression of the modularity is set to 1.

References

- 1. C. Wayne Gordon and Nicholas Babchuk. A Typology of Voluntary Associations. American Sociological Review, 24(1):22, February 1959.
- John P. Heinz, Edward O. Laumann, Robert L. Nelson, and Robert H. Salisbury. The Hollow Core: Private Interests in National Policy Making. Harvard University Press, Cambridge, MA, September 1997.
- Meredith I. Honig. The New Middle Management: Intermediary Organizations in Education Policy Implementation. *Educational Evaluation and Policy Analysis*, 26(1):65–87, March 2004.

- 4. Marvin E. Olsen. Social Participation and Voting Turnout: A Multivariate Analysis. American Sociological Review, 37(3):317, June 1972.
- Jan W. van Deth. Private Groups and Public Life: Social Participation and Political Involvement in Representative Democracies. Routledge, December 2003. Google-Books-ID: fMWEAgAAQBAJ.
- 6. David Lazer. Networks in Political Science: Back to the Future. *PS: Political Science & Politics*, 44(1):61–68, January 2011.
- David E. Campbell. Social Networks and Political Participation. Annual Review of Political Science, 16(1):33–48, May 2013.
- 8. Noah E. Friedkin. A Structural Theory of Social Influence. Cambridge University Press, 1 edition, September 1998.
- David Knoke. Networks of Political Action: Toward Theory Construction. Social Forces, 68(4):1041–1063, 1990. Publisher: Oxford University Press.
- M. Stephen Weatherford. Interpersonal Networks and Political Behavior. American Journal of Political Science, 26(1):117, February 1982.
- Ellen Quintelier, Dietlind Stolle, and Allison Harell. Politics in Peer Groups: Exploring the Causal Relationship between Network Diversity and Political Participation. *Political Re*search Quarterly, 65(4):868–881, 2012. Publisher: [University of Utah, Sage Publications, Inc.].
- 12. Matthew O. Jackson. Social and Economic Networks. Princeton University Press, 2008.
- Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. Dynamical Processes on Complex Networks. Cambridge University Press, Cambridge, 2008.
- 14. Carole Jean Uhlaner. Politics and Participation. In International Encyclopedia of the Social & Behavioral Sciences, pages 504–508. Elsevier, 2015.
- Mark S. Granovetter. The Strength of Weak Ties. American Journal of Sociology, 78(6):1360–1380, 1973. Publisher: University of Chicago Press.
- 16. David Knoke, Mario Diani, James Hollway, and Dimitris Christopoulos. *Multimodal Political Networks*. Cambridge University Press, 1 edition, May 2021.
- 17. Noe Gaumont, Maziyar Panahi, and David Chavalarias. Methods for the reconstruction of the socio-semantic dynamics of political activist Twitter networks.
- 18. Ronald S. Burt. *Structural holes: the social structure of competition*. Harvard University Press, Cambridge, Mass, 1992.
- 19. Robert D. Putnam. Bowling alone: The collapse and revival of American community. Bowling alone: The collapse and revival of American community. Touchstone Books/Simon & Schuster, New York, NY, US, 2000. Pages: 541.
- William P. Eveland and Steven B. Kleinman. Comparing General and Political Discussion Networks Within Voluntary Organizations Using Social Network Analysis. *Political Behavior*, 35(1):65–87, March 2013.
- Matthew A. Crenson. Social Networks and Political Processes in Urban Neighborhoods. American Journal of Political Science, 22(3):578, August 1978.
- David A. Siegel. Social Networks and Collective Action. American Journal of Political Science, 53(1):122–138, January 2009.
- 23. Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, February 2010. arXiv:0906.0612 [cond-mat, physics:physics, q-bio].
- 24. Michael J. Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76(6):066102, December 2007. arXiv:0707.1616 [cond-mat, physics:physics].
- M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, February 2004. arXiv:cond-mat/0308217.
- 26. Francesco Tudisco and Desmond J. Higham. Node and Edge Nonlinear Eigenvector Centrality for Hypergraphs, August 2021. Issue: arXiv:2101.06215 arXiv: 2101.06215 [physics].
- 27. Elisabetta Bucolo, Philippe Eynaud, and Joseph Haeringer. La gouvernance des associations en pratiques | Avise.org, 2014.
- 28. L. Freeman. Finding Social Groups: A Meta-Analysis of the Southern Women Data. 2003.
- Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008. arXiv: 0803.0476 [cond-mat, physics:physics].
- 30. Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, July 2006. Publisher: American Physical Society.
- Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. Proceedings of the National Academy of Sciences, 104(1):36–41, January 2007.



Opinion dynamics model revealing yet undetected cognitive biases

Guillaume Deffuant

Abstract This paper synthesises a recent research that includes opinion dynamics models and experiments suggested by the model results. The mathematical analysis establishes that the model's emergent behaviours derive from cognitive biases that should appear in quite general conditions. However, it seems that psychologists have not yet detected these biases. The experimental side of the research therefore includes specifically designed experiments that detect these biases in human subjects. The paper discusses the role of the model in this case, which is revealing phenomena that are almost impossible to observe without its simulations.

1 Introduction

The research on opinion dynamics assumes rules of interactions among agents, influencing each other by pairs or in larger groups, and then explores the effect of these hypotheses by simulation. The tested rules of interaction are diverse. Some take their inspiration in physics (e.g. [1]), others derive from precise experimental results in social psychology (e.g. [2]), others express intuitive or common sense assumptions (e.g. [3]). As pointed out in a recent review [4], few models aim at reproducing empirically observed patterns of opinion distribution.

Indeed, the model is more often considered interesting when its evolution leads to patterns that were difficult or even almost impossible to predict from the mere knowledge of the interaction function. In this case, understanding the emergence process becomes a research question and often a challenging one. When answering successfully this question, the research produces new knowledge about complex phenomena that may take place in social dynamics.

G. Deffuant

Université Clermont-Auvergne, INRAE, UR LISC, LAPSCO, France. E-mail: guil-laume.deffuant@inrae.fr

A typical example is the model of cultures from Axelrod [5] whose emergent properties (transition from small to large cultural domains in particular) have been the subject of an intense theoretical work (e.g. [6,7]).

In this general line of research, we consider a simple model involving noisy interactions with rather unexpected emergent properties, which could be understood only with non-obvious approximations of its average behaviour. Moreover, the analysis of this approximate model reveals the crucial influence of two emerging biases: a positive bias on self-opinions and a negative bias on the opinions about others. The effect of these biases is quite small at each interaction, but the simulations suggest that it can add-up and become very significant over time. This paper summarises this work which is reported in details in [8,9], with the goal to show the connection between this modelling approach and the experimental research that it induced.

Indeed, it appears that the biases revealed by the model have not been observed by psychologists yet. Psychologists robustly observed a bias of selfoverestimation which can be related to the positive bias on self-opinion, but its mechanism is very different. Thus the next arising question is: can the biases revealed by the model be observed in experiments with human subjects? Indeed, mathematically, the biases derive directly from general hypotheses relating influence and self-opinion that seem reasonable.

In order to check these hypotheses, we designed an online questionnaire in which the participants receive evaluations of their performance at a task and then are asked to self-evaluate several times in reaction to these feedbacks. The results confirm the existence of a positive bias on the self-opinion coming from the decreasing sensitivity to the feedback. This paper reports the main features of the experiment and its results. The details can be found in [10].

The rest of the paper is organised as follows. In the next section, we describe the opinion dynamics model and its emerging effects. Then, we define mathematically the positive bias on self-opinion from decreasing sensitivity to feedback. Then, we describe the experiment aiming at observing this bias and its results. Finally, we discuss the whole approach in a broader perspective.

2 The model revealing biases

2.1 State and dynamics

The model is presented in [8,9] and it is a simplified version of an earlier model [11]. It includes N_a agents. Each agent $i \in \{1, \ldots, N_a\}$ has an opinion a_{ij} about each agent $j \in \{1, \ldots, N_a\}$ including themselves. The opinions are real values between -1 and +1. In most simulations, at the initialisation, all opinions are set to 0: agents have a neutral opinion about themselves and all the others at the beginning of the simulations.

Graphically, the agents' opinions can be represented as a matrix (see examples on Figure 1) in which row i, with $1 \leq i \leq N_a$, represents the array of N_a opinions of agent i about the agents j. Column j, with $1 \leq j \leq N_a$,

represents the opinions all agents i about j. Positive opinions are represented with red shades and negative opinions with blue shades. Lighter shades are used for opinions of weak intensity (close to 0).

The dynamics consists in repeating:

- choose randomly two distinct agents i and j;
- -i and j interact: j influences i's opinions and i influences j's opinions.

In this interaction, $a_{ii}(t)$, *i*'s self-opinion, is influenced by $a_{ji}(t)$, the opinion of *j* about *i*. As a result of this influence, $a_{ii}(t)$ gets closer to a noisy evaluation of $a_{ji}(t)$. The modification of $a_{ii}(t)$, denoted by $\Delta a_{ii}(t)$, is ruled by the following equation, in which $\theta_{ii}(t)$ designates a number that is uniformly drawn between $-\delta$ and δ (δ being a parameter of the model):

$$\Delta a_{ii}(t) = h_{ij}(t)(a_{ji}(t) - a_{ii}(t) + \theta_{ii}(t)), \tag{1}$$

This equation expresses that the self-opinion of agent i is influenced by i's perception of the opinion of j about i. The noise models the mistakes in the perception of other's opinions.

Similarly, the change of opinion $a_{ji}(t)$, is:

$$\Delta a_{ji}(t) = h_{ji}(t) \left(a_{ii}(t) - a_{ji}(t) + \theta_{ji}(t) \right).$$
(2)

where $\theta_{ji}(t)$, is a uniformly drawn number between $-\delta$ and δ . The function of influence $h_{ij}(t)$ is given by equation 3, expressing that the more *i* perceives *j* as superior, the more *j* is influential on *i*.

$$h_{ij}(t) = H(a_{ii}(t) - a_{ij}(t)) = \frac{1}{1 + \exp\left(\frac{a_{ii}(t) - a_{ij}(t)}{\sigma}\right)}.$$
 (3)

In this model, self-opinions measure how well agents think they are perceived by others, with a stronger weight attributed to agents perceived as superior. This is in line with the hypothesis considering self-opinion as a sociometer [12].

When activating gossip, agents j and i influence their opinions about k agents g_p , $p \in \{1, \ldots, k\}$ drawn at random such that $g_p \neq i$ and $g_p \neq j$. The changes of the opinion of i about agents g_p are:

$$\Delta a_{ig_p}(t) = h_{ij}(t)(a_{jg_p}(t) - a_{ig_p}(t) + \theta_{ig_p}(t)), \text{ for } p \in \{1, \dots, k\},$$
(4)

where $\theta_{ig_p}(t)$ is a uniformly drawn number between $-\delta$ and δ . The changes of the opinion of j about these agents follow the same equations where j and i are inverted.

Overall, after the encounter between i and j, the opinions about i change as follows:

$$a_{ii}(t+1) = a_{ii}(t) + \Delta a_{ii}(t),$$
 (5)

$$a_{ji}(t+1) = a_{ji}(t) + \Delta a_{ji}(t).$$
(6)

The opinions about j change similarly (inverting j and i in the equations). If there is gossip (k > 0), k agents g_p are randomly chosen with $p \in \{1, \ldots, k\}$, $g_p \neq i$ and $g_p \neq j$, and the opinions about g_p , for $p \in \{1, \ldots, k\}$ change as follows:

$$a_{ig_{p}}(t+1) = a_{ig_{p}}(t) + \Delta a_{ig_{p}}(t), \tag{7}$$

$$a_{jg_{p}}(t+1) = a_{jg_{p}}(t) + \Delta a_{jg_{p}}(t).$$
(8)

The opinions are updated synchronously: at each encounter all the changes of opinions are first computed and then the opinions are modified simultaneously.

2.2 The main patterns

Figure 1 illustrates the main patterns of evolution of the opinions.

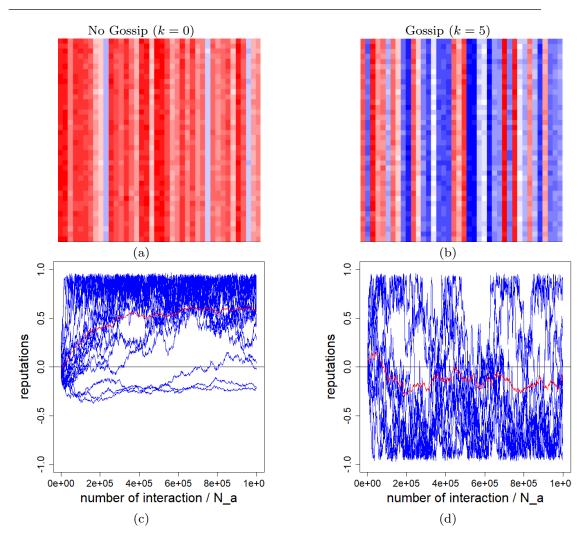
Panels (a) and (c) illustrate the pattern obtained without gossip (k = 0). Panel (a) shows a typical opinion matrix after a large number of iterations. In each matrix column the opinions are close and the differences between the matrix columns are stronger than the differences of opinions within each column. This is explained by the attractive dynamics which tends to align the opinions about a given individual. Note that most of the columns are red, indicating that the opinions about most agents are positive. On panel (c) the red curve shows the evolution of the average opinion. The blue curves are the evolution of the agents' reputations (the average opinion about an agent). Starting from 0, the average opinion increases and then fluctuates around a significantly positive value (close to 0.5). As already noticed in [8], this pattern is surprising because, at a first glance, the equations do not privilege changing opinions upward and the noise is symmetric around 0.

Panels (b) and (d), illustrate the pattern taking place when gossip is activated (in this case, k = 5). The matrix of opinions after a large number of interactions (panel b) shows numerous blue columns. On panel (d), the evolution of the average opinion (red curve) remains negative with significant fluctuations while the reputations (blue curves) are more dispersed than without gossip, with a larger density in the low part of the opinion axis.

2.3 Biases appearing in equations of average evolution.

In [9], it is shown that the patterns are related to two biases which are observed in a simplified setting where only one opinion varies, between two interacting agents:

- when the self-opinion of ego varies, it is on average slightly higher than the opinion of alter about ego. There is a positive bias on the self-opinion.
- when the opinion of ego about alter varies, symmetrically, it is on average slightly lower than alter's self-opinion. There is a negative bias on the opinion about others.



French Regional Conference on Complex Systems – Le Havre, May 31 June 2, 2023

Fig. 1: Typical patterns, with $\delta = 0.1$ (noise), $\sigma = 0.3$ (influence function parameter) and $N_a = 40$ agents. Panels (a) and (b) show the matrix of opinions after 1 million $\times N_a$ pair interactions. Panels (c) and (d) show the evolution of the average opinion (in red) and the evolution of the agent reputations (in blue, the reputation of agent *i* being the average of the opinions about *i*).

We present this simplified setting in details further in the paper. Similar biases are present when all opinions vary and more than two agents interact. Moreover, the analysis suggests that the drift to positive or negative opinions in the observed patterns is due to the dominance of one bias on the other:

- Without gossip, the positive bias on self-opinion dominates the negative bias on the opinions about others, which explains why the positive drift arises;
- Gossip increases the noise on the opinions about others, which increases the negative bias on opinion about others, leading to its possible domination over the positive bias on self-opinion.

This conclusion is established by deriving the evolution of the average opinions using a moment approximation and a first order approximation of the function of influence h. It is expressed by the following equation of the evolution the first order equilibrium opinion offset $e_i(t)$ of agent i, or equilibrium opinion for short. Indeed this evolution captures the common trend of all opinions about an agent. The definition of $e_i(t)$ is the following:

$$e_i(t) = \frac{1}{1 + S_i(t)} \left(\overline{x_{ii}}(t) + \sum_{j \neq i} \frac{\widehat{h_{ij}}(t)}{\widehat{h_{ji}}(t)} \overline{x_{ji}}(t) \right), \tag{9}$$

where $\overline{x_{ij}}(t)$ is the average of $a_{ij}(t) - a_{ij}(0)$ over an infinite number of replicas of simulations from the same initial conditions. Moreover:

$$\widehat{h}_{ij}(t) = h(\overline{a_{ii}}(t) - \overline{a_{ij}}(t)) - h'(\overline{a_{ii}}(t) - \overline{a_{ij}}(t))(\overline{x_{ii}}(t) - \overline{x_{ij}}(t)), \quad (10)$$

and:

$$S_i(t) = \sum_{j \neq i} \frac{\widehat{h_{ij}}(t)}{\widehat{h_{ji}}(t)}.$$
(11)

The evolution of $e_i(t)$ over time is given by:

$$e_{i}(t+1) = e_{i}(t) + \frac{2}{N_{c}(1+S_{i}(t))} \sum_{i \neq j} \overline{h'_{ji}}(t) \left(\overline{x_{ii}(t).x_{ji}(t)} - \overline{x_{ii}^{2}}(t) + \frac{\widehat{h_{ij}}(t)}{\widehat{h_{ji}}(t)} \left(\overline{x_{ji}^{2}}(t) - \overline{x_{ii}(t).x_{ji}(t)} \right) \right),$$
(12)

where $N_c = N_a(N_a - 1)$.

At any time step t, $e_i(t)$ is the value to which all the opinions about iwould converge over time if the $h_{ij}(t)$ were frozen. More precisely, imagining that from a given time t_0 , for all $t > t_0$ and for all $(i, j) \in \{1, \ldots, N_a\}$, $h_{ij}(t) = h_{ij}(t_0)$, then $\overline{x_{ji}}(t)$ for all j would converge to $e_i(t_0)$ and remain at this value. Therefore, equation 12 determines the second order effect applied to an opinion which is at the equilibrium of the first order effects. In the long run, this trend is common to all opinions about i (see [9]).

Moreover, the following terms appearing in equation 12:

$$\overline{h'_{ij}}(t)\left(\overline{x_{ii}(t).x_{ji}(t)} - \overline{x_{ii}^2}(t)\right),\tag{13}$$

are positive as $\overline{h'_{ij}}(t)$ is assumed strictly negative. Therefore, the effect of these terms is to increase the opinions about *i* and their sum can be seen as a positive bias on self-opinions. Similarly, the terms of 12:

$$\overline{h'_{ij}}(t)\left(\overline{x_{ji}^2}(t) - \overline{x_{ii}(t).x_{ji}(t)}\right),\tag{14}$$

are negative and tend to decrease the opinions about agent i. They can be interpreted as negative biases coming from the changes of opinions about i.

The overall trend is thus expressed as a weighted sum of the positive bias on i's self-opinion and the negative biases on the opinions about i. The negative biases are multiplied by the factor $\frac{\widehat{h_{ij}(t)}}{\widehat{h_{ji}(t)}}$, which is high when $\overline{a_{ii}} < \overline{a_{ij}}$ and $\overline{a_{jj}} > \overline{a_{ji}}$. Therefore, when the agents are in a consensual hierarchy, these factors are higher for agents i of low status. Hence, the opinions about the agents of low status grow less (or even can decrease) than the opinions about the agents of high status.

When gossip is activated, a term of first order coming from gossip is added to the equation of $e_i(t+1)$ but simulations show that the effect of this term is negligible. Therefore, like in the case without gossip, $e_i(t)$ provides the common trend of the evolution of the opinions about i.

However, gossip modifies the negative bias on the opinion about others because the opinions vary more strongly, which increases $\overline{x_{ji}^2}(t+1)$ and consequently the negative bias on x_{ji} in the following time steps.

When the agents are in a consensual hierarchy, the additional negative bias is stronger for agents of low status (see details in [9]).

2.4 Explanation of the patterns

In addition to the explanations that mathematical expressions can bring, the moment approximation provides a means to explore the average behaviour of the agent based model, without running millions of simulations. Such an exploration for 10 agents, when varying the width of the interval of initial self-opinions, reveals the following features:

- The opinions about the agents of high status (except the top status) tend to grow in roughly the same way, with or without gossip, because the opinions about them are not much affected by the negative biases;
- The opinions about agents of low status tend to grow when the differences between low and high status are small. Without gossip, this growth progressively decreases and even becomes close to zero when the initial inequalities increase. Indeed, this increase of inequalities increases the weights of the negative biases on the opinions about them. With gossip, these opinions grow only when the inequalities are very low and then the opinions about low status agents start decreasing as the inequalities increase, because the negative biases are increased by gossips.

The same explorations conducted with 20 and 40 agents yield similar results.

These observations provide some explanations to the patterns presented in section 2.2. Indeed, initially, in these patterns, all the opinions are the same, therefore, both with and without gossip, all the average opinions tend to grow together in a first period of a few thousand steps. However, because of the noise, more or less dispersion of the opinions takes place, introducing inequalities between agents:

- Without gossip, since all opinions tend to grow at a similar pace, the opinion inequalities remain moderate for a while and all opinions grow on average. When the inequalities reach a threshold though, the opinions about agents of low status grow more and more slowly or stagnate, while the opinions about agents of high status fluctuate when reaching the opinion limit at +1. Overall, the distribution of opinions is therefore significantly positive on average;
- When there is gossip, because the opinions about agents of low status grow much more slowly than the opinions about agents of high status, the inequalities of opinions increase more rapidly and easily reach a level in which the opinions about the lowest status agent starts decreasing. This further increases the opinion inequalities, and the opinions about agents of low status start decreasing, which further increases the opinion inequalities. Ultimately, when the inequalities are maximum, the opinions about a majority of agents tend to decrease. This explains why the distribution of opinions becomes negative on average.

Overall, our analysis reveals that the positive bias on the self-opinions and a negative bias on the opinion about others are combined in a way that is detrimental to the agents of low status, when the inequalities are significant and when there is gossip.

It seems therefore interesting to better understand the mechanisms behind the observed biases. In the following, we focus on the positive bias on selfopinion. In the next section, we try to express it in isolation and in a simplified setting that seems more appropriate for trying to detect it in an experiment. In the following section, we describe the experiment.

3 Positive bias on self-opinion from decreasing sensitivity

We now consider a simplified version of previously described agent model, involving a single agent who is interacting with a single source sending opinions that we call feedbacks. This section is partly extracted from [10]. We often use evaluation instead of opinion in this context.

3.1 General definition of the positive bias from decreasing sensitivity.

Consider an agent with self-evaluation a_t at time t when receiving feedback f_t (i.e. an evaluation coming from an outside source), when the agent has a constant opinion s about the source. The hypotheses of the previous model become in this particular cases:

- the change of self-evaluation due to this feedback is proportional to the difference between the feedback and the self-evaluation;
- Moreover, the coefficient of proportionality decreases when the self-evaluation a_t increases.

These hypotheses are thus expressed by equation 15:

$$a_{t+1} - a_t = h(a_t - s)(f_t - a_t), \tag{15}$$

where $h(a_t - s)$ is the same positive and decreasing function as previously and s is the opinion of the agent about the source which is assumed constant. In the following, for simplicity we write $h(a_t)$ instead of $h(a_t - s)$ and we assume that function h is derivable, thus its derivative h' is negative: $h'(a_t) < 0$ for all a_t .

When removing the reference to s, equation 15 can also address cases where the feedback does not come directly from another agent but is a general evaluation from the environment, like a failure or a success. In this case, a possible justification for assuming that h decreases is that agents with a high self-evaluation tend to be more confident and this makes them less prone to change their mind. The general hypothesis is that people having a high selfevaluation are less easily influenced than people having a low self-evaluation.

The fact that function h is decreasing induces a general positive bias that we now define mathematically. Assume that the feedback is a random distribution of average a_1 , which is also the initial self-evaluation. The first feedback is $f_1 = a_1 + \theta_1$, θ_1 being randomly drawn from the distribution of average 0, and the self evaluation after receiving this feedback is:

$$a_2 = a_1 + h(a_1)\theta_1. \tag{16}$$

Then, after the second feedback $f_2 = a_2 + \theta_2$, θ_2 being randomly drawn from the distribution of average 0, the self-evaluation a_3 after receiving this feedback is:

$$a_3 = a_2 + h(a_2)(a_1 + \theta_2 - a_2) \tag{17}$$

Assuming that θ_1 is small, the sensitivity $h(a_2)$ at a_2 can be approximated at the first order as:

$$h(a_2) = h(a_1) + h'(a_1)h(a_1)\theta_1.$$
(18)

Replacing a_2 by its value and $h(a_2)$ by this approximation yields:

$$a_{3} = a_{1} + h(a_{1})\theta_{1} + (h(a_{1}) + h'(a_{1})h(a_{1})\theta_{1})(\theta_{2} - h(a_{1})\theta_{1}),$$
(19)
= $a_{1} + h(a_{1})\theta_{1} + h(a_{1})(\theta_{2} - h(a_{1})\theta_{1}) + h'(a_{1})h(a_{1}))(\theta_{2}\theta_{1} - h(a_{1})\theta_{1}^{2}).$ (20)

Because we assume the averages of θ_1 and of θ_2 are 0, the average $\overline{a_3}$ of a_3 over all possible draws of θ_1 and θ_2 is:

$$\overline{a_3} = a_1 - h'(a_1)h^2(a_1)\overline{\theta_2^2}.$$
(21)

As we assume $h'(a_1) < 0$, we always have:

$$-h'(a_1)h^2(a_1)\overline{\theta_2^2} > 0.$$
(22)

This value defines the positive bias. The second evaluation a_3 is on average higher than the average feedback a_1 because of this bias.

This result extends to longer series of feedbacks [9]. The positive bias increases with the length of the series to an asymptotic value, which remains of the second order (in $\overline{\theta^2}$).

Our aim is to check experimentally the existence of this bias. If we directly derive the experiment from the previous formulas, we face a hard problem: we need a huge number of random draws of feedbacks in order to get their average close to 0 and get a chance to detect the bias. To overcome this difficulty, we consider particular series of feedbacks in which the bias appears without averaging over many trials.

3.2 Positive bias from decreasing sensitivity with alternating positive and negative feedbacks

Let $f_t - a_t$ be the intensity of feedback f_t . We say that a feedback is positive when its intensity is positive and negative otherwise. We show now that the previous model generates a positive bias when receiving a series of feedbacks of opposite intensities. We consider the simple example of an agent receiving two consecutive feedbacks of opposite intensities $\pm \delta$.

Assume that the agent starts with self-evaluation a_1 and receives first the positive feedback $f_1 = a_1 + \delta$. Applying equation 15, the self-evaluation of the agent becomes a_2 :

$$a_2 = a_1 + h(a_1)\delta. (23)$$

Then the agent receives the negative feedback $f_2 = a_2 - \delta$ and its selfevaluation a_3 becomes:

$$a_3 = a_2 - h(a_2)\delta. \tag{24}$$

The difference of self-evaluation between before and after receiving the couple of feedbacks is:

$$a_3 - a_1 = a_1 + h(a_1)\delta - h(a_2)\delta - a_1 = (h(a_1) - h(a_2))\delta.$$
(25)

As we assume that at any time t, $h(a_t) > 0$, we have $a_1 < a_2$ and, as h is decreasing, we have: $h(a_1) - h(a_2) > 0$, hence $a_3 - a_1 > 0$.

Now, if we invert the order of the feedbacks $(f_1 = a_1 - \delta \text{ and } f_2 = a_2 + \delta)$, we have:

$$a_3 - a_1 = (h(a_2) - h(a_1))\delta.$$
(26)

Now $a_2 < a_1$, therefore again, because h is decreasing $a_3 - a_1 > 0$.

Therefore, after receiving two feedbacks of opposite intensities, the selfevaluation tends to increase.

Developing $h(a_2)$ at the first order like previously, we can approximate the value of the bias:

$$h(a_2) \approx h(a_1) + h'(a_1)h(a_1)\delta$$
, if $f_1 = a_1 + \delta$; (27)

$$h(a_2) \approx h(a_1) - h'(a_1)h(a_1)\delta$$
, if $f_1 = a_1 - \delta$. (28)

Therefore, for both sequences of feedbacks we get:

$$S(a_1) = a_3 - a_1 \approx -h'(a_1)h(a_1)\delta^2.$$
(29)

This positive bias is thus expected to be of the second order of the intensity of the feedback, hence rather small.

With a series of feedbacks of opposite intensities, the positive bias appears directly, without requiring to average on a large number of trials. In an experiment, the participants processing such a series of feedbacks of opposite intensities are expected to provide a noisy value of function h(a) for each selfevaluation a in the series. We expect to approximate the average value of h(a)and the related bias when computing them from data collected on a sufficient number of participants.

3.3 Self-enhancement

The literature suggests that the participants to an experiment generally show a significant self-enhancement and we should adapt our model to this case. The self-enhancement is the usual positive bias on self-opinion, which comes from a tendency to give much credit to positive feedbacks and to dismiss negative ones [13–16]. In the framework of this model, self-enhancement takes place when the sensitivity $h_p(a_t)$ to positive and $h_n(a_t)$ to negative feedbacks are different :

$$a_{t+1} - a_t = h_p(a_t)\delta, \text{ if } f_t = a_t + \delta, \tag{30}$$

$$a_{t+1} - a_t = -h_n(a_t)\delta$$
, if $f_t = a_t - \delta$. (31)

Considering feedbacks of intensity $\pm \delta$, the bias of self-enhancement E(a)at a given self-evaluation a can be expressed as the difference between the reaction to the positive feedback $f_p = a + \delta$ and the reaction to the negative feedback $f_n = a - \delta$:

$$E(a) = (h_p(a) - h_n(a))\delta.$$
(32)

3.4 Combining self-enhancement and positive bias from sensitivity

Now, assume that the agent's self-evaluation is a_1 and that the agent receives a positive and then a negative feedback. Repeating the previous calculations, we get:

$$a_2 = a_1 + h_p(a_1)\delta, (33)$$

$$a_3 = a_2 - h_n(a_2)\delta. (34)$$

The total bias $B(a_1)$ from these successive feedbacks is:

$$B(a_1) = a_3 - a_1 \tag{35}$$

$$= (h_p(a_1) - h_n(a_1))\delta - h'_n(a_1)h_p(a_1)\delta^2.$$
(36)

We recognise the self-enhancement bias (equation 32) in the first term and the bias from decreasing sensitivity (equation 29) in the second term. For this sequence of feedbacks, the bias from decreasing sensitivity is thus:

$$S(a_1) = -h'_n(a_1)h_p(a_1)\delta^2.$$
(37)

This value is positive when $h'_n(a_1)$ is negative and we have:

$$B(a_1) = E(a_1) + S(a_1).$$
(38)

Moreover, if we have a series of 2 positive and 2 negative feedbacks in a random order (as it will be the case in the experiment), the average bias from decreasing sensitivity is:

$$S(a) = \frac{1}{4} \left(-h'_n(a)h_p(a) - h'_p(a)h_n(a) - h'_p(a)h_p(a) - h'_n(a)h_n(a) \right) \delta^2, \quad (39)$$

$$S(a) = -h'_m(a)h_m(a)\delta^2,\tag{40}$$

where h_m is the average of h_p and h_n : $h_m(a) = \frac{1}{2}(h_p(a) + h_n(a))$. In the following experiments, we derive average values of functions h_n and h_p from data collected on several participants and then we evaluate the biases from self-enhancement and decreasing sensitivity using the above formulas.

4 Experiment

4.1 Design

The experiment design has been approved by the committee of ethics from Clermont Auvergne Université (reference number IRB00011540-2020-39). The participants live in France and were recruited online by a specialised company which verifies that they are not bots. The participants receive a series of 4 feedbacks, two positive, two negative, of same intensity in absolute value, starting from different self-evaluations. The main objective is to collect data about the sensitivities to positive and negative feedbacks (functions h_p and h_n in the model) and about the different biases.

4.1.1 Online questionnaire

The questionnaire includes the following steps:

- The participants are requested to assess the size of the coloured surface in the 3 different 2D images. An example of image is shown on figure 2.
- The participants are told that the experimenters can compute exactly their error of surface assessment on these three images and can do the same for a large number of other people who already performed the task. Moreover, the participants are told that the experimenters gathered the errors (G_0 to G_5) from 6 groups of randomly chosen 100 people and that the error of the participant will be compared to the errors of these groups. This comparison provides an evaluation, between 1 and 100, of the participant with respect to the group. We tested two evaluation scales: rank and score which are described further.
- The participants are asked their expected evaluation a_2 in group G_2 . They are requested to express this self-evaluation between their previous expectation a_1 and the feedback f_1 that they just received.
- The same process is repeated again three times, with feedbacks f_2 , f_3 and f_4 that are presented as the evaluation of the participant in groups G_2 , G_3 and G_4 , and requesting the participant's expected evaluations a_3 , a_4 and a_5 in groups G_3 , G_4 and G_5 (interpreted as successive self-evaluations). Actually, each time, the feedbacks are computed as:

$$f_t = a_t \pm \delta \pm \epsilon, \tag{41}$$

where a_t is the expected evaluation of the participant in group G_t given the last feedback f_{t-1} which is (allegedly) their evaluation in group G_{t-1} .

- Finally, the participants are asked if they believed that the feedbacks were really the evaluation of their error with respect of the errors from real groups of 100 persons or if they believed that these feedbacks were manipulated by the experimenters. The participants are requested to rate their belief between 0 (the feedbacks are fake) to 10 (the feedbacks are real). In the following, we call this answer: "trust in feedback" or sometimes simply "trust" of the participant.

The sequence of positive and negative feedbacks is chosen at random in the six possible sequences that contain two positive and two negative feedbacks (see table in Fig: 2). However, in some cases, when the self-evaluation a_t is close to the limit 1 or 100, the chosen feedback would leave the [1,100] interval. In these cases, the feedback is truncated in order to remain in [1,100]. This might lead to some sequences where the positive and negative feedbacks are not balanced. We removed these sequences from the treated results.

Finally, the experiment also includes a questionnaire evaluating the selfesteem of the participants using Rosenberg's scale [17].

4.1.2 Protocol

The main variables of the experimental protocol are:

French Regional Conference on Complex Systems - Le Havre, May 31 June 2, 2023

f_0	f_1	f_2	f_3
+	+	-	-
+	-	+	-
+	-	-	+
-	+	+	-
-	+	-	+
-	-	+	+

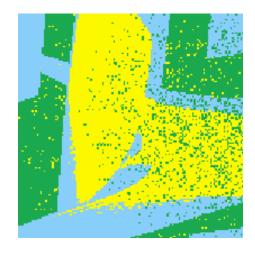


Fig. 2: Left panel: the 6 possible sequences of 4 feedbacks. Right panel: Example of image used in the task. The participants are requested to evaluate the percentage of surface in green on three images similar to this one.

- The anchor (first feedback that influences the first self-evaluation of the participant), which is an integer in [1,100];
- The type of evaluation which can be:
 - Rank: The rank of the participant's performance at the task within the group of 100 persons. 1 is the best rank, 101 is the worst:
 - Score: The number of persons in the group whose performance at the task is worse than the one of the participant. 100 is the best score, 0 is the worst.

The anchor and the evaluation scale (rank or score) are distributed as follows:

- One third of the participants were given a low anchor (randomly chosen in [15, 40]) two third were given a high anchor (randomly chosen in [60, 85]). Indeed, a pilot experiment suggested that sensitivity to feedbacks decreases more strongly when the anchor is high and we wanted to check this on more data.
- For half the participants, the evaluation is by rank and for the other half, the evaluation is by score. Indeed, it seems important to check how the biases depend on the evaluation scale.

4.2 Results

The experiment involves 1509 participants (803 females, 706 males, age between 17 and 79). We removed 155 participants because their series of selfevaluations got too close to 0 or 100 and we could not apply the planned feedback, or because they filled the questionnaire in less than 3 minutes. In total, after these exclusions the data set includes 5416 triples $(a_t^i, f_t^i, a_{t+1}^i)$ for 1354 participants (723 females and 631 males). The results allow us to compute an evaluation of the functions of sensitivity h_p to positive feedbacks, h_n to negative feedbacks and h to all feedbacks by making linear regressions of the self-evaluation change $a_{t+1} - a_t$ as a function of the evaluation a_t .

We compute the sensitivities to feedbacks and the biases when mixing participants and several time steps. The details of the results (not reported here) suggest that the self-evaluation change at the fourth time step is probably less reliable and that there is a temporal trend on the slope of the sensitivity to positive feedbacks, when all values of trust are mixed. Therefore, we focus on mixing times steps 1 and 2 and mixing times steps 1, 2 and 3.

Table 2 shows the slope of the sensitivity to feedbacks c. The following observations seem particularly noticeable:

- The values of c generally remain rather stable between $t \in \{1, 2\}$ and $t \in \{1, 2, 3\}$, especially for the sets mixing all the values of trust.
- All the values of the slope of the sensitivity c are negative except for one subset (low self-esteem and rank evaluation). The values are not high but they are significant in most cases.
- For sets mixing all the values of trust in feedbacks:
 - The slope of sensitivity to feedbacks c is more negative (and more significant) for high self-esteem.
 - The values of c are rather similar for males and females, except for rank evaluation where males show a more negative slope.
 - For high self-esteem and for males, the slope c is more negative for rank than for score.
- For sets of high trust $(T \ge 7)$:
 - The slope c is more negative than for subsets mixing all trust values except for females and score,
 - The slope c is particularly more negative for high self-esteem and rank.
 - For score, males show a more negative slope c than females while the slopes for males and females are similar for rank.

$T \ge 7$	High trust T in feedback
SE > 3	High self-esteem
$SE \leq 3$	Low self-esteem
1:2	Time step $t \in \{1, 2\}$
1:3	Time step $t \in \{1, 2, 3\}$

Table 1: Signification of the abbreviations used in tables.

Now, we focus on the measures of biases described in the method section. Unfortunately, we can compute the total bias and the bias from sensitivity only on sets including complete series of four time steps. Indeed, the definition of the total bias assumes a series of equal number of negative and positive feedbacks. This condition is not guaranteed on the first two time steps and it is never fulfilled on the three first time steps.

-			
	Slope of sensitivit	ty to feedbacks (c)	
ĺ	All	Rank	Score

French Regional Conference on Complex Systems - Le Havre, May 31 June 2, 2023

		All		Rank		Score	
	t	1:2	1:3	1:2	1:3	1:2	1:3
1	411	-0.08^{**}	-0.08^{***}	-0.08^{*}	-0.09^{**}	-0.09^{*}	-0.08^{**}
SE	7 > 3	-0.15^{***}	-0.13^{***}	-0.17^{**}	-0.17^{***}	-0.12^{*}	-0.09^{*}
SE	$C \leq 3$	-0.03	-0.04	0.01	-0.01	-0.06	-0.08 .
Female		-0.07^{*}	-0.08^{**}	-0.05	-0.06	-0.09 .	-0.09^{*}
Male		-0.09^{*}	-0.09^{**}	-0.11 .	-0.11^{*}	-0.08	-0.07
	All	-0.12^{**}	-0.12^{***}	-0.13^{*}	-0.15^{**}	-0.12^{*}	-0.11^{*}
	SE > 3	-0.2^{**}	-0.2^{***}	-0.24^{*}	-0.26^{**}	-0.17^{*}	-0.16^{*}
$T \ge 7$	$SE \leq 3$	-0.06	-0.05	-0.03	-0.05	-0.08	-0.06
	Female	-0.1	-0.09 .	-0.14	-0.14 .	-0.06	-0.06
	Male	-0.14^{**}	-0.15^{**}	-0.13	-0.16^{*}	-0.17^{*}	-0.15^{*}

Table 2: Slope of the sensitivity to feedback linear approximation (c). See the meaning of the variables in the left column in table 1.

Enhancement blas (E)							
		All		Rank		Score	
	t	1:2	1:3	1:2	1:3	1:2	1:3
1	411	0.31	1.09	8.01	8.87	-6.57	-6
SE	T > 3	2.74	4.05	8.02	10.22	-1.66	-1.16
SE	$T \leq 3$	-2.32	-2.03	7.95	7.55	-12.27	-11.45
Female		-3.01	-1.91	6.56	7.48	-11.58	-10.42
Male		4.14	4.58	9.64	10.56	-0.76	-0.88
	All	-1.82	-0.09	7.36	8.69	-9.12	-7.11
	SE > 3	1.62	2.82	8.71	9.43	-3.96	-2.5
$T \geq 7$	$SE \leq 3$	-5.68	-3.4	6.02	7.75	-14.77	-12.16
	Female	-4.56	-2.79	6.87	7.05	-12.87	-10
	Male	1.06	2.77	7.87	10.29	-4.98	-3.88

Enhancement	bias	(E)
-------------	------	-----

All		Rank		Score			
	t	1:2	1:3	1:2	1:3	1:2	1:3
1	411	0.6	0.58	0.21	0.28	0.91	0.87
SE > 3		0.96	0.77	0.88	0.81	0.97	0.72
$SE \leq 3$		0.31	0.4	-0.39	-0.22	0.98	1.02
Female		0.64	0.65	0.08	0.17	1.11	1.13
Male		0.52	0.45	0.36	0.35	0.64	0.53
	All	0.98	0.94	0.68	0.75	1.28	1.15
	SE > 3	1.41	1.41	1.39	1.48	1.41	1.44
$T \ge 7$	$SE \leq 3$	0.67	0.54	-0.05	0.07	1.3	0.99
	Female	0.79	0.8	0.67	0.78	0.9	0.86
	Male	1.02	0.99	0.61	0.72	1.5	1.36

Theoretical sensitivity bias (S')

Table 3: Enhancement and sensitivity biases (E and S'). Both are expressed as a percentage of δ , the absolute value of the difference between the self-evaluation and the feedback. See the meaning of the variables in the left column in table 1.

Table 3, shows the theoretical bias from sensitivity, as a proxy for the sensitivity bias, and the enhancement bias. In the following comments, we simply call the theoretical bias from sensitivity as the bias from sensitivity or the sensitivity bias. The complete paper shows that this approximation is rather reliable (see [10]). Remember that the biases are expressed as a percentage of the absolute value of the feedback intensity. The following observations about the table of enhancement biases (top of table 3) seem noticeable:

- The self-enhancement bias is positive for rank and negative for score in all cases.
- The self-enhancement bias is higher for males and for high self-esteem participants than for females and for low self-esteem participants. The average self-esteem of males (3.08) is only slightly higher than the average self-esteem of females (2.98) and this difference of self-esteem seems insufficient to explain the strong difference of self-enhancement.
- For score, the participants of high trust show a more negative self-enhancement than when mixing all trust values. There is no clear difference for rank between high trust and all values of trust.

In the table of sensitivity biases (bottom of table 3) the following points appear:

- The sensitivity bias S' is in most cases between 0.5% and 1.5%, and is significantly smaller than the absolute value of the enhancement bias.
- The sensitivity bias S' is higher for score and for high trust. For high trust and score, the sensitivity bias of males is higher than the one of females. For all trusts and score, on the contrary, the sensitivity bias of females is higher than the one of males.
- Participants of high self-esteem have a similar high sensitivity bias (about 1.4%) for rank and for score. However, participants of low self-esteem show a much smaller and even negative sensitivity bias for rank evaluation while their sensitivity bias is only moderately smaller than the one of high self-esteem participants for score.

Discussion

Our main objective in this paper is to show an example of theoretical opinion dynamics model leading to an experimental work that would not have been envisaged without the model. Indeed, the presented opinion dynamics model suggests the existence of yet undetected cognitive biases. These biases were identified because their effects were easily observed in long lasting simulations, involving millions of virtual interactions. We could then detect their much smaller effect, which we had initially overlooked, on short simulations. Then understanding these observations requited a significant effort of mathematical analysis.

In this paper, we focus on the positive bias from decreasing sensitivity to feedback. In its simplified version, this bias appears if the reactions to feedbacks decrease when the self-evaluation increases. Again, it seems almost impossible to observe this bias in real life without looking for it using specific measurements on data from a specific experiment.

The main results of the experiment confirm that sensitivities to feedbacks computed from the experimental data decrease when the self-evaluation increases. As expected, we measured a positive bias from sensitivity, its value being around 1% of the feedback intensity. This relatively small value explains the difficulty to detect it. Moreover, this effect is generally combined with the self-enhancement which tends to be higher (in our experiment). Nevertheless, the long term simulations suggest that, in certain conditions, adding the small impact of the bias overtime can lead to very significant effects. This potential cumulative effect as well as the relations between the self-enhancement and the bias from sensitivity are discussed in more details in [10].

The main point that we wish to underline here, is that this work is an example, common in physics but not so much in social sciences, of an initially purely theoretical concept whose existence is confirmed experimentally. Moreover, the opinion dynamics model may suggest other experiments.

Indeed, we also observed a negative bias on the opinion about others in the simulations, which is also due to the decreasing sensitivity to the feedbacks. Again, this negative bias seems yet undetected. We are currently designing experiments in order to detect it.

Moreover, the analysis of the model dynamics shows that the positive and the negative biases combine each other with different outcomes depending on the position of the agents in the hierarchy: for agents located high in the hierarchy, the positive bias dominates and the average opinion about these agents tends to increase. On the contrary, for agents situated low in the hierarchy, especially when there is gossip, the negative bias tends to dominate and the average opinion about these agents tends to decrease. In our view, designing experiments checking these model predictions is another exciting scientific challenge.

Acknowledgements This research has been partially funded by the Agence Nationale de la Recherche in the ToRealSim project (ANR-18-ORAR-0003-01).

References

- 1. S. Galam, The European Physical Journal B 25(4), 403 (2002)
- 2. S. Huet, G. Deffuant, Advances in Complex Systems 13(3), 405 (2010)
- 3. M. Granovetter, American Journal Of Sociology 86(6), 1420 (1978)
- A. Flache, M. Maes, T. Feliciani, E. Chattoe-Brown, G. Deffuant, S. Huet, J. Lorenz, Journal of Artificial Societies and Social Simulation 20(4) (2017)
- 5. R. Axelrod, The Journal of Conflict Resolution 41(2), 203 (1997)
- D. Vilone, A. Vespignani, C. Castellano, Physics of Condensed Matter 30, 399 (2002). DOI 10.1140/epjb/e2002-00395-2
- 7. K. Klemm, V. Eguíluz, R. Toral, M. San-Miguel, Physical Review E 67(4) (2003)
- 8. G. Deffuant, I. Bertazzi, S. Huet, Advances in Complex Systems 21, 1 (2018)
- G. Deffuant, T. Roubin, Physica A: Statistical Mechanics and its Applications 592, 126780 (2022). DOI https://doi.org/10.1016/j.physa.2021.126780. URL https://www.sciencedirect.com/science/article/pii/S0378437121009626
- 10. G. Deffuant, T. Roubin, A. Nugier, S. Guimond, HAL-open science (2022)
- G. Deffuant, T. Carletti, S. Huet, Journal of Artificial Societies and Social Simulation 16(23) (2013)
- M.R. Leary, E. Tambor, S. Terdal, D. Downs, Journal of Personality and Social Psychology 16(11), 76 (2005)
- B. Fischhoff, P. Slovic, S. Lichtenstein, Journal of Experimental Psychology: Human Perception and Performance 3, 552 (1977)

- R. Buchler, D. Griffin, M. Ross, Journal of Personality and Social Psychology 67, 366 (1991)
- R. Vallone, D. Griffin, S. Lin, L. Ross, Journal of Personality and Social Psychology 58, 582 (1990)
- D. Griffin, D. Dunning, I. Ross, Journal of Personality and Social Psychology 59, 1128 (1990)
- 17. E. Vallières, R. Vallerand, International Journal of Psychology 25(2), 305 (1990)

Dynamics & Self-Organization



A toy model for approaching volcanic plumbing systems as com- plex systems
Remy Cazabet, Catherine Annen, Jean-Françcois Moyen and Roberto Weinberg
Reconstruction of variables of interest in nonlinear complex systems: application to a C. elegans biological neural network Nathalie Verdière, Sébastien Orange and Loïs Naudin
POSTER: Agent-based modelling to simulate realistic self- organizing development of the mammalian cerebral cortex Umar Abubacar and Roman Bauer
How to Grasp the Complexity of Self-Organised Robot Swarms? Jérémy Rivière, Aymeric Henard, Etienne Peillard, Sébastien Kubicki and Gilles Coppin
 Analysis of a Network of Hodgkin-Huxley Excitatory and In- hibitory Neurons B. Ambrosio, M.A. Aziz-Alaoui, M. Maama and S.M. Mintchev
Hopf Bifurcation in Oncolytic Therapeutic Model with Viral Lytic Cycle Fatiha Najm, Radouane Yafia and M.A. Aziz Alaoui 136
Poster presentation: Preterm birth indicates higher neural rich club organisation than term counterparts <i>Katherine Birch, Dafnis Batalle and Roman Bauer</i> 140



A toy model for approaching volcanic plumbing systems as complex systems

Rémy Cazabet \cdot Catherine Annen \cdot Jean-François Moyen \cdot Roberto Weinberg \cdot

Magmas form at depth, move upwards and evolve chemically through a combination of processes. Magmatic processes are investigated by means of field work combined with geophysics, geochemistry, analogue and numerical models, and many other approaches (e.g., [1-3]). Volcanism is studied through a combination of monitoring active volcanoes, geophysical imaging, time series of eruptions, as well as investigating eruption products. However, scientists in the field still struggle to understand how the variety of magmatic products arises and there is not consensus yet on models of volcanic plumbing systems. This is because eruptions result from the integration of multiple processes beneath the eruption centre, rooted in the magma source either in the mantle or lower crust that feeds a complex network of magma bodies linking source and volcano.

Power-law relationships describe eruption magnitudes, frequencies, and durations [4]. Such relationships are typical of non-linear and self-organised systems with critical points. A growing body of evidence indicates that several magma chambers connect before and during eruptions [5]. This raises new questions about the role of the magmatic network in triggering eruptions and controlling their magnitude and duration. For example, how is the magnitude of an eruption controlled by magma chamber size and network connectivity? How do such networks focus magma migration to one place? Are volcanic erup-

C. Annen

J.F. Moyen

R. Weinberg

R. Cazabet

Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France

Institute of Geophysics of the Czech Academy of Sciences, Prague, Czech Republic

Université Jean-Monnet, Laboratoire Magmas et Volcans, UCA-CNRS-IRD, F-63170 Aubière

School of Earth, Atmosphere and Environment, Monash University, Clayton, 3800 Victoria, Australia

tions the result of cascading events where perturbation in one magma chamber propagates across the network?

To answer these questions, it is necessary to develop a dynamic model of the volcanic plumbing system as an interconnected network, and discover the laws that rule the behaviour of both the nodes of the networks and their connections.

Model definition In this work, we investigate the potential of the network approach through a prototype of magma pool interaction and magma transfer across the crust. In network terms, it describes a diffusion process on a dynamic spatial network, in which diffusion and network evolution are intertwined: the diffusion affects the network structure, and reciprocally. The diffusion process and network evolution mechanisms come from rules of behaviour derived from rock mechanics and melting processes. Nodes represent magma pools and edges physical connections between them, e.g., dykes or veinlets.

Diffusion process Nodes are described by an individual constant c, their capacity; variable v, the volume of liquid magma inside them; and their pressure p = f(v, c). Edges are described by one constant global parameter w, the volume of magma they can transfer per time unit. Top nodes represent volcanoes, outputs of the system. They have infinite c. The diffusion process is defined as: a) magma is introduced in the system to one node at the bottom, one unit per simulation step, i.e., v = v + 1, b) as an effect of time, the volume of magma in the pools decreases, due to solidification, proportionally to their volume, v, c) when two pools are connected, magma flows between them until their pressure equilibrates.

Network evolution process The network evolution is defined as follows: a) when a critical pressure is reached in one pool, an edge links it to a neighbouring pool, with a preference for those located above (distance on the upward vertical axis is reduced by a fact α , parameter), to account for the role of melt buoyancy, b) when pressure equilibrates between pools, no magma flows, thus the connection solidifies, i.e. the edge is removed. As more magma enters the system from below, pressure in the system increases and more connections are created, and magma migrates upwards, until it exits the system through top nodes (e.g. volcano). The initial distribution of pools (spatial distribution, maximum volume capacity), α , the cooling behaviour, the time allowed for magma migration, are the variables that control system behaviour.

Experiments

Two modes of behaviour are shown in Fig. 1. All parameters are fixed, except the initial distribution of pools and the maximum volume allowed in each node. Despite a linear magma volume input at the bottom of the system, non-linear behaviours emerge: the homogeneous case (Fig. 1a) produces relatively continuous 'volcanic' output (green line on right diagram), and the heterogeneous one (Fig. 1b) leads to cyclic 'volcanic' output. We can track the topology of the network; the activity at each magma pool and connection over time; the average volume of magma as a function of depth, etc.

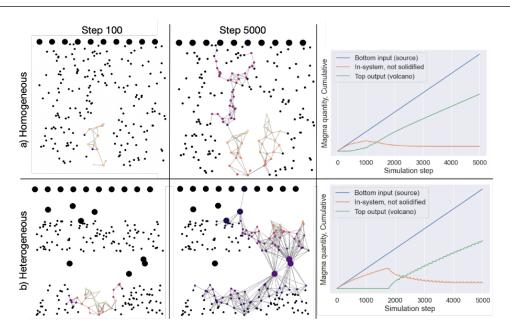


Fig. 1 Example of the toy model outputs with a) a homogeneous and b) a heterogeneous random distribution of magma pools. Node colours represent pressure, using a relative heat scale (from low to high: black, red, orange, yellow). In b) the four horizontal layers are composed alternatively of either many small capacity pools, or a few large capacity pools. Node size corresponds to the capacity of pools c. Note linear versus pulsating volcano output recorded by the green line (right-hand-side diagrams).

Conclusion

This model is a proof of concept where mechanisms are based on analogies: space and time are not realistic, the system is isothermal, etc.

Nevertheless, it succeed in showing that a system governed by the same mechanical rules and fed by a linear input of magma at the bottom, can result in nonlinear behaviors at the surface (e.g., episodic volcanic eruptions).

In future works, we plan to propose a more realistic model, mimicking more closely the natural magmatic plumbing system. We will automatically explore the space of possible node organizations, in order to discover the necessary and sufficient conditions for complex, non-linear behaviors to emerge: are a small number of magma chambers enough? Is the heterogeneity of chambers size more or less important than their physical locations?

References

- H. Schmeling, G. Marquart, R. Weinberg, H. Wallner, Geophysical Journal International 217(1), 422 (2019)
- J.F. Moyen, V. Janouek, O. Laurent, O. Bachmann, J.B. Jacob, F. Farina, P. Fiannacca, A. Villaros, Lithos 402, 106291 (2021)
- R. Sparks, C. Annen, J. Blundy, K. Cashman, A. Rust, M. Jackson, Philosophical Transactions of the Royal society A 377(2139), 20180019 (2019)
- N. Deligne, S. Coles, R. Sparks, Journal of Geophysical Research: Solid Earth 115(B6) (2010)
- G.A. Gualda, M.S. Ghiorso, A.A. Hurst, M.C. Allen, R.W. Bradshaw, Frontiers in Earth Science 10, 798387 (2022)



Reconstruction of variables of interest in nonlinear complex systems: application to a *C. elegans* biological neural network

Nathalie Verdière · Sébastien Orange · Loïs Naudin

Abstract Reconstructing the variables of interest in complex networks is an important process consisting in the capability of inferring the values of the relevant variable of some target nodes from the knowledge of some other node states. However, few approaches deal with the case of networks of nonlinear systems and nonlinear couplings. Our purpose is to present a new one based on specific local relations obtained from equations of each node and to apply it to the study of a biological neural network of *C. elegans*.

Keywords Reconstruction \cdot Variable of interest \cdot Complex networks \cdot Nonlinear couplings \cdot Neurons \cdot *C. elegans*

1 Introduction

The best way to reconstruct the state of variables of interest of target nodes would be to monitor them directly. However, it may be infeasible in practice to measure the state of these nodes due to practical and experimental impediments. Actually, some nodes in the network can provide sufficient information to be able to retrieve the state of other nodes. One of the reconstruction problems aims precisely to determine which of the nodes of the network must be observed to deduce the state of the target nodes.

Methods for reconstructing state variables can be encountered in the literature. For example, the one proposed in [1] is based on observability and seeks to infer the state of all the nodes of the network while the authors in [2] deal with the target reconstruction in the case of linear nodal dynamics and linear couplings.

Loïs Naudin

Nathalie Verdière and Sébastien Orange

LMAH, FR-CNRS-3335, Université Le Havre Normandie,

E-mail: nathalie.verdiere@univ-lehavre.fr, sebastien.orange@univ-lehavre.fr

Sorbonne Université, INSERM, CNRS, Institut de la Vision, F-75012 Paris, France E-mail: lois.naudin@gmail.com

In this work, we propose a new approach to study the reconstruction of relevant variables of nonlinear target nodes coupled with linear and nonlinear terms in complex networks. This approach is based on specific local relations from which propagating reconstruction properties are deduced. These theoretical results constitute the basis of the algorithm TargetReconstruction which returns sets of nodes from which the relevant variables of target nodes can be reconstructed. We apply it to determine sets of neurons from which the *C. elegans* muscle dynamics involved in a chemotaxis behavior [3] can be inferred.

2 Target reconstruction approach

In this work, we focus on nonlinear systems network with nonlinear couplings. The dynamics of each node is governed by a known nonlinear dynamical system. The system corresponding to one node may differ from one node to another. Let denote $X_i = (x_{i,1}, \ldots, x_{i,n})^T$ the state vector of the *i*th node $(i = 1, \ldots, N)$. We assume that its components satisfy a system of ordinary differential equations of the form:

$$\begin{cases}
\dot{x}_{i,1} = f_{i,1}(X_i, \Theta_i) + \sum_{j \in \mathcal{N}_i^-} c_j(x_{i,1}, x_{j,1}) + u_i \\
\dot{x}_{i,2} = f_{i,2}(X_i, \Theta_i), \\
\vdots \\
\dot{x}_{i,n} = f_{i,n}(X_i, \Theta_i),
\end{cases}$$
(1)

where

- $f_{i,j}(X_i, \Theta_i)$ are linear combinations of the state variables $x_{i,2}, \ldots, x_{i,n}$ whose coefficients are analytical functions in the parameter vector Θ_i , the state variable $x_{i,1}$ and its derivatives;
- $\sum_{j \in \mathcal{N}_i^-} c_j(x_{i,1}, x_{j,1})$ is a coupling term with \mathcal{N}_i^- the in-neighbors set of the node *i* (that is, the adjacent nodes whose an edge comes into the node *i*), and $c_j : \mathbb{R}^2 \to \mathbb{R}$ is an infinitely differentiable function such that, for all $x_{i,1} \in \mathbb{R}$, $x_{j,1} \to c_j(x_{i,1}, x_{j,1})$ is a one to one function.

Note that the coupling term involves only the first variables of the nodes which are, in our study, the variables of interest.

The target reconstruction method that will be presented is based on local relations obtained from the differential equations corresponding to one node. These relations involve only the variables of interest of the nodes, the coupling terms and some eventual controls. From them, two reconstruction properties propagating in the network are deduced and constitute the basis of the algorithm TargetReconstruction. The latter determines sets of nodes which should be observed to deduce the state of a given target set of nodes.

3 Target reconstruction of C. elegans muscles

The *C. elegans* worm is a well-known model organism in neuroscience due to its simple nervous system, made up of 302 neurons and about 7000 synaptic connections,

and its fully mapped connectome [4]. Using data from this connectome, together with powerful computational tools, some neuronal networks underlying specific behaviors in *C. elegans* have been cracked. Here, we consider a neuronal network associated with a chemotaxis behavior (Figure 1) [3]. Each neuron is arbitrarily named with three capital letters for convention, and a fourth letter L (left) or R (right) [4]. The coupling between these neurons is determined by the biological connectome. Chemical synapses are nonlinear couplings, while electrical synapses are linear ones.

Furthermore, conductance-based models [5] reproducing the experimental dynamics of individual neurons in the worm have been recently build [6][7]. These models take the form (1) and are coupled between them following the biological connectome. Here, we deal with the target reconstruction of the muscle node, whose activity form the basis of the locomotor behavior. Further, using the algorithm TargetReconstruction, we will determine which neurons in the network should be observed to infer the state of the muscle.

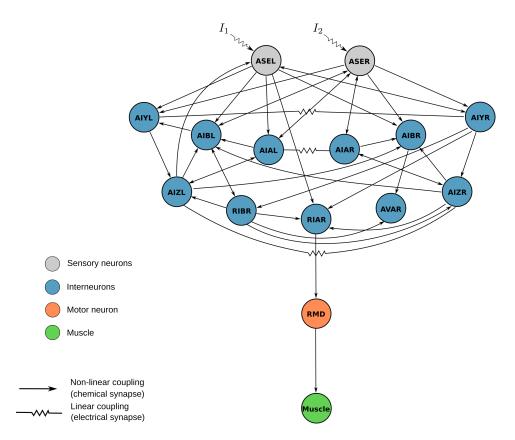


Fig. 1: A neural network underlying a chemotaxis behavior in *C. elegans* [3].

References

1. I. Sendiña-Nadal and C. Letellier, "Observability analysis and state reconstruction for networks of nonlinear systems," *Chaos*, vol. 32, 2022.

- A. N. Montanari, C. Duan, L. A. Aguirre, and A. E. Motter, "Functional observability and target state estimation in large-scale networks," *Proceedings of the National Academy of Sciences*, vol. 119, no. 1, p. e2113750119, 2022.
- A. Costalago-Meruelo, P. Machado, K. Appiah, A. Mujika, P. Leskovsky, R. Alvarez, G. Epelde, and T. M. McGinnity, "Emulation of chemical stimulus triggered head movement in the c. elegans nematode," *Neurocomputing*, vol. 290, pp. 60–73, 2018.
- J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, "The structure of the nervous system of the nematode caenorhabditis elegans: the mind of a worm," *Phil. Trans. R. Soc. Lond*, vol. 314, no. 1, p. 340, 1986.
- 5. A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of physiology*, vol. 117, no. 4, pp. 500–544, 1952.
- L. Naudin, N. Corson, M. Aziz-Alaoui, J. L. Jimenez Laredo, and T. Démare, "On the modeling of the three types of non-spiking neurons of the caenorhabditis elegans," *International Journal of Neural Systems*, vol. 31, no. 02, p. 2050063, 2021.
- L. Naudin, J. L. Jiménez Laredo, Q. Liu, and N. Corson, "Systematic generation of biophysically detailed models with generalization capability for non-spiking neurons," *PloS one*, vol. 17, no. 5, p. e0268380, 2022.



POSTER: Agent-based modelling to simulate realistic self-organizing development of the mammalian cerebral cortex

Umar Abubacar · Dr. Roman Bauer

Abstract The neocortex is a highly complex and intricate structure that plays a crucial role in the cognitive abilities of humans and other vertebrates. Its computational capabilities are enabled by its unique layered cytoarchitecture, consisting of six distinct layers, which are responsible for a vast array of cognitive functions, including perception, attention, memory, and decision-making [1]. However, the developmental mechanisms that govern the formation of the neocortex remain poorly understood [5].

To address this knowledge gap, we employed a computational agent-based model (ABM) in the high-performance simulation platform BioDynaMo to simulate the development of the neocortex starting with a pool of neuroepithelial cells. This platform was chosen for its exceptional performance as an agent-based modelling software, improving single-threaded performance for reference neuroscience models by a factor of 72 when compared to the previous state-of-the-art Cortex3D. This performance gain is even more significant for complex simulations and is further augmented by BioDynaMo's parallelized architecture relative to the single-threaded Cortex3D [2][3]. This unique architecture enables a highly scalable simulation that reduces simulation time by utilizing multiple physical CPU cores, thus enabling modelling of much larger systems than was previously achievable.

Our study aimed to investigate how gene-type rules generate the highly complex structure of the neocortex in a self-organizing manner. Our model was designed to mimic the sequential development of the neocortex through a gene regulatory network (GRN) that sequentially simulates the differentiation and migration of neuroepithelial cells into distinct layers [6]. This cell differen-

U. Abubacar, Dr. Roman Bauer

University of Surrey

Stag Hill

University Campus Guildford

GU2 7XH E-mail: ua00104@surrey.ac.uk

tiation process and its parametrization are derived from [7][1]. The simulated cells formed a naturally occurring layered structure in accordance with the mammalian cortex. We can reproduce the number of neurons in each layer through manipulating the probabilistic differentiation distribution of each cell type in the GRN. Through this process, realistic numbers as observed in specific areas of the human cortex or in other species are reproduced.

Our simulation provides insights into the complex molecular and cellular mechanisms that underlie the formation of the neocortex. It considers the interactions between different types of cells, gradients of signalling molecules, and the physical environment, allowing for highly realistic simulation of the neocortex's development. Our work demonstrates that ABM's with high-performance computer simulations can reproduce crucial aspects of human cortical development and organization, with potential applications in a wide range of problems in computational neuroscience.

Future research will involve optimizing differentiation probabilities using a stochastic tuning regime. Additionally, we plan to simulate various cortical areas, with a particular focus on the auditory and visual cortices. Moreover, we intend to investigate more complex neuron models, and include peripheral cells such as glial cells, oligodendrocytes, and astrocytes in the simulation. The inclusion of these elements may reveal the mechanistic impact of a more diverse environment.

In summary, our simulation of the developing neocortex provides a proofof-concept for the use of computational methods to investigate the mechanisms of brain development. Our work demonstrates that it is possible to use BioDynaMo to create a flexible generic model for generating different layer thicknesses with the correct neuronal numbers. The resulting model, which is species-independent, is scalable and can be refined with further experimental data. We anticipate that this model will be valuable in future studies of the cortex, in the investigation of the healthy and diseased brain [4], as well as in the development of novel *in silico* neural networks for AI tasks.

Keywords self-organising \cdot neuroscience \cdot BioDynaMo \cdot cerebral cortex development \cdot neurogenesis \cdot ABM

Acknowledgements This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) DTP Studentship 2753922 for the University of Surrey.

References

 Roman Bauer, Gavin J. Clowry, and Marcus Kaiser. "Creative Destruction: A Basic Computational Model of Cortical Layer Formation". In: *Cerebral Cortex* 31 (7 July 2021), pp. 3237–3253. ISSN: 14602199. DOI: 10.1093/cercor/bhab003.

- Lukas Breitwieser et al. "BioDynaMo: an agent-based simulation platform for scalable computational biology research". In: (2020). DOI: 10.1101/ 2020.06.08.139949. URL: https://doi.org/10.1101/2020.06.08. 139949.
- [3] Lukas Breitwieser et al. "High-Performance and Scalable Agent-Based Simulation with BioDynaMo". In: (Jan. 2023). DOI: 10.48550/arxiv. 2301.06984. URL: https://arxiv.org/abs/2301.06984.
- [4] Ji Yeoun Lee. "Normal and Disordered Formation of the Cerebral Cortex : Normal Embryology, Related Molecules, Types of Migration, Migration Disorders". In: *Journal of Korean Neurosurgical Society* 62 (3 May 2019), pp. 265–271. ISSN: 2005-3711. DOI: 10.3340/jkns.2019.0098.
- [5] Baptiste Libé-Philippot and Pierre Vanderhaeghen. "Annual Review of Genetics Cellular and Molecular Mechanisms Linking Human Cortical Development and Evolution". In: (2021). DOI: 10.1146/annurev-genet-071719. URL: https://doi.org/10.1146/annurev-genet-071719-.
- [6] Pierre Vanderhaeghen and Franck Polleux. "Developmental mechanisms underlying the evolution of human cortical circuits". In: *Nature Reviews Neuroscience* (Feb. 2023). ISSN: 1471-003X. DOI: 10.1038/s41583-023-00675-z. URL: https://www.nature.com/articles/s41583-023-00675-z.
- [7] Frederic Zubler et al. "An instruction language for self-construction in the context of neural networks". In: *Frontiers in Computational Neuroscience* 5 (Dec. 2011). ISSN: 16625188. DOI: 10.3389/fncom.2011.00057.



How to Grasp the Complexity of Self-Organised Robot Swarms?

Jérémy Rivière · Aymeric Hénard · Étienne Peillard · Sébastien Kubicki · Gilles Coppin

Abstract Robot swarms consist of large numbers of autonomous robots, whose behaviour has been greatly inspired by existing complex biological, physical or chemical systems. This is especially the case for behaviours that involve mechanisms leading to spatial self-organisation of robots. The complex nature of these behaviours prevents a human operator from keeping a mental model of them, which makes it difficult to interact with them, even though this is necessary in certain cases: prediction of a loss of stability, detection of blocking situations, etc. How to allow an operator to grasp the complexity of self-organised robot swarms? This article aims at providing leads to answer this question, by investigating what humans are capable of perceiving of a complex system, and what additional information could be needed to enable them to understand its dynamics and state, and to predict the effects of their control. We first present what an operator is able to perceive from a large number of agents, self-organised or not, through a state of the art of existing works in cognitive sciences, vision and swarm robotics. Secondly, we identify in the literature the different types of information on robot swarms that are transmitted to the operator, with the aim of facilitating his perception and improving his understanding. Finally, we discuss what could be the information needed to build a mental model of the operator, the avenues being explored and the possible challenges to be taken into account.

Keywords Robot Swarm \cdot Self-Organisation \cdot Perception \cdot Interaction

É. Peillard, G. Coppin IMT Atlantique, CNRS, Lab-STICC, 29238 Brest, France S. Kubicki

J. Rivière, A. Hénard

Univ Brest, CNRS, Lab-STICC, 29238 Brest, France E-mail: jriviere@univ-brest.fr

ENIB, CNRS, Lab-STICC, 29238 Brest, France

1 Introduction

Robot swarms are decentralised systems composed of a large number of small, autonomous robots capable of collectively exploring a space, coordinating their movements, or dynamically self-allocating tasks [25,4]. These capacities come from algorithms and models inspired by complex biological, physical or chemical systems, that define the behaviour of a robot according to its state at time t and its perception and action capabilities [25]. The overall behaviour of the swarm emerges from local interactions between these robots. One of the most widely used model is the Reynolds model [22], that allows robots to move in a coordinated way, as flocks of birds or some fish species, through the use of three simple social forces stated as rules: repulsion, attraction and alignment (cf. Figure 1). This coordinated movement (also known as *flocking*) is one of the elementary behaviours of robot swarms, as well as aggregation, pattern formation and area coverage [25, 4]. These elementary behaviours can be identified as the building blocks used to construct higher-level behaviours such as collective exploration, target tracking or surveillance of an area of interest.

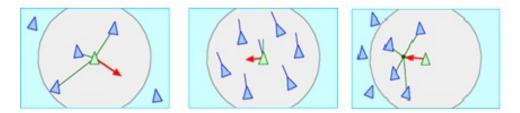


Fig. 1 The three rules of the Reynolds model [22]. From left to right: avoid collision with too close neighbours (repulsion), match heading and speed with neighbours (orientation), stay close to neighbours (attraction).

By their nature, swarms of robots exhibit properties of robustness, self-adaptation, resilience, and scalability that are very useful, but make visualisation, understanding and control by a human operator very difficult [19], especially when the robot's behaviour is unknown, but not only. The very source of these properties, the emergence of the swarm's behaviour through the interactions between the robots, has been preventing the operator from keeping a *mental model* of the swarm behaviour over time [19,23].

A mental model is defined in cognitive sciences as an internal and simplified representation of reality, and is dynamically constructed from observations of this reality [17]. To effectively control a dynamic process, an operator must base his actions on the mental model of this process that allows him to describe, explain and predict its form, function and state [17]. The example of the *Neglect Benevolence* [33], highlighted in robot swarms, is a good illustration of this need: in order to get a swarm to carry out a task as efficiently as possible, it is sometimes necessary to wait until the self-organisation has stabilised *before* exercising control. Figure 2 illustrates this counter-intuitive phenomenon: it is only after the robots have self-organised to the same orientation (b) that sending a coordinated motion command will be most effective in finding targets (c).

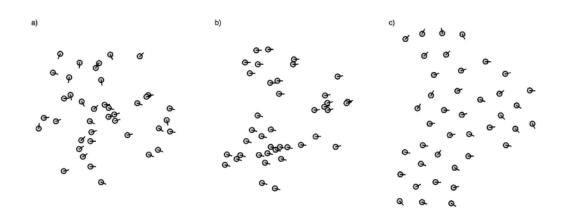


Fig. 2 Robot swarm simulation used in the evaluation of *Neglect Benevolence* (from [33]). At the initial state (a), each robot follows its own randomly generated heading. The middle section (b) shows the robots after they have self-organised to follow the same direction. Section (c) shows the robots after the reception of the operator's command to coordinate their movement.

When interacting with a complex system such as swarms of self-organising robots, however, humans are able to perceive a number of properties exhibited by a set of moving agents, that help them to build mental models. In the following section, we thus start by presenting a state of the art of human perception of swarms in cognitive science, vision and swarm robotics. Section 3 reviews the literature in the field of human-swarm interactions (HSI) to present the different types of information that are transmitted to the operator, in order to enable him to better perceive, understand and even predict the swarm's dynamics. Finally, we discuss in section 4 the information that we believe is necessary to communicate to an operator to enable him to understand and correctly predict the complex dynamics of robot swarms.

2 Human perception of swarms

One of the first properties of human perception that can be highlighted is the ability to identify the movement of one or more discrete elements (agents) in a noisy environment. Thus, humans observing a set of moving agents are able to quickly identify an agent moving in a fixed direction among a set of agents going in random directions [34]. According to the cognitive notion of *visual inertia* [1], human perception seems to prefer to see movement in one direction, rather than movements that change direction abruptly.

Human perception also automatically constructs groupings of agents that form coherent sets. This ability depends on a certain number of properties carried by the agents, that were introduced by the Gestalt theory [35] in cognitive sciences. Among these agent properties that influence this ability to perceive groups, a few are particularly interesting in the context of the perception of dynamic systems (see [29] for more details):

- proximity, *i.e.* the agents that are closest to each other
- similarity, *i.e.* the agents that are more similar to each other (in terms of size, colour, orientation, movement etc.), relatively to other elements of the system

- common fate, *i.e.* the moving agents that are heading toward the same point
- synchrony, *i.e.* the agents that exhibit simultaneously changes (in a broad sense).

All these properties are more or less present in the observed system, and seem to combine in a synergistic way in perception: humans will unconsciously use these cues in such a way as to form groups. This process may go so far as to *abstract* the individual elements of the system from the perception [29].

These properties are particularly interesting because they can be carried by the agents of a complex system, self-organising in space, as robots in swarms. Humans, when interacting with such systems, seem to be able to perceive the swarm (or several swarms) as a *coherent* set of grouped agents, as stated by Seiffert et al. [26]. The authors have shown that humans are able to discriminate more effectively between agents moving in a coordinated way as a school of fish, and agents whose movement lacks the coherence of biological movement, even in the presence of noise in both cases. According to the authors, the human perception of biological flocking is more efficient because it benefits from a more coherent organisation of agents among themselves. In flocking, we clearly find the properties of common fate, similarity and proximity which help in the perception of groups. It is also interesting to note that one result of this study is that individual movements of agents are abstracted by cognitive processes to be replaced by a single global movement of an entity seen as coherent, the swarm [26]. Another study, in the area of robot swarms, was conducted to identify the properties influencing the human perception of a robot swarm cohesion and stability [27]. The results agree with the previous conclusions, and have shown that the perceived cohesion of a swarm of Zooids robots [20] depends on three parameters: the tendencies of the robots to synchronise their movements, to stay in a group, and to follow one of their number. In this study, the robots' behaviour is a coordinated movement of flocking.

To resume, humans are therefore able to perceive one or more groups of agents, acting coherently, within a noisy environment. According to several previous works, humans are also able to categorise different self-organised behaviours of robot swarms, quickly, in the presence of noise or occlusion of part of the swarm members. Thus, Walker et al. [31] have shown that humans are able to identify and differentiate between flocking, aggregation and dispersion behaviours in the presence of noise, in a swarm of 2048 simulated robots. Furthermore, with maximum noise (i.e. agents that do not take in account other agents in their behaviour), humans are able to identify that there is no self-organisation in the swarm. Feedback from participants shows that they rely on properties directly related to Gestalt theory, such as proximity and similarities in orientation and speed, to identify those behaviours. Another work by Harvey et al. [13] studied the perception of different behaviours generated by a simulation of the Reynolds model when one or more rules are omitted. The aim of this study was to evaluate the links between the perception of common direction, the perception of grouping and the perception of a biologically realistic flocking. The results show that the perception of realistic flocking is more often present when the members of the swarm are judged to be grouped but not oriented in the same direction. Moreover, flocking is perceived and

identified as such in the majority of cases, showing the ability of the participants to correctly categorise different behaviours.

3 Augmented Perception through Improved Transparency

If humans are able to perceive and categorise self-organised swarm behaviour, nothing in the literature, to our knowledge, has shown that they are able to explain complex behaviour, or to predict its evolution. These natural perceptions alone are therefore not sufficient to build a mental model that could enable an effective control of swarm.

In order to build a mental model, one of the solutions to bring more knowledge and information to humans is based on learning. The three following learning approaches can be identified [14]. The first approach is model exploration. It consists in implementing the control algorithm oneself and testing the influence of each parameter on the outcome behaviour. Model exploration is time consuming but can lead to the creation of an effective mental model. Bottom-up understanding is the second approach, and involves considering the swarm as a black box that must be analysed. To do this, one must first study the behaviour of one single agent, then add a second agent and study their interactions, then increase the number of agents until the entire swarm can be obtained. Finally, the last approach is embracing complexity: interacting directly with the swarm in its entirety and allowing time for the human to experiment, proceed by trial and error to control it, and study its various reactions. Learning is a relevant and essential method for obtaining knowledge about a dynamical system and evolving its mental model [17]. However, the learning process can be too long, incomplete, and has to be started over again with another system.

Another solution to improve the quality, accuracy and quickness of the establishment of a mental model is to improve the system *transparency*. Transparency can be broadly defined as the means of conveying additional *relevant information* about the system to human [23]. Finding the right level of transparency is challenging: too much transparency, for example, can overload the operator with information, and have negative effects on the usability and understanding of the system. This is particularly true when the system is a swarm made up of several dozen, or even hundreds, of robots (Human Capability Limitations), whose behaviour emerges from the interactions between these robots (Emergent Behaviour), and has limited communication abilities [23].

In the swarm robotics literature, what types of additional information are provided to the operator? First, most works have focused on the transmission of certain atomic information to the operator. Displaying the direction of each agent [6,32,23] makes it possible to visually represent their movement, especially when their speed is too low for their movement to be perceptible. This is usually done by adding an arrow or a line on each agent pointing towards its direction, as in Figure 2, or on each real robot through the use of Augmented Reality [6]. Other atomic information can concern the state of the robots (*e.g.* id, position, type / role [2]) or inter-robot communication (*e.g.* messages sent / received, communication links [18] or textual information such as individual logs [10]).

Global information about the swarm can also be communicated to the operator. This type of information is most often computed from atomic information about the robots, such as the average direction of the swarm or its centre of mass. In [32], the authors evaluate the effects of visualising atomic or global information (see Figure 3), in a task where participants had to predict the final position in space of a robot swarm. The results showed that participants obtained similar low average accuracies, between 38% and 45%, when predicting the final position of flocking or aggregating swarms, independently of the type of information conveyed. The prediction accuracy was better, between 62% and 65%, when the swarm was exhibiting a dispersion behaviour. Both the *centroid* (global information) and the *full information* (atomic information) displays were found significantly better than the two other displays. However, it would have been interesting from our article's perspective to compare these results with the prediction accuracy of participants when non additional information is communicated.

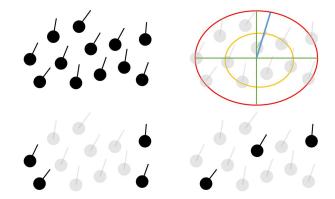


Fig. 3 The display types of the swarm (from [32]): in the top left, the *full information* display, *i.e.* each agent's direction; in the top right, the swarm's centroid (at the intersection of the green cross), along with a bounding ellipse (in red) and another ellipse (in yellow) representing the robot positions median; and in the bottom left and right, only some leaders' direction.

Other examples can be found in the field of teleoperation, where the operator controls the movements of a leader robot to influence the speed and direction of an autonomous swarm (able to avoid obstacles, maintain a formation, etc.). Thus, in [21,9], two global information about the swarm are communicated to the operator: the difference in velocity between the leader robot and the position desired by the operator, and the proximity of obstacles expressed by means of an average of repulsive forces. This information is provided by simple haptic feedback, *i.e.* forces applied to the joystick used by the operator.

In general, such global information can therefore enable the operator to better perceive characteristics of the group as a whole, such as speed, direction, spatial distribution, and so on. Global information can also focus on the swarm dynamic behaviour state. For example, Haas et al. [12] proposed to convey to the operator four indicators about this state, which they called "health": the swarm speed, strength, capability, and dispersion (see Figure 4). It is not specified how these indicators are constructed, but several of them seem to be related to the potential field approach used to implement the swarm behaviour. These indicators are nonetheless interesting, because they communicate certain dynamic aspects of the state of the swarm.



Fig. 4 Control and supervision display for an operator, from [12]. The four indicators of the swarm's health are in the bottom left. Their value is conveyed by status lights (green, yellow, red).

Finally, an important role of a mental model is the ability to predict, in the more or less short term, the evolution of the system. As detailed above, Walker et al. [32] have studied the effect of different types of visualisation on the prediction accuracy. In another article [33], these authors also proposed to display, as an additional information, the future position of agents, and to evaluate this display efficiency in the context of *Neglect benevolence*. The future position of each agent corresponds to its position twenty seconds later: it is automatically calculated from its orientation and speed, to be displayed by transparency in the interface. This display has been shown to be useful to counterbalance the potential latency of the control transmitted by the operator. However, the effects of direct interactions between agents do not seem to be taken into account in the calculation of their future position, and this type of visualisation does not seem to be generalisable, especially in real situations.

4 Discussion

To summarise, in the literature, two types of information are communicated to the operator in order to improve the swarm transparency: atomic, local information, coming from the robots composing the swarm, and global information, concerning the characteristics of the swarm. Information at these two levels is in our opinion

not sufficient to complete the perception of an operator in order to allow him to create a mental model of the swarm's behaviour.

In this section, we discuss the following hypothesis, already put forward by Kolling et al. [19] in 2016, but that has been not much taken up to our knowledge:

H1: The swarm models (e.g. bio-inspired models) may offer suitable metaphors to facilitate the understanding of swarm dynamics as well as the impact of control inputs.

This hypothesis means in our view that the relevant type of information to be communicated to the operator is on the level of the swarm behaviour selforganisation mechanisms, *i.e.* the basic elements participating in self-organisation, such as attraction or random movement.

Let us take the famous example of ants and their ability to self-organise to efficiently bring back resources to the nest by communicating through the environment with pheromones (stigmergy). With the right perspective (typically a view from above, as in many multi-agent simulation platforms), humans should be able to perceive and identify the movements of groups of ants between the food and the nest, and by completion the paths and patterns formed by the movements of these groups. Occasionally, they could also be able to account for the dynamics of these paths at time t, based on the past evolution of the number of ants travelling on them. However, without prior knowledge on stigmergy and pheromones, it should be difficult to *explain* how groups are formed and how patterns appear, and to *predict* the evolution of these dynamics: the disappearance or enlargement of a path, its stability, the emergence of a new path between the nest and a food source, etc.

The aforementioned hypothesis H1 could take the form, in this particular example, of displaying the pheromones deposited by the ants, to obtain additional information that could allow a faster and more complete construction of the observer's mental model. Indeed, the paths that human perception completed based on the movements of ants are now explicitly visible thanks to the visualisation of pheromones, as shown by Figure 5. For the prediction of the path dynamics, it also becomes possible to observe the increase / decrease of the quantity of pheromones *before* the appearance / disappearance of the ant column. Finally, in this specific case, the role of pheromones in self-organisation, highlighted by the visualisation and inferred from the mental model, could allow the operator to consider their use as a mean of control: depositing pheromones in the environment to influence the formation of paths to designated food sources, for example.

With robot swarms exhibiting spatial self-organising behaviour, following H1 requires:

- 1. to identify the self-organisation mechanisms in each elementary behaviour;
- 2. to convey to the operator their dynamics during the behaviour execution of the robots.

With regard to the first point, Brambilla et al. [4] have listed for each of the elementary behaviours their source of inspiration, as well as the models used to implement them. These inspirations are real biological (*e.g.* colonies of bees, ants, birds, fish), physical or chemical (*e.g.* forces, crystal patterns, molecules)

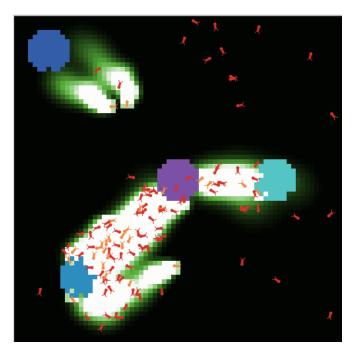


Fig. 5 Screen capture of a simulation run of the NetLogo Ants model. The pheromone concentration is shown in a green-to-white gradient, from a top view.

systems that all have the particularity of being complex: composed of a large number of simple interacting parts, with limited communication between these parts, and with no central control or leader, they show emergent capacities for self-organisation and self-adaptation. These capacities come from mechanisms that are abstracted by the concrete implementation of these behaviours, mainly based on finite state machines or virtual forces [4].

A preliminary work [15] allowed us to identify the spatial self-organisation mechanisms of these behaviours, based on direct interactions between individuals [11] and involving positive and negative feedback loops [28, 16, 3]. Different mechanisms, used together, form several self-organisation methods. One method can lead to the appearance of one or more collective behaviours, and one behaviour can be obtained through one or more method. The Sankey diagram in Figure 6 summarises these relations between methods and behaviours.

For example, the Reynolds model [22] is based on the "Attraction, Alignment and Repulsion" method, whose main self-organisation mechanisms are:

- attraction and repulsion, that form two feedback loops balancing the distance between agents and allowing them to remain aggregated;
- the alignment of speeds and directions of aggregated agents;
- and random movement, that allows the agents to encounter each other by exploring the space, and that gives the direction followed by the swarm when the agents are aggregated.

This method not only can lead to coordinated movement, but also to aggregation, area coverage and pattern formation.

Regarding the second point, constructing indicators of these mechanism dynamics and finding suitable metaphors to visualise them would constitute an im-

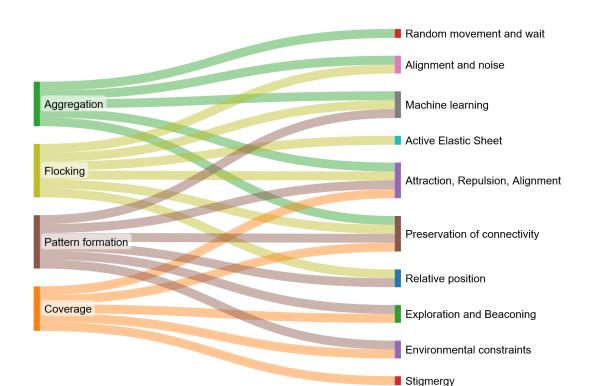


Fig. 6 Sankey diagram of four emergent robot swarms' elementary behaviours (left) that can be obtained through the use of self-organisation methods (right). From [15].

portant challenge. The data communicated by the robots in real time (mainly coming from their interactions and behaviour) are the basis for the construction of these indicators: we can only follow the evolution of the mechanisms if we know what is happening at the *micro* level. The challenges linked to the recovery of these data and their communication concern [19, 10]:

- the possible incompleteness of the received data (bandwidth, latency, sensor errors);
- a risk of cognitive overload for the operator;
- the need to relate this information to the swarm collective.

On these issues, the main conclusion of the literature is that it may be necessary to summarise in the display the state of the swarm from aggregated, merged and noise-reduced local information. Thus, one solution could be to collect, through communication with the robots and their tracking, data from the *micro* level, and to pass these data through a process of aggregation, fusion and noise reduction, with the aim of giving the operator intuitive real-time information on the behaviour of the swarm in terms of mechanisms and interactions. The concept of "Macroscope", proposed by de Rosnay [7], was already built on this principle: "the Macroscope filters the details, amplifies what connects, brings out what brings together, to observe the infinitely complex".

Another solution could be inspired by, for example, the work of Escobedo et al. [8]: identify and characterise the interactions between individual robots from these data, and develop a mathematical model of the mechanisms. Finding ways

to convey indicators of these mechanisms in real-time, through an intuitive visualisation, remains an open challenge.

The proposal developed in this discussion, to identify the complex mechanisms of self-organisation, and then to retrieve from the *micro* level their state in real time to update a visualisation of these mechanisms, is still at an embryonic stage. Its purpose is first of all to initiate a reflection and a discussion on its soundness, the expected challenges and outcomes. This proposal is made here in the context of swarm robotics, but it goes beyond this framework and could be generalised to many complex artificial systems of human design.

Acknowledgements This work is part of the ANR ARTUISIS¹ project, which received funding under the reference ANR-21-CE33-0006.

References

- 1. Anstis, S., Ramachandran, V.S. (1987). Visual inertia in apparent motion. Vision Research, vol. 27, no 5, p. 755-764.
- 2. Batra, S., Klingner, J., Correll N. (2019). Augmented Reality for Human-Swarm Interaction in a Swarm-Robotic Chemistry Simulation. preprint.
- 3. Bonabeau, E., Dorigo, M., Theraulaz, G. (1999). Swarm Intelligence : From Natural to Artificial Systems. Oxford University Press, Inc., USA.
- 4. Brambilla, M., Ferrante, E., Birattari, M., Dorigo, M. (2013). Swarm robotics: a review from the swarm engineering perspective. Swarm Intelligence, vol. 7, no 1, p. 1-41.
- 5. Buchmüller, J., Jäckle, D., Cakmak, E., Brandes, U., Keim, D. A. (2019). *MotionRugs : Visualizing Collective Trends in Space and Time*, in: IEEE Transactions on Visualization and Computer Graphics, vol. 25, no 1, p. 76-86.
- Daily, M., Cho, Y., Martin, K., Payton, D. (2003). World embedded interfaces for humanrobot interaction. In Proceedings of the IEEE 36th Annual Hawaii International Conference on System Sciences.
- 7. De Rosnay, J. (1975). Le macroscope : vers une version globale. Editions du Seuil.
- Escobedo, R., Lecheval, V., Papaspyros, V., Bonnet, F., Mondada, F., Sire, C., Theraulaz, G. (2020). A data-driven method for reconstructing and modelling social interactions in moving animal groups. Philosophical Transactions of the Royal Society B, vol.375, no 1807, 20190380.
- Franchi, A., Robuffo Giordano, P., Secchi, C., Son H. I., Bülthoff, H. H. (2011). A passivitybased decentralized approach for the bilateral teleoperation of a group of UAVs with switching topology, IEEE International Conference on Robotics and Automation, Shanghai, China, p. 898-905
- Ghiringhelli, F., Guzzi, J., Di Caro, G. A., Caglioti, V., Gambardella, L. M., Giusti A. (2014). Interactive augmented reality for understanding and analyzing multi-robot systems in IEEE/RSJ International Conference on Intelligent Robots and Systems, p. 1195–1201.
- 11. Gorodetskii, V.I. (2012). Self-organization and multiagent systems: I. Models of multiagent self-organization. J. Comput. Syst. Sci. Int. 51, p. 256–281.
- 12. Haas, E., Fields, MA., Hill, S., Stachowiak, C. (2009). Extreme Scalability Designing Interfaces and Algorithms for Soldier-Robotic Swarm Interaction. Army Research Laboratory, public report.
- 13. Harvey, J., Merrick, Kathryn E., Abbass Hussein A. (2018). Assessing Human Judgment of Computationally Generated Swarming Behavior. Frontiers in Robotics and AI, vol. 5.
- 14. Hasbach, Jonas D., Witte, Thomas E.F., Bennewitz, M. (2020). On the importance of adaptive operator training in human-swarm interaction. International Conference on Human-Computer Interaction, Springer, p. 311-329.
- 15. Hénard, A., Rivière, J., Peillard, É., Kubicki, S., Coppin, G. (2023). A Unifying Methodbased Classification of Robot Swarm Spatial Self-Organisation Behaviours. Adaptive Behaviour, in press.

¹ https://siia.univ-brest.fr/artuisis/

- 16. Heylighen, F. (2008). *Complexity and Self-organization*. Encyclopedia of Library and Information Sciences.
- 17. Jones, N. A., Ross, H., Lynam, T., Perez, P., Leitch., A. (2011). Mental models: an interdisciplinary synthesis of theory and methods. Ecology and Society, vol. 16, no 1.
- Kolling, A., Sycara, K., Nunnally, S., Lewis, M. (2013). Human-swarm interaction: an experimental study of two types of interaction with foraging swarms. J. Hum.-Robot Interact. 2, 2 (June 2013), 103–129.
- Kolling, A., Walker, P., Chakraborty, N., Sycara, K., Lewis, M. (2016). Human Interaction With Robot Swarms: A Survey in IEEE Transactions on Human-Machine Systems, vol. 46, no. 1, p. 9-26.
- Le Goc, M., Kim, Lawrence H., Parsaei, A., Fekete, J-D., Dragicevic, P., Follmer, S. (2016). Zooids: Building blocks for swarm user interfaces. In : Proceedings of the 29th annual symposium on user interface software and technology, p. 97-109.
- Nunnally, S., Walker, P., Lewis, M., Chakraborty, N., Sycara, K. (2013). Using Haptic Feedback in Human Robotic Swarms Interaction. In proceedings of the Human Factors and Ergonomics Society Annual Meeting. 57(1), p. 1047-1051.
- 22. Reynolds, Craig W. (1987). Flocks, herds and schools: A distributed behavioral model, ACM SIGGRAPH Computer Graphics, vol. 21, no 4, p. 25-34.
- Roundtree, KA, Goodrich, MA, Adams JA. (2019). Transparency: Transitioning From Human-Machine Systems to Human-Swarm Systems. Journal of Cognitive Engineering and Decision Making, 13(3), p. 171-195.
- 24. Saint-Martin, F. (2011). La théorie de la gestalt et l'art visuel: essai sur les fondements de la sémiotique visuelle. Presses de l'Université du Québec.
- 25. Schranz, M., Umlauft, M., Sende, M., Elmenreich, W. (2020). Swarm Robotic Behaviors and Current Applications. Frontiers in Robotics and AI, 7, art. no. 36.
- 26. Seiffert, A., Hayes, S., Harriott, C., Adams, J. (2015). *Motion perception of biological swarms*. In proceedings of the 37th Annual Meeting of the Cognitive Science Society.
- St-Onge, D., Levillain, F., Zibetti, E., et al. (2019). Collective expression: how robotic swarms convey information with group motion. Paladyn, Journal of Behavioral Robotics, vol. 10, no 1, p. 418-435.
- Sumpter, D.J.T (2006). The principles of collective animal behaviour. Phil. Trans. R. Soc. B 361:5–22
- Wagemans, J., Elder, James H., Kubovy, M., et al. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. Psychological bulletin, vol. 138, no 6, p. 1172.
- Wagemans, J., Feldman, J., Gepshtein, S., et al. (2012) A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. Psychological bulletin, vol. 138, no 6, p. 1218.
- Walker, P., Lewis, M., Sycara, K. (2016). Characterizing human perception of emergent swarm behaviors, in 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), p. 2436-2441.
- Walker, P., Lewis, M., Sycara, K. (2016). The effect of display type on operator prediction of future swarm states, in 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), p. 2521-2526.
- 33. Walker, P., Nunnaly, S., Lewis, M., et al. (2012). Neglect benevolence in human control of swarms in the presence of latency. In : IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, p. 3009-3014
- 34. Watamaniuk, Scott N.J., McKee, Suzanne P., Grzywacz, Norberto M. (1995). Detecting a trajectory embedded in random-direction motion noise. Vision Research, vol. 35, no 1, p. 65-77.
- 35. Wertheimer, M. (1922). Untersuchungen zur Lehre von der Gestalt, I: Prinzipielle Bemerkungen. Psychologische Forschung, 1, 47-58. (Translated extract reprinted as "The general theoretical situation." In W. D. Ellis (Ed.), (1938). A source book of Gestalt psychology (p. 12-16). London, U. K.: Routledge & Kegan Paul Ltd.).



Analysis of a Network of Hodgkin-Huxley Excitatory and Inhibitory Neurons

B. Ambrosio \cdot M.A. Aziz-Alaoui \cdot M. Maama \cdot S. M. Mintchev

Abstract This articles deals with the analysis of the dynamics of a network of Hodgkin-Huxley (HH) ordinary differential equations (ODEs) inspired by the visual cortex V1. The model includes a stochastic drive for each neuron and recurrent inputs coming from the network activity from which emergent properties arise.

Keywords Complex Systems \cdot Dynamical Systems \cdot Emergent Properties \cdot Neuroscience \cdot Visual Cortex \cdot Synchronization \cdot Hodgkin-Huxley

This article deals with an analysis of the dynamics of a network of Hodgkin-Huxley (HH) ordinary differential equations (ODEs). The network topology is of random type and inspired by the visual cortex V1. The model includes a stochastic drive for each neuron and recurrent inputs coming from the network activity. The equation for the network reads as

$$\begin{cases} CV_{it} = \overline{g}_{Na}m_{i}^{3}h_{i}(E_{Na} - V_{i}) + \overline{g}_{K}n_{i}^{4}(E_{K} - V_{i}) + \overline{g}_{L}(E_{L} - V_{i}) \\ +g_{Ei}(E_{E} - V_{i}) + g_{Ii}(E_{I} - V_{i}) \\ n_{it} = \alpha_{n}(V_{i})(1 - n_{i}) - \beta_{n}(V_{i})n_{i} \\ m_{it} = \alpha_{m}(V_{i})(1 - m_{i}) - \beta_{m}(V_{i})m_{i} \\ h_{it} = \alpha_{h}(V_{i})(1 - h_{i}) - \beta_{h}(V_{i})h_{i} \\ \tau_{E}g_{Eit} = -g_{Ei} + S^{dr} \sum_{s \in \mathcal{D}(i)} \delta(t - s) + S^{QE} \sum_{j \in \Gamma_{E}(i), s \in \mathcal{N}(j)} \delta(t - s) \\ \tau_{I}g_{Iit} = -g_{Ii} + S^{QI} \sum_{j \in \Gamma_{I}(i), s \in \mathcal{N}(j)} \delta(t - s) \end{cases}$$
(1)

where Q is equal to E for E-neurons and I for I-neurons. The first four equations in (1) correspond to the standard HH-equations; the variable V stands for the voltage, and m, n, h are gating variables for ionic fluxes. We refer to [1-4] for textbooks presenting this classical model. Now,

B. Ambrosio

Normandie Univ, UNIHAVRE, LMAH, FR-CNRS-3335, ISCN, 25 rue Philippe Lebon, Le Havre 76600, France

The Hudson School of Mathematics, 244 Fifth Avenue, Suite Q224, New York, 10001, New York, USA

E-mail: benjamin.ambrosio @univ-lehavre.fr

for each node in the network, we add two equations to the original HH ODE system. Those are the equations which contain coupling terms inputs coming from:

- 1. the network (presynaptic E and I-neurons)
- 2. a stochastic input drive, only for g_E .

The two variables g_E and g_I stand also here for gating variables and are added as such to the first equation. As a result of the values set for E_E and E_I , an increase of g_E induces excitation whereas an increase of g_I induces inhibition. It is important to note here that network models of ODEs can be used to address typical questions prevalent in the physiological literature such as to what extent are the responses of neuronal ensembles shaped by feedforward vs. recurrent circuitry? A good example is provided by the neurons in layer IV of the primary visual cortex, which receive sparse afferent inputs from the lateral geniculate nucleus (LGN) concurrent with abundant inputs from other cortical neurons; see [5]. This idea of leveraging the interplay between feed-forward and recurrent connection has indeed been used successfully in a series of recent papers aiming to describe the electrical activity of the visual cortex V1, see [6] and references therein cited for a global depiction of the brain as a complex network with a particular focus on V1. One of the main differences with the description above is that in those references the single neurons are represented by leaky integrate and fire equations.

Below, we provide an overview of results concerning the dynamics of the network; we illustrate therein some numerical simulations of the network and discuss the emergence of organized behavior subject to the variation of appropriate parameters. For a more detailed analysis and description of the phenomena, we refer to [7]

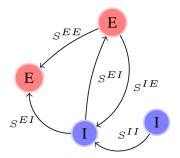


Fig. 1 Schematic representation of the coupling in the network. Each E neuron receives kicks with coupling strength S^{EE} from E neurons, coupling strength S^{EI} from I neurons. Each I neuron receives kicks with coupling strength S^{IE} from E neurons, coupling strength S^{II} from I neurons.

Results

We considered a network of N = 500 neurons with Ne = 375 E neurons and Ni = 125 I neurons. We explored the parameter space around

$$S^{EI} = S^{IE} = S^{II} = S^{EE} = 0.01,$$

where we took in turn varying each of these parameters separately within the range [0.002; 0.03]. Among these four parameters, S^{EE} appeared to be the most effective for establishing a path from stochastic homogeneity to synchronization in the network. Our aim was to illustrate how variation of the model parameters led to emergent properties in the network. Our investigation allowed to highlight the following emergent phenomena:

- 1. a path from homogeneity to partial synchronization and synchronization;
- 2. the correlation between g_E and g_I ;
- 3. emergence of the γ -rhythm at some point the network has its own rhythm of oscillation consistent with the so called gamma frequency, which may be different from individual neuronal rhythms;
- 4. variations in the mean spiking rates of E and I neurons that result from changes to the parameters.

Figure 3 captures a significant proportion of the observed emergent phenomena. Overall, our study identifies an original path in the parameters space to reach synchronization and gamma rhythms. It relies on a detailed analysis of the mechanisms leading to these emergent phenomena.

References

- C. Borgers, An Introduction to Modeling Neuronal Dynamics (Springer, Springer Nature, 2017)
- J. Cronin, Mathematical aspects of Hodgkin-Huxley neural theory (Cambridge University Press, Cambridge Cambridgeshire New York, 1987)

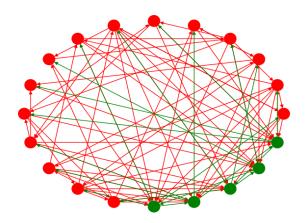
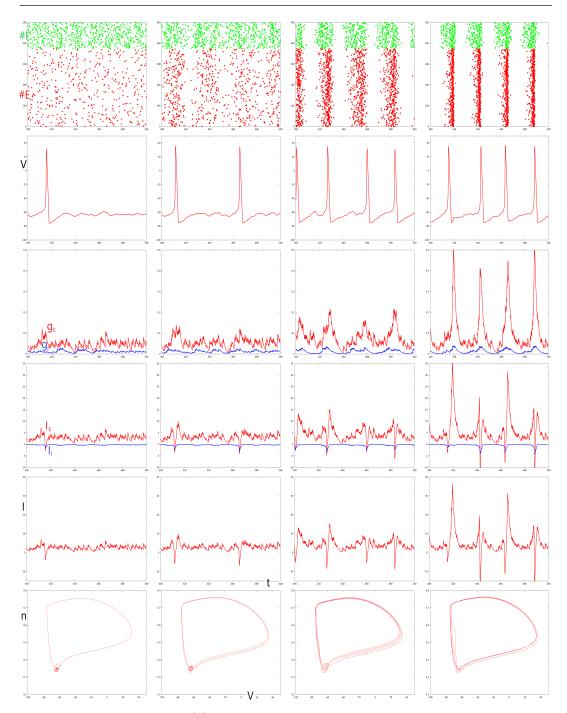


Fig. 2 Illustration of a network with fifteen E neurons and five I neurons with a similar topology than the network considered in the article (the network considered in the article is 25 times larger than the network represented here).

- G.B. Ermentrout, D.H. Terman, Mathematical Foundations of Neuroscience (Springer, New York, 2010). DOI 10.1007/978-0-387-87708-2. URL https://doi.org/10.1007/978-0-387-87708-2
- E.M. Izhikevich, Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting (Computational Neuroscience) (The MIT Press, Cambridge, MA, 2006). URL https://www.xarg.org/ref/a/0262090430/
- P.E. Tibbetts, Principles of Neural Science edited by Eric R. Kandel, James H. Schwartz, Thomas M. Jessell, Steven A. Siegelbaum, A. J. Hudspeth, and Sarah Mack, vol. 88 (2013). DOI 10.1086/670559. URL https://doi.org/10.1086/670559
- L.S. Young, Journal of Statistical Physics 180(1-6), 612 (2020). DOI 10.1007/s10955-019-02483-1. URL https://doi.org/10.1007/s10955-019-02483-1
- M. Maama, B. Ambrosio, M.A. Aziz-Alaoui, S.M. Mintchev. Emergent properties in a v1inspired network of hodgkin-huxley neurons (2020). DOI 10.48550/ARXIV.2004.10656. URL https://arxiv.org/abs/2004.10656



French Regional Conference on Complex Systems - Le Havre, May 31 June 2, 2023

Fig. 3 Simulation of system (1). This figure illustrates a path from random homogeneity to synchronization as the parameter S^{EE} is increased. In this picture, the parameters S^{II}, S^{EI} , and S^{IE} are set to 0.01 and each column from left to right corresponds to a specific value of S^{EE} ; respectively: $S^{EE} = 0.01, 0.017, 0.02$ and 0.03. The first row illustrates rasterplots: each time a specific neuron spikes, a dot is plotted along the y-axis. On the top of the figure, in green, the *I*-neurons are plotted. *E*-neurons are plotted in red below. On the left side, we observe a state where the dots do not to seem to concentrate around a specific time. We call it random homogeneous activity. Increasing S^{EE} induces synchronization. The second row represents the potential V_1 of neuron #1 as a function of a time. The third row, the *E*-conductance g_E in red and the *I*-conductance g_I in blue for the same neuron. The fourth row illustrates the *E*-current denoted by I_E in red and the *I*-current denoted by I_I in blue. The fifth row illustrates the sum of *E* and *I* currents which plays the role of I(t) in a single HH equation. The last row illustrates the projection of the trajectory of neuron #1 in the (V, n) phase space.



Hopf Bifurcation in Oncolytic Therapeutic Model with Viral Lytic Cycle

Fatiha Naj
m $\,\cdot\,$ Radouane Yafia $\,\cdot\,$ M. A. Aziz Alaoui

Abstract In this paper, we propose a delayed mathematical model describing oncolytic virotherapy treatment of a tumour that proliferates according to the logistic growth function and incorporating viral lytic cycle. The tumour population cells is divided into uninfected and infected cells sub-populations and the virus spreading is supposed to be in a direct mode (i.e. from cell to cell). Depending on the time delay, we analyze the positivity and boundedness of solutions and the stability of tumour uninfected-infected equilibrium (UIE) is established. We prove that, delay can lead to "Jeff's phenomenon" observed in laboratory which causes oscillations in tumour size whose phase and period changes over time. Some numerical simulations are carried out to illustrate our theoretical results.

Keywords Anti tumour virus \cdot delay differential equation \cdot Jeff's phenomenon \cdot Hopf bifurcation

1 Mathematical model

it is known that, in oncolytic virotherapy, the mode of transmission of virus infection is an important factor that specifies the treatment efficacy [1]. We suppose that the spread of virus into the tumour site by a direct transmission (cell to cell) and the tumour cells grow following the logistic low at a rate r (for uninfected cells) and s (for infected cells). The maximum sizes of the

F. Najm \cdot R. Yafia

Department of Mathematics Faculty of Sciences Ibn Tofail University Campus Universitaire, BP 133, Kénitra, Morocco

E-mail: fatiha.najm@uit.ac.ma, radouane.yafia@uit.ac.ma

M. A. Aziz Alaoui

Normandie Univ, France; ULH, LMAH F-76600 Le Havre; FR-CNRS-3335, ISCN 25 rue Ph. Lebon, 76600 Le Havre, France

E-mail: aziz.alaoui @univ-lehavre.fr

two tumour populations u and v are given by the same carrying capacity k. β is the spread rate of the virus into the tumour site. Infected tumour cells population is killed by the virus at a rate a. τ is the viral lytic cycle. $e^{-d\tau}$ models the survival function. The mathematical model is given by (see [2])

$$\begin{cases} \frac{du(t)}{dt} = ru(t)(1 - \frac{N(t)}{k}) - du(t) - \beta u(t)v(t) \\ \frac{dv(t)}{dt} = \beta e^{-d\tau}u(t - \tau)v(t - \tau) + sv(t)(1 - \frac{N(t)}{k}) - av(t) \\ N = u + v \\ u(s) = \varphi(s) \ge 0, v(s) = \psi(s) \ge 0, s \in [-\tau, 0]. \end{cases}$$
(1)

u:= Tumor cells that are not infected by the virus, v:=Tumor cells that are infected by the virus, N = u + v:=Total number of cells in the tumor micro-environment.

2 Properties of solutions and steady states

Proposition 1 Let $\varphi(0) > 0$ and $\psi(0) > 0$, then there exist a constant $\sigma > 0$, for $t \in [0, \sigma[$, such that

i) All solutions of system (1) with positive initial conditions uniquely exist and are positive.

ii) $\lim_{t \to +\infty} \sup u(t) \le k \text{ and } \limsup_{t \to +\infty} v(t) \le L, \text{ where } L = \frac{k(r+a)^2}{4ra} + \frac{kse^{d\tau}}{4a}.$ *iii)* $\liminf_{t \to +\infty} u(t) \ge M, \text{ where } M = k\left(1 - \frac{Lr + k(d + \beta L)}{kr}\right), \text{ with } kr > Lr + k(d + \beta L).$

Theorem 1 The solution of system (1) with positive initial condition is existent, unique, positive and bounded on $[0, +\infty)$ and $\Upsilon = \{(\varphi(q), \psi(q)) \in \mathcal{C}\}$ $M \in \varphi(q) \leq h, 0 \leq \psi(q) \leq L\}$ is positively invariant

 $\Upsilon = \{(\varphi(s), \psi(s)) \in \mathcal{C} \setminus M \le \varphi(s) \le k, 0 \le \psi(s) \le L\}$ is positively invariant set for system (1).

Let $R_1 = \beta k(a-s) + ar - sd$, $R(\tau) = \beta (e^{-d\tau}(r+\beta k) - s)$, $R_2(\tau) = \beta ke^{-d\tau}(r-d) - ar + sd$.

The possible steady states are given by $E_0 = (0,0)$ (Tumour free equilibrium TFE), $E_1 = (u_1,0) = (\frac{k}{r}(r-d),0)$ (Infected free equilibrium IFE), $E_2 = (0,v_2) = (0,\frac{k(s-a)}{s})$ (Uninfected free equilibrium UFE), $E^*(\tau) = (u^*(\tau), v^*(\tau)) = (\frac{R_1}{R(\tau)}, \frac{R_2(\tau)}{R(\tau)})$ (Uninfected-Infected equilibrium UIE). Define $\overline{\tau} = \frac{1}{d} \ln \left(\frac{r+\beta k}{s}\right)$, $\widehat{\tau} = \frac{1}{d} \ln \left(\frac{\beta k(r-d)}{ar-sd}\right)$ and $\tau_{min} = \min(\overline{\tau}, \widehat{\tau})$, $\tau_{max} = \max(\overline{\tau}, \widehat{\tau})$. Let the hypotheses: $(\mathbf{H})_0 : \frac{r+\beta k}{s} > 1$; $(\mathbf{H})_1 : \frac{\beta k(r-d)}{ar-sd} > 1$; $(\mathbf{H})_2 : 0 < \tau < \tau_{min}$ and $\beta k(a-s) > sd - ar$; $(\mathbf{H})_3 : \tau > \tau_{max}$ and $\beta k(a-s) < sd - ar$. Note that, the hypotheses $(\mathbf{H})_0$ and $(\mathbf{H})_1$ guarantee the existing the set of t

 $s_1 < s_2 - ar$. Note that, the hypotheses $(\mathbf{H})_0$ and $(\mathbf{H})_1$ guarantee the existence and positivity of $\overline{\tau}$ and $\widehat{\tau}$ respectively. The hypotheses $(\mathbf{H})_2$ and $(\mathbf{H})_3$ guarantee the positivity and non-positivity respectively of R, $R(\tau)$ and $R_1(\tau)$.

3 Occurrence of Hopf bifurcation at UIE equilibrium

The characteristic equation associated to $E^*(\tau)$ is given as follows

$$\Delta_1(\lambda,\tau) = D_1(\lambda,\tau) + D_2(\lambda,\tau)e^{-\lambda\tau}$$
(2)

where $A(\tau) = \frac{r}{k}u^* + \beta e^{-d\tau}u^* + \frac{s}{k}v^*$, $B(\tau) = -u^*(\frac{r}{k} + \beta)\frac{s}{k}v^*$, $C(\tau) = -\beta e^{-d\tau}u^*$, $D(\tau) = (\frac{r}{k} + \beta)u^*\beta e^{-d\tau}v^*$ and has a purely imaginary roots defined by $w_{\pm}^2 = \frac{1}{2}\left(-\chi_1(\tau) \pm \sqrt{\delta(\tau)}\right)$, where $\chi_1(\tau) = A^2(\tau) - 2B(\tau) - C^2(\tau)$, $\chi_2(\tau) = B^2(\tau) - D^2(\tau) \ \delta(\tau) = \chi_1^2(\tau) - 4\chi_2(\tau)$. Define $\tau_l = \frac{1}{d}\ln\left(\frac{k\beta}{s}\right)$ and the set $\alpha_\tau = \{\tau \in \mathbb{R}/\tau \in [0, \min(\tau_{min}, \tau_l))\}$. Consider $\tau \in \alpha_\tau$, and suppose $\varphi_+(\tau) \in [0, 2\pi)$ defined by

$$\begin{cases} \sin(\varphi_{+}(\tau)) = \frac{(w_{+}^{2} - B(\tau))C(\tau)w_{+} + w_{+}A(\tau)D(\tau)}{w_{+}^{2}C^{2}(\tau) + D^{2}(\tau)}\\ \cos(\varphi_{+}(\tau)) = -\frac{(B(\tau) - w_{+}^{2})D(\tau) + w_{+}^{2}A(\tau)C(\tau)}{w_{+}^{2}C^{2}(\tau) + D^{2}(\tau)} \end{cases}$$

and the function $\tau_n(\tau) : \alpha_{\tau} \longrightarrow \mathbb{R}_+$ defined as follows $\tau_n(\tau) := \frac{\varphi_+(\tau)+2n\pi}{w_+(\tau)}, n \in \mathbb{N}$. Let us introduce the following continuous and differentiable function S_n defined by $S_n(\tau) = \tau - \tau_n(\tau), \tau \in \alpha_{\tau}, n \in \mathbb{N}$. Then we have the following theorems.

Theorem 2 Equation (2), has a pair of purely imaginary roots $\lambda = \pm iw_+$, w_+ is real for $\tau \in \alpha_{\tau}$ and at some $\tau_c \in \alpha_{\tau}$, such that $S_n(\tau_c) = 0$ for some $n \in \mathbb{N}$. This pair roots cross the imaginary axis from left (resp. right) to the right (resp. left) if $sign(\Re'(\lambda(\tau))_{\neq \tau=\tau_c}) > 0$ (resp. $sign(\Re'(\lambda(\tau))_{\neq \tau=\tau_c}) < 0$).

Define $\tau_{cmin} = \min \{ \tau \in \mathbb{R}_+ / S_n(\tau) = 0 \}$ and $\tau_{cmax} = \max \{ \tau \in \mathbb{R}_+ / S_n(\tau) = 0 \}$.

Theorem 3 Assume that $(H)_2$ and $\beta > \frac{s}{k}$ are verified, system (1) has the following properties

i) if $\alpha_{\tau} = \emptyset$ or $\alpha_{\tau} \neq \emptyset$, but $S_n(\tau) = 0$ has no positive roots in α_{τ} , E^* is asymptotically stable for all $\tau \in [0, \tau_{min})$.

ii) If $\alpha_{\tau} \neq \emptyset$ and $S_n(\tau) = 0$ has positive roots $\tau_c \in \alpha_{\tau}$ such that, $sign(\Re'(\lambda(\tau))_{\neq \tau=\tau_c}) > 0$ for some $n \in \mathbb{N}$, then E^* is asymptotically stable for all $\tau \in [0, \tau_{cmin}) \cup (\tau_{cmax}, \tau_{min})$ and unstable for $\tau \in (\tau_{cmin}, \tau_{cmax})$, where τ_{cmin} and τ_{cmax} are the Hopf bifurcation values.

References

- 1. A.S.Novozhilov, F.S. Berezovskaya, E.V.Koonin, G.P. Karev, Mathematical modeling of tumor therapy with oncolytic viruses: regimes with complete tumor elimination within the framework of deterministic models. Biol. Direct 1, 6(2006).
- F. Najm, R. Yafia, M. A. Aziz-Alaoui, Hopf Bifurcation in Oncolytic Therapeutic Modeling: Viruses as Anti-Tumor Means with Viral Lytic Cycle, International Journal of Bifurcation and Chaos, 2022, 32(11), 2250171.

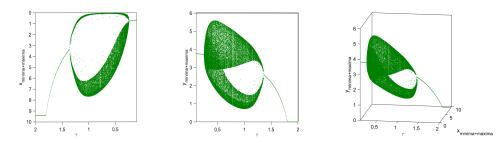


Fig. 1 Bifurcation diagrams in 2d and 3d when taking τ as a parameter of bifurcation with parameters values r = 15; s = 5; $\beta = 2$; d = 0.9; k = 10; a = 4.

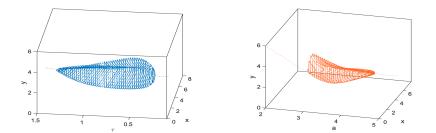


Fig. 2 The occurrence of Hopf bifurcation around the UIE equilibirm point (i.e. existence of periodic solution) by varying τ and a with parameters values r = 2; s = 5; $\beta = 2$; d = 0.9; a = 4; $k = 15 \tau = 1.1$.

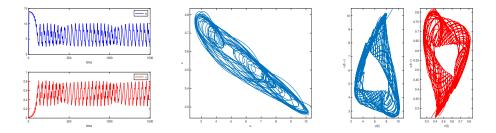


Fig. 3 The existence of choatic solutions for the delay bigger than the critical value τ_c with parameters values r = 2; s = 4; $\beta = 1.9$; d = 0.1; a = 5.5; k = 15; $\tau = 12.08$.



POSTER: Preterm birth indicates higher neural rich club organisation than term counterparts

Katherine Birch $\,\cdot\,$ Dr Dafnis Batalle $\,\cdot\,$ Dr Roman Bauer

Abstract Preterm birth now accounts for around 11% of births [5]. Hence, it is important to investigate the implications of premature birth and its impact on brain development. However, current studies into the neurostructural differences between infants born at different gestational ages have been largely inconclusive. This project aims to better understand structural differences between the preterm and term infant brain. To this end, we employ graph theory measures to compare structural brain imaging data.

Rich club (RC) organisation describes a complex network in which highly connected nodes (hubs) are more likely to connect to other hubs than would be expected of randomly connected nodes [6]. Rich-club coefficient is a measure to quantify this and establish whether such RC organisation exists in a complex network.

Previous studies have used RC coefficient to assess brain structural development, with contradictory findings. For example, one study [1] found weakened RC connectivity in preterm subjects, while another [4] found stronger RC connectivity in preterm infants. The broader implications regarding the structural development of the brain and its clinical significance remain inconclusive.

The present research uses diffusion tensor imaging (DTI) data from the Developing Human Connectome Project (dHCP) [7] [9]. The data was processed using the SIFT2 method [8] to generate structural connectivity matrices. These mathematically represent the white matter tracts in the brain as complex networks and allow for RC coefficients to be calculated.

Dr Dafnis Batalle

Dr Roman Bauer

Katherine Birch

Department of Computer Science, University of Surrey, Guildford GU2 7XH E-mail: k.birch@surrey.ac.uk

Department of Forensic and Neurodevelopmental Science, King's College London, SE1 7EH

Department of Computer Science, University of Surrey, Guildford GU2 7XH

In the dataset, DTI scans from 426 infants at term normalised age (postmenstrual age (PMA) between 37-44 weeks) were included. The participants were categorised into age groups based on gestational age (GA). These were preterm (< 37 weeks GA) and term (\geq 37 weeks GA). In the statistical analysis, both PMA and sex are included as covariates.

RC organisation was observed across thresholds in both preterm and term infants, particularly at high thresholds. Moreover, preliminary analysis suggests increased RC coefficients for preterm infants than term. This is similar to previous research including [2] [3] [4]. In order to explore why contradictory findings may have come about, next stages of this research involve further statistical analysis into the nature of the differences in structural connectivity between preterm and term born infants. This will include considering other complex network features and regional comparisons. Beyond this, an investigation into the various normalisation techniques will be important in order to enable easier comparisons between current literature.

Keywords Complex Networks \cdot Developmental neuroscience \cdot biological complexity \cdot rich-club organization \cdot network hubs

Acknowledgements This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) DTP Studentship 2753824 for the University of Surrey.

Data were provided by the developing Human Connectome Project,KCL-Imperial-Oxford Consortium funded by the European Research Council under the European Union Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement no. [319456]. We are grateful to the families who generously supported this trial.

References

- Joana Sa de Almeida et al. "Preterm birth leads to impaired rich-club organization and fronto-paralimbic/limbic structural connectivity in newborns". In: *NeuroImage* 225 (Jan. 2021). ISSN: 10959572. DOI: 10.1016/ J.NEUROIMAGE.2020.117440.
- [2] Dafnis Batalle et al. "Early development of structural networks and the impact of prematurity on brain connectivity". In: *NeuroImage* 149 (Apr. 2017), pp. 379–392. ISSN: 10959572. DOI: 10.1016/J.NEUROIMAGE.2017.01.065.
- [3] R Bauer and M Kaiser. "Nonlinear growth: an origin of hub organization in complex networks". In: R. Soc. open sci 4 (2017), p. 160691.
 DOI: 10.1098/rsos.160691. URL: http://dx.doi.org/10.1098/rsos.
 160691Electronicsupplementarymaterialisavailableonlineathttps://dx.doi.org/10.6084/m9.figshare.c.3711967..
- [4] Vyacheslav R Karolis et al. "Reinforcement of the Brain's Rich-Club Architecture Following Early Neurodevelopmental Disruption Caused by Very Preterm Birth". In: (2016). DOI: 10.1093/cercor/bhv305. URL: https://academic.oup.com/cercor/article/26/3/1322/2367425.
- [5] George Mandy. "Incidence and mortality of the preterm infant". In: (2018).

- [6] Tore Opsahl et al. "Prominence and control: The weighted rich-club effect". In: *Physical Review Letters* 101 (16 Oct. 2008), p. 168702. ISSN: 00319007. DOI: 10.1103/PHYSREVLETT.101.168702/FIGURES/3/MEDIUM. URL: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett. 101.168702.
- [7] Mikail Rubinov and Olaf Sporns. "Complex network measures of brain connectivity: Uses and interpretations". In: *NeuroImage* 52 (3 Sept. 2010), pp. 1059–1069. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2009.10. 003.
- [8] Robert E. Smith et al. "SIFT2: Enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography". In: *NeuroImage* 119 (Oct. 2015), pp. 338–351. ISSN: 1053-8119. DOI: 10.1016/ J.NEUROIMAGE.2015.06.092.
- [9] Yassine Taoudi-Benchekroun et al. "Predicting age and clinical risk from the neonatal connectome". In: *NeuroImage* 257 (Aug. 2022). ISSN: 10959572. DOI: 10.1016/J.NEUROIMAGE.2022.119319.

Diffusion & Epidemics



Supply, demand and spreading of news during COVID-19 and assessment of questionable sources production
Pietro Gravino, Emanuele Brugnoli, Giulio Prevedello, Martina Galletti and Vittorio Loreto . 144
A local Agent-Based Model of COVID-19 Spreading and Inter- ventions Perrette Benjamin, Cruz Christophe and Cherifi
Hocine
Towards a Generic Agent Based Vector-Host Model Cyrine Chenaoui, Nicolas Marilleau and Slimane Ben Miled
Exploring and optimising infectious disease policies with a stylised agent-based model Jeonghwa Kang and Juste Raimbault
Contact networks in daily-life pedestrian crowds and risks of viral transmission Alexandre Nicolas and Simon Mendez
Minimizing epidemic spread through mitigation of anti-vaccine opinion propagation Sarah Alahmadi, Markus Brede and Rebecca Hoyle 201



Supply, demand and spreading of news during COVID-19 and assessment of questionable sources production

Pietro Gravino^{1,2,3} · Emanuele Brugnoli² · Giulio Prevedello^{2,3} · Martina Galletti^{2,3} · Vittorio Loreto^{1,2,3,4}

Abstract We exploit the burst of news production triggered by the COVID-19 outbreak through an Italian database partially annotated for questionable sources [1]. We compare news supply with news demand, as captured by Google Trends data. We identify the Granger causal relationships between supply and demand for the most searched keywords, quantifying the inertial behaviour of the news supply. Focusing on COVID-19 news, we find that questionable sources are more sensitive than general news production to people's interests, especially when news supply and demand are mismatched. We introduce an index assessing the level of questionable news production solely based on the available volumes of news and searches. Furthermore, we introduce an analysis of the spreading layer of the dynamics. Measured from social media data, it can represent the bridge of interplay between supply and demand. We contend these results can be a powerful asset in informing campaigns against disinformation and providing news outlets and institutions with potentially relevant strategies.

Keywords Information \cdot Disinformation \cdot News Dynamics

References

1. Gravino, P., Prevedello, G., Galletti, M., Loreto, V. (2022). "The supply and demand of news during COVID-19 and assessment of questionable sources production." Nature Human Behaviour, 6(8), 1069–1078.

¹ Sony Computer Science Laboratories Rome, Joint Initiative CREF-SONY, Centro Ricerche Enrico Fermi, Rome, Italy

 $^{{\}bf 2}$ Centro Ricerche Enrico Fermi, Rome, Italy

 $^{{\}bf 3}$ Sony Computer Science Laboratories Paris, Paris, France

⁴ Physics Department, Sapienza University of Rome, Rome, Italy

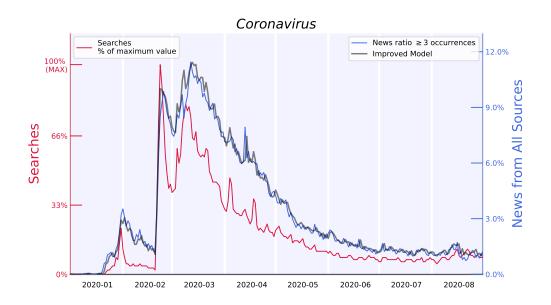


Fig. 1 Fractions of Searches (red), News from All Sources (blue) for "coronavirus" in Italy. Searches the percentage of the maximum observed in the monitored period. News is the daily fraction of articles containing at least three keyword occurrences.

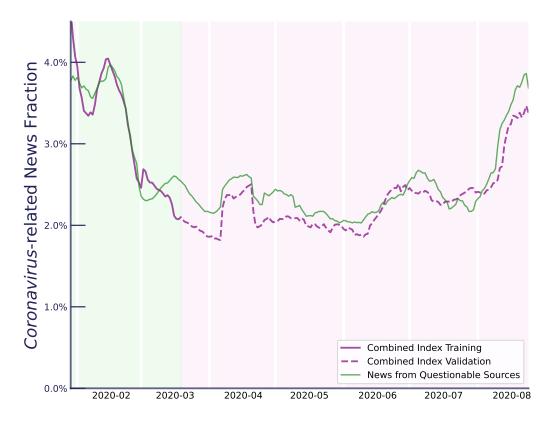


Fig. 2 Timeseries of news from questionable sources compared with the Combined Index.



A local Agent-Based Model of COVID-19 spreading and interventions

Benjamin Perrette · Christophe Cruz · Hocine Cherifi

Abstract This research presents a simulation model that combines metapopulation geospatial data with the SEIR epidemiological model to simulate a city of up to 250,000 residents while considering various factors such as virus transmission rate, disease severity, and prevention and control measures. This model can assist decision-makers in exploring different pandemic response strategies, including lockdowns, social distancing, mass testing, contact tracing, and vaccination. This simulation aims to provide decision-makers with a better understanding of the implications of their choices and enable them to make informed real-time decisions to manage a health crisis.

1 Introduction

The COVID-19 pandemic is a global challenge that significantly impacts health and the economy. Compartmental and metapopulation models [1,2] are crucial in evaluating the effects of containment policies like facility closures [3–8]. More sophisticated models consider the spatial and social structure and the people's evolving awareness. However, they are unsuitable for analyzing the transmission dynamics in small populations with complex behavior. Agentbased modeling offers the advantage of observing global behavioral patterns by modeling the interactions between individuals with distinct characteristics [9]. This study presents an agent-based model to evaluate the transmission process of COVID-19. The model relies on collected meta-population geospatial data and the SEIR epidemiological model. The framework enables simulations of various scenarios, such as different levels of virulence, confinement consideration, and the number of available intensive care units. Initially limited to

ICB Institut Interdisciplinaire Carnot de Bourgogne

⁹ avenue Alain Savary, BP 47870, 21078 Dijon Cedex, France E-mail: christophe.cruz@u-bourgogne.fr

the city center of Dijon, the study considers a greater Dijon area allowing simulations of up to 250,000 residents[10].

2 Compartmental model

In the SEIR (Susceptible-Exposed-Infectious-Recovered) compartmental model, any individual can be in one of the following states:

S: Susceptible: The individual is healthy;

E: Exposed: The individual is incubating the virus and is characterized by the incubation period;

I: Infectious: The individual is contagious and in the period of infection;

R: Recovered: The individual has recovered or has been removed (dead). The infectious state period includes three cases: the classic Us, the asymptomatic A, and the rarer Ua. Patients in the Us class present symptoms of the virus and are contagious. Patients in the A category are contagious but have no symptoms. Ua is an intermediate case in which the patient begins to be infectious before the end of the incubation period but does not present any symptoms yet. One can integrate individual conditions, such as age, gender, and co-morbidity, into the model by defining the transition rates at the personal level. The simulations reported in the experiments use generic parametrization from [11,9]. However, one can use parameters obtained by an epidemiologist from on-site experimentation. In this study, we use the vaccination strategies applied in France for the vaccine model. According to data on vaccine efficacy, the vaccine reduces severity by 90%, infectivity by 30%, and susceptibility by 90%. The vaccination is initially disabled, and these data are modifiable during the simulation.

3 Experimentation

This section presents our experimentation using the GAMA platform [12]. It combines explicit multi-agent simulations with GIS data management, multi-level modeling, and the capacity to implement Belief-Desire-Intention (BDI) and reactive agents. Thus, GAMA is powerful for prototyping and automatic simulation through its agent-oriented language GAML.

We collected data from the Dijon metropolitan area from OpenStreetMap. We extracted a Shapefile containing all of the buildings throughout Burgundy and a file containing all the roads. The simulation starts when the agents go to work between 6 and 8 am. In the evening, they stop between 4 and 8 pm and can then go to a place of leisure. The principle is the same for those under 18, but they go to school rather than work. One can change all these parameters. As shown below, one can tune many parameters on the agents to adapt the model to its needs. One of the agents is infected at the beginning of the simulation. Depending on their state of health, the agents have a different color. The healthy agents are yellow, those in the incubation period are orange. The symptomatic infected are red, the asymptomatic infected are red with purple borders. The hospitalized infected are red dots with green contours, the agents in specific states are orange with red contours, the recovered are green. The dead are black with red contours. It is possible to change the virus settings, as reported in (fig. 1). One can close schools so that young people stay at home. The suspension of leisure activities prevents agents from going to public places such as large shopping centers, restaurants, or places of worship.

In the case of a complete lockdown, the agents stay at home and no longer move. All of these parameters are aimed at reducing the spread of the virus. It is also possible to implement preventive measures to decrease virus transmission. Hospitalization of infected individuals can be activated to decrease the likelihood of dying from the virus. One can change the Parameters linked to the hospitals, such as the number of beds. It is also possible to change the entire vaccination model by deciding when the vaccinations should be carried out and how many vaccine doses are available in the city. One can deal with many variants in the simulation. It can also save information about agents and their environments to a CSV file to create a dataset. Afterward, it is possible to search for information on a particular agent (for example, patient 0) and save it to another CSV file. The simulation can be manually paused at any time or automatically. All these parameters are tunable during the simulation.

The different figures illustrate the infection development in the population. It shows the main places where infections occur and the evolution of available hospital beds. They also report the growth of the variant compared to the strain, the variation of the vaccinated population of the first and the second dose, and the distribution of deaths related to the virus according to agents' age. In the first curve (fig. 3a), we can observe the evolution of the number of infections over time in different locations. The gray line represents infections at home. It continues to rise throughout the simulation, even during a lockdown. The dark blue line represents workplace infections, which remain stable during the lockdown but experience a sharp spike after its lifting. The light blue line represents infections at leisure venues, which stay stable when these places are closed. Finally, the purple line represents infections in schools, which remain stable as long as they are closed but increase upon reopening. In the second graph (fig. 3b), we can observe the evolution of the agent states during the simulation. The number of individuals in the incubation period (state E) is orange. It varies according to the restrictions, decreasing during confinement and increasing afterward. The number of infected agents (state I) is shown in red. It falls with improved regulations. The number of agents who have recovered from the disease is shown in green, while the number of agents who died is black (state R). Finally, the number of agents in the specific state (state Ua) is violet.

4 Conclusion

This model could be a valuable tool for decision-makers in the field of public health to determine when it is the right time to impose restrictions, predict the rise of an epidemic, and act accordingly to stop it. It can be significantly improved, particularly regarding computational speed, to facilitate larger-scale simulations and use knowledge graphs to refine agent behavior.

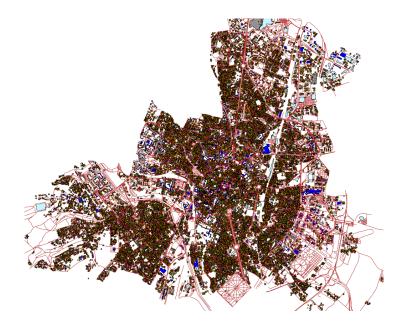


Fig. 1: Dijon metropolitan area with 250,000 inhabitants

References

- 1. A. Singh, H. Cherifi, et al., IEEE Access 8, 1945 (2019)
- 2. M. Arquam, A. Singh, H. Cherifi, IEEE Access 8, 94510 (2020)
- Z. Wang, C.T. Bauch, S. Bhattacharyya, A. d'Onofrio, P. Manfredi, M. Perc, N. Perra, M. Salathé, D. Zhao, Physics Reports 664, 1 (2016). Statistical physics of vaccination
- N. Gupta, A. Singh, H. Cherifi, in 2015 7th international conference on communication systems and networks (COMSNETS) (IEEE, 2015), pp. 1–6
- D. Chakraborty, A. Singh, H. Cherifi, in Computational Social Networks: 5th International Conference, CSoNet 2016, Ho Chi Minh City, Vietnam, August 2-4, 2016, Proceedings 5 (Springer International Publishing, 2016), pp. 62–73
- M. Kumar, A. Singh, H. Cherifi, in Companion Proceedings of the The Web Conference 2018 (2018), pp. 1269–1275
- 7. M. Messadi, H. Cherifi, A. Bessaid, arXiv preprint arXiv:2106.04372 (2021)
- Z. Ghalmane, M.E. Hassouni, H. Cherifi, Social Network Analysis and Mining 9, 1 (2019)
- 9. E. Cuevas, Computers in biology and medicine 121, 103827 (2020)
- C. Prudhomme, C. Cruz, H. Cherifi, in MARAMI 2020-Modèles & Analyse des Réseaux: Approches Mathématiques & Informatiques-The 11th Conference on Network Modeling and Analysis (2020)
- 11. E. Dong, H. Du, L. Gardner, The Lancet infectious diseases 20(5), 533 (2020)



French Regional Conference on Complex Systems - Le Havre, May 31 June 2, 2023

Fig. 2: Dynamic parameters tunable during simulation

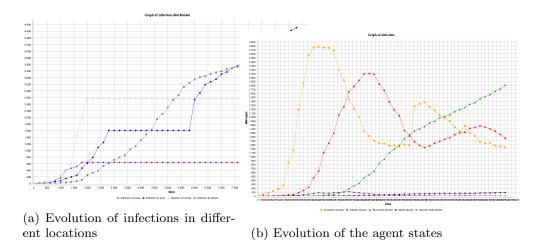


Fig. 3: Evolution of variables during the simulation

12. P. Taillandier, B. Gaudou, A. Grignard, Q.N. Huynh, N. Marilleau, P. Caillou, D. Philippon, A. Drogoul, GeoInformatica **23**, 299 (2019)



Towards a Generic Agent Based Vector-Host Model

Cyrine Chenaoui · Nicolas Marilleau · Slimane Ben Miled

Abstract The aim of our work is to develop a conceptual generic agent-based model to formalize the interaction of vector and host in view of climate change. The model consists in creating a hypothetical example of a vector-host system. It simulates the vector's life cycle while considering interactions, respectively, with hosts and the temperature. It is presented following the ODD protocol and is based on a set of parameters and processes to conceptualize the vector-host complex and could be accommodated to a wide spectrum of vector species and different biogeographic regions. The model's primary goal is to evaluate the overall effects of temperature variations and host dispersion patterns on tick population dynamics while considering temperature effects on development. We show that seasonality is a primary determinant of the synchronization of distinct physiological stages. Our model can be extended to more ecologically complex systems with multiple species and real-world landscape complexity to test different host- and/or vector-targeted control strategies and identify effective approaches in managing vectors population and dispersion patterns.

Keywords Agent based models \cdot Vector-host system \cdot ODD protocol

C.Chenaoui

Ecole Nationale d'Ingénieurs de Tunis, Université Tunis El-Manar, Tunis, Tunisia Institut Pasteur de Tunis, Université Tunis El-Manar, Tunis, Tunisia

Unité de Modélisation Mathématique et Informatique des Systèmes Complexes IRD,Sorbonne Université,Île-de-France, France

E-mail: cyrine.chenaoui@gmail.com

N.Marilleau

S.Ben Miled Institut Pasteur de Tunis,Université Tunis El-Manar, Tunis, Tunisia

Unité de Modélisation Mathématique et Informatique des Systèmes Complexes IRD,Île-de-France, France

1 Introduction

Ticks are ectoparasitic acarians of vertebrates (e.g. cattle, sheep, etc.), vectors of pathogens responsible for zoonoses and important economic losses for livestock.

The control of vector-borne diseases mainly requires the understanding of the role of ectoparasites in the transmission of pathogens of medical and veterinary importance by studying their population dynamics and the relationship with possible hosts as well as the climatic factors that drive these dynamics. Vector-borne diseases remain a prominent source of morbidity and mortality worldwide, despite significant advances in public health. Simulation models of infectious disease transmission and vector-host systems give a new degree of knowledge to support public health strategies for disease control. In this regard, the study of tick population dynamics and their interactions with the environment seems essential for the surveillance of vector-borne diseases. One of the factors of primary interest is environmental change. These changes include climatic variations, grazing and plant cover use, causing possible ecosystem modifications. Therefore, environmental factors induce great alterations in the distribution of vectors in nature and in particular in ticks.

To further understand how ecological mechanisms and processes underlying the vector-host systems, computer simulations are used to address questions that controlled experiments or observations cannot solely answer and to simulate more complicated ecological systems with explicit details, implementing individual-level interactions and spatial structures. The relationships between biological processes, the environment, and ecological patterns across different scales can be revealed. Computer-based ecological models can further predict the future of evolution. Developing, optimizing, and applying computer simulations for ecological modelling has become the central task for computational ecology.

In recent decades, the deployment of simulation models got in importance. Models are being used to investigate the relevance of biological and ecological features on disease transmission and expansion, enabling the examination of systems to explain previously observed data, forecast the future, and investigate the impact of potential mitigation and management approaches. Many IBM models have been developed (e.q. [1,2,3,4,5]) but most of them are used for specific ecological systems and husbandry systems or species of hosts, or considering well-determined diseases. Since simulation models are increasingly being used to solve problems and to aid in decision-making. Our model consists in creating a hypothetical example of a semi-intensive livestock production system. This system is closed and composed of only three entities: ticks, dairy cows and rodents. We consider that the dairy cows are grazing and then return to a barn, we do not take into account in our model the diurnal activity of ticks, nor that of rodents. The host population is closed without intra-interactions, while the tick population is an open monospecific population whose recruitment is done through egg laying. It is stratified in cohorts of biological life stages, respectively, egg, larva, nymph and adult. Interstadial development ticks are temperature-dependent processes. The only interaction between the hosts and the ticks is through a blood meal.

The aim of our work is to develop a generic agent-based model to formalize the interaction of vector and host in view of temperature change. The model consists of creating a hypothetical example of a vector-host system. It stimulates the vector's life cycle while considering interactions, respectively, with hosts and the temperature. We compare two scenarios of host dispersion; random movement of hosts while following a dominant individual vs predefined path followed in herd movement.

This paper is organised as follows: In section 2 we describe the background knowledge and modelling assumptions we considered. In section 3 we explain our model using the overview, design concepts, and details (ODD) protocol [6]. Then in section 4 we present our simulation's results by comparing two dispersion approaches of cattle agents. Finally, in section 7, we discuss our results and pinpoint some limits and perspectives to generalise and promote our model to fill in the gap between understanding the ecology of the vector and hosts on the one hand and temperature effects on the ecology on the other hand.

2 Ecology of the vector and background knowledge

The vector model in this study is species of hard ticks. Ticks have four developmental stages (figure 1); an egg and three active obligatory haematophagous ectoparasitic stages characterized by only one larval and nymphal stage before reaching the adult stage and discontinuous feeding, single blood meal to get through the next life stage. Freshly laid eggs are brown due to the thin brown shell covering the colourless inner mass, where the nucleus is embedded and surrounded by the cytoplasm [7].

At the end of the embryo's development, the eggs hatch into larvae. As soon as the incubation debris are removed, and the integuments harden, the larva begins to seek a host. Ixodid ticks survive long periods of fasting and then ingest large amounts of blood in a few days. Two known host finding strategies are described in literature [8,9]: a passive strategy called "questing" for encountering a host, and an active seeking host strategy adopted by different ticks species. In the first strategy, by the passing-by host, ticks cling to the surface of their hosts using their hypostome and secreting cement-like substance into the skin permitting its anchorage to be completely fixed and start feeding by sucking fluids and blood that are stored in the midgut diverticula for gradual digestion [10].

Ticks will climb plants (such as grass or shrubs) to obtain greater access to a moving potential host while they are searching for one. Despite the fact that ticks spend much of their time moving vertically, they lose water vapour through the cuticle due to two factors: their spiracles must be opened often, particularly during locomotion, to allow for respiration, and increased exposure to driver conditions [11].

The meal on the host for larvae lasts from 3 to 6 days and then detaches from the host and falls on the ground. After a period of one to one and a half months, the metamorphosis/moulting ends and gives a nymph. The adult females, usually fertilized, feed to repletion for a 3 to 10 days period to lay eggs. After attaching to a host, they start by a slow-feeding phase (4–7 days) phase where the tick sucks blood and increases up to 10 times her unfed weight. Final engorgement arises during the last 24–36 hours of rapid feeding, where it can multiply its unfed body mass approximately 100-fold with protein and lipid-rich nutrients for the production of eggs [12].

During this rapid engorgement period, the midgut lumen acts as a blood reservoir, where the digestion of the blood meal continues gradually [10]. If the environmental factors are unfavourable for questing, feeding or moulting, the vector starts a phase of dormancy, for instance, behavioural diapause is a state of torpidity as a response to hazardous environmental conditions, usually, low-temperature [13].

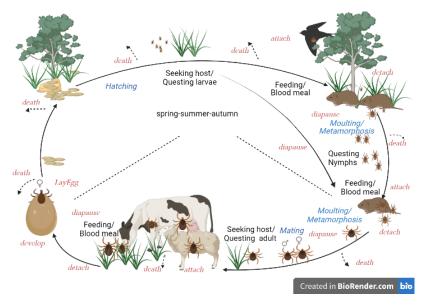


Fig. 1 Schematic life cycle of ticks

3 The conceptual model implementation

The description of the model follows the "Overview, design concepts, and Details (ODD)" protocol developed by [14,6] to standardize the use of ABMs. It was intended to provide a protocol for characterizing ABMs in ecology and has seven elements, each described by questions to answer and explain, and organized into three categories. Note that "Adaptation", "Objective", "Learning" and "prediction" ODD part design concepts should not be applied to our model.

3.1 Overview

3.1.1 Purpose and patterns

The purpose of the model is to compare host dispersion on the vector's life cycle and population dynamics while considering temperature effects on development. Our goal is to develop a generic ABM model following the ODD protocol to model the interaction of vector-host, considering the ecological dynamics of vectors given climate change.

Entities, state variables, and scales

The model combines two agents: Simulated host and vector populations. The interactions between agents are ruled by three ecological processes: (1) vector population dynamics, (2) host movement, and (3) temperature fluctuation. **The vector agent:**

The vector agents are characterized mainly by the life stage attribute state respectively; egg, larva, nymph, and adult. The activity status of the vector agents is characterized by three behavioural states (1)questing: this is a generic term to designate the activity of the vector before attaching to a possible host and while it is waiting for the passing-by of a host or the actively seeking the host, (2) feeding: while the vector attaches to the host for taking the blood meal and finally (3) moulting: the behavioural state of the after completing the blood meal and detaching from the host.

The host agent:

The Host agents are characterized mainly by the list of vectors attached to it, vector_of_parasite. The Host agent also includes two sub-agents: cattle and rodent agents: They inherit the same attributes from the Host agents. Rodent agents have a random movement in the environment characterized by a random velocity fixed at the beginning of the simulation. We developed two designs to identify the eventual effects of cattle dispersion on vector population dynamics.

Attribute	Type	Dim	Desc.	Freq
Vector				
location	\mathbb{R}^2	_	Position	1 H
speed	\mathbb{R}^2	km/h	Speed of vector agent	1 H
BehState	Category	_	Behavioural state	1 D
State	Category	_	Life stages; egg, larva, nymph, adult	1 D
AttachDetachDate	Date	-	attachment/detachment date	1 D
LayDate	Date	-	Date of egg hatching	1 D
QuestDate	Date	-	Questing start date	1 D
NeighHosts	List	-	Neighbouring hosts	1 H
TargetHost	Host	-	Target host	1 H
LayingEgg	Boolean	-	Fitness of adults to lay eggs	1 H
Diapause	Boolean	-	Diapause	1 D
LaidToIncubation	\mathbb{R}	-	Incubation duration	1 D
PreoviToOvi	\mathbb{R}		Preoviposition duration	1 D
MoultingDuration	\mathbb{R}	-	Moutling duration	1 D
oviposition	Boolean		Oviposition	1 D
moulted	Boolean	-	True to change the life state	1 D
incubation	Boolean	-	incubation	
ProbIndAttach	\mathbb{R}	-	Probability of attachment	
counter	\mathbb{R}	-	Counter for development rate	
Hosts				
location	\mathbb{R}^2	-	Position of the agent	1 H
speed	\mathbb{R}^2	km/h	Speed of host agent	1 H

Table 1: Summary of attributes and variables with definitions

French Regional Conference on Complex Systems - Le Havre, May 31 June 2, 2023

Attribute	Type	Dim	Desc.	Freq
VectorOfParasite	\mathbb{N}^n	-	List of vectors	1 H

Table 1: Summary of attributes and variables with definitions

In the first design, cattle agents are endowed with random movements that do not follow a particular trajectory and with no herding. In the second design, cattle agents follow a herd dispersion pattern and follow a dominant cattle agent considered *leader*. In both designs, cattle are characterized by a diurnal activity whether they are active or resting by regaining the *barn*.

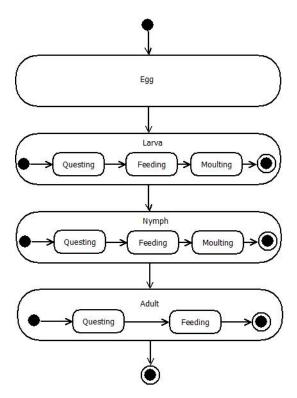


Fig. 2 Vector life stages (state) and behavioural states BehState

The environment

The environment is set up as a hypothetical square world with a dimension of 1 km x 1 km with unwrapping boundaries movement and does not cause Host agents to jump to the other side of the world and to have more realistic host movement. The environment is considered a grazing area where all Host agents can move around. In the centre of the environment, a hypothetical barn of square geometric shape and a dimension of 50m x 50m. As for the climatic parameters, only the temperature is considered a climatic factor affecting the vector agents.

3.1.2 Process overview and scheduling

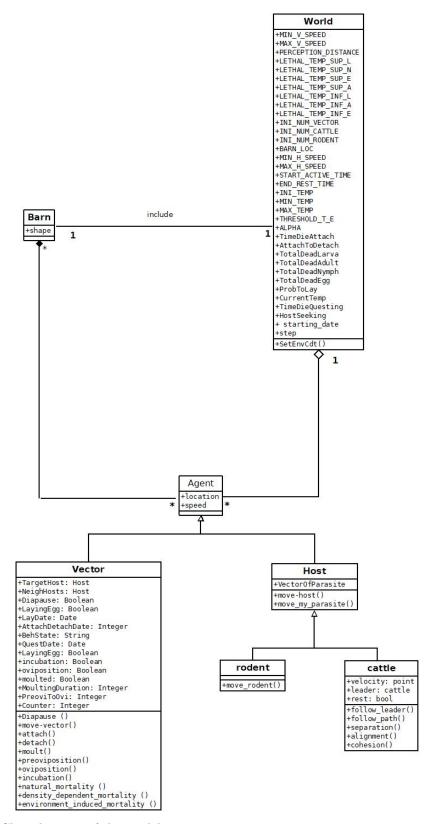


Fig. 3 Class diagram of the model

The model simulates the interactions between the host agent and the vector agent on the one hand and between environmental changes (temperature) host availability within the perception range of the vector, and the vector agent on the other hand. In our model we consider seven submodels that happen in the following order (figure 4); Every hour, vector agents are stationary and waiting for the passing-by of a possible host within a specific distance *PER*-*CEPTION_DISTANCE*. When finding a *Host*, vector agents start a blood meal after attaching for a fixed number of days *AttachToDetach* to hosts based on the *attach* submodel. As for the *Host*, agents move according to the *move_host* submodel.

Each day, vectors detach from hosts following the *detach* submodel after finishing the blood meal. The dynamics of the vector's population is a result of the *develop* submodel. It is a cluster of four sub-models that cover *preoviposition*, *oviposition*, *incubation* and *moulting*. The *death* submodel is composed of two submodels; *natural_mortality* and *environment_induced_mortality*. The temperature is updated every day.

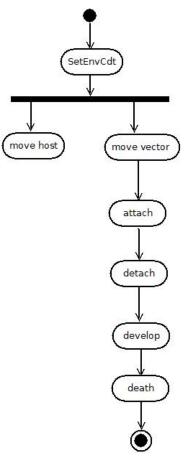


Fig. 4 General processes and submodels activity diagram

3.2 Design concepts

3.2.1 Basic principles

The model involves a mechanistic description of vector-host interactions. The conceptualization of the vector life cycle involves four processes: First we consider questing vector behavior, vector attachment to hosts, vector detachment from hosts, and vector mortality processes. Moulting duration, preoviposition, incubation and diapause are temperature-dependent processes. Mating is implicitly included in our model, but vectors are assumed to mate on the host, so adult vectors are able to lay eggs after a successful blood meal. Adult vectors are all females and would give offspring. We have considered the host population a closed system with no variation.

In the first design, Host agents make movements randomly with a random speed fixed after every resting period. While in the second design, Host agents move in a herd according to the boid movement principles. Both designs do not include trophic resources for hosts or population dynamics processes for *Host* agents.

3.2.2 Emergence

The traits of vector population emerge from the properties of empirical relationships between temperature and moulting duration, and also from host location and probability of attaching to a possible host permitting the fulfilment of the transition into the next behavioural stage (figure 2). The combined actions and states of the *Host* agents and *vector* in response to the environment produce the total number of vectors questing, feeding, or moulting in any particular location at any given moment.

3.2.3 Sensing

Moulting and diapause behaviours are temperature-dependent processes, vector agents sense the temperature variation and enter diapause when the daily temperature reaches a stage-dependent threshold. larva and nymph states sense the temperature when moulting. Eq. vector agents sense the temperature when incubating, as for Adult vector sense it when, either in preoviposition or oviposition phase. Intraspecific interactions are not taken into account in our model; Vectors do not sense each other or communicate. Except within the range of their distance of perception *PERCEPTION_DISTANCE*, vectors do not perceive *Host* agents or interact with them. Hosts do not sense the presence of the vectors when detached, but they can sense the presence of vectors attached to them through the submodel *move_my_parasite* by changing the attached vectors' location to theirs. Both cattle and rodents do not sense the temperature. Rodents do not sense each other, nor sense cattle agents. In the first design, cattle do not sense each other, meanwhile, in the second design, they sense the movement of each other and move in a herd pattern while following the leader cattle agent.

3.2.4 Interaction

Vector agents interact with *Host* agent through blood meals. *Larva vector* agents can only have a blood meal *rodent* agents for blood meal. *Nymph* and *adult vector* agents can attach to both *rodent* and *cattle* agents. The probability of attachment depends on the number of hosts in the range of perception distance of *vector* agents.

Host agents interact with *vector* agents only when they are attached to them. Otherwise, they don't interact with them. *cattle* agents interact with each other when herding. The two *Host* subagents do not interact with each other.

3.2.5 Stochasticity

Stochasticity forms an essential drive in three submodels of the model. First, rodent movement is random, and their speed is fixed randomly between minimum and maximum speed, at the start of the simulation (table 2). As mentioned in section 3, stochasticity constitutes the main drive of *cattle* agents movement. Moreover, environment-induced mortality is a stochastic submodel, at every time step a stage-dependent and temperature-dependent probability of death is computed. Furthermore, the *attach* submodel is based on stochastic processes, since the probability of attachment and the selection of the host is computed randomly at every time step.

3.2.6 Collectives

Vectors do not form an aggregation that affects their behaviour, nor do the rodents in both designs. In the first design, cattle do not form any collective behaviour. In the case of the second design, cattle agents establish a herd movement pattern whose mobility depends on the location of a dominating cattle; a *leader* cattle. Hosts and vectors attached to them, form a collective of vector-host, crucial for the fulfilment of the feeding behavioural state and the transition to moulting behaviour.

3.2.7 Observation

The main target observations are the vector's population density, including the stage class densities. Additionally, we tracked the life stage class activities (questing, feeding, moulting, and diapause).

3.3 Details

3.3.1 Initialization

Each simulation starts with an environment wherein cattle are situated in the centre of the barn. The dominant cattle agent is selected randomly. Rodent agents are located randomly in the environment. 70% of the vectors are adults in the early preoviposition stage, and the rest are vectors in the early moulting larval stage. The initial date of the simulation is January 16, 1990. The initial daily temperature corresponds to that of the initialization date. The initial values of all parameters are in table 2.

Param.	Sym.	Val.	Dim.	Ref.
World				
The current daily value of T [°]	CurrentTemp	-	°C	[15]
Maximum T° Minimum T° Coeff.of	MAX_TEMP MIN_TEMP ALPHA	40 -30 0.2	°C °C ℝ	
Environment- induced mortality				
Time step of the model	step	1	R	-
World size POsition of the barn	shape BARN_LOC	1 (500,500)	\mathbb{R}^2	-
Vector				
Initial number of vectors	INI_NUM_VECTOR	150	N	-
Distance of percep- tion	PERCEPTION_DISTANCE	7	N	-
Larva highest Lethal T°	LETHAL_TEMP_SUP_L	40	°C	[16]
Nymph highest Lethal T [°]	LETHAL_TEMP_SUP_N	45	°C	[16]
Egg highest Lethal T°	LETHAL_TEMP_SUP_E	32	°C	[16]
Adult highest Lethal T°	LETHAL_TEMP_SUP_A	35	°C	[17]
Nymph minimum Lethal T°	LETHAL_TEMP_INF_N	-18	°C	[18]
Adult minimum Lethal T°	LETHAL_TEMP_INF_A	-20	°C	[18]
Larva minimum Lethal T°	LETHAL_TEMP_INF_L	-18	°C	[19]
Egg minimum Lethal T°	LETHAL_TEMP_INF_E	-30	°C	[20]
Threshold T [°] of di- apause for L.	THRESHOLD_T_L	8	°C	[19]
Threshold T [°] of di- apause for N.	THRESHOLD_T_N	11	°C	[18]
Threshold T [°] of di- apause for E.	THRESHOLD_T_E	8	°C	[20]
Threshold T [°] of di- apause for A.	THRESHOLD_T_A	8	°C	[17]
Proba. to lay eggs Larva Natural Mortality Proba.	ProbToLay P_NAT_MOR_N	$\begin{array}{c}1\\0.006\end{array}$	[0,1] [0,1]	-
Nymph Natural Mortality Proba.	P_NAT_MOR_L	0.006	[0,1]	-
Adult Natural Mortality Proba.	P_NAT_MOR_A	0.006	[0,1]	-
Egg Natural Mor- tality Proba.	P_NAT_MOR_E	0.006	[0,1]	-
Proba. of attach- ment	P_ATTACH	0.9	[0,1]	-
1110110	I	I	I	I

Table 2: Global Parameters and input values of the vector agent and host agent attributes. For the dimension, the following symbols are used: "-" indicates no dimensions

Param.	Sym.	Val.	Dim.	Ref.
Duration of attach-	AttachToDetach	7	days	[21]
ment				
Time to die when	$\operatorname{TimeDieQuesting}$		N	-
questing				
Host				
Initial number of	INI_NUM_CATTLE	50	N	-
cattle				
Initial number of	INI_NUM_RODENT	100	N	-
rodents				
Minimum Host	MIN_H_SPEED		\mathbb{R}^2	-
speed				
Maximum Host	MAX_H_SPEED		\mathbb{R}^2	-
speed				
Cattle's activity	START_ACTIVE_TIME	9	hour	-
start time				
Cattle's activity	END_ACTIVE_TIME	16	hour	-
end time			2	
New position of	velocity	(0,0)	\mathbb{R}^2	-
boid movement				
Resting status	rest	False	Boolean	-
Minimal distance	$minimal_distance$	50	meter	-
of perception				

Table 2: Global Parameters and input values of the vector agent and host agent attributes. For the dimension, the following symbols are used: "-" indicates no dimensions

3.3.2 Input data

The model is parameterized for the species *Ixodes scapularis*. Climate data used in this model are the Climatic Research Unit (CRU) TS Time Series datasets 4.04 [15]. Such a dataset are monthly estimates of Temperature recorded between the years 1990 and 2000 with a spatial resolution of $(0.5 \times 0.5 \text{ degree})$ grids for the region of Wisconsin (). The daily estimates are generated by fitting data points to a polynomial function of the order 3.

The parameters used in this model come from bibliographic data and/or the modeller's expertise as detailed in table 2.

3.3.3 Submodels

Host's agent submodels

1 move_my_parasite submodel

Since the vectors attached to the hosts, no longer control their position, we assume that the hosts will move the attached hosts on themselves according to the *move_my_parasite* submodel. This submodel is common for both *rodent* and *cattle* agents. *Vector* agents attached to a *Host* agent, change their location as per the host they are fixed on until they detach.

$2 move_cattle$ submodel

The move_cattle sub-model simulates the movement and activity patterns of cattle sub-agents. In the first scenario, cattle and rodent agents move randomly according to the move_host_random. In the second scenario, unlike the first design, cattle agents follow a leader agent according to the follow_leader process, which itself follows a predefined path according to the follow_path starting from the centre of the barn and returning to it. Cattle agents, besides the leader, are moving in a herd cohesion according to a boid movement composed of the "separation", "alignment", and "cohesion" processes as described by [?], cattle agents flocks to the centre of a mass of agents within the mini-mal_distance, and avoiding other agents while trying to match the position of other cattle agent.

If a current hour of the day is in the grazing time range, the leader will move with a random speed following the path, otherwise, the leader cattle subagent will move to the center of the barn *BARN_LOC*.

Vector's agent submodels

3 The *attach* submodel

The *attach* submodel describes the attachment of the vector to one host agent within the *PERCEPTION_DISTANCE* when the vector's attribute *BehState* is equal to "q". Vectors which have their state equal to "larva" attach to *rodent* agent, *nymph* and *adult* attach both *cattle* and *rodent*. The *attach* submodel does not apply to *egg* since it is a free life stage. The process of attachment has a probability *ProbIndAttach* dependent on the number of neighbouring hosts *NeighHosts* computed as follows:

 $ProbIndAttach = 1 - (0.8)^{NeighHosts}$ (1) with $NeighHosts \in \mathbb{R}$. This probability is given by the fact that as long as the number of hosts in the range of the vector's perception distance is large, the vector is more likely to attach to one of its hosts (figure 5).

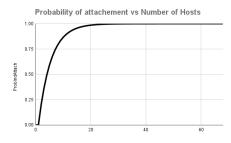


Fig. 5 attach submodel activity diagram

4 detach submodel

This process describes the detachment of vectors from the host after a suc-

cessful blood meal. A successful blood meal is controled by the duration of the blood meal $current_date=AttachDetachDate + AttachToDetach$, the vector detach from the Host agent at a random proximate location and BehState will be updated to moulting.

5 develop submodel

The *develop* submodel considers the life stage transitions and diapause process. In this submodel we consider the development of the vectors as daily temperature functions. The blood meal is not a source of energy and acts on the development only as a condition for the transition to the next instar. We divided the "develop" sub-model into five processes, respectively, *diapause*, *moulting*, *preoviposition*, *oviposition* and finally *incubation*. The transition between life state is controlled according to the previous state and behaviour of the vector(see figure 6), the transition between states is controlled by the execution of the "attach", "detach" and "moult" processes.

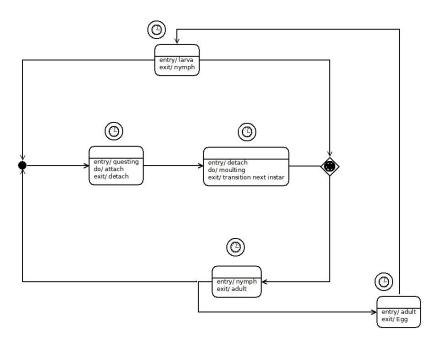


Fig. 6 Life state transition as a state machine diagram

diapause submodel

The *diapause* process describes the temperature-dependent ecology of the diapause. When the diapause process is activated, the *larva* and *nymph* agents can no longer attach or detach, nor can the *adult* agent, which in turn can no longer produce offspring. The incubation of *egg* agents is also in stand-by, and the counters for *preoviposition* duration and *oviposition* are paused until the diapause process is deactivated.

moult preoviposition and incubation submodels

The moult preoviposition and incubation submodel are temperature-dependent (figure 8) following the finding of [16]. We denote f(t) the duration function

of every behaviour, respectively, $m_{larva \to nymph}$, $m_{nymph \to adult}$, p (figure 8 and equations 2 and Inc and each behaviour is structured by development stage i, respectively i_{moult} , $i_{preoviposition}$ and ihatching.

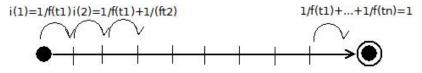


Fig. 7 A schematic diagram for the duration function incrementation

The ending of every behaviour occurs when i = 1, and we assume at every time step Every time step, the duration of moulting for larva into a nymph, and nymph into an adult, the preoviposition and the incubation duration are computed according to, respectively, the equations 2, where m is respectively is the moulting duration respectively for larva and nymph, preoviposition duration, and $T_{current}$ is the daily temperature (see figure 7, and equations 2, table 3).

$$m_{j(stage1 \to stage2)} = a * T_{current}^{-b}$$

$$i_{j(stage1 \to stage2)} = i_{j(stage1 \leftarrow stage2)} + 1/m_{j(stage1 \to stage2)}$$
(2)

Table 3: Parameters of the equation 2, respectively for, moult, preoviposition and incubation submodels

j	stage 1	stage 2	а	b
moult	larva	nymph	101181	2.9
moult	nymph	adult	1300	1.23
preoviposition	-	-	1200	1.3
incubation	-	-	34600	2.35

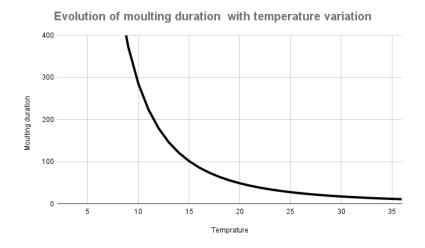


Fig. 8 Evolution of larva moulting duration with temperature variation

6 death submodel

The *death* submodel describes mortality processes. Two types of mortality are considered in this submodel; *natural_mortality*, *environment_induced_mortality*.

Natural mortality

It is life-stage dependent. In each time step, each life stage cohort vector had a given probability of dying, respectively $P_NAT_MOR_E$ for egg state, $(P_NAT_MOR_L$ for larva state, $P_NAT_MOR_N$ for nymph state and $(P_NAT_MOR_A$ for the adult state. Since Hard ticks are semelparous species, then, after laying eggs, all adult agents will die.

Environment-induced mortality process

This mortality process is temperature-dependent. Every day, the death probability is switched based on stage-dependent lethal mortality (see equation 3 and table 1).

$$p_{env} = \begin{cases} p_{env-i}^{sup} & \text{if } T_{current} >= T_{lethal_i}^{sup} \\ p_{env-i}^{inf} & \text{if } T_{current} < T_{lethal_i}^{inf} \end{cases}$$
(3)

4 Effect of simulations replication

Simulations were performed using the Gama 1.8.2 [22] platform on headless mode to develop the agent-based simulator. All simulations were run over a period of 10 years. As mentioned before (see section 3.1.1), we have tested two scenarios, S1 and S2, regarding the movement patterns of *cattle* agent. Both simulation sets were run using the same parameter values and initial state. For both scenarios, we analyzed the total and per-stage population and the total abundance and per-stage abundance.

For the first set of simulations, the total population has an increasing trend based on the Mann-Kendall Trend test ($p_{value} < 0.05$). In the figure below we plot the total population abundance per replication for both temperature sets. Figure 9, shows a yearly fluctuation of the total population for all the replication, but also, a remarkable fluctuation in amplitude between them.

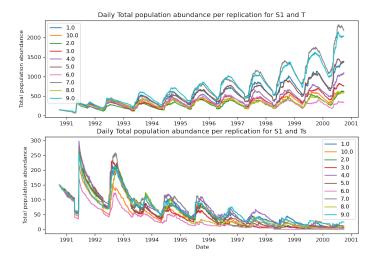


Fig. 9 Total population abundance per replication for both simulations sets (T and T_s)

Graphically, we notice four replication outliers (figure 9 for S1 and T set). We try to fit the population abundance into a polynomial regression of degree 3 (see figure 10. We proceeded to compute the residuals, and normalize the variations, in order to attenuate the amplitude of each bifurcation. We computed the residuals σ_i between raw outputs and the regression estimate for every replication as follows (equation 4):

$$\sigma_{i} = \frac{y_{i}(t) - y_{i}^{*}(t)}{y_{i}^{*}(t)}$$
(4)

Where $y_i^*(t)$ is the regression estimate, $y_i(t)$ is the simulation output and σ_i is the difference between both data series. The plot of σ_i shows steady seasonal fluctuations(figure 11), but starting from the seventh year of the simulation period, we notice a sharp increase in the amplitude of the total population of some replications.

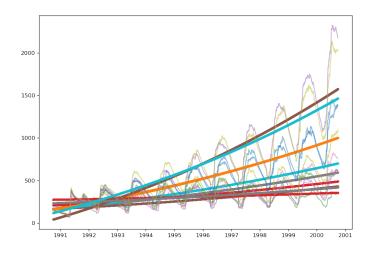


Fig. 10 Fitting of the replication into a polynomial regression for S1 and T set

After computing residuals between the regression curve and replication outputs of the real temperature set for every time step (figure 11), we explore graphically σ_i ; Overall, the residuals have steady fluctuations, with a slight decrease between the years 1997-1999 and we notice a peak of one replication in the first two years of the simulation. We also notice no major differences between simulation replications apart from this outlier.

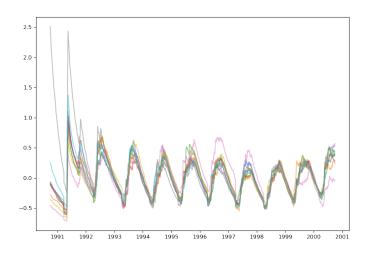


Fig. 11 Residuals σ_i between the regression estimates and the total population abundance for every replication for the simulation set S1-T

In fact, the standard deviation of residuals varies between 0.81 and 0.02 (see figure 12). The maximum value of σ_i is registered at the initial period of the simulation then residuals are slightly fluctuating around zero, apart from the first year of the simulation. Mean residuals know a decreasing sinusoidal fluctuation but we notice a slight increase in the last year of the simulation.

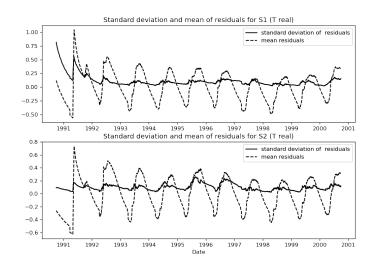


Fig. 12 Variation and means of residuals for S1 and S2 while testing real temperature data set

Regarding the outputs of simulations of smoothed temperature, we followed the same steps to analyse them. We proceeded to fit the replication into polynomial regression. Since the population is decreasing, the residual graph knows no remarkable fluctuations (see figures 9 and 18).

As for the first scenario, we tested the same temperature data sets and the same parameter sets listed in table 1. We notice, as for the random movement, the population abundance declines for T_s and increases for the temperature data set T (see figure 13) for 9 replication. Only one replication has a decreasing trend (the p_value of Mann-Kendall Trend is over 0.05)

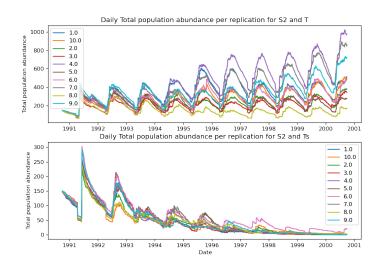


Fig. 13 Total population abundance per replication for both simulations sets (T and T_s) in simulation S2

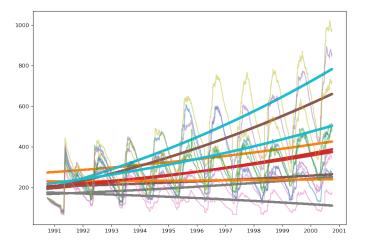


Fig. 14 Fitting of the total population for the second scenario in the case of real temperature data set

As for the first scenario, we proceed to fit the replication into a polynomial regression (figures 14 and 15) and we computed the residuals (equation 4, figure 12). The standard deviation of the residuals is characterized by slight fluctuating values varying between 0.26 and 0.21.

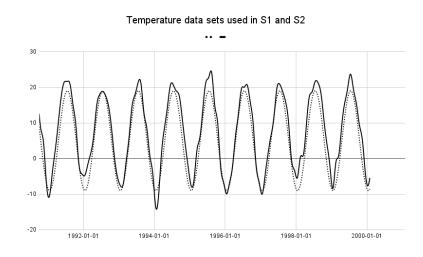


Fig. 16 Real temperature (solid line) and smoothed temperature(dashed line)

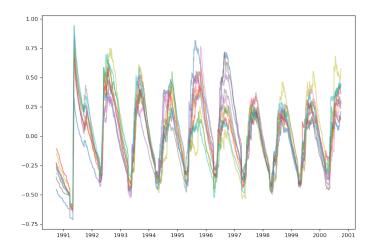


Fig. 15 Standard deviation and mean Residuals after fitting simulations replications for the second scenario in the case of smoothed temperature set

5 Real Temperature versus smoothed temperature sets simulations

For every scenario, we tested two temperature sets, real and smoothed temperatures (see equation 5 and figure 16). The smoothed temperature T_s is a sinusoidal function computed based on the maximum value of temperature during the period 1990-2001. For every day we computed the corresponding smoothed temperature as follows:

$$T_s = T_{max} \cos\left(2\pi T_{real}/365\right) + 5$$
(5)

Where T_{max} is the maximal temperature registered over the period 1990-2001 and T_{real} is the daily real temperature. We run ten replications for every scenario and temperature set (see figures 9 and 13). We display on the same graph (figure 17), the evolution of the average abundance of the population versus,

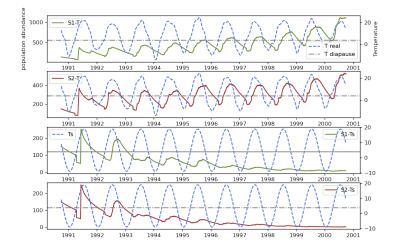


Fig. 17 Mean total population abundance versus, respectively, real temperature and smoothed temperature

respectively, the real temperature and the smoothed temperature, we mark the average temperature of diapause of the sharks. We notice that during the first year, the response of the model to the real temperature, gives an important decrease in the population after the first summer, followed by a slight increase in the total population abundance (explained by the oviposition and the hatching of the eggs). In the second simulation set, an abrupt decrease in the population is observed during the first summer of the simulation. According to figure 16, the winter of the real temperature series of the first year of the simulation is colder than that of the smoothed series, as well as the summer, which is warmer.

6 Random dispersion versus Herd movement simulations

The purpose of testing the random dispersion pattern of *cattle* agents, is to highlight whether the dispersion stochasticity could impact the vectors' temporal dynamics. We denote the total abundance of the *vector* population at time t and $\mathbf{N}(t)$ the yearly mean population abundance and $N_{state(i)}(t)$, the life stage population abundance, where i is, respectively to "egg", "larva", "nymph" and "adult" stage.

$$\bar{\mathbf{N}(t)} = \sum_{i} \bar{N}_{state(i)}(t) \tag{6}$$

Then, we plot the mean over the replication of both, the total and the perstage population abundance, respectively for the two temperature sets (see figure 18 and 20). By comparing graphically the average populations, respectively, for the two scenarios and the two sets of temperatures, we notice that the average abundances of the total population as well as the population per stage of the random movement strategy simulations are greater in magnitude than the herd movement strategy. Also, we plot the mean monthly per stage

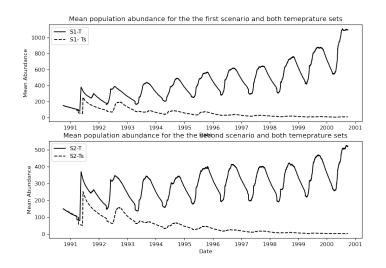


Fig. 18 Mean total population for both scenarios and temperature sets

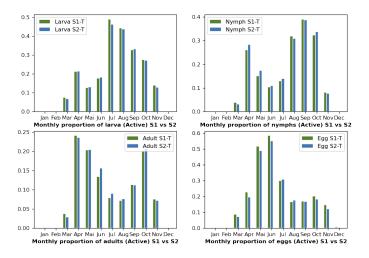


Fig. 19 Monthly per stage active population for S1 and S2 in real temperature data set

population proportion in figure 19. This plot shows the monthly proportions of active vectors in different stages. The results shown in figure 18 are confirmed by the average monthly distribution by stage as plotted in figure 19. There is a slight difference between the two scenarios but the monthly distribution is almost the same between both scenarios.

In the figure 20, we plot the life-stage population $\mathbf{N}_{state\{i\}}(t)$. Figure 20, shows the different temporal variations of respectively, egg, larva, nymph and adult populations for both S1 and S2. In S1, the egg population knows 6 important production peaks over 10 years of simulations, each characterized by two small peaks one in late autumn and the second occurring in early spring. Larva are characterized by sharp peaks occurring in early/mid-summer,

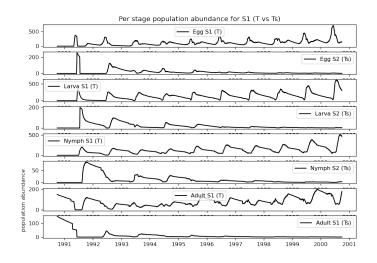


Fig. 20 Per-stage total population abundance for S1 for both temperature sets

then know a steady behaviour for merely a year marking their overwintering behaviour. The graph relative to nymph knows more slight pics over the years as well as for the adult population. The figure 21, represents the per stage mean total population abundance $\bar{\mathbf{N}}(t)$ for the second scenario of the second scenario S2, where *cattle* agents follow a predefined path, while running, respectively, the simulation with temperature sets T and T_s .

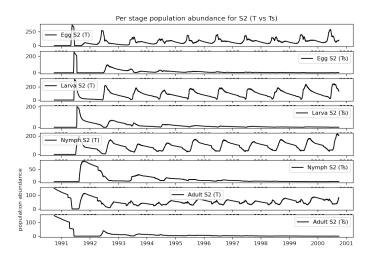


Fig. 21 Per stage mean population dynamics for S2 for both temperature sets (T and T_s

As for the first scenario, the seasonality of the cycle begins to set in from the second year. Larva and egg abundance are always higher than in nymphs and adult stages. This is directly due to mortality and lack of recruitment through the introduction of other host populations that are infected by vectors. The fluctuations of the per-stage populations of the second scenario are similar to first scenario, respectively, for both temperature sets.

7 Discussion

We developed an agent-based model to investigate the dynamics of tick populations. The model's primary goal is to evaluate the overall effects of temperature variations and host dispersion patterns on tick population dynamics. We tested different host behaviour under a set of two different temperature data sets. Over ten years of simulations we note that the population shows a decreasing trend at a smoothed temperature without interannual fluctuations for the different host behaviors. Nevertheless, the population shows a decreasing trend at smoothed temperatures. This suggests that interannual fluctuations allow the population to be maintained.

Model validation seeks to show that the computational model is a sufficiently reliable representation of the real model within the scope of use set for the simulation. The model was parameterized using empirical and estimated data, as well as expert input. First, we verified that the code programming on GAMA and the implementation of the conceptual model are correct. The validity of a conceptual system is assessed by evaluating that the theories and assumptions underlying the model are proper and that the model's representation of the problem entity, as well as the model's structure, logic and causal interactions, are "reasonable" for the model purpose [23]. In our model, we have simplified the ecology of ticks, without taking into account vertical diurnal translations and explicitly introducing intraspecific interactions. The states and actions of the model's agents (i.e. vector agent) are traced through the model to assess whether the model's logic is true We have tried to get as close to reality as possible to give more credibility to the model and formalize the natural processes of the "Host-Vector" complex.

The objective of the replication analysis of the simulations is to show that the normalized error remains small between replications (see figures 11 and 15). We also notice that the normalized error remains small and stochasticity no longer plays a role. This allowed us to calculate the average replications of the simulations for the interpretation of the effects of temperature on the population in the first place and the validation of the model in the second place. The differences found between the four simulations replications can be explained by the absence of a driving force on the population in our model (e.j. the definition of a carrying capacity for the environment) allowing to stabilize the population (figures 9 and 13). The motivation behind the choice of perfect sinusoidal temperature and comparing its effects on the population dynamics with real temperatures characterized by stochastic annual fluctuations is to detect the population response to temperature variability. The results found are contradictory to what we expected, since the stability of the temperature between the two scenarios caused the population to become extinct, which al-

lows us to conclude that the dispersion pattern of the host is not a determining factor in our model (figure 18). The smooth and stable temperatures did not allow the population to have inter-annual variability, since ecological processes such as diapause or development are temperature-dependent and therefore the said temperature created a more or less identical repetitive inter-annual pattern that led to losses in the population without remediation by forcing a gain in the population since individual fecundity is fixed and total fecundity is directly related to the number of adults present in the population.

The larger magnitudes of the results (total average abundance) of the simulations of the first scenario strategy stipulating the random movement of agent cattle, could be explained by the fact that this pattern of dispersal allows a larger space in which they may have a higher probability of encountering a vector. This will allow for greater population growth due to lower starvation mortality.

As was pointed out in the introduction (section 1), tick biology is complex and exhibits a high level of biological polymorphism observed between the different genera and even between the various species within a given genus.For instance, Our goal behind choosing the genus Ixodes is based on the fact that the genus Ixodes has various tick species that are medically significant all around the world as well as their large biogeographical distribution [10] and data availability to parameterize and validate the model. Also, the importance of our model is to be able to predict population densities as well as the seasonality of the tick activity peaks and to be able to predict population dynamics of understudied species or species we can't control their life cycle on laboratory conditions. Indeed, we show that seasonality is a primary determinant of the synchronization of distinct physiological classes.

It seems possible that these results are due to the different experimental protocols used to identify the monthly proportions of every life stage. Our reasoning was based on the fact that we can collect only active ticks, i.e. in the state guesting or moulting, which we could not be sure if the work of [16] and [24] have followed the same reasoning.

Our model is simple but realistic and is based on a multitude of essential parameters and processes deemed necessary to formalize the *vector-host* complex and could be accommodated to a wide spectrum of vector and host species and different biogeographic regions. It has a limited representation of life histories, energy allocation, and behaviour. Coupling it with a DEB model [25] in the next step, thus, would be an opportunity to explain the biological and life history traits of ticks and provide a better understanding of the physiological state of the tick and its interactions with its ecosystem, combining the deterministic aspects of DEB theory and the stochasticity of IBMs to study the effects of different factors at the population level.

Thus, our model can be extended to more ecologically complex systems with multiple species and real-world landscape complexity to test different host- and/or environment-targeted control strategies and identify effective approaches in managing vector population dynamics and dispersion.

References

- 1. S.J. Halsey, J.R. Miller, Ecological Modelling 387, 96 (2018). DOI 10.1016/j.ecolmodel.2018.09.005. URL https://doi.org/10.1016/j. ecolmodel.2018.09.005https://experts.illinois.edu/en/publications/ a-spatial-agent-based-model-of-the-disease-vector-ixodes-scapularhttps: //linkinghub.elsevier.com/retrieve/pii/S0304380018303016
- O. Tardy, C.E. Vincenot, C. Bouchard, N.H. Ogden, P.A. Leighton, Royal Society Open Science 9 (2022). DOI 10.1098/rsos.220245. URL https://royalsocietypublishing. org/doi/10.1098/rsos.220245
- H. Gaff, R. Nadolny, Mathematical Biosciences and Engineering 10, 625 (2013). DOI 10.3934/mbe.2013.10.625
- R.M. Nadolny, H.D. Gaff, Letters in Biomathematics 5, 2 (2018). DOI 10.1080/ 23737867.2017.1412811. URL https://doi.org/10.1080/23737867.2017.1412811
- C. Healy, P.J. Pekins, S. Atallah, R.G. Congalton, Ecological Complexity 41, 100813 (2020). DOI 10.1016/j.ecocom.2020.100813. URL https://linkinghub.elsevier.com/ retrieve/pii/S1476945X19300376
- 6. V. Grimm, S.F. Railsback, C.E. Vincenot, U. Berger, C. Gallagher, D.L. Deangelis, B. Edmonds, J. Ge, J. Giske, J. Groeneveld, A.S. Johnston, A. Milles, J. Nabe-Nielsen, J.G. Polhill, V. Radchuk, M.S. Rohwäder, R.A. Stillman, J.C. Thiele, D. Ayllón, Journal of Artificial Societies and Social Simulation 23 (2020). DOI 10.18564/jasss.4259. URL https://www.jasss.org/23/2/7.html
- 7. D.R. Arthur, Parasitology 39, 53 (1948). DOI 10.1017/S0031182000083554. URL https://www.cambridge.org/core/product/identifier/S0031182000083554/type/journal_article
- R.S. Lane, J. Mun, H.A. Stubbs, 34 (2009). URL http://doi.wiley.com/10.1111/j. 1948-7134.2009.00034.x
- 9. H.A. Mejlon, T.G.T. Jaenson, Experimental and Applied Acarology 21, 747 (1997). DOI 10.1023/A:1018421105231. URL https://link.springer.com/article/10.1023/ A:1018421105231
- D.E. Sonenshine, R.M. Roe (eds.), Biology of Ticks Volume 1, vol. 1, 2nd edn. (Oxford University Press, 2014). DOI 9780199744053. URL https://global.oup.com/academic/product/biology-of-ticks-volume-1-9780199744053?cc=us&lang=en&
- 11. S.G. Vail, G. Smith. Vertical movement and posture of blacklegged tick (acari: Ixodidae) nymphs as a function of temperature and relative humidity in laboratory experiments (2002). URL https://academic.oup.com/jme/article/39/6/842/860923
- W.R. Kaufman, P.C. Flynn, S.E. Reynolds, Journal of Experimental Biology 213, 2820 (2010). DOI 10.1242/jeb.044412
- J.S. Gray, O. Kahl, R.S. Lane, M.L. Levin, J.I. Tsao, Ticks and Tick-borne Diseases 7, 992 (2016). DOI 10.1016/j.ttbdis.2016.05.006. URL https://pubmed.ncbi.nlm.nih. gov/27263092/pmc/articles/PMC5659180
- V. Grimm, U. Berger, D.L. DeAngelis, J.G. Polhill, J. Giske, S.F. Railsback, Ecological Modelling 221, 2760 (2010). DOI 10.1016/j.ecolmodel.2010.08.019. URL http://linkinghub.elsevier.com/retrieve/pii/S030438001000414X
- 15. I. Harris, T.J. Osborn, P. Jones, D. Lister, Scientific Data 2020 7:1 7, 1 (2020). DOI 10.1038/s41597-020-0453-3. URL https://www.nature.com/articles/ s41597-020-0453-3https://www.nature.com/articles/s41597-020-0453-3/
- N.H. Ogden, L.R. Lindsay, G. Beauchamp, D. Charron, A. Maarouf, C.J. O'Callaghan, D. Waltner-Toews, I.K. Barker, Journal of Medical Entomology 41, 622 (2004). DOI 10.1603/0022-2585-41.4.622. URL https://academic.oup.com/jme/article-lookup/ doi/10.1603/0022-2585-41.4.622
- J.K. Vandyk, D.M. Bartholomew, W.A. Rowley, K.B. Platt, Journal of Medical Entomology 33, 6 (1996). DOI 10.1093/jmedent/33.1.6. URL https://academic.oup.com/ jme/article/33/1/6/883174
- J.L. Brunner, M. Killilea, R.S. Ostfeld, Journal of Medical Entomology 49, 981 (2012). DOI 10.1603/ME12060
- T.J. Daniels, R.C. Falco, K.L. Curran, D. Fish, Journal of Medical Entomology 33, 140 (1996). DOI 10.1093/jmedent/33.1.140. URL https://academic.oup.com/jme/ article-lookup/doi/10.1093/jmedent/33.1.140

- 20. H. Dautel, W. Knülle, Embryonic diapause and cold hardiness of Ixodes ricinus eggs (Acari: Ixodidae) (Springer Netherlands, 2010). DOI 10.1007/978-90-481-9837-5. URL http://link.springer.com/10.1007/978-90-481-9837-5
- K.M. Kocan, J.D.L. Fuente, L.A. Coburn, Parasites and Vectors 8 (2015). DOI 10. 1186/s13071-015-1185-7
- P. Taillandier, B. Gaudou, A. Grignard, Q.N. Huynh, N. Marilleau, P. Caillou, D. Philippon, A. Drogoul, GeoInformatica 23, 299 (2019). DOI 10.1007/S10707-018-00339-6/FIGURES/6. URL https://link.springer.com/article/10.1007/s10707-018-00339-6
- R.G. Sargent, Proceedings Winter Simulation Conference pp. 166–183 (2010). DOI 10.1109/WSC.2010.5679166
- R.S. Ostfeld, K.R. Hazler, O.M. Cepeda. Temporal and spatial dynamics of ixodes scapularis (acari: Ixodidae) in a rural landscape (1996). URL https://academic.oup. com/jme/article/33/1/90/883260
- 25. S. Kooijman, Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki p. 534 (2010). DOI 10. 1098/rstb.2010.0167. URL http://scholar.google.com/scholar?hl=en&btnG=Search& q=intitle:No+Title#0



Exploring and optimising infectious disease policies with a stylised agent-based model

Jeonghwa Kang • Juste Raimbault

Abstract The quantitative study of the spread of infectious diseases is a crucial aspect to design health policies and foster responsiveness, as the recent COVID-19 pandemic showed at an unprecedented scale. In-between abstract theoretical models and large-scale data driven microsimulation models lie a broad set of modelling tools, which may suffer from various issues such as parameter uncertainties or the lack of data. We introduce in this paper a stylised ABM for infectious disease spreading, based on the SIRV compartmental model. We account for a certain level of geographical detail, including commuting modes and workplaces. We apply to it a set of model validation methods, including global sensitivity analysis, surrogates, and multi-objective optimisation. This shows how such methods could be a new tool for more robust design and optimisation of infectious disease policies.

Keywords Infectious disease \cdot Agent-based modelling \cdot Model exploration and validation \cdot Multi-objective optimisation

1 Introduction

The impact of one of the fatal epidemics in history, the Black Death, has led to the death of almost one-third of the population of Europe, and the most recent epidemic in the world today, COVID-19, is currently costing millions of valuable human lives [1]. Before the introduction of mathematical and computational modelling, humans had little knowledge on how to effectively control and minimise the spread of infectious diseases. Furthermore, the importance of

J. Kang CASA, UCL

J. Raimbault LASTIG, IGN-ENSG E-mail: juste.raimbault@ign.fr

forecasting the spread of diseases in recent years has become even more crucial as the world is rapidly globalising with increasing travel rates and distances [2]. Beyond the health burden of epidemics, many dimensions of social systems are affected, with for example broad economic impacts [3]. The current world is a highly complex system in which numerous dimensions play a role in decisionmaking [4]. Designing disease control strategies imply accounting for various interacting factors and finding a compromise between multiple objectives [5].

The first modelling approach to disease spreading goes back to the 17th century as John Graunt conducted an empirical study on infections affecting individuals in various regions in Britain [6]. Bernouilli introduced in the 18th century an equation-based approach to studying the outbreak of the smallpox epidemic in Europe [7]. Most recent modelling approaches range from elaborated mathematical models to data-driven microsimulation models. The recent pandemic has shown the usefulness of quantitative modelling to evaluate policies, such as how to vaccinate the population [8] or to evaluate which degree of lockdown is necessary.

Agent-based modelling are proposed by some researchers as a powerful tool to explore practical scenarios for decision-making when managing the spread of infectious diseases [9]. The classical epidemiological modelling relies on compartmental models, such as the SIR model (Susceptible, Infected, Recovered), with differential equations governing the evolution of the different compartments, or in some cases discrete-time equations [10]. Recent extensions, such as the SEIRD model [11], include a more detailed description of disease stages. [12] describe an agent-based model based on similar states for agents. The advantage of the ABM approach is that one can define for each agent characteristics such as location, age, and many other factors believed necessary in the model. Once the agents are defined with their own set of characteristics, they proceed through a set of rules which enable them to interact with other groups of agents within the simulation. Age groups have contact probabilities to simulate the interactions and infections. More details can be included in an ABM, such as wearing face masks or respecting social distancing [13]. Furthermore, the infectious disease may interact with other chronic conditions, what can be accounted for in an ABM [14].

We propose in this paper to explore the potential of model validation methods to design and optimise epidemiological policies. We introduce a stylised model for the spread of infectious diseases, which is hybrid in the sense that it lies halfway large data-driven models and abstract theoretical models. We account for geographical factors, in particular commuting and workplaces. We focus, as an illustration, for two policy factors, which are social distancing and face masks [15], and vaccination [16]. Our main contribution is the application of state-of-the-art model sensitivity analysis and validation techniques [17] to such a stylised model, opening research perspectives to cases where data is lacking, model parametrisation is uncertain, or actors have not the ressources to implement a large scale model, but where accounting for spatial and social structure remains crucial. We do not provide directly actionable policy insights, but rather provide a proof-of-concept of the use of such techniques to extend model capabilities: since our model lies between simple but scalable compartment models and very complicated data-driven microsimulation models, it could in principle not be used to investigate scenarios and decision-making. Our approach allows reducing uncertainty on parameters and providing multi-objective optima for example, improving thus the applicability of the model.

A few examples of the application of such techniques to epidemiological models can be found in the literature. [18] utilise a multi-objective optimisation algorithm for the distribution of limited vaccines resources. [19] review sensitivity analysis of infectious disease models. No systematic application of an umbrella of model validation methods can however be found.

The rest of this paper is organised as follows: we first describe the stylised agent-based model, we then apply model exploration and validation methods to it, and finally discuss the implications of our results.

2 Model description

2.1 Rationale

Our simulation model projects a real-life situation where agents commute from home to their designated workplaces via different modes of transportation on a fixed daily schedule. At the beginning of the simulation, the entire population is set to susceptible. Then, an infected agent is introduced daily for the first 30 days because if none of the infected agents gets introduced into the environment, the infection process does not happen. The infection then spreads through localised interactions between agents.

2.2 Model setup and dynamics

One main assumption of the model is that most human interactions occur while people travel and work (while home infections play a role in disease transmission, our model is built to focus on parameters linked to social distancing in transport and at work). Hence, the scenario is primarily about people commuting from home to their designated workplaces via different modes of transportation, closely depicting ordinary people's daily routine. The simulation of this model considers several essential assumptions. The entire duration of the simulation is set to 365 days. Each day is equivalent to 24 ticks. Hence, each tick is equivalent to an hour. A single day comprises three pre-defined phases: the 'Home-hour', 'Commute-hour and 'Work-hour' phases, as show in Fig. 1.

During these phases, the movement of agents is completely randomised. No learning behaviour alternates the agents' movement in the simulation since it would be difficult for an individual to be fully aware of another agent's infection status. During the interaction, agents fall into one of the four states defined by

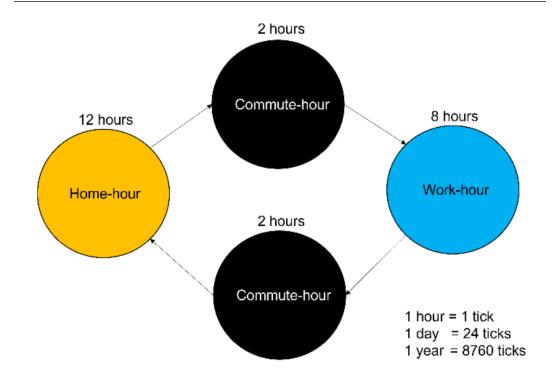


Fig. 1 Transitions between phases in the model.

the SIRV compartmental model. The SIRV model is an extended SIR model that accounts for the vaccination of the susceptible population [20]. The agent states are shown in Fig. 2.

Agents can interact with each other based on the state in which they fall. However, the interaction and infection processes are set to occur between the agents within the same patch only. The infection does not happen during the 'Home-hour' phase as this model does not consider social activities or household infections. During the 'Commute-hour' phase, an infected agent can only infect the susceptible agents that use the same mode of transportation as itself. It would be logically sensible that an infected agent commuting via train can only interact with and infect other susceptible agents commuting via the same mode of transportation. This same logic applies during the 'Work-hour' phase, where an infected agent can only interact with and infect other susceptible agents working in the same company. We include four modes of transportation in the model (metro, train, walking, car). The stylised way transportation is described allows avoiding the complexity of transport microsimulation (see e.g. [21] for an adaptation of the MATSim framework to epidemiological modeling) while still accounting for transport processes. We furthermore consider stylised company size distributions (7 large companies, 4 medium sized, 4 small). There are total of fifteen companies which are randomly categorised into small, mid, and big-sized companies. The central assumption is that the bigger the company size, the more the human interactions involved. Hence, the agents working in a big-sized companies are more likely to get infected than the

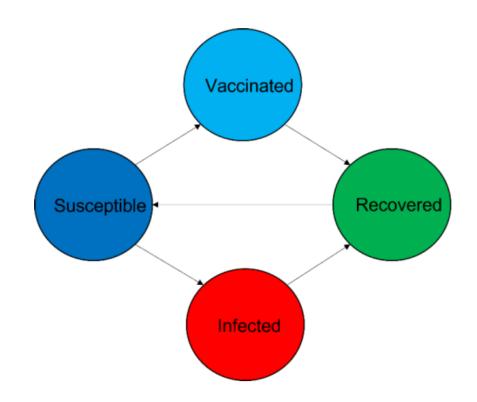


Fig. 2 Agent states in the model

agents working in mid and small-sized companies (in practice, this assumption could be refined, as infections also depends on the size of the premises and on ventilation; simulating density and real-time interaction remains however out of the scope of our stylised approach).

At the beginning of the simulation, N = 1500 susceptible agents are randomly distributed in the simulation environment. The variable values for the agents, such as the transportation and company, are randomly assigned during the setup stage. An infected agent is then constantly introduced at a random location for the simulation's first month (30 days), to take into account possible transient dynamics. Every day, a random percentage value from zero to one percent of the total population gets replaced. This is based on the assumption that the population is dynamic (considering thus an open system, consistently with the intermediate complexity of our model). A certain number of people leave the area, and others come to stay in the area. Hence, the total population is not fixed to the initial population throughout the simulation.

Model dynamics are summarised in Fig. 3.

2.3 Model parametrisation

We give in this section more details on the parametrisation of the stylised model.

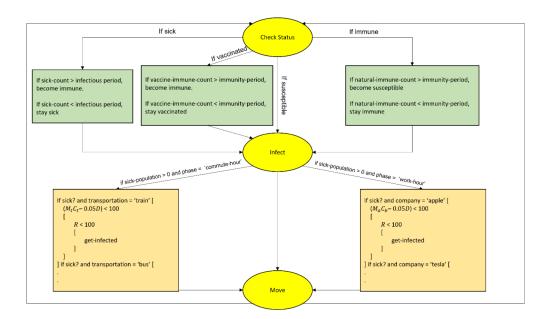


Fig. 3 Description of model dynamics. Processes are described using the NetLogo syntax, as used in the model implementation.

2.3.1 Global variables

Infection rates are calculated based on global and local parameters. The variables that define contact rates for each mode of transportation and company are pre-defined with specific values.

For transportation contact rates, we take stylised values decreasing with the density one can expect in each transport mode: $c_{bus} = 40\%$, $c_{metro} = 35\%$, $c_{train} = 30\%$, $c_{walk} = 10\%$, $c_{car} = 0\%$.

No discrete values define the accurate contact rates for different modes of transportation, but the general assumption was made based on the understanding that the lower the capacity of transportation, the higher the density; hence, the higher contact rates between the passengers. The contact rate specifically for the car is set to zero because it is assumed that an agent who commutes by a car makes no contact with other agents during the 'Commutehour' phase. The same logic was applied when defining the contact rates for different companies. The assumption was based on the understanding that the bigger the company size, the more the agents interact.

For contact rates on work site, we therefore take the following values: $c_{Big} = 40\%$, $c_{Medium} = 35\%$, $c_{Small} = 30\%$. These values do not have an empirical basis, and should be integrated in the sensitivity analysis, would the model be applied to a real-world decision making problem (in the following, we provide only a proof-of-concept with few other parameters).

2.3.2 Agent variables

The importance of wearing a mask while interacting with an infected agent can significantly reduce the chance of infection. Furthermore, patients with cardiovascular diseases are possibly with a higher risk of getting an infection [22] (this aspect could also capture individuals with a deficient immune system for example). Due to these reasons, we have pre-defined values for infection risk rates based on the combination of these two variables: No mask & Cardio = 50% risk rate; Mask & Cardio = 45%; No mask & No cardio = 15%; Mask & No Cardio = 10%. We do not consider the fatality of the disease, but these would follow similar patterns.

2.4 Model parameters

We list below main model parameters which will vary during numerical experiments, corresponding to the processes on which the model focuses.

- Infectious period, ranging from 2 to 4 weeks; the longevity of the infectious period of an agent is believed to play an essential role in the spread of diseases because the longer the infected agent stays contagious, the more the chance of infection during an extended amount of time [23].
- Immunity period, ranging from 9 to 11 months. This is selected as one of the models input parameters because it is believed to play a vital role in protecting the agents from infection [24]. Hence, we are interested to know to what extent the immunity period significantly affects the spread of diseases.
- Vaccination rate: as discussed in the literature review, vaccines are the most effective preventive measure which can trigger a biological immune response to fight disease-causing organisms: beside strongly reducing the fatal outcomes, they also induce a decrease in infection rates (variable depending on diseases and vaccines, for example some estimates of effectiveness against infection on particular Covid-19 variants and vaccines were mainly above 50% 30 days after injection [25]). The value ranges from 0 to 0.5 %, meaning from 0 to 0.5 % of the susceptible population is set to be vaccinated daily. A high vaccination rate is strongly believed to be the key to a successful control of epidemics; hence, finding out the impacts of different vaccination rates on the overall infection process is meaningful.
- Social distancing level: a certain level of social distancing is being applied by governments worldwide, hoping to stop the spread of diseases like COVID-19. However, the effect of social distancing is not yet well known. Due to this, it would be an interesting experiment to find out the effectiveness of such measures on disease transmission dynamics. The level of social distancing ranges from 1 to 3. Each level refers to a percentage value that reduces the contact rate by 5. Hence, the actual percentage value of the social distancing ranges from 5 to 15 percent. The higher the intensity of social distancing, the lesser the chance people contact each other. In

practice, quantifying social distancing is complicated [26], thus this stylised mechanism which should be refined in case of a practical policy application of our model.

2.5 Model indicators

As mentioned previously, the ABM is constructed following a structure similar to the SIRV compartmental model. Therefore, the indicators consist of proportional numbers of agents in each compartment, including susceptible, infected, recovered, and vaccinated.

2.6 Model implementation and exploration

The model is implemented in NetLogo, which is a reasonable choice for such hybrid models in which visualisation is important. It is integrated into the Open-MOLE platform for model exploration [27] to carry out the numerical experiments described below. Source code of the model and exploration scripts are openly available on a git repository at https://github.com/henry-kang-7/CASA0004.

The pre-defined input parameter values of the model are the following: infectious-period $\in \{2, 3, 4\}$ (weeks); immunity-period $\in \{4, 5, 6\}$ (months); vaccination-rate $\in \{0, 0.25, 0.5\}$ (%); social-distancing-levels $\in \{1, 2, 3\}$ (levels).

This results in 81 different combinations of the input parameter values for a grid sampling. The model is set to run over 125 times for every combination of the input parameter values, and the output parameters' average values are obtained at the end. The size of the 95% confidence interval around the estimated average of a normal distribution of standard deviation σ , with nreplications, is given by $2 \cdot 1.96 \cdot \sigma / \sqrt{n}$, and therefore n = 125 ensures that the error on indicators is more than ten times smaller than the standard deviation.

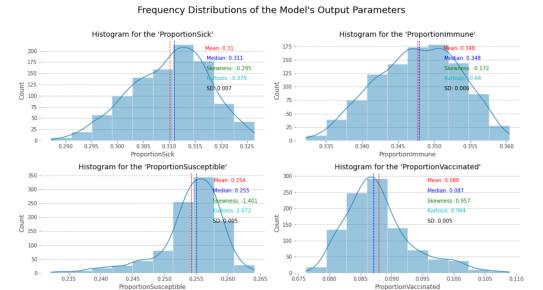
3 Results

We describe now numerical results obtained through the extensive exploration with the OpenMOLE platform.

3.1 Statistical distribution of indicators

The LHS (Latin hypercube Sampling) method was used to explore the input space of the model. A total of 10,000 samples were drawn during the process (corresponding to 4 days and 4h of CPU time for model execution), used to study statistical properties of model indicators.

The histograms in Fig. 4 illustrate the four different output indicators of the model. The histograms in the upper row comprise the 'ProportionSick' and the



0.240 0.245 0.250 0.260 0.255

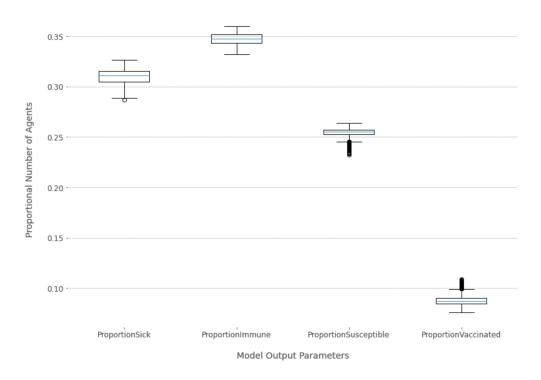
Fig. 4 Histograms of model outputs.

'ProportionImmune' histograms, which appear symmetrical, but the skewness values indicate the traits of negative skewness on each histogram plot. Both kurtosis values are negative, suggesting that the distributions are thin-tailed. From this finding, we can assume that they generally follow characteristics of the platykurtic distribution. The 'ProportionSusceptible' and the 'ProportionVaccinated' histograms in the lower row appear negatively and positively skewed, respectively. The 'ProportionSusceptible' histogram is highly skewed to the left, which is indicated by the skewness value lesser than -1, while the 'ProportionVaccinated' histogram appears moderately skewed to the right as its skewness value is lower than 1. Furthermore, the kurtosis values of the two histograms are positive, which indicates that they both have heavier tails and are likely to follow the shape of the leptokurtic distribution.

The boxplots presented in Fig. 5 show different data ranges and medians of the four different output parameters of the model. The boxplot representing the 'ProportionImmune' appears to have the highest median among the groups, followed by the 'ProportionSick', 'ProportionSusceptible' and 'ProportionVaccinated', respectively. Outliers in the boxplots are present regarding the 'ProportionSusceptible' and 'ProportionVaccinated'. Based on the histograms shown in Fig. 4, these distributions have positive kurtosis values and are disproportionally distributed, indicating low standard deviation; hence, high chance of producing outliers.

3.2 Grid exploration

The total combination of the model's input parameters with predefined values is 81. The model run for each combination of the inputs is set to 125 times as



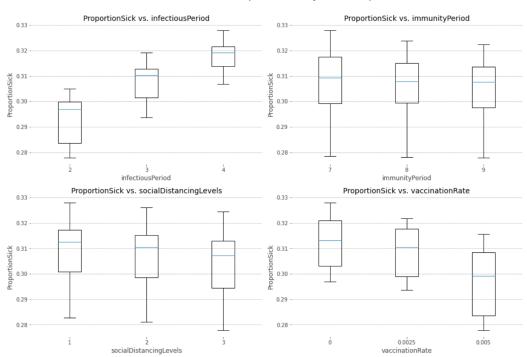
Statistical Distributions of Agents by the Model's Ouput Parameters

Fig. 5 Box plots of model outputs

explained previously. This results in 10,125 model runs, for a total execution time of 4 days and 5 hours.

The box plots in Fig. 6 show statistical distributions of the 'Proportion-Sick', which are grouped by each input parameter value of the model. The range of each input parameter is indicated on the horizontal axis of the graphs. The first graph represents the grouped values of the 'ProportionSick' by the values of the 'infectiousPeriod'. It indicates that a unit increase in the infectiousPeriod' value increases the boxplot's overall median and max values. The boxplots related to the 'immunityPeriod' and the 'socialDistancingLevels' show a marginal decrease in the median of the boxplots per unit increase. The last graph is related to the 'vaccinationRate', which indicates that a 0.0025 (0.25 %) increase in the rate decreases the boxplots overall values. It indicates that vaccines have an apparent effect on successfully controlling the spread of diseases.

The correlation matrix illustrated in Fig. 7 displays the correlation between different model parameters. The 'ProportionSick' and the 'infectiousPeriod' appear to be highly correlated. According to the box plots shown in Fig. 6, every unit increase in the 'infectiousPeriod' also increased the 'ProportionSick' values. The rest of the input parameters, including the 'immunityPeriod', 'so-cialDistancingLevels' and 'vaccinationRate' are negatively correlated with the 'ProportionSick' at different intensity levels.



Statistical Distributions of the 'ProportionSick' by Model's Input Parameters

Fig. 6 Box plots showing aggregate values of model outputs based on each value of inputs.

An OLS regression with the simulated data shows that the spread of infection mainly depends on the longevity of the infectious period. The longevity of the infectious period plays a vital role in disease transmission because a more extended infectious period allows an infected agent to interact with a more significant number of susceptible agents for an extended time. For example, an infected agent who is infectious for a single day is unlikely to infect a more significant number of susceptible agents than an infected agent who is infectious for a week. A more extended infectious period is, therefore, a dangerous factor that may rapidly increase the chances of infections, often creating a mass infection wave.

3.3 Model surrogate

Then, a random forest model is trained with the original data obtained by the LHS method during the first step of the analysis to forecast the proportional number of sick agents. This supervised machine learning method is versatile, easy to use and train, but also provides interpretable results on the relative importance of features, and we use it for these reasons.

In order to train the random forest model to predict the values of the 'ProportionSick', the original data is categorised into train, test, and validation data in the ratio of 7:1.5:1.5. Before fitting the model with the train data, various numbers of trees with their corresponding accuracy scores are calculated.

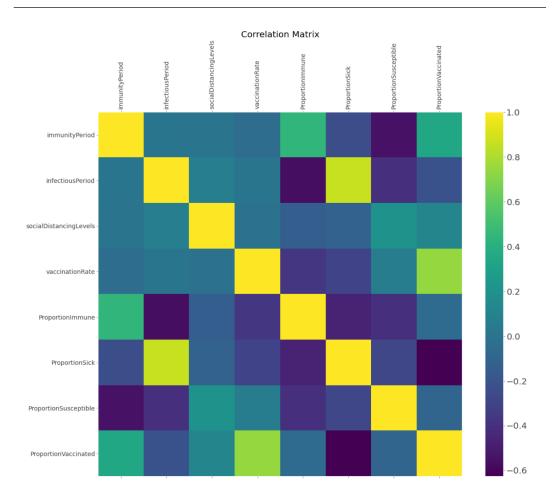
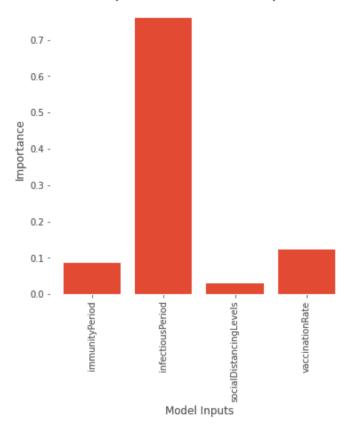


Fig. 7 Correlation matrix of model parameters and indicators.

Then, this hyper-parameter for the model was optimised using the grid search method for a given set of estimator values, including 800, 900, 950, 1000 and 1050. As a result, 900 is the best value for optimising the random forest model. The random forest with the 900 decision trees was developed by fitting the training data to predict the values of the 'ProportionSick' and including all four input parameters of the model.

We obtain a MAE (how big of an error we can expect from the forecast on average) of 0.001, which implies that the average errors between the predicted and actual values are marginal; hence, the predictions are considered highly accurate. Moving on to the MAPE, the value of 0.363 indicates that, on average, the forecast is off by 0.363%. This means that the accuracy of the forecast is as high as 99.637%. Lastly, the value of the RMSE is also 0.001, which indicates that the difference between the predicted and actual values is marginal.

The random forest model is furthermore useful to quantify the importance of the model's input parameters, shown in Fig. 8. The 'infectiousPeriod' is by far the most significant input parameter of the model, with a value of approximately 0.75, followed by the 'vaccinationRate' with a value of slightly over 0.1.



Importance of the Model Inputs

Fig. 8 Feature importance of input parameters obtained with the random forest surrogate.

Table 1	1	Saltelli	total	order	indices

Outputs	infectiousPeriod	immunityPeriod	vaccinationRate	socialDistancingLevels
ProportionImmune	0.72	0.333	0.665	0.7
ProportionSick	0.692	0.09	0.365	0.143
ProportionVaccinated	0.516	0.319	0.587	0.337
ProportionSusceptible	0.341	0.066	0.92	0.277

The 'immunityPeriod' and the 'socialDistancingLevels' have an importance rate below 0.1, showing less importance than the previous two parameters.

3.4 Global sensitivity analysis

The previous results of parameter importance can be compared with an other method to quantify parameter's influence on indicators. The global sensitivity analysis method, introduced by [28], gives a broad summary of such an influence. In order to generate an accurate result for the sensitivity analysis, a total of 60,000 samples were run (25 days 7 hours of model runtime).

The total-order indices shown above indicate that the 'infectiousPeriod' has the most significant impact on the values of the 'ProportionSick' followed

by the 'vaccinationRate', 'socialDistancingLevels' and 'immunityPeriod' respectively. We find a very low effect of the immunity period on the proportion of sick and susceptible, meaning that uncertainties on this value will have a limited impact. The vaccination rate is also a critical input parameter of the model that plays a vital role in controlling the spread of diseases because the vaccine is a significant source of protective measures that can effectively stop viruses from transmitting one agent to another [29].

3.5 Multi-objective model optimisation

In attempting to control the spread of epidemics, many problems involve multiple objectives which cannot simply be described as the more, the better or the lesser, the better; instead, each objective has an ideal target value, and the main goal is to be as close as possible of the targeted value. To optimise the values of the input parameters, we define two objectives that are expected to be minimised. Firstly, we expect the proportional number of sick agents to be optimised, which is directly linked to the main interest of our research question. Secondly, we expect the proportional number of vaccinated agents to be optimised. Vaccination is a primary preventive measure that plays a vital role in successfully controlling the spread of diseases and even has the potential to achieve herd immunity, while other measures such as social distancing can only slow down the process of disease spreading instead of stopping them [16]. However, vaccines are not always available for multiple reasons, such as the high manufacturing and transportation costs [30]. This is where decisionmakers often face a dilemma because attempting to control epidemics will cost money and the best option which they might have at this point is to find a good balance between the use of vaccines and the proportional number of sick agents, respecting the economic budget.

The way to find the results that provide a good approximation of the Pareto frontier with acceptable trade-offs between the identified objectives is by using the NSGAII algorithm. The NSGAII is a commonly used type of multi-objective optimisation algorithm which includes a fast non-dominated sorting, crowding distance assignment and a sorting procedure [31].

The Pareto Front consisting of non-dominated solutions relative to the 'ProportionSick' and 'ProportionVaccinated' is drawn after running a total of 10,000 generations of the OpenMOLE implementation of the algorithm.

The Pareto Front shown in Fig. 9 illustrates the optimal solution set of the 'ProportionSick' and the 'vaccinationRate' as two objectives. The value of the 'ProportionSick' appears to decrease rapidly as the value of the 'vaccinationRate' increases. The slope of the line is generally steep throughout the graph. Plus, the difference between the values of the 'ProportionSick' with and without the vaccination is clear. The non-linearity of the front is an interesting feature, witnessing the underlying complexity when exploring policy trade-offs.

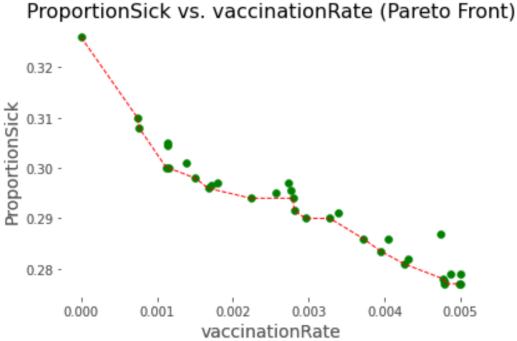


Fig. 9 Pareto front of proportion of sick vs. vaccination rate.

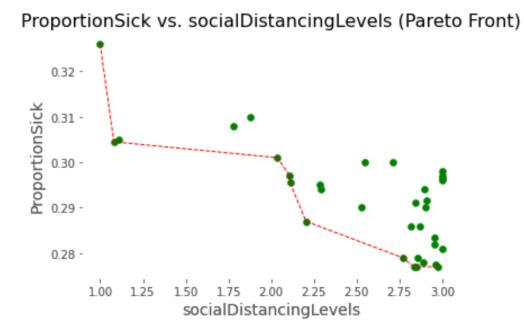


Fig. 10 Pareto front of proportion of sick vs. social distancing levels.

The Pareto Front, visualised in Fig. 10, is formed with the optimal values of the 'ProportionSick' and the 'socialDistancingLevels' as two objectives. It indicates a clear sign of a decrease in the value of the 'ProportionSick' as the value of the 'socialDistancingLevels' increases. In some regions of the front, a quasi-vertical increase means that much control may be gained at a very low social cost, what has implications when finding the compromise between the epidemic impact and the social impact of lockdowns. The social distancing level is indeed an essential input parameter that acts as a preventive measure against infectious diseases in our model. Social distancing must be implemented with care because if the intensity is too high, it prohibits necessary human interactions, which reduces the productivity rate, negatively impacting the economy in general [32]. Due to this, finding a good balance between the social distancing level and the proportional number of sick agents would be essential in ensuring a healthy economy while keeping the number of infected cases low.

4 Discussion

The diverse numerical experiments, applying various model exploration and validation methods, can be used to provide useful policy insights in this stylised case. They are not directly applicable to real world situations, as a slightly more data-driven approach would be needed, but this already suggests how such methods can compensate for model simplifications. Thus, we suggest that our work is a proof-of-concept of how such systematic model exploration may hinder issues due to parameter uncertainty, or the lack of data, for similar epidemiological models.

Some limitations of this work can be given. First, the population size of the simulation had to be limited to a small number, mainly due to the expensive computational cost and time. In order to fasten the process, these executions have been distributed using OpenMOLE, but the limitation in the level of detail and granularity is a recurring issue with ABMs. We still need to investigate if the trade-off chosen here is reasonable.

The current number of input parameters is limited to only four parameters. However, there can be more numbers of input parameters introduced in our model to generate more dynamic results which can better reflect reallife situations. Furthermore, the current model can be extended with more compartments that can describe various states of an agent in the infection process, such as death or exposure [33]. The model is also flexible in that it can add many more characteristics of agents, such as age, occupation, vaccination record and gender, which can be considered in the process of interaction and infection.

As a future work, regarding the stylised parametrisation taken here for several aspects of model setup, a more advanced sensitivity analysis relaxing these parameters could be done. More particularly, quantifying the role of geography (transport geography, but also company size distribution and locations), is an interesting prospect, as new methods to achieve this have recently been developed [34].

5 Conclusion

This study proposed various model assessment methods to explore results obtained with a stylised ABM we introduced, based on the SIRV compartmental model to study the spread of diseases. Through the model exploration, this study analysed the impacts of the pre-defined input parameters, which included the infectious period, immunity period, daily vaccination rate and social distancing level, on the spread of infectious diseases. We provide thus a proof-of-concept of how model exploration and validation methods can be a powerful tool to design and optimise health policies.

References

- 1. K.A. Glatter, P. Finkelman, The American journal of medicine 134(2), 176 (2021)
- 2. L. Saker, K. Lee, B. Cannito, A. Gilmore, D.H. Campbell-Lendrum, et al., (2004)
- R. Boucekkine, A. Carvajal, S. Chakraborty, A. Goenka, Journal of Mathematical Economics 93, 102498 (2021)
- S.J. Bickley, H.F. Chan, A. Skali, D. Stadelmann, B. Torgler, Globalization and health 17(1), 57 (2021)
- 5. J. Kaszowska-Mojsa, P. Włodarczyk, A. Szymańska, Entropy 24(1), 126 (2022)
- 6. A. Morabia, Epidemiology (Cambridge, Mass.) 24(2), 179 (2013)
- 7. K. Dietz, J. Heesterbeek, Mathematical biosciences 180(1-2), 1 (2002)
- P. Bosetti, C. Tran Kiem, A. Andronico, V. Colizza, Y. Yazdanpanah, A. Fontanet, D. Benamouzig, S. Cauchemez, BMC medicine 20(1), 33 (2022)
- F. Miksch, B. Jahn, K.J. Espinosa, J. Chhatwal, U. Siebert, N. Popper, PloS one 14(8), e0221564 (2019)
- A. Ramani, A. Carstea, R. Willox, B. Grammaticos, Physica A: Statistical Mechanics and its Applications 333, 278 (2004)
- 11. G. Chowell, M. Miller, C. Viboud, Epidemiology & Infection 136(6), 852 (2008)
- D. Chumachenko, V. Dobriak, M. Mazorchuk, I. Meniailov, K. Bazilevych, in 2018 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET) (IEEE, 2018), pp. 192–195
- 13. J.M.A. Minoza, V.P. Bongolan, J.F. Rayo, arXiv preprint arXiv:2101.11400 (2021)
- G. Fekadu, F. Bekele, T. Tolossa, G. Fetensa, E. Turi, M. Getachew, E. Abdisa, L. Assefa, M. Afeta, W. Demisew, et al., International Journal of Physiology, Pathophysiology and Pharmacology 13(3), 86 (2021)
- S. Kwon, A.D. Joshi, C.H. Lo, D.A. Drew, L.H. Nguyen, C.G. Guo, W. Ma, R.S. Mehta, F.M. Shebl, E.T. Warner, et al., Nature Communications 12(1), 3737 (2021)
- M. Bicher, C. Rippinger, G. Schneckenreither, N. Weibrecht, C. Urach, M. Zechmeister, D. Brunmeir, W. Huf, N. Popper, Scientific Reports 12(1), 2872 (2022)
- 17. J. Raimbault, D. Pumain, Geographical Modeling: Cities and Territories 2, 125 (2019)
- E.G. Baquela, A.C. Olivera, in Humanitarian Logistics from the Disaster Risk Reduction Perspective: Theory and Applications (Springer, 2022), pp. 273–291
- J. Wu, R. Dhingra, M. Gambhir, J.V. Remais, Journal of The Royal Society Interface 10(86), 20121018 (2013)
- R.C. Poonia, A.K.J. Saudagar, A. Altameem, M. Alkhathami, M.B. Khan, M.H.A. Hasanat, Life 12(5), 647 (2022)
- 21. K.W. Axhausen, in Institute of Space and Earth Information Science (ISEIS) seminar at the Chinese University of Hong Kong (IVT, ETH Zurich, 2021)

- 22. N. Ielapi, N. Licastro, M. Provenzano, M. Andreucci, S.d. Franciscis, R. Serra. Cardiovascular disease as a biomarker for an increased risk of covid-19 infection and related poor prognosis (2020)
- 23. R.R. Wilkinson, K.J. Sharkey, Physical Review E **97**(5), 052403 (2018)
- 24. J. Reyes-Silveyra, A.R. Mikler, Theoretical Biology and Medical Modelling 13, 1 (2016)
- I. Mohammed, A. Nauman, P. Paul, S. Ganesan, K.H. Chen, S.M.S. Jalil, S.H. Jaouni, H. Kawas, W.A. Khan, A.L. Vattoth, et al., Human vaccines & immunotherapeutics 18(1), 2027160 (2022)
- P. Caley, D.J. Philp, K. McCracken, Journal of the Royal Society Interface 5(23), 631 (2008)
- R. Reuillon, M. Leclaire, S. Rey-Coyrehourcq, Future Generation Computer Systems 29(8), 1981 (2013)
- A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, *Global sensitivity analysis: the primer* (John Wiley & Sons, 2008)
- C.B. Storlie, B.D. Pollock, R.L. Rojas, G.O. Demuth, P.W. Johnson, P.M. Wilson, E.P. Heinzen, H. Liu, R.E. Carter, E.B. Habermann, et al., in *Mayo Clinic Proceedings*, vol. 96 (Elsevier, 2021), vol. 96, pp. 1890–1895
- S. Plotkin, J.M. Robinson, G. Cunningham, R. Iqbal, S. Larsen, Vaccine 35(33), 4064 (2017)
- K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, IEEE transactions on evolutionary computation 6(2), 182 (2002)
- 32. S. DeLuca, N. Papageorge, E. Kalish, Johns Hopkins coronavirus resource Center (2020)
- B. Reyné, N. Saby, M.T. Sofonea, Anaesthesia, Critical Care & Pain Medicine 41(1), 101017 (2022)
- J. Raimbault, C. Cottineau, M. Le Texier, F. Le Nechet, R. Reuillon, Journal of Artificial Societies and Social Simulation 22(4) (2019)



Contact networks in daily-life pedestrian crowds and risks of viral transmission

Alexandre NICOLAS · Simon MENDEZ

Abstract In order to assess the risks of short-range airborne transmission of a virus in pedestrian crowds, detailed information must be gathered about the network of complex interactions between people in the crowd. We have collected such detailed data on the field, in various situations, and developed a methodology to estimate the rate of new infections in the crowd. The method relies on coarse-graining simulations of aerosol trajectories at the microscale, in numerous ambient flows, into spatio-temporal maps of viral concentration around a potential emitter. Through this coupling, we are able to rank the situations that we have investigated by the risks of viral transmission that they raise; our study also highlights that even the most modest air flows dramatically lower the quantitative rates of new infections.

Keywords pedestrian crowds · epidemiology · COVID-19

1 Introduction

The COVID-19 pandemic has put in the limelight the need for epidemiological models capable of predicting the spread of respiratory diseases. While largescale models may rely on a coarse description of the patterns of contacts in a population, in order to get a more accurate view of the transmission risks within a given setting, a finer description of the complex network of interactions between people and of the propagation of the viral carriers (namely, aerosols) is needed.

S. MENDEZ

A. NICOLAS

Institut Lumière Matière, CNRS, Univ. Lyon 1, Villeurbanne, F-69622, France Tel.: +33-4.72.44.82.37. E-mail: alexandre.nicolas@cnrs.fr

IMAG, Univ. Montpellier, CNRS, Montpellier, F-34095, France E-mail: simon.mendez@umontpellier.fr

Taking up this challenge, we have collected a set of field data about crowds in daily-life settings and extracted the information relevant for the estimation of viral transmission (in particular, the positions and head orientations of the people) [1]. In parallel, we have run extensive Computational Fluid Dynamics (CFD) simulations of the propagation of exhaled respiratory micro-droplets and aerosols in various ambient conditions (notably varying the wind speed) and derived from these simulations coarse-grained maps of transmission risks between an emitter and a receiver [2]. Coupling the field data with these coarsegrained maps of transmission risks provides a particularly efficient means to estimate the risks of respiratory disease transmission within a given crowd.

Here, we will insist mostly on the detailed contact networks obtained from our empirical data and the results obtained by coupling them to coarse-grained CFD results; details pertaining to the simulations can be found in [2].

2 Empirical interactions and contact networks in pedestrian crowds

In 2020 and 2021, during the COVID-19 pandemic, we collected an original set of empirical data about crowds in non-confined settings, at various locations: streets, river banks train and metro stations, an outdoor market, a screening centre, as well as street cafés. The positions of pedestrians, together with their head orientations, were manually extracted; for each scenario, several hundreds of pedestrian trajectories were typically gathered. These data are freely available at https://zenodo.org/record/4527462.

From these data, networks of 'contacts', often arbitrarily defined by the criterion that two people should stand within a given distance R_c (set to 1.5m here), are obtained. Figure 1 displays some part of these contact networks for two scenarios, which underlines the difference in their structure. Still, we find that the topology of these networks is quite sensitive to the R_c -criterion used to define it.

3 Propagation of virus-laden aerosols and coarse-grained viral concentration maps

Epidemiological models would typically try to predict the spread of the disease on the basis of contact networks such as those of Fig. 1. This prompts questions regarding the robustness of these models. Indeed, as emphasised above, empirical contact networks are sensitive to the arbitrary distance criteria used to define them and, perhaps more alarmingly, they often overlook the contact duration [represented by the edge widths in Fig 1(b-d)].

Moreover, the underlying assumption of isotropic transmission around the emitter is opposed by the actual propagation of the disease carriers (i.e., virusladen aerosols), illustrated in the CFD simulations of Fig. 2, which exhibits rich spatio-temporal dynamics. Instead, we decided to take into account the fine spatio-temporal details of the interactions between people and to resort

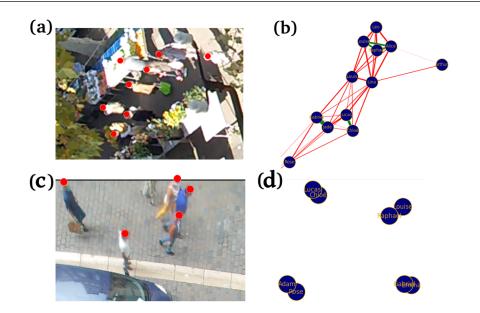


Fig. 1 (a) Snapshot of the crowd at an outdoor market in Lyon, France (January 2021). (b) Zoom on the associated contact network established over 1 minute of recording. (c) Snapshot of a 'pedestrian' street in the Old Town of Lyon, France (September 2020). (d) Zoom on the associated contact network established over 2 minutes of recording. The edge widths are proportional to the contact duration, defined as the time spent by two people within $R_c = 1.5$ m of each other; green edges connect people identified as belonging to the same group, and given names were of course chosen arbitrarily.

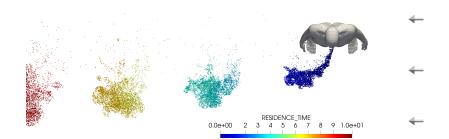


Fig. 2 CFD simulations of the respiratory micro-droplets and aerosols exhaled by a static pedestrian, in an ambient wind blowing at 0.3 m/s.

to extensive microscopic CFD simulations to derive the risks of transmission around a contagious person, depending on the external wind, the walking speed, and the expiratory activity (mouth-breathing, speaking, ...). This would be computationally intractable if one CFD simulation had to be performed for each and every situation, which is why we relied on coarse-grained maps describing the average transmission risks around an emitter, depending on his or her relative speed with respect to the ambient air (Fig. 3).

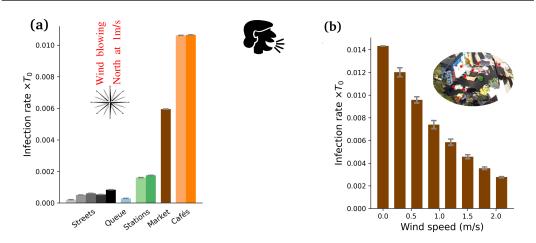


Fig. 3 (a) Estimated risks of new infections in the crowd, depending on the scenario and the activity of the supposedly contagious individual, in a *very* light breeze (1 m/s). (b) Variations of the risks of new infections with the wind speed in the outdoor-market scenario. The expiratory activity under study is 'normal speech'.

4 Results: Estimated transmission risks in the crowd

Coupling these coarse-grained dynamic maps with our field data, we were able to estimate the rate of new infections (assuming that one individual in the crowd is contagious) in diverse situations, for different wind conditions. Figure 3(a) shows that, among the situations under study, street cafés clearly raise the highest risks, even though these risks are dramatically lowered in the presence of (even moderate) wind [Fig. 3(b)]. These are our major findings.

The light-weight strategy¹ mediated by coarse-graining is computationally appealing for screening and guidance purposes, but it naturally has limitations in terms of accuracy, for several reasons: coarse-graining and interpolating viral concentration maps are not exact operations; assuming that exhalation and inhalation are decoupled processes is a fairly crude approximation; the assumption that each virus has an equal chance to cause an infection, regardless of the size of the aerosol in which it is encased may be found inaccurate.

References

- 1. W. Garcia, S. Mendez, B. Fray, A. Nicolas, Safety Science 144, 105453 (2021)
- 2. S. Mendez, W. Garcia, A. Nicolas, Advanced Science [arXiv:2208.03147] (in press)

¹ A Python implementation is available under: https://github.com/an363/InfectiousRisksAcrossScales



Minimizing epidemic spread through mitigation of anti-vaccine opinion propagation

Sarah Alahmadi · Rebecca Hoyle · Markus Brede

Abstract In this study we investigate the impact of vaccination attitudes, as a social contagion, on disease dynamics. Vaccine-related negative information poses a significant concern since it is a major motivator of vaccination hesitancy. This information may cause social contagion of anti-vaccination opinions, resulting in clusters of unprotected people and potentially larger epidemic outbreaks. Therefore, this research aims to minimize the spread of an epidemic by mitigating anti-vaccine social contagion with an effective countercampaign that promotes vaccination. In a coupled-dynamic model describing processes of opinion diffusion and disease spreading, we propose a number of techniques to mitigate the propagation of anti-vaccine opinions and prevent the expansion of anti-vaccine communities. We observed that the existence of pro-vaccine sentiments has a varying impact on the incidence of epidemics. Additionally, our research demonstrates that targeted positive campaigns can effectively reduce epidemics compared to a random approach.

Keywords epidemic dynamics \cdot opinion diffusion \cdot complex networks \cdot anti-vaccine sentiments

1 Introduction

Diseases have long threatened human health, and COVID-19 is a recent example. Vaccination is a vital tool for combating disease prevalence. However, in many instances, the availability of vaccinations does not necessarily lead to their full uptake owing to vaccination hesitancy. The analysis and investigation of this dilemma has been studied extensively using different mechanisms. For example, the vaccination decision behavior has been studied with respect

Rebecca Hoyle School of Mathematical Sciences, University of Southampton

Sarah Alahmadi, Markus Brede

School of Electronics and Computer Science, University of Southampton E-mail: sha1a21@soton.ac.uk

to vaccine-related information diffusion [6-8,2,9-11], with respect to rationality in social imitation using evolutionary game theory [12, 13, 4, 5, 1], and with respect to both approaches [14,11]. Moreover, mitigating misinformation propagation has received researchers' attention in influence maximization literature. Researchers have investigated the problem of mitigating negative influence either by blocking nodes [16] or edges [17], or applying true information propagation methods as in [18–22]. Unlike existing work on mitigating misinformation propagation, the primary purpose of this study is not reducing the number of anti-vaccine opinion adopters or increasing the number of pro-vaccine opinion adopters. Our purpose is instead altering the negative diffusion behaviour and mitigating the emergence of anti-vaccine communities that comprise unvaccinated people. Thus, this study explores strategies to mitigate the impact of anti-vaccine opinion propagation by modelling the impact of an effective counter-campaign that propagates positive information about vaccinations. The primary research questions are: to what extent can the spread of anti-vaccine opinion be mitigated? How can an effective positive counter-campaign be implemented to change the dynamic of anti-vaccine opinions and prevent or mitigate the emergence of anti-vaccine communities? Simulations reveal that the dynamics of coexisting anti- and pro-vaccine opinions have a varying impact on the incidence of epidemics. Certain strategies are shown to be effective in monitoring and mitigating the development of anti-vaccination opinions, while others are less effective.

2 Model description and methods

Diffusion model

We propose an agent-based model consisting of two stages, in which the first stage is opinion diffusion, followed by vaccination of all non-negative opinion adopters, and the second stage is the disease spread among unvaccinated individuals. For the disease spread, we use the SIR model previously described in [15], and we develop a model for the dual propagation of positive and negative vaccine opinion by extending the model presented in [2]. Each agent has one of three opinion states: negative, positive, or neutral. Agents can be exposed to positive and/or negative vaccine sentiments from two different sources at each time step: a general exposure (e.g. media) and peer influence. Initially, all nodes are neutral, and the general exposure triggers the social contagion (negative or positive) in the network. To adopt a negative opinion, an agent needs to have two more negative exposures than positive exposures, and vice versa for a positive opinion.

Positive information spread methods

- 1. Random positive spread (RandStratgy): In this method the negative and positive general exposures target all individuals randomly at each time step at a predefined rate.
- 2. Campaigning positive spread In this method positive exposure targets a set of individuals picked from the whole population based on certain criteria. We consider static and dynamic campaigning control. In **static control**, the set of targets is selected prior to the campaign's launch, and

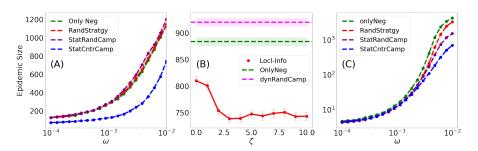


Fig. 1 Epidemic size resulting from the coexistence of pro- and anti-vaccine opinions compared to the scenario where only anti-vaccine opinions are present (green lines). Panel (A) demonstrates random and static campaigns for spreading positive-vaccine information with target set T = 500 as a function of the peer influence rate ω , using neg-limit setting with $N_{neg} = 2000$. Panel (B) demonstrates the dynamic campaigns with t = 1, T = 50, and $\omega = 0.006$ as a function of the target number of anti-vaccine neighbours ζ , using neg-limit setting with $N_{neg} = 2000$. Panel (C) demonstrates random and static campaigns with target set T = 500 as a function of ω , using time-limit setting with $\tau = 400$. The influence rate of general exposures is $\mu_{-} = 0.001$ for negative and $\mu_{+} = 0.0005$ for positive. We utilize a small-world network with size N = 5000, rewiring probability p = 0.01, and average degree $\langle k \rangle = 10$. SIR parameters: infection rate $\beta = 0.1$, recovery rate $\gamma = 0.1$, and seed set $I_0 = 1$. For each scenario we generate X different networks and for each network we run X SIR infection simulations, X = 500 in (A) and (C), and X = 1300 in (B).

remains unchanged. In **dynamic control**, the set of targets is initially selected at random, and then at each time t, they are replaced by new neutrals that meet a certain criteria. We consider the following heuristic methods to pick the target set:

- (a) **Static random strategy (StatRandCamp)**: the set of targets is selected at random from the entire population.
- (b) **Static central-based strategy (StatCntrCamp)**: the set of targets is selected with the highest values of betweenness centrality in the network.
- (c) **Dynamic random strategy (DynRandCamp)**: the set of targets is replaced by new neutrals chosen at random.
- (d) **Dynamic local negative information-based strategy (Locl-Info)**: the set of targets is replaced by new neutrals selected based on local information. The selection seeks targets whose number of negative neighbors is as close as possible to a target number ζ .

Experiment settings: We consider two distinct settings as stopping criteria for the opinions diffusion phase. In the first setting, the opinion dissemination stage proceeds until a certain number of negative adopters N_{neg} is reached (neg-limit). This is motivated by our interest to evaluate the influence of the presence of pro-vaccine dynamics on the structure of anti-vaccine communities, which has previously been investigated using this setting with the existence of only anti-vaccine opinion [2]. In the second setting, the opinion dissemination stage continues for a specific duration of time τ (time-limit).

3 Main findings

Figure 1 demonstrates the size of the epidemic resulting from the coexistence of pro- and anti-vaccine opinions compared to the scenario in which only antivaccine opinion exists. In general, the random spread of vaccine-positive sentiment (RandStratgy, StatRandCamp, and DynRandCamp) reduces the extent of an epidemic, but it is less effective than central-based and Locl-Info strategies. Utilizing the neg-limit setting yields no effect for the random methods, as shown in figure 1(A), or exacerbates the epidemic, as shown in both figure 1(A) and 1(B). However, relaxing this setting and adopting the time-limit one reveals a more conspicuous impact, as illustrated in figure 1(C). The static central campaign significantly reduces the epidemic spread as it obstructs the merging of anti-vaccine communities and prevents the growth of their size. Furthermore, the (Locl-Info) dynamic campaign with more frequent updates (where t=1) of the target set leads to a significant reduction in epidemic spread, as shown in figure 1(B). Targeting candidate neutrals who are more vulnerable to negative infection further improved the strategy's effectiveness.

Acknowledgements We would like to express our gratitude to Dr. Michael Head for his valuable contributions to this work. His expertise and insights were helpful in understanding and shaping our study.

References

- 1. Zhang, H., Shu, P., Wang, Z., Tang, M. & Small, M. Appl. Math. Comput. 294 pp. 332-342 (2017)
- 2. Campbell, E. & Salathé, M. Sci. Rep. 3, 1-6 (2013)
- Zhang, H., Zhang, J., Zhou, C., Small, M. & Wang, B. New J. Phys.. 12, 023015 (2010)
 Zhang, H., Wu, Z., Tang, M. & Lai, Y. Sci. Rep.. 4, 1-10 (2014)
- 5. Ichinose, G. & Kurisaku, T. Phys. A: Stat. Mech. Appl.. 468 pp. 84-90 (2017)
- 6. Salathé, M. & Bonhoeffer, S. J R Soc Interface. 5, 1505-1508 (2008)
- 7. Funk, S., Gilad, E., Watkins, C. & Jansen, V. PNAS. 106, 6872-6877 (2009)
- Ruan, Z., Tang, M. & Liu, Z. Phys. Rev. E. 86, 036117 (2012)
 Silva, P., Velásquez-Rojas, F., Connaughton, C., Vazquez, F., Moreno, Y. & Rodrigues, F. Phys. Rev. E. 100, 032313 (2019)
- 10. Mehta, R. & Rosenberg, N. Evol. Hum. Sci.. 2 (2020)
- 11. Yin, Q., Wang, Z., Xia, C. & Bauch, C. Commun Nonlinear Sci Numer. 109 pp. 106312 (2022)
- 12. Fu, F., Rosenbloom, D., Wang, L. & Nowak, M. Proc. Royal Soc. B. 278, 42-49 (2011)
- 13. Cardillo, A., Reyes-Suárez, C., Naranjo, F. & Gómez-Gardenes, J. Phys. Rev. E. 88, 032803 (2013)
- 14. Meng, X., Han, S., Wu, L., Si, S. & Cai, Z. Reliab. Eng. Syst. Saf.. 219 pp. 108256 (2022)
- 15. Kermack, W. & McKendrick, A. Proc. R. soc. Lond. Ser. A-Contain. Pap. Math. Phys. Character. 115, 700-721 (1927) Vaccines. 9, 607 (2021)
- 16. Pham, D., Nguyen, G., Nguyen, T., Pham, C. & Nguyen, A. IEEE Access. 8 pp. 78879-78889 (2020)
- 17. Zareie, A. & Sakellariou, R. Online Soc. Netw. Media. 29 pp. 100206 (2022)
- 18. Budak, C., Agrawal, D. & El Abbadi, A. Proceedings Of The 20th International Conference On World Wide Web. pp. 665-674 (2011)
- 19. Liu, W., Yue, K., Wu, H., Li, J., Liu, D. & Tang, D. Knowl Based Syst. 109 pp. 266-275 (2016)
- 20. Tong, A., Du, D. & Wu, W. Adv Neural Inf Process Syst
- . **31** (2018)
- 21. Yang, L., Li, Z. & Giua, A. 2019 American Control Conference (ACC). pp. 5608-5613 (2019)
- 22. Yang, L., Li, Z. & Giua, A. Inf. Sci. 506 pp. 113-130 (2020)

Linguistics & Multilayer



Imbalanced Multi-label Classification for Business related Text with Moderately Large Label Spaces
Muhammad Arslan and Christophe Cruz 206
SINr: a python package to train interpretable word and graph embeddings
Thibault Prouteau, Nicolas Dugué, Simon Guillot
and Anthony Perez
Topics evolution through multilayer networks
Andrea Russo, Antonio Picone and Vincenzo
Miracula
Towards efficient multilayer network data management
Georgios Panayiotou, Matteo Magnani and Bruno
Pinaud
Knowledge Graph for NLG in the context of conversational agents
Hussam Ghanem, Massinissa Atmani and
Christophe Cruz



Imbalanced Multi-label Classification for Business-related Text with Moderately Large Label Spaces

Muhammad Arslan · Christophe Cruz

Abstract In this study, we compared the performance of four different methods for multi-label text classification using a specific imbalanced business dataset. The four methods we evaluated were fine-tuned BERT, Binary Relevance, Classifier Chains, and Label Powerset. The results show that fine-tuned BERT outperforms the other three methods by a significant margin, achieving high values of accuracy, F1-Score, Precision, and Recall. Binary Relevance also performs well on this dataset, while Classifier Chains and Label Powerset demonstrate relatively poor performance. These findings highlight the effectiveness of fine-tuned BERT for multi-label text classification tasks, and suggest that it may be a useful tool for businesses seeking to analyze complex and multi-faceted texts.

Keywords Business domain \cdot Multi-lingual \cdot Performance comparison \cdot News documents \cdot Fine-tuned BERT

1 Introduction

In today's digital age, businesses are generating vast amounts of data through various channels, such as online news articles, press releases, and company web-sites [1]. This data contains valuable information regarding the launch of new products, services, and business initiatives, among other things. However,

Christophe Cruz

Muhammad Arslan

Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB), Université de Bourgogne, 9 Av. Alain Savary, 21000 Dijon, France

Tel.: +33 3 80 39 50 00

E-mail: muhammad.arslan@u-bourgogne.fr

Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB), Université de Bourgogne, 9 Av. Alain Savary, 21000 Dijon, France

analyzing this data manually can be a tedious and error-prone process, especially when it comes to identifying key insights and trends, especially in cases of moderately large label spaces.

Multi-label text classification is the task of assigning one or more labels to a text document [2]. Multilabel text classification using BERT [3] and Problem Transformation approaches [4] can greatly enhance the efficiency and accuracy of analyzing textual business data, including moderately large label spaces. These machine learning models are designed to process large amounts of text data and can automatically identify key topics and themes present in the text. With the ability to classify text into multiple categories, these models can handle more complex datasets and provide a more nuanced understanding of the information contained within.

For instance, a business seeking to understand customer feedback on a new product can use multilabel text classification to analyze reviews and identify which aspects of the product are most highly praised or criticized. Similarly, a company seeking to expand its operations to new geographic regions can analyze news articles and identify the sentiment towards the company in different regions.

In this study, we have compared the performance of four different methods for multi-label text classification using a specific imbalanced business-related dataset. Imbalanced dataset refers to a dataset in which the distribution of text examples among the different categories are unequal. The first method, Fine-tuned BERT, is a popular pre-trained language model that has been fine-tuned for a specific task. The other three methods evaluated in the study were Binary Relevance, Classifier Chains, and Label Powerset, which are all traditional multi-label classification methods.

The structure of this paper is as follows. In Section 2, we provide a review of the background of the multi-label text classification approaches used in this article. Section 3 introduces the proposed work. In Section 4, we discuss the results and implications of our approach. Finally, we conclude the paper in Section 5.

2 Background

Problem Transformation approaches are a family of techniques used in multilabel classification tasks to transform the original multi-label problem into one or more simpler, more manageable classification tasks [5]. These approaches are often used when there are a large number of possible labels, which can make the multi-label classification problem computationally expensive and challenging. The most common Problem Transformation approaches include Binary Relevance, Classifier Chains, and Label Powerset [5].

In Binary Relevance method [6], a separate binary classifier is trained for each label, and each classifier predicts whether or not the input belongs to that particular label. The main advantage of the Binary Relevance method is its simplicity and flexibility. It can work with any binary classifier, and the classifiers can be trained independently, making it easy to add or remove labels without affecting the performance of other classifiers. However, the method does not consider any correlations between the labels, which may affect the overall accuracy of the multilabel classification task.

The Classifier Chain method [7] use a chain of binary classifiers to predict the labels. In this method, the labels are treated as a sequence, and the classifiers are trained in the order of the label sequence. The main advantage of the Classifier Chain method is its ability to model the correlations between labels, which can lead to improved accuracy in the multilabel classification task. However, the method can be computationally expensive, especially if there are many labels in the dataset.

The Label Powerset method [8] involves transforming the multilabel problem into a multiclass problem. In this method, each unique combination of labels is treated as a separate class, and a multiclass classifier is trained to predict the class for each input. The main advantage of the Label Powerset method is its ability to handle any number of labels, and it can capture complex dependencies between labels. However, the method suffers from the curse of dimensionality, as the number of classes grows exponentially with the number of labels in the dataset.

Apart from Problem Transformation approaches, fine-tuning an existing pre-trained BERT (Bidirectional Encoder Representations from Transformers) model for multi-label text classification is a popular and effective approach in the existing literature [9] for imbalanced datasets having large label spaces (i.e. high number of categories i.e. labels). Fine-tuning a pre-trained BERT model for multi-label text classification involves training the model on a specific dataset, with the labels and associated text inputs. During training, the weights of the BERT model are updated to optimize the performance of the model on the specific multi-label text classification task.

There are numerous studies exist covering advanced methods for multilabel classification [10] [11] [12]. However, their applicability to business-related imbalanced dataset with moderately large label spaces has not been extensively explored in the literature. In order to bridge this gap, this article presents a comparative analysis of four different techniques using a business-related dataset. This study aims to provide valuable insights into the effectiveness of these methods for multi-label classification tasks in a business context, and to inform future research in this area.

3 Analyzing models for business-related text

In this section, we will provide an overview of the dataset used for our multilabel classification tasks. We will then use this dataset to train and evaluate several multi-label classification models.

3.1 Dataset

In preparation for the analysis of various multi-label classification models, a business-related dataset was obtained from the French company FirstECO. This dataset comprises 28,941 business-related texts in French, each of which can be classified into one or more of 80 possible categories. These categories are basically grouped into seven parent domains, which include Intangible development, activity, products, Material investment, Increased standby, Financial development, Company life, Geographical development, and Public finances. For data confidentiality reasons, only the 1st level of categories (labels) are reported. Each category in the dataset contains at least 50 to 100 text examples. However, the dataset remains imbalanced because the distribution of text examples among the different categories are unequal (see Fig. 1). Prior to analysis, the dataset was subjected to pre-processing procedures.

The text preprocessing is used as to clean and prepare text data before it is used for text classification. It takes the text data as input and applies several pre-processing steps to each text row. First, it removes punctuation and digits from the text using regular expressions. Then, it converts the text to lowercase and tokenizes it into words using the tokenizer. Next, it removes stop words from the text using a predefined set of stop words. Finally, it joins the preprocessed words back into a string and appends it to a new list. The preprocessing function is a crucial step in preparing text data for machine learning tasks, as it can help to reduce noise and improve the accuracy of the resulting models.

3.2 Implementation

The implementation involves the execution of Problem Transformation approaches and fine-tuning BERT model for imbalanced multi-label classification for business-related text with moderately large label spaces.

a) Problem Transformation approaches: The process starts by importing necessary modules for data preparation, model training, and evaluation. The imported modules include GaussianNB and MultinomialNB for Naive Bayes classification, and Accuracy_score for evaluation metrics, train_test_split for splitting the data into training and testing sets, and TfidfVectorizer for transforming the text data into feature vectors. The scikit-multilearn library is also imported to support multi-label classification problems. The process then creates an instance of TfidfVectorizer to convert the text data into feature vectors. The TfidfVectorizer is set to use inverse document frequency and normalization.

Next, the process creates an instance of MultiLabelBinarizer and applies it to the labels. The MultiLabelBinarizer transforms the list of labels into a binary matrix where each row corresponds to an instance and each column corresponds to a unique label. Then, the data is split into training and testing sets using train_test_split, with a test size of 20%. Finally, the process

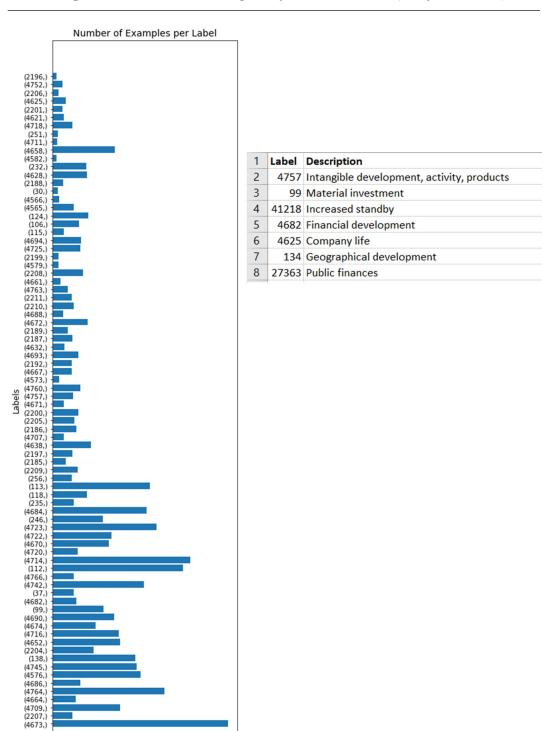


Fig. 1 a) Number of examples per label (left), b) Description of each label (right)

4000

3000 Number of examples

1000

ő

2000

returns X_train, X_test, Y_train, and Y_test, which are the feature matrices and label matrices for the training and testing sets, respectively. These matrices are used to train and evaluate multi-label classification models based on Problem Transformation approaches, which are; Binary Relevance, Classifier Chain, and Label Powerset in our case. Finally, the performance of these approaches is evaluated on the testing dataset using various metrics such as accuracy, precision, recall, and F1-score. Accuracy is the percentage of correctly classified samples out of the total number of samples. F1-Score is a weighted average of Precision and Recall, where the F1-Score gives equal importance to both Precision and Recall. Precision measures how precise the model's positive predictions are. Lastly, Recall measures how well the model can identify all positive samples.

b) Fine-tuning BERT: To fine-tune existing BERT-based model for text classification, the model "bert-base-multilingual-cased" [13] is chosen as it supports the French language. The process of fine-tuning starts with importing the necessary libraries such as NumPy, Pandas, Scikit-learn, PyTorch, and Transformers. Then, a number of hyperparameters are set, including Max_Len, which is set to 80 and represents the maximum length of input sequences. Train_Batch_Size is set to 16, and Valid_Batch_Size is set to 8. The process also specifies the number of Epochs to train the model, which is set to 5, and sets the learning rate to 1e-05. Additionally, a pre-trained BERT tokenizer using the BertTokenizer class is used from the transformer's library.

Furthermore, the data is split into training and testing datasets using a Train_Size of 0.8. The training dataset is created by randomly sampling 80% of the data from the original dataset, while the testing dataset is created by dropping the samples in the training dataset from the original dataset. The final training dataset has 23,153 samples, while the testing dataset has 5,788 samples. The BERT model is trained on the training dataset by feeding batches of input sequences to the model, computing the loss, and optimizing the weights using backpropagation. Finally, the model's performance is evaluated on the testing dataset using accuracy, precision, recall, and F1-score.

Accuracy, F1-score, precision, and recall are commonly used performance metrics that can be used to evaluate the effectiveness of a classifier. Accuracy measures the fraction of instances that are correctly classified by the classifier. Precision measures the fraction of correctly identified positive instances among all instances predicted as positive, while recall measures the fraction of correctly identified positive instances among all positive instances in the data. Using these definitions, we can compute accuracy, precision and recall.

True positives (TP) are instances that are positive and are correctly classified as positive by the classifier. False positives (FP) are instances that are negative but are incorrectly classified as positive by the classifier. True negatives (TN) are instances that are negative and are correctly classified as negative by the classifier and False negatives (FN) are instances that are positive but are incorrectly classified as negative by the classifier. However, accuracy may not be a suitable metric to use when the classes are imbalanced. This is because a classifier that simply predicts the majority class for all instances would achieve high accuracy even if it performs poorly on the minority class. To address this problem, we can use the F1-Score, which is a harmonic mean of precision and recall. It combines both precision and recall into a single metric that balances the trade-off between them. The F1-Score ranges between 0 and 1, with a value of 1 indicating perfect precision and recall. Note that precision measures the accuracy of the positive predictions made by the classifier, while recall measures the completeness of the positive predictions made by the classifier.

4 Results

The first method, Binary Relevance, achieves an accuracy of 0.730, F1-Score of 0.936, Precision of 0.952, and Recall of 0.922. This method creates a separate binary classifier for each label and assigns a label to each text independently. The second method, Classifier Chains, achieves an accuracy of 0.103, F1-Score of 0.539, Precision of 0.590, and Recall of 0.495. This method builds a chain of classifiers where each classifier considers the predictions of the previous classifiers in the chain. The third method, Label Powerset, achieves an accuracy of 0.143, F1-Score of 0.278, Precision of 0.350, and Recall of 0.230. This method transforms the multilabel classification problem into a multi-class classification problem by assigning each unique combination of labels to a single class. The fourth method, fine-tuned BERT, achieves the highest accuracy of 0.895, F1-Score of 0.978, Precision of 0.948, and Recall of 0.988.

One important factor that has impacted the performance of a multilabel classification model is the imbalanced distribution of labels in the dataset. In other words, some labels may have significantly more instances than others, making it challenging for the model to learn to classify the minority labels correctly. In this context, the provided results suggests that the dataset used to train and evaluate the model is imbalanced, with some labels having significantly fewer instances than others. To address the problem of imbalanced labels, various techniques can be used, such as oversampling or undersampling, class weighting, or ensemble methods. These techniques aim to balance the distribution of labels in the training set, which can improve the model's performance on the minority labels. However, choosing the appropriate technique depends on the specific characteristics of the dataset and the algorithm used. Nonetheless, the discussed model's (i.e. fine-tuned BERT in our case) performance is still competitive, and further investigation may be necessary to identify the factors that affect its performance on different labels.

5 Discussion and Conclusion

The study focused on the problem of imbalanced multi-label classification for business-related text with moderately large label spaces. The experiment compared the performance of four methods, namely Binary Relevance, Classifier Chains, Label Powerset, and Fine-tuned BERT, and evaluated their effectiveness based on accuracy, F1-Score, Precision, and Recall values. The results revealed that the fine-tuned BERT method significantly outperformed the other three methods, achieving high accuracy, F1-Score, Precision, and Recall values. Binary Relevance also performed well, but Classifier Chains and Label Powerset exhibited relatively poor performance on this imbalanced dataset.

Overall, the findings suggest that fine-tuning the pre-trained BERT model is a good idea because it enables the model to adapt to a specific application. BERT is a powerful language model that has been pre-trained on a large corpus of text, allowing it to learn the intricacies of language and syntax. However, while pre-training provides a strong foundation, it does not necessarily optimize the model for specific tasks such as sentiment analysis, question answering, or text classification. Fine-tuning allows us to take the pre-trained BERT model and tailor it to our specific needs by training it on a smaller dataset that is specific to the task at hand.

Reproducing the results of a fine-tuned BERT model requires the same preprocessing, architecture, and hyperparameters used in the original experiment. To reproduce the results, the same evaluation metrics should be used, and the results should be compared to the original experiment. However, the dataset used for fine-tuning the BERT model is highly dependent on the application. Different tasks require different datasets, and the quality and size of the dataset can greatly affect the model's performance. A small, poorly labeled dataset may not provide enough information for the model to learn, while a large, well-labeled dataset may enable the model to generalize better. Therefore, it is essential to choose an appropriate dataset for the specific task at hand to achieve optimal results.

The primary objective of this paper was not to provide a step-by-step guide for reproducing a specific model. Instead, the paper aims to introduce and advocate the idea of fine tuning BERT models on imbalanced datasets to achieve superior performance.

Acknowledgements The authors thank the French company FirstECO for providing the dataset, the French government for the plan France Relance funding, and Cyril Nguyen Van for his assistance.

References

- 1. Arslan, Muhammad, and Christophe Cruz, Semantic taxonomy enrichment to improve business text classification for dynamic environments, In 2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), pp.1-6, IEEE (2022).
- Read, Jesse, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank, Classifier chains for multi-label classification, Machine learning 85, pp.333-359 (2011).
- González-Carvajal, Santiago, and Eduardo C. Garrido-Merchán, Comparing BERT against traditional machine learning text classification, arXiv preprint arXiv:2005.13012 (2020).
- Liu, Jingzhou, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang, Deep learning for extreme multi-label text classification, In Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, pp.115-124 (2017).

- 5. Spolaôr, Newton, Everton Alvares Cherman, Maria Carolina Monard, and Huei Diana Lee, A comparison of multi-label feature selection methods using the problem transformation approach, Electronic notes in theoretical computer science 292, pp.135-151 (2013).
- Luaces, Oscar, Jorge Díez, José Barranquero, Juan José del Coz, and Antonio Bahamonde, Binary relevance efficacy for multilabel classification, Progress in Artificial Intelligence 1, pp.303-313 (2012).
- 7. Read, Jesse, Bernhard Pfahringer, Geoffrey Holmes, and Eibe Frank, Classifier chains: a review and perspectives, Journal of Artificial Intelligence Research 70, pp.683-718 (2021)
- 8. Read, Jesse, Antti Puurula, and Albert Bifet, Multi-label classification with meta-labels, In 2014 IEEE international conference on data mining, pp.941-946 IEEE (2014).
- 9. Lee, Jieh-Sheng, and Jieh Hsiang, Patent classification by fine-tuning BERT language model, World Patent Information 61, 101965 (2020).
- Bogatinovski, Jasmin, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev, Comprehensive comparative study of multi-label classification methods, Expert Systems with Applications 203, 117215 (2022).
- 11. Haghighian Roudsari, Arousha, Jafar Afshar, Wookey Lee, and Suan Lee, PatentNet: multi-label classification of patent documents using deep learning-based language understanding, Scientometrics, pp.1-25 (2022).
- Huang, Anzhong, Rui Xu, Yu Chen, and Meiwen Guo, Research on multi-label user classification of social media based on ML-KNN algorithm, Technological Forecasting and Social Change 188, 122271 (2023).
- 13. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, Bert: Pretraining of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).



SINr: a python package to train interpretable word and graph embeddings

Thibault Prouteau · Nicolas Dugué · Simon Guillot · Anthony Perez

Abstract In this paper, we introduce the SINr Python package to train word and graph embeddings. The SINr approach is based on community detection: a vector for a node is built upon the distribution of its connections through the communities detected on the graph at hand. Because of this, the algorithm runs very fast, and does not require GPUs to proceed. Furthermore, the dimensions of the embedding space are interpretable, those are based on the communities extracted. The package is distributed under Cecill-2.1 license and is available on Github and pypi.

Keywords Graph embedding \cdot Word embedding \cdot Interpretability \cdot Frugal machine learning

1 Introducing SINr

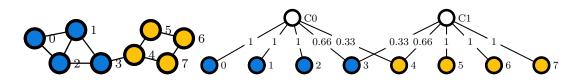
With neural approaches, tremendous progress was made in natural language processing, notably to represent the vocabulary of the language at hand, those representations are then used as input for machine learning algorithms. These representations are dense numeric vectors named word embeddings. Some examples of approaches to train such vectors using large textual corpora are Word2vec [6], Glove [8], and the Transformer-based approached for contextualized representations, Camembert [5] or Flaubert [4] in French. This progress was transfered to the graph universe, allowing the emergence of graph embedding, a whole field of research with Word2vec inspired approaches such as Node2vec [3], matrix factorization methods like HOPE [7] and auto-encoding paradigms [2].

Anthony

Thibault, Nicolas and Simon

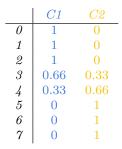
Le Mans Université, LIUM, EA 4023, Laboratoire d'Informatique de l'Université du Mans E-mail: first.last@univ-lemans.fr

Univ. Orléans, INSA Centre Val de Loire, LIFO EA 4022, Orléans E-mail: anthony.perez@univ-orleans.fr



(a) A graph G = (V, E) partitioned in two communities.

(b) Bipartite projection of G into graph $G' = (\top, \bot, E)$ along the communities. Weight on the edges is the proportion of neighbors in that community.



(c) Adjacency matrix of GI, each row is a SINr embedding.

Fig. 1: Illustration of SINr, vertices are represented based on the communities they are linked to.

SINr was introduced to take advantage of this progress, it allows to embed words and nodes just like the aforementioned methods. However, it is based on community detection: for each node, the vector of embedding is calculated as the proportion of its links going to the communities as described Figure 1. This approach allows to avoid some flaws inherent to the usual approaches:

- As far as we know, SINr is the first approach specifically designed to deal with both word and graph embeddings. Textual corpora are represented as graphs, and with the adequate preprocessing provided by the package, word embeddings can easily be extracted with SINr. For graph embedding, no specific pre-processing is required.
- Contrary to the neural approaches that require complex GPU calculations,
 SINr is based on the Louvain [1] algorithm to detect community and thus runs in linear-time, it can be executed on standalone laptops.
- Contrary to the usual approaches, because dimensions are based on the communities, the space in which words and graphs are embedded with SINr is interpretable.

The performances of SINr were evaluated on several tasks, including link prediction on graphs, and pair of words similarities for textual data [9]. While providing good performances, it runs faster than most of the other embedding approaches. Furthermore, the interpretability of the model was also demonstrated to be comparable to the state-of-the-art when considering word embedding [10]. In this paper, we consider the SINr package that we distribute on https://github.com/SINr-Embeddings/sinr and that can also be found on Pypi to be installed with pip.

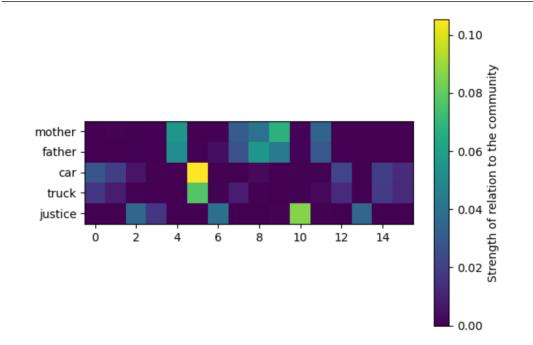


Fig. 2: Because of the sparsity of the embedding space, one can see that related words have non-zero values for the same dimensions (abscissa), *mother* and *father* for dimensions 4, 7, 8, 9 and 11 for instance. The non-zero dimensions are distinct when comparing *mother* and *car* that are unrelated.

2 Playing with SINr

One can setup SINr using pip install sinr, and train word and graph embeddings using the package features. While these embedding features are shared among models, SINr brings the option to probe and get an understanding of the resulting embedding space as one can see Figure 2. Using SINr leads to sparse vectors: a node is not connected to all the communities of a graph, and similarly, a word is not related to all the topics of a corpus. As shown by [11], sparsity is one of the features required to enforce interpretability. The other one is to embed data in spaces larger than the classic 128 dimensions ones. The γ resolution parameter of the Louvain [1] algorithm allows to vary the number of communities, thus controlling the number of dimensions.

3 Conclusion

We presented SINr, a package to train word and graph embeddings based on community detection. It runs in linear-time on standalone laptops and leads to interpretable spaces that can be inspected using features provided in the package. In the future, we plan to deal with temporal graphs and thus temporal embeddings with SINr and to include more features to visualize vectors.

Acknowledgements This work was funded by ANR-21-CE23-0010 DIGING.

References

- 1. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory* and experiment, 2008(10):P10008, 2008.
- Shaosheng Cao, Wei Lu, and Qiongkai Xu. Deep neural networks for learning graph representations. In Proceedings of the AAAI conference on artificial intelligence, volume 30, 2016.
- 3. Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.
- Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. Flaubert: Unsupervised language model pre-training for french. arXiv preprint arXiv:1912.05372, 2019.
- 5. Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894, 2019.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
- Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. Asymmetric transitivity preserving graph embedding. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1105–1114, 2016.
- 8. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.
- Thibault Prouteau, Victor Connes, Nicolas Dugué, Anthony Perez, Jean-Charles Lamirel, Nathalie Camelin, and Sylvain Meignier. Sinr: Fast computing of sparse interpretable node representations is not a sin! In *IDA*, pages 325–337, 2021.
- 10. Thibault Prouteau, Nicolas Dugué, Nathalie Camelin, and Sylvain Meignier. Are embedding spaces interpretable? results of an intrusion detection evaluation on a large french corpus. In *LREC*, 2022.
- 11. Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. Spine: Sparse interpretable neural embeddings. In *Proceedings of the* AAAI conference on artificial intelligence, volume 32, 2018.



Topics evolution through multilayer networks Analysing 2M tweets from 2022 Qatar FIFA World Cup

Andrea Russo · Vincenzo Miracula · Antonio Picone

Abstract In this study, we conducted a comprehensive data collection on the 2022 Qatar FIFA World Cup event and used a multilayer network approach to visualize the main topics, while considering their context and meaning relationship. We structured the data into layers that corresponded with the stages of the tournament and utilized Gephi software to generate the multilayer networks. Our visualizations displayed both the relationships between topics and words showing words-context relationship, as well as the dynamics and changes over time by layer, of the most frequently discussed topics.

PACS 05.45.a · 05.65.+b · 01.75.+m

Keywords Multilayer \cdot NLP \cdot Social networks analysis \cdot Football \cdot Data Visualization

1 Introduction

In complexity sciences, given the massive amount of data from the sample under observation, and given the difference in the data collected, a simple layer network cannot suffice to express and show the interactions of many interdependent components in a whole system. Such systems – and the self-organization and emergent phenomena they manifest – lie at the heart of many challenges of global importance for the future of the worldwide knowledge society [1]. For these reasons, in this paper we decided to create a multilayer network [2] with the goal of showing how in a social sample,

Vincenzo Miracula University of Catania

Antonio Picone University of Catania

Andrea Russo University of Catania E-mail: Andrea.Russo@phd.unict.it

people may change topics depending on various events and time in a given event. The World football cup in Qatar in 2022, was a great opportunity to collect a huge amount of data about about various topics and fans that support national teams. We have collected almost two millions (1.923.283) tweets coming from 2022 Qatar FIFA World Cup (the entire event counts 2.2M tweets)¹, with the #Hashtag "#FIFAWorld-Cup" and "#QATAR2022". In this paper, therefore, we tried to compare how different layers differ from each other, and to get more information about the "context" of the most important topics.

2 Data & Method

We collected the Twitter data thanks to the Twitter API. To create the multilayer network and accurately represent the desired information [3], we began by analyzing the data and then separated the database based on the time of data collection. This enabled us to obtain temporally correct data for each stage of the tournament and visualize it in a 3D dimension [4]. The dates for the tournament stages are listed in Table 1. After separating the various layers, we cleaned the dataset using stop-words. Then to obtain a word network, which could depict the topics but also the context and connected meaning of them over time, we analyzed the data using an algorithm called Bigram. A bigram is a sequence of two adjacent elements from a string of

Stages	Start dates (2022)	Ending dates (2022)
Group stage	20 November	2 December
Round of 16	3 December	6 December
Quarter-finals	9 December	10 December
Semi-finals	13 December	14 December
Final	17 December	18 December

 Table 1
 Dates used by Stages-Layers

tokens, which are typically letters, syllables, or words. A bigram is an n-gram for n=2. The frequency distribution of every bigram in a string is commonly used for simple statistical analysis of text in many applications, including in computational linguistics, cryptography, speech recognition, and so on. We did not use *TD-IDF* or similar algorithm because we opted for Bigram as the best software for our words-context relation goal. In our code we selected the most used words by users (nodes), and linked them (edges) with the seconds most used words in the same sentence.

We took the most used words that appear for each time-layer, and after several trials², we had to choose a limit of 300 words, because it is a good compromise between technical limitations of Gephi³, and the challenge to posed by a large amount of words with too many nodes and edges, which can make it difficult to understand and not provide useful information such as the texts-contexts.

We used as the layer pillar-words "World", "fifa" and "Team," since they are present in all layers.

¹ https://getdaytrends.com/trend/%23FIFAWorldCup/

² https://github.com/AndreaRussoAgid/Multilayer_FRCCS_2023.git

³ https://answers.launchpad.net/gephi/+question/107399

There are various tools for graphical visualization of multilayers, such as Pymnet⁴, but at the graphical level we believe that Gephi is definitely better.

To differentiate layers, with Gephi it is a bit difficult despite having MultiViz [5], in fact even though this tool gives the possibility to make Multilayer starting from some parameters, it is difficult to create a multilayer from the obtained data, because many words are repeated (both at Nodes and Edges level in the various layer). For this reason, we modified the word ID, adding symbols that ransom both the layer and edges of reference between nodes. For the Group stage we did not add any symbols, while for the others (for the purpose of recognition described above) we chose "^" for the Round of 16, "*" for the quarters, "†" for the semifinals, and "‡" for the finals.

In conclusion, we initiated the Gephi community algorithm (modularity class) to learn about "word communities" and thus amplify more of the context and meaning of the reference hub and related words.

3 Results & Conclusion

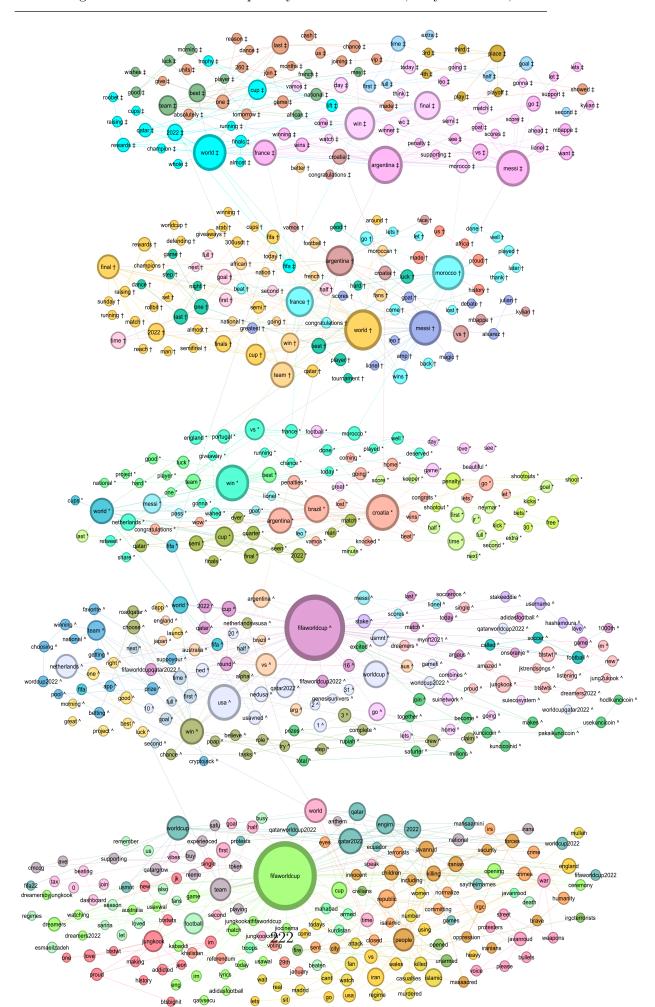
After the cleaning and bigram process, and also the exclusion of nodes and edges that are not part of the main network (gigantic component), we obtained a multilayer network with 858 Nodes and 1041 Edges.

The network depicted in Figure 1 shows that there are changes in the topics discussed across different layers. During the Group stage, word connections that created a discussion context related to crime and social injustices (orange and yellow) and the ceremony's performer, the *BTS* group (in red) were observed. However, as we moved to the subsequent layers (quarters, semifinals, and finals), there was a shift in the focus of the discussion, with less emphasis on the event itself and more on the teams and players. By using Gephi's community algorithm, we identified prominent topics (similar to those of social injustice and BTS) as discussion context. For example, the pride of the Moroccan team (in red \dagger) and the challenge between Messi and Mbappe (in dark pink \ddagger). We think that the combination of information visualization and multilayer networks could be an effective combination and method for studying the temporal evolution of topics in contexts with multiple levels of interaction.

References

- S. Boccaletti, G. Bianconi, R. Criado, C.I. Del Genio, J. Gómez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, M. Zanin. The structure and dynamics of multilayer networks (2014)
- M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M.A. Porter, S. Gómez, A. Arenas. Mathematical formulation of multilayer networks (2013). DOI 10.1103/PhysRevX.3.041022. URL https://link.aps.org/doi/10.1103/PhysRevX.3.041022
- M. Kivelä, A. Arenas, M. Barthelemy, J.P. Gleeson, Y. Moreno, M.A. Porter, Journal of complex networks 2(3), 203 (2014)
- F. Mcgee, M. Ghoniem, G. Melançon, B. Otjacques, B. Pinaud, in *Computer Graphics Forum*, vol. 38 (Wiley Online Library, 2019), vol. 38, pp. 125–149
- J.P.C. S., A. Chatterjee, G. M., A. Mukherjee. Multiviz: A gephi plugin for scalable visualization of multi-layer networks (2022). DOI 10.48550/ARXIV.2209.03149. URL https://arxiv.org/abs/ 2209.03149

⁴ http://www.mkivela.com/pymnet/





Towards efficient multilayer network data management

Georgios Panayiotou · Matteo Magnani · Bruno Pinaud

Abstract Real-world multilayer networks can be very large and there can be multiple choices regarding what should be modeled as a layer. Therefore, there is a need for their effective storage and manipulation. Currently, multilayer network analysis software use different data structures and manipulation operators. We aim to categorize operators in order to assess which structures work best for certain operator classes and data features. In this work, we propose a preliminary taxonomy of layer and data manipulation operators. We also design and execute a benchmark of select software and operators to identify potential for optimization.

Keywords Multilayer networks \cdot Multiplex networks \cdot Data management

1 Introduction

Multilayer networks have become increasingly popular across various disciplines for representing, manipulating and analyzing complex systems, such as brain [1] and ecological [2] networks. Typical applications also include social networks, which often consist of millions of actors and associated relationships when collected from online sources [3]. With large multilayer network data availability increasing, an open challenge is to provide a framework for effective storage and access, which will help reducing data preprocessing costs [4] and support development of interactive analysis systems [5].

Using established database management systems for multilayer network data storage is not ideal, as the state-of-the-art standards, i.e. relational and graph database systems, lack methods to represent, manipulate and analyze multilayer network data. On the other hand, there is a lack of consensus on the appropriate data structure for multilayer network data storage, as quite a few different alternatives exist within analysis and visualization libraries. More importantly, there

G. Panayiotou · M. Magnani

InfoLab, Dept. of Information Technology, Uppsala University, Sweden

B. Pinaud Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, France is an imbalance between scalability of a software's underlying data structures and process efficiency; depending on the chosen data model, the performance of even simple layer manipulation operators vastly differs (cf. Sect. 3).

Despite increasing interest in multilayer networks, work on multilayer network data management remains scarce. Previous work by the data engineering community on heterogeneous information networks [6] neither includes the concept of layer nor has looked into specialized storage solutions to optimize layer operations. In the recently proposed multilayer graphs, an extension for property graphs [7], the term layer is used, but instead refers to levels of nesting in the network. Also recently, a data mining approach was proposed, based on converting EER-diagrams into multilayer networks [8]; however, the previous does not directly address the impact of different data structures for multilayer networks.

With the previous issues in mind, this work provides basis for a data management framework natively supporting multilayer networks, by considering storage, access and manipulation of both vertices and edges in the network, and the layers themselves. We propose a taxonomy covering and extending layer and data manipulation operators found in popular multilayer network software [9–12]. Finally, we provide a benchmark of select operators' performance on currently available libraries and database models able to represent a multilayer network. We aim to both compare alternative data management approaches from a scalability and efficiency perspective, and spotlight processes in need of further research.

2 Data management taxonomy

We provide a taxonomy of tasks for multilayer network representation and manipulation, covering and extending operators found in major multilayer network software previously cited, which introduces a common basis for comparing various layer-supporting data structures. Our taxonomy differs from the recently introduced visualization-centered one [5]; our proposal focuses on data management operators, as typically provided by database management systems.

A preliminary taxonomy can be seen in Table 1. Our operator examples consider a generic multilayer network model similar to the one by Kivelä et al. [13], for a multilayer network $M = (V_M, E_M, V, \mathbf{L})$ with d dimensions. We use δ to denote a dimension of the network, v to denote a vertex, l to denote a single layer and σ to denote a predicate.

Tasks in the layer definition category focus on redefining the layer structure by creating and deleting dimensions. Layer manipulation operators also create new layers (or views thereof), but they instead derive new layers based on existing layers' topological features, e.g. layer flattening, projection and difference.

We consider layer query as a special class, as it arguably falls under both layer and data manipulation categories. Operators here essentially obtain a subset of the nodes and edges that satisfy a condition related to either attributes or topological features.

Finally, operators in the data manipulation category, similarly to manipulating a single-layer graph database, include adding and removing nodes or edges, while also considering the layers they associate to.

Operator class	Examples
Layer definition	$ ext{create/delete-dimension}(\delta) \\ ext{create/delete-layer}(\delta) \\$
Layer manipulation	flatten-layer (δ, l_s, l_t) project-layer (δ, l_s, l_t) diff-layer (δ, l_s, l_t)
Layer query	filter-layer _{σ} (l)
Data manipulation	${ m add/update/remove-node}(v,l) \ { m add/update/remove-edge}(v_s,l_s,v_t,l_t)$

Table 1 Preliminary taxonomy of layer-supporting data model operators

3 Benchmark

Multilayer networks can be implemented using data structures as different as tensors, dictionary-based adjacency lists, tables in relational databases, or combinations of the previous, as with graph database systems. We perform a comparison of these structures on the operators in our taxonomy, aiming to discover processes in need of optimization. Our primary focus is on layer manipulation and query operators, as we expect their performance within different systems to vary with respect to the underlying data structure and whether that is layer-native.

As an example, we consider the performance of the layer aggregation operator in MuxViz [9], multinet [10] and Pymnet [11] libraries. This experiment is done for randomly generated two-layer Erdős-Rényi multilayer networks, with network size (nodes associated to each layer) ranging between 100–100,000 and average node degree ≈ 4 . Note that multinet and MuxViz are libraries for R, while Pymnet is a library for Python, which can slightly affect performance.

Fig. 1 confirms the aforementioned imbalance between data structure efficiency and scalability, as systems exhibit various behaviours related to their underlying data structures; notice that some systems can handle larger networks, but at the expense of efficiency.

To expand our benchmark, we will also compare our findings with equivalent database models able to represent multilayer networks. Namely, we consider a relational database model representing each element of the multilayer network quadruple as its own relation and a graph database model operating on the network defined by (V_M, E_M) .

4 Conclusion and future work

Despite only dealing with preliminary results, we can already stress the importance of the chosen data structure, as this choice can determine which networks we can practically handle. These results should be complemented with a comparison of additional established multilayer network operators on both the aforementioned libraries and database models, in order to explore the strengths and weaknesses of various multilayer network management approaches. Finally, future work includes defining a minimal set of necessary operators for multilayer network data and a rule-based approach to transform multilayer networks into database models.

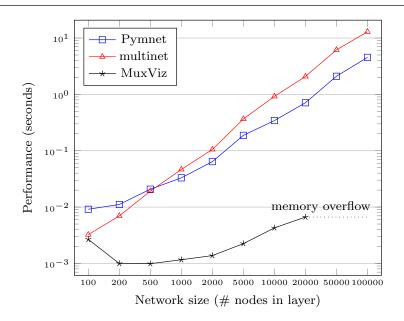


Fig. 1 Comparison of three multilayer network libraries for performance of two layer network aggregation over network size.

Acknowledgements This work has been partly funded by eSSENCE, an e-Science collaboration funded as a strategic research area of Sweden, and the FRÖ program by the French Institute in Sweden.

References

- 1. M. De Domenico, GigaScience 6(5), 1 (2017)
- S. Timóteo, M. Correia, S. Rodríguez-Echeverría, H. Freitas, R. Heleno, Nature Communications 9, 140 (2018)
- M.E. Dickison, M. Magnani, L. Rossi, *Multilayer Social Networks* (Cambridge University Press, 2016)
- R. Interdonato, M. Magnani, D. Perna, A. Tagarelli, D. Vega, Computer Science Review 36, 100246 (2020)
- F. McGee, M. Ghoniem, B. Otjacques, B. Renoust, D. Archambault, A. Kerren, B. Pinaud, G. Melançon, M. Pohl, T. von Landesberger, *Visual Analysis of Multilayer Networks*. Synthesis Lectures on Visualization (Springer International Publishing, Cham, 2021)
- C. Shi, Y. Li, J. Zhang, Y. Sun, P.S. Yu, IEEE Transactions on Knowledge and Data Engineering 29(1), 17 (2017)
- R. Angles, A. Hogan, O. Lassila, C. Rojas, D. Schwabe, P. Szekely, D. Vrgoč, in Proceedings of the 5th ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA) (Association for Computing Machinery, New York, NY, USA, 2022), GRADES-NDA '22, pp. 1–6
- A. Santra, K. Komar, S. Bhowmick, S. Chakravarthy, Data & Knowledge Engineering 141, 102058 (2022)
- 9. M. De Domenico, M.A. Porter, A. Arenas, Journal of Complex Networks 3(2), 159 (2015)
- 10. M. Magnani, L. Rossi, D. Vega, Journal of Statistical Software 98, 1 (2021)
- 11. Multilayer Networks Library for Python (Pymnet) Multilayer Networks Library 0.1 documentation. URL http://www.mkivela.com/pymnet/
- D. Auber, D. Archambault, R. Bourqui, M. Delest, J. Dubois, A. Lambert, P. Mary, M. Mathiaut, G. Melançon, B. Pinaud, B. Renoust, J. Vallet, in *Encyclopedia of Social Network Analysis and Mining*, ed. by R. Alhajj, J. Rokne (Springer, New York, NY, 2017), pp. 1–28
- M. Kivela, A. Arenas, M. Barthelemy, J.P. Gleeson, Y. Moreno, M.A. Porter, Journal of Complex Networks 2(3), 203 (2014)



Knowledge Graph for NLG in the context of conversational agents

Hussam Ghanem¹ · Massinissa Atmani¹ · Christophe Cruz¹

Abstract The use of knowledge graphs (KGs) enhances the accuracy and comprehensiveness of the responses provided by a conversational agent. While generating answers during conversations consists in generating text from these KGs, it is still regarded as a challenging task that has gained significant attention in recent years. In this document, we provide a review of different architectures used for knowledge graph-to-text generation including: Graph Neural Networks, the Graph Transformer, and linearization with seq2seq models. We discuss the advantages and limitations of each architecture and conclude that the choice of architecture will depend on the specific requirements of the task at hand. We also highlight the importance of considering constraints such as execution time and model validity, particularly in the context of conversational agents. Based on these constraints and the availability of labeled data for the domains of DAVI, we choose to use seq2seq Transformer-based models (PLMs) for the Knowledge Graph-to-Text Generation task. We aim to refine benchmark datasets of kg-to-text generation on PLMs and to explore the emotional and multilingual dimensions in our future work. Overall, this review provides insights into the different approaches for knowledge graph-to-text generation and outlines future directions for research in this area.

Keywords Conversational agents \cdot Knowledge graphs \cdot Natural Language Generation

1 Introduction

Conversational agents, also known as chatbots, are computer programs designed to simulate conversation with human users [55]. These agents can be

ICB, UMR 6306, CNRS, Université de Bourgogne 21000 Dijon, France

integrated with messaging platforms, mobile applications, and websites to provide instant support to customers and handle simple tasks, such as answering questions or helping with bookings. Conversational agents using knowledge graphs (KG) [9] are a type of chatbot that leverages structured data stored in a knowledge graph to generate human-like responses. The knowledge graph is a graph-based representation of entities and their relationships, providing a structured source of information for the chatbot to access and use during conversation. This enables the chatbot to provide more accurate and comprehensive answers to user's questions. The use of knowledge graphs can greatly enhance the capabilities of conversational agents and make interactions more informative and useful [57].

Generating answers during conversations consists in generating text from data. Data-to-text processes require algorithms that generate linguistically correct sentences for humans and express the semantics and structure of nonlinguistic data (sequence, tree, graph, etc.). In addition, the generation of textual answers requires supporting several languages. And for a better interaction with a conversational agent, the emotional context of the conversation (Common Ground) is fundamental. The aim of this work in collaboration with the company DAVI is to integrate a socio-emotional dimension into humanmachine interactions which complement the technical and "business" skills linked to professional expertise. The company DAVI is a software publisher in SaaS mode which has expertise in the fields of AI, Affective Computing, and Human Machine Interactions (HMIs). The following picture presents the composite AI of DAVI's solution including Natural Language Understanding, Emotion detection, skills modeling, Natural Language Generation, and Body Language Generation.

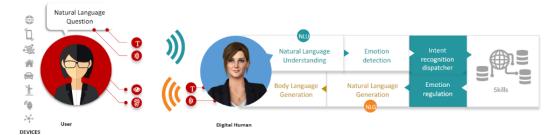


Fig. 1 Composite AI pipeline of virtual agents at DAVI

For now, the natural language generation (NLG) of a conversational engine does not benefit from the latest technological advances in natural language processing (NLP). The Natural Language processing step is based on a manual process to define the template of the answer. This process requires a costly amount of time. Thus, the purpose of this project is to automate and reduce the burden of the generation of template-based responses (as the responses are manually written through a set of rules) in the implementation of conversational agents. The template-based responses are modellized and stored in the skills' database. Regarding the emotional dimension, emotional responses are injected automatically in the answer depending the emotional analysis of the user. To automate NLG answers from Skills, knowledge graphs were identified.

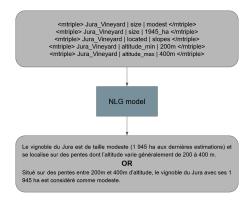


Fig. 2 Automatic text generation from the knowledge graph

In NLG, Two criteria [33] are used to assess the quality of the produced answers. The first criterion is semantic consistency (Semantic Fidelity) which quantifies the fidelity of the data produced against the input data. The most common indicators are 1/ Hallucination: It is manifested by the presence of information (facts) in the generated text that is not present in the input data; 2/ Omission: It is manifested by the omission of one of the pieces of information (facts) in the generated text; 3/ **Redundancy**: This is manifested by the repetition of information in the generated text; 4/ Accuracy: The lack of accuracy is manifested by the modification of information such as the inversion of the subject and the direct object complement in the generated text; 5/ Ordering: It occurs when the sequence of information is different from the input data. The second criterion is linguistic coherence (Output Fluency) to evaluate the fluidity of the text and the linguistic constructions of the generated text, the segmentation of the text into different sentences, the use of anaphoric pronouns to reference entities and to have linguistically correct sentences.

Today, neural approaches offer performances exceeding all classical methods for linguistic coherence. However, limits are still present to maintain semantic consistency, and their performance deteriorates even more on long texts [30]. Another limitation due to the complexity of neural approaches is that text generation is non-parameterized with no control over the structure of the generated text. Thus, most of the current neural approaches arrive behind template-based approaches on the criterion of semantic consistency [31], but they are far superior to them on the criterion of linguistic consistency. This can be explained by the fact that large language models manage to capture certain syntactic and semantic properties of the language.

The present review is organized as follows, Section 2 presents a comprehensive overview of the current state-of-the-art approaches for knowledge graphto-text generation. In Section 3, we present the latest architectures and techniques that have been proposed in this field. Finally, Section 4 critically examines the strengths and limitations of these techniques in the context of conversational agents

2 Knowledge Graph-to-Text Generation

KG-to-text generation aims at producing easy-to-understand sentences in natural language from knowledge graphs (KGs) while maintaining semantic consistency between the generated sentences and the KG triplets. Compared to the traditional text generation task (Seq2Seq), generating text from a knowledge graph is an additional challenge to guarantee the authenticity of the words in the generated sentences. The existing methods can be classified according to three categories (Figure 3) and will be detailed later:

- Linearisation with Sequence-to-sequence (Seq2Seq): convert the graph to a sequence which is the fed to a sequence-to-sequence model;
- Graph Neural Networks (GNNs): encode topological structures of a graph and learn the representation of an entity by the aggregation of the features of the entities and neighbors. They are not used as a standalone model and require a decoder to complete the encoder-decoder architecture;
- Graph Transformer (GT): the enhanced version of the original transformer adapted to handle graphs.

The term "knowledge graph" has been around since 1972, but its current definition can be traced back to Google's 2012 announcement of their Knowledge Graph. This was followed by similar announcements from companies such as Airbnb, Amazon, eBay, Facebook, IBM, LinkedIn, Microsoft, and Uber, among others, leading to an increase in the adoption of knowledge graphs by various industries [4]. As a result, academic research in this field has seen a surge in recent years, with an increasing number of scientific publications on knowledge graphs [4]. These graphs utilize a graph-based data model to effectively manage, integrate, and extract valuable insights from large and diverse datasets [5].

Knowledge graphs, which are composed of nodes that represent different types of entities and edges that denote various types of relationships between those entities, are known as heterogeneous graphs. The integration of information from multiple sources and domains in knowledge graphs leads to an even greater degree of heterogeneity. To address this, recent research has applied heterogeneous graph embedding methods to represent knowledge graphs

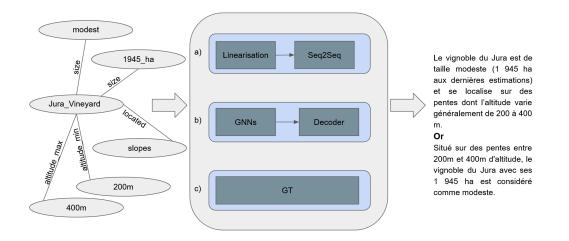


Fig. 3 The architecture of KG-to-text generation with the three categories of representation, a) Linearization + Seq2Seq, b) GNNs with decoder (e.g. LSTM), and c) Graph Transformer (GT)

effectively. For example, ERNIE [40] and KnowBERT [41] employ knowledge graph embedding techniques such as TransE [42] and TuckER [58] to encode knowledge graphs.

Generating text and learning alignments between source entities/relationships and target tokens from scratch is a challenging task for standard language models because of the limited amount of parallel graph-text data [17,34]. To overcome this limitation, recent research has focused on developing generalist pre-trained language models for KG-to-text generation. A common approach is to linearize input graphs into text sequences and fine-tune pre-trained seq2seq Transformer models such as GPT [35], BART [10], or T5 [36] based on KG-to-text datasets [1,37]. These pre-trained language models can generate high-quality texts with a simple fine-tuning to the target task, thanks to their self-supervised pre-training on large-scale corpora of unlabeled texts. In fact, pre-trained language models outperform other models with sophisticated structures in KG-to-text generation tasks. This type of approach will be detailed in section 3.

According to [14], text generation tasks using KG-to-text models mainly fall under three aspects:

- Encoder modification: To reduce the loss of structural information in sequence encoders with graph inputs that have been linearized [24,6,32], proposals concentrate on constructing more intricate encoder structures to improve the representation of graphs, including GNNs and GTs;
- Unsupervised training: These proposals consist in designing unsupervised training methods to jointly learn graph-to-text and text-to-graph con-

version tasks with non-parallel graph-to-text data [39,17,18]. This makes it possible to compare the final result of the process with the input data;

- **Build pre-trained models**: With the development of pre-trained Natural Language Generation (NLG) models such as GPT, BART, and T5, some recent work directly refines these models on graph-to-text datasets and reports significant performance [1,37,15,2].

Compared to existing work on pre-trained models for KG-text generation, the JoinGT model [14] uses pre-training methods to explicitly learn graphtext alignments instead of directly tuning the pre-trained models seq2seq on KG-to-text datasets.

3 Architectures

In this section, the different architectures used in data-to-text tasks will be presented. As the nature of the data greatly influences the choice of the architecture of the neural approaches, most works either try to adapt the inputs to the architectures of the models or propose new architectures better adapted to the types of input data. Due to the nature of sentences and the tree structure of its representations, several works have proposed to model data structures of this type to enhance performance.

3.1 Graph linearisation and sequence to sequence models (Seq2seq)

Recent years have been marked by significant achievements in the field of PLMs (Pretrained Language Models) [38,44]. Pre-trained on massive corpora, PLMs exhibit good generalization ability to solve related NLG (Natural Language Generation) downstream tasks [43]. However, most existing PLMs were trained on textual data [44,10] without ingesting any structured data input. The seq-to-seq category consists of linearizing the KG [1, 25, 24, 3] and then formulating a Seq2seq generation task using PLMs like GPT [35], BART [10] or T5 [36] with linearized KG nodes as input to generate sentences. The use of pre-trained language models (PLMs) in KG-to-text generation has shown superior performance, but still faces two major challenges: 1) loss of structural information during encoding, as existing models like BERT do not explicitly take into account the relationship between input entities; and 2) lack of explicit graph-text alignments, as complex knowledge graph structures make it difficult to learn graph-text alignments through text reconstruction-based pretraining tasks. Despite attempts to retain as much of the graph topology as possible with seq2seq methods, the Transformer-based seq2seq models' cost is not cheap (especially in the pretraining phase). Also, the computational cost of linearization can be high for large knowledge graphs. Hence, and so to better keep the graph topology, Graph Neural Networks (GNNs) have been proposed, which will be discussed in the next section.

3.2 Graph Neural Networks (GNNs)

Different approaches use different variants of GNNs architectures such as GCNs (Graph Convolutional Networks) [27] or extensions of GCNs such as Syn-GCNs [45] or DCGCNs [11]. Other approaches use the variant GATs (Graph Attention Networks) [8]. Or approaches that use the GGNNs (Gated Graph Neural Networks) variant [16,21,47]. Graph Neural Networks (GNNs) are a type of neural network that are well-suited for processing graph-structured data. In the context of knowledge graph-to-text generation, GNNs can be used to model the relationships between entities in a knowledge graph and generate text based on those relationships. Recent research on using GNNs for knowledge graph to text generation has shown promising results. Some studies have used graph convolutional networks (GCNs) [27] to encode the relationships between entities in a knowledge graph into a low-dimensional representation, or extensions of GCNs such as Syn-GCNs [45] and DCGCNs [11]. Other studies have used graph attention networks (GATs) [8] to dynamically weight the importance of different entities and relationships in the knowledge graph when generating text. Other studies have used a gating mechanism (Gated Graph Neural Networks or GGNNs) that allows for effectively controlling the flow of information between nodes in the graph, which is useful for incorporating contextual information [16,21,47]. Additionally, some researchers have combined GNNs with reinforcement learning to generate text that maximizes a reward function defined over the generated text and the knowledge graph [21].

Overall, the use of GNNs for knowledge graph to text generation is an active area of research, with many recent studies exploring different architectures and training methods. The results so far suggest that GNNs can effectively capture the relationships between entities in a knowledge graph and generate high-quality text based on that information. A limitation relied to KG-to-Text generation with GNNs, is that GNNs can be computationally expensive and may struggle to handle large knowledge graphs. Additionally, their performance may degrade for graphs with complex relationships or structures. Despite these limitations, GNNs remain a promising direction for knowledge graph-to-text generation.

3.3 Graphs Transformers (GTs)

In order to benefit from the power of models based on Transformer and to be able to model tree or graph-type data structures as with GNNs, and to overcome the limitations of local neighborhood aggregation while avoiding strict structural inductive biases, recent works have proposed to adapt the Transformer architecture. As Graph Transformers are equipped with self-attention mechanisms, they can capture global context information by applying these mechanisms to the nodes of the graph. According to [12], GT differs from GNNs in that it allows direct modeling of dependencies between any pair of nodes regardless of their distance in the input graph. An undesirable consequence is that it essentially treats any graph as a fully connected graph, greatly reducing the explicit structure of the graph. To maintain a structure-aware view of the graph, their proposed model introduces an explicit relationship encoding and integrates it into the pairwise attention score computation as a dynamic parameter.

From the GNNs pipeline, if we make several parallel heads of neighborhood aggregation and replace the sum on the neighbors by the attention mechanism, e.g. a weighted sum, we would get the Graph Attention Network (GAT). Adding normalization and MLP feed-forward, we have a Graph Transformer [46]. For the same reasons as Graph Transformer, [13] presents the K-BERT model, they introduce four components to augment the Transformer architecture and to be able to handle a graph as input. The first knowledge layer component takes a sentence and a set of triplets as input and outputs the sentence tree by expanding the sentence entities with their corresponding triplets. They also add the Seeing Layer component to preserve the original sentence structure and model the relationships of the triples by building a Visibility Matrix. Another component is the Mask-Transformer where they modify the self-attention layer to consider the visibility matrix when calculating attention.

The use of Graph Transformers for Knowledge graph to text generation has gained popularity in recent years due to their ability to effectively handle graph structures and capture the relationships between nodes in the graph. Additionally, Graph Transformers can handle large graphs and have the ability to model long-range dependencies between nodes in the graph. Despite the advantages, the training of Graph Transformers can be computationally expensive and the interpretability of the model remains a challenge. Overall, the use of Graph Transformers for Knowledge graph to text generation is a promising area of research and can lead to significant improvements in the generation of text from knowledge graphs.

4 Discusion

The Graph Neural Network (GNN), the Graph Transformer, and linearization with seq2seq models are three different architectures for knowledge graph to text generation, a task that involves generating natural language text from structured knowledge representations like knowledge graphs. GNNs are a type of deep learning model that are well-suited for processing graph-structured data. They provide a flexible and scalable way to model the graph structure and relationships, but they may not be able to efficiently handle large and complex graphs. The Graph Transformer is a specialized version of the Transformer, designed for graph-to-sequence learning tasks. It provides a more direct and efficient way to process the graph structure compared to linearization with seq2seq models, but it may require more training data and computation resources. Linearization with seq2seq models is a simpler and easier to implement approach that involves converting the knowledge graph into a linear sequence, such as a sentence, and then using a seq2seq model to generate text from the linearized representation. However, this approach can lose some of the structural information and relationships in the knowledge graph during the linearization process.

In summary, each of the three architectures has its own advantages and disadvantages, and the choice of architecture will depend on the specific requirements of the actual task.

As our project is part of a context of conversational agents, we must take into account all the consequent constraints such as the execution time (to respect the instant conversation constraint) and the validity of the model where the answer must not be incorrect or ambiguous.

If most of the current models have satisfactory validity performances, the inference time of models based on GNN and GraphTransformer exceeds the limit threshold found in a fluid and natural conversation and requires a huge memory load that violates the standards of current industrialization (MLOps) of neural models.

In light of these elements, and with the constraint of the data labellisation for the domains of DAVI, we choose to go further with seq2seq Transformer based models (PLMs) in our Knowledge Graph-to-Text Generation. We also want to shed light on the fact that DAVI already handles such models in their pipeline and they have the knowledge and infrastructure to optimize the integration and deployment of the seq2seq models. Hence, DAVI should still remain in control of the time to market of the NLG solution.

5 Conclusion

In conclusion, the document discusses different data-to-text architectures and highlights their advantages and limitations in the context of graph-to-text project with DAVI. The Graph Neural Network (GNN), Graph Transformer, and seq2seq models are three approaches that have been applied to the task of generating natural language text from structured knowledge representations like knowledge graphs. Each approach has its own advantages and disadvantages, and the choice of architecture will depend on the specific requirements of the task. Considering the constraints of our project, which includes developing a conversational agent that must generate valid responses in real-time, we have decided to move forward with seq2seq Transformer-based models (PLMs) as they have satisfactory performance on validity and execution time. Additionally, DAVI already handles such models in their pipeline and can optimize their integration and deployment. Our next step will be to explore state-of-the-art approaches that take into account the emotional and multilingual dimensions to achieve the objectives of the graph-to-text project.

References

- 1. Ribeiro, L. F., Schmitt, M., Schütze, H., & Gurevych, I. (2020). Investigating pretrained language models for graph-to-text generation. arXiv preprint arXiv:2007.08426.
- Mager, M., Astudillo, R. F., Naseem, T., Sultan, M. A., Lee, Y. S., Florian, R., & Roukos, S. (2020). Gpt-too: A language-model-first approach for amr-to-text generation. arXiv preprint arXiv:2005.09123.
- Hoyle, A., Marasović, A., Smith, N. (2020). Promoting graph awareness in linearized graph-to-text generation. arXiv preprint arXiv:2012.15793.
- Hogan A, Blomqvist E, Cochez M, d'Amato C, Melo GD, Gutierrez C, Kirrane S, Gayo JE, Navigli R, Neumaier S, Ngomo AC. Knowledge graphs. ACM Computing Surveys (CSUR). 2021 Jul 2;54(4):1-37.
- 5. N. F. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor. 2019. Industryscale knowledge graphs: Lessons and challenges. ACM Queue 17, 2 (2019).
- Distiawan, B., Qi, J., Zhang, R., & Wang, W. (2018, July). GTR-LSTM: A triple encoder for sentence generation from RDF data. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1627-1637).
- Ribeiro, L. F., Zhang, Y.,& Gurevych, I. (2021). Structural adapters in pretrained language models for amr-to-text generation. arXiv preprint arXiv:2103.09120.
- 8. Ribeiro, L. F., Zhang, Y., Gardent, C.,& Gurevych, I. (2020). Modeling global and local node contexts for text generation from knowledge graphs. Transactions of the Association for Computational Linguistics, 8, 589-604.
- 9. Moon, S., Shah, P., Kumar, A., & Subba, R. (2019, July). Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 845-854).
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ...& Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461. (not graph)
- Guo, Z., Zhang, Y., Teng, Z., & Lu, W. (2019). Densely connected graph convolutional networks for graph-to-sequence learning. Transactions of the Association for Computational Linguistics, 7, 297-312.
- Cai, D., & Lam, W. (2020, April). Graph transformer for graph-to-sequence learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 05, pp. 7464-7471).
- Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., Wang, P. (2020, April). K-bert: Enabling language representation with knowledge graph. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 03, pp. 2901-2908).
- Ke, P., Ji, H., Ran, Y., Cui, X., Wang, L., Song, L., ...& Huang, M. (2021). Jointgt: Graph-text joint representation learning for text generation from knowledge graphs. arXiv preprint arXiv:2106.10502.
- 15. Chen, W., Su, Y., Yan, X., & Wang, W. Y. (2020). Kgpt: Knowledge-grounded pretraining for data-to-text generation. arXiv preprint arXiv:2010.02307.
- 16. Chen, Y., Wu, L., & Zaki, M. J. (2020). Toward subgraph guided knowledge graph question generation with graph neural networks. arXiv preprint arXiv:2004.06015.
- 17. Guo, Q., Jin, Z., Qiu, X., Zhang, W., Wipf, D., & Zhang, Z. (2020). Cyclegt: Unsupervised graph-to-text and text-to-graph generation via cycle training. arXiv preprint arXiv:2006.04702
- Jin, Z., Guo, Q., Qiu, X., & Zhang, Z. (2020, December). Genwiki: A dataset of 1.3 million content-sharing text and graphs for unsupervised graph-to-text generation. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 2398-2409).
- Wang, Q., Yavuz, S., Lin, V., Ji, H., & Rajani, N. (2021). Stage-wise Fine-tuning for Graph-to-Text Generation. arXiv preprint arXiv:2105.08021.

- 20. Song, L., Wang, A., Su, J., Zhang, Y., Xu, K., Ge, Y., & Yu, D. (2021). Structural information preserving for graph-to-text generation. arXiv preprint arXiv:2102.06749.
- Chen, Y., Wu, L., Zaki, M. J. (2019). Reinforcement learning based graph-to-sequence model for natural question generation. arXiv preprint arXiv:1908.04942.
- 22. Harkous, H., Groves, I., & Saffari, A. (2020). Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. arXiv preprint arXiv:2004.06577.
- 23. Koncel-Kedziorski, R., Bekal, D., Luan, Y., Lapata, M.,& Hajishirzi, H. (2019). Text generation from knowledge graphs with graph transformers. arXiv preprint arXiv:1904.02342.
- 24. Gardent, C., Shimorina, A., Narayan, S.,& Perez-Beltrachini, L. (2017, September). The WebNLG challenge: Generating text from RDF data. In Proceedings of the 10th International Conference on Natural Language Generation (pp. 124-133).
- 25. Yang, Z., Einolghozati, A., Inan, H., Diedrick, K., Fan, A., Donmez, P.,& Gupta, S. (2020). Improving text-to-text pre-trained models for the graph-to-text task. In Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+) (pp. 107-116).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ...& Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
- Kipf, T. N., Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
- M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., vol. 2, pp. 729–734, IEEE, 2005.
- F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," IEEE Transactions on Neural Networks, vol. 20, no. 1, pp. 61–80, 2008.
- 30. Ratish Puduppully, Li Dong, and Mirella Lapata. 2018. Data-to-text generation with content selection and planning.
- 31. Yevgeniy Puzikov and Iryna Gurevych. 2018. E2e nlg challenge: Neural models vs. templates.
- 32. Moryossef, A., Goldberg, Y.,& Dagan, I. (2019). Step-by-step: Separating planning from realization in neural data-to-text generation. arXiv preprint arXiv:1904.03396.
- Ferreira, T. C., van der Lee, C., Van Miltenburg, E., & Krahmer, E. (2019). Neural datato-text generation: A comparison between pipeline and end-to-end architectures. arXiv preprint arXiv:1908.09022.
- 34. Fu, Z., Shi, B., Lam, W., Bing, L., & Liu, Z. (2020). Partially-aligned data-to-text generation with distant supervision. arXiv preprint arXiv:2010.01268.
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving language understanding by generative pre-training.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ...& Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140), 1-67.
- Kale, M.,& Rastogi, A. (2020). Text-to-text pre-training for data-to-text tasks. arXiv preprint arXiv:2005.10433.
- Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Schmitt, M., Sharifzadeh, S., Tresp, V., & Schütze, H. (2019). An unsupervised joint system for text generation from knowledge graphs and semantic parsing. arXiv preprint arXiv:1904.09447.
- Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129.
- Peters, M. E., Neumann, M., Logan IV, R. L., Schwartz, R., Joshi, V., Singh, S., & Smith, N. A. (2019). Knowledge enhanced contextual word representations. arXiv preprint arXiv:1909.04164.
- Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. Advances in neural information processing systems, 26.

- 43. Li, J., Tang, T., Zhao, W. X., Wen, J. R. (2021). Pretrained language models for text generation: A survey. arXiv preprint arXiv:2105.10311.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
- 45. Marcheggiani, D., Frolov, A.,& Titov, I. (2017). A simple and accurate syntaxagnostic neural model for dependency-based semantic role labeling. arXiv preprint arXiv:1701.02593.
- 46. Chaitanya K.Joshi. Transformers are Graph Neural Networks. The Gradient.
- Li, Y., Tarlow, D., Brockschmidt, M., Zemel, R. (2015). Gated graph sequence neural networks. arXiv preprint arXiv:1511.05493.
- 48. Dwivedi, V. P.,& Bresson, X. (2020). A generalization of transformer networks to graphs. arXiv preprint arXiv:2012.09699.
- 49. Dušek, O., & Kasner, Z. (2020). Evaluating semantic accuracy of data-to-text generation with natural language inference. arXiv preprint arXiv:2011.10819.
- 50. Tian, R., Narayan, S., Sellam, T., & Parikh, A. P. (2019). Sticking to the facts: Confident decoding for faithful data-to-text generation. arXiv preprint arXiv:1910.08684.
- Xiao, Y., & Wang, W. Y. (2021). On hallucination and predictive uncertainty in conditional language generation. arXiv preprint arXiv:2103.15025.
- 52. Sam Wiseman , Stuart Shieber , and Alexander Rush . 2017 . Challenges in data-todocument generation . In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2253 – 2263 ,Copenhagen, Denmark. Association for Computational Linguistics .https://doi.org/10.18653/v1/D17-1239
- 53. Durmus, E., He, H.,& Diab, M. (2020). FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. arXiv preprint arXiv:2005.03754.
- Katja Filippova. 2020. Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. 864\u00e1870.
- 55. Wahde, M., & Virgolin, M. (2022). Conversational agents: Theory and applications. In HANDBOOK ON COMPUTER LEARNING AND INTELLIGENCE: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation (pp. 497-544).
- 56. Hussain, S., Ameri Sianaki, O.,& Ababneh, N. (2019). A survey on conversational agents/chatbots classification and design techniques. In Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33 (pp. 946-956). Springer International Publishing.
- 57. Ait-Mlouk, A., & Jiang, L. (2020). KBot: A Knowledge graph based chatBot for natural language understanding over linked data. IEEE Access, 8, 149220-149230.
- Balažević, I., Allen, C., & Hospedales, T. M. (2019). Tucker: Tensor factorization for knowledge graph completion. arXiv preprint arXiv:1901.09590.

Urban



How to Reduce Streets-Network Sprawl? Supharoek Chattanachott, Frédéric Guinand and Kittichai Lavangnananda
A hybrid network: Sea-land connectivity in the global system of cities Cesar Ducruet, Barbara Polo and Bruno Marnot 244
 Agent-based modelling of urban expansion and land cover change: a prototype for the analysis of commuting patterns in Geneva, Switzerland. Flann Chambers, Christophe Cruz and Giovanna Di Marzo Serugendo
Radial analysis and scaling law of housing prices in French urban areas using DVF data Gaëtan Laziou, Rémi Lemoy and Marion Le Texier 256
Towards a geographical theory different from that of the natural sciences: foundations for a relational complexity model Olivier Bonin



How to Reduce Streets-Network Sprawl?

Supharoek Chattanachott · Frédéric Guinand · Kittichai Lavangnananda

Abstract In a near future, because of climate change and the necessary reduction of soil artificialization, cities will likely search for more and more free spaces for their development by reusing already artificialized areas. One possibility is to look at infrastructures dedicated to car mobility, infrastructures that sprawl over very wide areas and that cover a substantial part of city surfaces. This work investigates the possibility of reducing city streets networks by converting several lanes of streets into one-way streets with only one lane.

Keywords Strong connectivity \cdot Streets-network \cdot Pareto Front

1 Introduction, Model and Problem formulation

The latest IPCC report paints a picture of a frightening future due to rising temperatures and their consequences. It has been shown that the rise in temperature results from an increase in the concentration of greenhouse gases (GHGs) in the atmosphere, among which 15% are due to transportation systems. But GHGs emissions and pollution are not the only critical problems. Another issue is the increase in land use for various human activities. A good balance must be striked among city development (housing, shops, schools, etc), new infrastructure construction and land degradation in order to reach the objectives of the Land Degradation Neutrality program of the UN. This work investigates a scenario where already degraded lands may be assigned to new projects. According to Adam Millard-Ball, based on previous studies,

S. Chattanachott and F. Guinand

LITIS Laboratory, Le Havre Normandie University, Le Havre (France)

 $E\text{-mail: supharoek.chattanachott} @etu.univ-lehavre.fr \ / \ frederic.guinand @univ-lehavre.fr \\ \\$

K. Lavangnananda

School of Information Technology (SIT), King Mongkut's University of Technology Thonburi (KMUTT), Bangkok (Thailand)

E-mail: kitt@sit.kmutt.ac.th

streets right-of-way represents an important percentage of cities area, between 13% and 30% in US and streets sprawling is still going on [1]. Based on these elements, we focus on the reduction of urban networks, by transforming the problems into graph-related problems. Reducing streets-network sprawl can be done in several ways. Several types of streets can be distinguished with respect to their direction and their number of lanes. Neglecting traffic density, the target goal becomes minimizing the sprawl of streets while keeping some streets-network properties with respect to the following constraints: (i) adding new lanes and new streets is not allowed, (ii) removing a street reduced to only one lane is not allowed (iii) removing a lane from a multi-lane street is allowed (iv) reversing the direction of a lane is not allowed (v) For each location, it must be reachable from any other location in the city The (v) constrain can be seen as having a strongly connected graph while the (iv) constraint is, probably the most important constraint. Streets-networks can be modelled by various types of graphs, from mixed graphs [4] to multi-directed graphs (digraphs). In such digraphs, each junction or roundabout is a node (or vertex). Nodes are linked by various types of streets, and each street lane in the network is associated with an arc in the graph model. Therefore all one-way streets can be represented as a set of arcs pointing in the same direction, and every two-way street is represented in the model by two sets of arcs with symmetric orientations. The original streets-network can be modelled as G = (V, A) and the digraph obtained after some removing arcs can be G' = (V', A'). Note that V' = V denotes u_V , the node u in G and $u_{V'}$ the same node in G'.

Two objectives are considered. The first objective is to minimize the number of arcs in the graph while maintaining the strong connectivity constraint such that if two vertices were connected in G they still have to be connected in G'. The second objective is to minimize the difference in the sum of eccentricities of nodes between G and G'. The eccentricity of a node is its longest distance to any other node in the graph. This problem belongs to the family of strong orientation optimization problems [2], where the goal consists in finding a strong orientation of a graph allowing some operations (arc removal or arc reversal for instance) and under some constraints, usually distance-based. However, the hypotheses considered in the present work do not allow a change in the nodes' neighbourhood, if two nodes are linked in the original graph, this property must remain in the modified graph and this constitutes a main variation with respect to state-of-the-art methods.

2 Method, Experiments and Analysis

According to the objectives, the proposed method aims at removing as many arcs as possible while keeping the eccentricity values as small as possible and under the strong connectivity constraint. First note that without loss of generality, from a strong connectivity point of view, every set of arcs pointing in the same direction can be reduced to only one arc. In addition, such operations have no impact on the values of node eccentricities. Thus, the focus of the proposed method will be on graphs for which between any two neighbour nodes there are either two arcs pointing in opposite directions or only one arc. Symmetric arcs in this work share the same extremities but point in opposite directions. The original graph is the graph with symmetric arcs and the reduced graph is the graph obtained after removing part or all symmetric arcs from the original graph. Both the original graph and the reduced graph are strongly connected.

Two objectives are considered: minimizing the number of symmetric arcs and minimizing the sum of the eccentricities. The first one is measured as the ratio between the number of conserved symmetric arcs and the total number of symmetric arcs in the original graph. The second objective is measured as the percentage of increase of the sum of eccentricities in the reduced graph compared to this sum in the original graph.

The problem is thus a bi-objective optimization problem, for which the solutions are described as a Pareto front (like the one in Figure 1 (b)) As the work is at a preliminary stage, a simple iterative method described by Algorithm 1 is proposed. Starting from an original graph with many symmetric arcs, the method consists in removing part of these symmetric arcs, modifying some arcs orientation in order to retain the strong connectivity constraint and then the sum of eccentricities is computed. This process satisfies the two objectives mentioned earlier.

Al	Algorithm 1: Method for reducing Streets-Network Sprawl					
I	Input: streets network G , m number of symmetric arcs					
1 M	ain():					
2	2 $S \leftarrow$ sum of eccentricities in G					
3	3 for $k \leftarrow 0$ to 100 do					
4	/* k is the percentage of removed arcs $*/$					
5	$G' \leftarrow \text{clone}(G)$					
6	remove randomly $k\%$ of symmetric arcs from G'					
7	make G' strongly connected					
8	$m' \leftarrow$ number of symmetric arcs of G'					
9	$S \leftarrow \text{sum of eccentricities}$					
10	P = (m'/m, S'/(S+S'))					
11	if P is non-dominated then					
12	add P to the Pareto front					
13	end					
14	end					
15						

For the experiments, we build a generic Manhattan-like street network generator based on a partial grid with one or two lanes of streets as illustrated in Figure 1 (a). According to Robbins' theorem [3], generated networks are bridgeless graphs, and according to the underlying application, graphs are already strongly connected. A random number of symmetric arcs are removed from the original graph while keeping the strong connectivity property.

For an original graph, many reduced graphs can be extracted by removing a variable number of symmetric arcs. For each of these reduced graphs, the

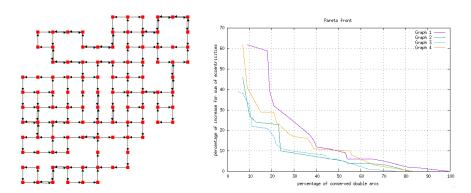


Fig. 1 (a) Manhattan-like streets-network model (b) The computed Pareto front .

sum of eccentricities is computed, thus a solution represented by a couple of values corresponding to the two objectives: the percentage of conserved symmetric arcs (the smaller the better) and the percentage of increase for the sum of eccentricities with respect to the original graph (the smaller the better). Considering two solutions $S = (o_1, o_2)$ and $S' = (o'_1, o'_2)$, if both $o_1 < o'_1$ and $o_2 < o'_2$, then S' is dominated by S and should be removed from the set of solutions. The set of non-dominated solutions builds a Pareto front. Therefore, from each original graph, it is possible to build such a Pareto front as illustrated in Figure 1 (b).

3 Conclusion and Perspectives

In this work, the problem of the minimization of streets-network sprawl is investigated by reducing in the number of lanes for existing streets. The variant of the considered problem does not allow arcs reversal. The preliminary results obtained from the analysis of instances of streets-networks produced by a generator designed on purpose, show that a reduction of almost 50% of symmetric arcs (two ways streets) entails a small increase of around (10%) for the maximum distance between any couple of places in the city. Future work lies in the implementation of more efficient optimization methods and applying them on real streets-networks in order to obtain a lower bound of the surface that can be saved by reducing streets-networks in small, medium and large cities from different countries.

References

- Christopher Barrington-Leigh and Adam Millard-Ball. Global trends toward urban street-network sprawl. Proceedings of the National Academy of Sciences, 117(4):1941– 1950, 2020.
- Rainer E. Burkard, Karin Feldbacher, Bettina Klinz, and Gerhard J. Woeginger. Minimum-cost strong network orientation problems: Classification, complexity, and algorithms. *Networks*, 33(1):57–70, January 1999.
- Herbert Ellis Robbins. A theorem on graphs, with an application to a problem of traffic control. The American Mathematical Monthly, 46(5):281–283, 1939.
- 4. Vincent Verbavatz and Marc Barthelemy. From one-way streets to percolation on random mixed graphs. *Physical Review E*, 103(4), April 2021.



A hybrid network: sea-land connectivity in the global system of cities

César DUCRUET · Barbara POLO MARTIN · Bruno MARNOT

Abstract The main objective of this research is to analyze the connectivity of cities in a coupled network made of planar (railways) and non-planar (maritime) topologies. It takes the state of the network during the period 1880-1925, namely the context of the First Globalization wave (1880-1914), when trade and urban development were closely tied to progress in communications systems and especially steam propulsion. Edges represent intercity physical infrastructure on land, and inter-port ship voyages at sea. We test several hypotheses in terms of inter-network specialization and port-city relationships. Main results underline a crucial influence of railway proximity on vessel traffic volume and steam specialization.

Keywords coupled networks ; globalization ; hinterlands ; ports ; maritime transport ; multilayer networks ; multigraph ; railway networks ; steam shipping ; urban network

1 Introduction

A relative consensus has been reached among scholars about the importance of both ports and railways in the rapid development of cities during the late 19th and early 20th centuries (Bretagnolle, 2015). Ports in particular had to "race for constant adaptation" to keep their position in an increasingly competitive environment (Marnot, 2005). Attracting maritime trade became more and more dependent on their ability to connect inland markets efficiently. Such

César DUCRUET and Barbara POLO MARTIN French National Centre for Scientific Research (CNRS) & UMR 7235 EconomiX Bruno MARNOT University of La Rochelle & UMR 7266 LIttoral ENvironnement et Sociétés (LIENSs)

This research acknowledges funding from the Agence Nationale de la Recherche (ANR) research project MAGNETICS (MAritime Globalization, Network Externalities, and Transport Impacts on CitieS), 2023-2026.

dynamics had the effect of expanding the hinterland boundaries of successful gateways, at the expense of numerous, less-equipped nodes, like French ports for instance (Merger, 2004).

While the growth or decline of port cities in this context has been well documented by historians, such across the Atlantic (Konvitz, 1994) and in Asia (Murphey, 1969), we still miss a "global picture" of such a global transportation system connecting cities of the world. Recent research has been done on the global maritime network in the age of steam, notably examining the relationship between maritime connectivity, technological innovation, and urban development (Ducruet and Itoh, 2022), but leaving aside the land-based network and ignoring inland cities. Some parent works investigated such dynamics in other contexts, such as inter-network externalities among ports, canals, and roads in England between 1760-1830 (Bogart, 2009), or the combination of airline and maritime global networks in recent years (Ducruet et al., 2011). Three main hypotheses are tested in the present paper:

- H1 : railway proximity to ports fosters maritime traffic volume;
- H2 : railway proximity to ports fosters steam shipping specialization;
- H3 : railway proximity to ports fosters global shipping connectivity.

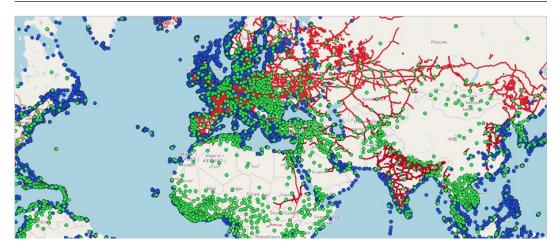
2 Data and methodology

We first reconstituted the global railway network between 1880-1920 on the basis of digitized historical maps¹ using QGIS. In this software, a manual work has been done to recreate the railway edges over the Open Street Map layer, together with two types of nodes: stations related to cities near railways, and intermediate junctions like crossroads. Ports and cities were attributed to this network using additional urban databases and the *Lloyd's Maritime Atlas* (see Figure 1). The global maritime network is derived from the *Lloyd's Shipping Index* on global inter-port vessel movements between 1880 and 1925, for both sailing and steam ships.

3 Preliminary results

As the Geographical Information System (GIS) environment allows calculating land distances between railway segments and port nodes, the first hypothesis can be verified with Figure 2. We observe that port-railway proximity has a noticeable influence on steam traffic size, as the volume of ship calls declines with distance. A similar distribution applies to the evolution of steam specialization (Figure 3). However, it is the fourth class (5 - 9.9 km) which keeps being more advanced in terms of shipping technology. Distance also influences steam specialization levels but in weaker ways than total traffic.

¹ The sources are David Rumsey collection, Bibliothèque Nationale de France (BNF), and Library of Congress.



French Regional Conference on Complex Systems - Le Havre, May 31 June 2, 2023

Fig. 1 Snapshot of railway connections in 1920 (red), ports (blue), and cities (green)

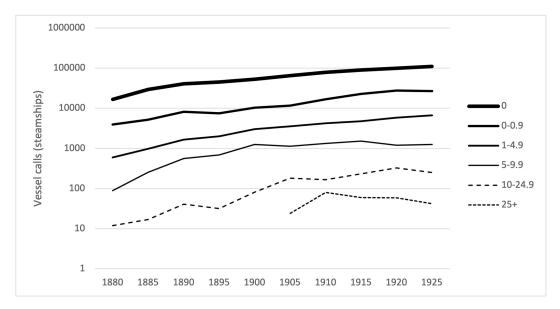
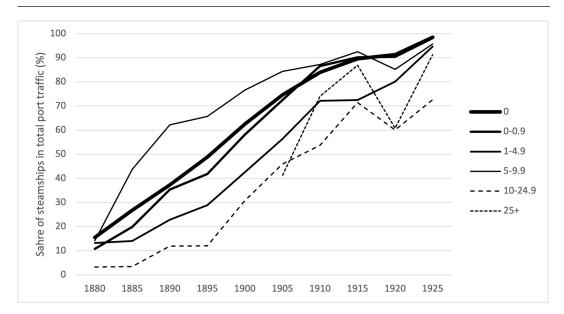


Fig. 2 Steam traffic evolution and distance from railway junctions (km)

The third step has been to shift the focus from port and railway nodes to urban nodes as the spatial unit of analysis, based on the delineation of cities proposed by Ducruet et al. (2018). We calculated the linear correlation between railway proximity (i.e. for each city, average of 1 / distance to all railway segments) and global maritime centrality in the steam and sailing ship networks (Table 1). Results are significant, again showing a stronger connection between steam shipping and railways. The sudden surge of correlation in 1925 for sailing comes from its near-disappearance (i.e. less than 3% global traffic) and its retreat around a few ports well-equipped with railways. The correlation with degree is always lower than with betweenness, while the latter is fading over the period.



French Regional Conference on Complex Systems - Le Havre, May 31 June 2, 2023

Fig. 3 Steam traffic specialization and distance from railway junctions (km)

Table 1 Correlation between (local) railway proximity and (global) maritime centrality

Network	Measure	1880	1885	1890	1895	1900	1905	1910	1915	1920	1925
Steam	Betweenness Degree										
Sail	Betweenness	0.287		0.273	0.263	0.266	0.194	0.200	0.223	0.114	0.443

4 Concluding thoughts and further research

This research confirmed the close relationship between railway proximity to ports and steam shipping distribution across the world between 1880 and 1925. The precision offered by the GIS calculations allow to conclude that not only the presence or proximity of railways matter for ports, but their connexion *within* ports.

Further research is ongoing for testing the relationship between city size, growth (i.e., population) and modal as well as intermodal centrality/accessibility. Another research avenue is the integration of the global road network backbone, navigable rivers and canals as additional layers. In such a multilayered system, attention will also be paid to the centrality and development of nonport, inland cities, including a view on the influence of maritime shipping on their overall connectivity. As in recent works focusing on the non-port capital cities of coastal countries (Ducruet and Guerrero, 2022), a variety of trade, logistical, and economic variables will be used to control for country-level effects, depending on their availability, as the present study is planned to cover a century of transport and urban development at the global scale.

References

Bogart D. (2009) Inter-modal network externalities and transport development: evidence from roads, canals, and ports during the English Industrial Revolution. *Networks and Spatial Economics*, 9: 309-338.

Bretagnolle A. (2015) City-systems and maritime transport in the long term. In: Ducruet C. (Ed.), *Maritime Networks: Spatial Structures and Time Dynamics*, London and New York: Routledge, pp. 27-36.

Ducruet C., Cuyala S., Hosni A. (2018) Maritime networks as systems of cities: The long-term interdependencies between global shipping flows and urban development (1890–2010). *Journal of Transport Geography*, 66: 340-355.

Ducruet C., Guerrero D. (2022) Inland cities, maritime gateways, and international trade. *Journal of Transport Geography*, 104: 103433.

Ducruet C., Ietri D., Rozenblat C. (2011) Cities in worldwide air and sea flows: A multiple networks analysis. *Cybergeo: European Journal of Geography*, 528: https://journals.openedition.org/cybergeo/23603

Ducruet C., Itoh H. (2022) The spatial determinants of innovation diffusion: evidence from global shipping networks. *Journal of Transport Geography*, 101: 103358.

Konvitz J.W. (1994) The crises of Atlantic port cities, 1880 to 1920. Comparative Studies in Society and History, 36(2): 293-318.

Marnot B. (2005) Interconnexion et reclassements : l'insertion des ports français dans la chaîne multimodale au XIXe siècle. *Flux*, 59(1): 10-21.

Merger M. (2004) Inner ports and railway networks in France or the 'history of a divorce' (1830-1914). In: Dienel H.L. (Ed.), Unconnected Transport Networks: European Intermodal Traffic Junctions 1800-2000. Frankfurt & New York: Campus Verlag.

Murphey R. (1969) Traditionalism and colonialism: Changing urban roles in Asia. The Journal of Asian Studies, 29(1): 67-84.



Agent-based modelling of urban expansion and land cover change: a prototype for the analysis of commuting patterns in Geneva, Switzerland.

Flann Chambers $\,\cdot\,$ Christophe Cruz $\,\cdot\,$ Giovanna Di Marzo Serugendo

Abstract Agent-based modelling has been used in many studies of urban expansion, land use and land cover change patterns. While representing a powerful tool for depicting and formulating predictions about the evolution of interconnected complex systems, this method also poses a series of challenges to the researcher community, most notably in terms of model calibration and validation, and output data visualisation. Based on these findings, we present an agent-based model developed in GAMA, coupled with a data exploration platform coded in python, for analysing commuting patterns in the canton of Geneva, Switzerland. Output datasets generated from a set of simple evolution rules for the agents, are distributed in open access together with the code for the associated data visualisation platform. This prototype is early work in developing a series of agent-based models for simulating urban expansion and land cover change dynamics, together with their own data exploration platforms for calibration, validation and output data analysis purposes. These toolboxes will be developed with the intent of addressing the various shortcomings in agent-based modelling research discussed in this paper.

Keywords Agent-based modelling \cdot Urban expansion \cdot Land use \cdot Land cover change \cdot Commuting patterns

Flann Chambers Centre Universitaire d'Informatique Université de Genève

Christophe Cruz Laboratoire Interdisciplinaire Carnot de Bourgogne Université de Bourgogne

Giovanna Di Marzo Serugendo Centre Universitaire d'Informatique Université de Genève

1 Introduction

Urban management decision-makers, policy makers, and the scientific community recognise the high importance of monitoring the dynamics of land cover change and the evolution of public transportation offer, before, during, and after, urban expansion projects are planned and achieved [2, 6, 12, 13].

In those complex social-ecological systems, Agent-Based Modelling (ABM) is proven to be particularly well suited to understanding, describing, but also predicting phenomena arising from the intricate interactions between the involved parties (such as transportation network, rural surfaces in competition with newly built areas) in a context of rapid urban consolidation [1,4,8].

However, ABM research has encountered many challenges and setbacks inherent to this method, such as the requirement of large amounts of data at a fine resolution [12, 13], rigorous model testing and validation [1, 12], and clear presentation of results in the form of calibration, validation, and output data exploration and visualisation [1]. We develop an agent-based model for land use and land cover change in urban areas, coupled with a data visualisation platform, that will help us calibrate and validate our model efficiently, as well as analyse and display its results in a versatile and interactive manner. For this task we benefit from open access to a large database of vector geomatic data¹, which both solves the data hungriness aspect of such a model, and provides highly relevant [12, 13] fine spatial detail to the simulation. We present a first version of the model and platform, and demonstrate that they are capable of producing and displaying measurements from data generated from the simulation. This version is early work towards using validated agent-based models to describe, predict and prescribe phenomena related to urban expansion, land use and land cover change in three different case studies, in a way that will be able to inform decision-making processes.

The following sections will navigate us through the context for this study (section 2), a discussion of our research methodology (section 3) and of planned future works (section 4) for the project.

2 Contextualisation

Urban planning occupies a central space in any government's agenda. Rapid urban consolidation bears tremendous impacts on the ecosystems' as well as the residents' well-being. Indeed, not only do negative side-effects build up on the local biodiversity, such as habitat loss from the conversion of rural areas and forests to built areas [9, 11], they also affect the population, who may suffer for instance from the emergence of isolated zones of abnormally high temperatures during heat spells in summer [7].

We focus on the use case of the canton of Geneva and its desire to develop large axes of public transportation in the form of tram lines. Already

¹ Système d'Information du Territoire Genevois.

expressed in policies enacted in 1988 2 , this resolution has been concretized more recently through the establishment of tram lines 14 and 18 connecting most notably the city center to Meyrin and CERN – located north-west, next to the international airport of Geneva. By itself, this decision already heavily impacts the urban dynamics along this high-traffic axis. However, it may also prompt the development of additional residential and commercial centers which would benefit from the increased accessibility of the area, resulting into accrued urban consolidation and land cover changes.

Motivated by the works of ABM researchers in the field of urban planning related to land use and land cover change [2, 6, 12, 13] and reviews from An et al. [1], Groeneveld et al. [4], and Rounsevell et al. [8], we develop an agent-based model for the Cornavin-Meyrin-CERN axis, one of the most important transportation axis in the canton of Geneva. By representing individual behaviors at the smallest scale (i.e. the commuter's perspective and decision-making), and by integrating complex interactions between heterogeneous agents, agent-based modelling proves to be particularly well suited in representing commuting patterns [5] and related urban consolidation processes. Indeed, agent-based modelling encapsulates emerging phenomena, that, despite being only witnessed at the global scale, can not be simply explained and represented by the sum of its parts [3]. It, in facts, remedies the lack of representation of the individuals' agency that can be found in other types of urban models [1].

3 Method

This section presents all methods used when developing the model, including the collection of data, the model design and building, and the data compilation and visualisation processes.

3.1 Data

We use data sourced from the SITG³ from which we gather vector data in the form of shapefiles. The spatial extent and useful attributes of buildings, tram lines, tram stops, roads and initial commuter places of residence⁴ are harnessed from these shapefiles. We also use statistics from the OCSTAT⁵ to divide the virtual population into an accurate representation of age classes.

² Votation populaire du 12 juin 1988.

³ Système d'Information du Territoire Genevois.

 $^{^4\,}$ See section 3.2 for agents details.

⁵ Office Cantonal des Statistiques.

Age class	Action plan	Time of day		
Children	1) Go to closest school.	Anytime between 7:00 and 7:59.		
	2) Leave school:	Anytime between 17:00 and 17:59.		
	a. Go to closest leisure (50%) .			
	b. Go home (50%).			
	3) If 2a., go home.	2 hours after arrival.		
Students	1) Go to university (Cornavin).	Anytime between 7:00 and 7:59.		
	2) Go home.	Anytime between 17:00 and 17:59.		
Adults	1) Go work.	Anytime between 6:00 and 8:59.		
	2) Go home.	10 hours later.		
Retired	1) Choose activity:	Anytime between 8:00 and 20:59.		
	a. Closest shop (20%) .	Spend one hour.		
	b. Closest leisure (60%).	Spend two hours.		
	c. Stay home (20%).			
	2) If relevant, go home	1 or 2 hours after arrival.		

 Table 1 Commuter agents action plan details.

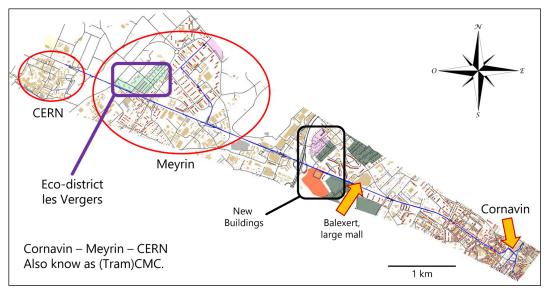


Fig. 1 Model visualisation for the Cornavin-Meyrin-CERN axis. The blue lines correspond to the tram lines 14 and 18.

3.2 Model building

We use the GAMA platform [10] to develop our agent-based model (see figure 1 for a visualisation of the model inside the GAMA platform). We define roads, tram lines and stops, buildings, tram vehicles and commuters as agent species, which is the terminology used in GAMA to construct groups of agents. Tram vehicles and commuters are able to move through space, and commuters have their own action plan and decision-making framework.

Whereas tram lines, tram stops, roads and buildings simply provide information about their locations and some attributes, such as building type (residential, business, school, leisure, shop), trams and commuters play a more active role by virtue of being able to move in the virtual environment. Trams service tram stops and follow the path given by the tram lines. They automatically reverse at terminal stations. Among the vast Geneva public transport offer, two different tram lines are represented: line 14 starts from Meyrin center and line 18 starts from CERN. They join shortly after their departure and service the same major axis. Despite continuing their journey past the station of Cornavin towards the communes of Onex and Carouge, respectively, for the interests of this case study, we simplify and set their second terminal station as the station of Cornavin, where they reverse and continue their journey the opposite way.

Commuters represent people who use some transportation method to go to work (or other building types) every day. At model initialisation, they are assigned a place of residence, and an age class (child, student, adult, or retired). Depending on their age class, they adopt different actions plans, which are depicted in Table 1. Based on their action plan and their current intention, commuters evaluate the transportation options and choose the option that fits their needs based on multiple factors: the distance between their origin and their destination, the availability of tram stops within 400 meters of their origin and their destination, and whether or not a tram line directly services both the corresponding stops of their origin and their destination. Commuters will simply walk from point A to point B if the Euclidean distance is less than 800m, and otherwise take the tram or the car. Among adult commuters, 50% will always journey by car. The remaining commuters will evaluate if the journey by tram is possible and if there are stops within 400m of their origin and destination, if yes, they will take the tram, otherwise they will journey by car.

The model proceeds in steps of 10 seconds and simulates whole days starting from 6:00 am. The day ends when all commuters have finished their action plans and reached home. Another day then starts. The model starts in the year 2010, and at the end of each day, may also advance to a new year, as defined by the user. In this case, based on their construction year, new buildings appear and may welcome new commuters inside the virtual world. These new commuters are initialised in the same way as the existing ones, and will also evolve in the system.

3.3 Output data exploration and visualisation

We develop our data exploration, analysis and visualisation platform in python, with the help of libraries such as *pandas*, *geopandas*, *plotly* and *Dash*. Example datasets generated from the model, as well as python code for the aforementioned platform, are available in open $access^6$ to anyone willing to analyse model results. The dataset contains information about every journey undertaken during the simulation: who made the commute, what was their age class, what transportation method (by foot, by car or by tram) did they use. The locations of each tram whenever their current number of passengers changes

⁶ https://gitlab.unige.ch/cas/TCMC_ABM_Data_Code

are also recorded. Finally, there is one shapefile containing the road network for each year simulated, and each road holds the information of the amount of unique commuters (journeying by car) it has seen for that year.

4 Future works

The ABM prototype presented in this paper is early work in a larger modelling effort, which includes a total of three different use cases, and a broader field of study. The spatial extent of our model will be expanded to the whole canton of Geneva, and the region of Greater Geneva, which also includes the neighboring areas of the canton of Vaud, Switzerland, as well as the French departments of Ain and Haute-Savoie. We will simultaneously continue developing the data visualisation platform to match the model's calibration, validation and output data analysis needs. The conceptual framework of the system will have its priority shifted towards the representation, understanding, and prediction and prescription of urban expansion and land cover change mechanisms, while retaining parts of the commuting behaviors model we presented in this paper.

5 Conclusion

We have presented an early version of an agent-based model for investigating urban expansion related mechanisms such as commuting patters and behaviors, applied to the use case of a major transportation axis in the canton of Geneva, Switzerland. A series of evolution rules have been implemented for the different agent types, and the output datasets are available for anyone to peruse. We introduced a data visualisation and exploration platform coded in python, which is and will be dedicated to the analysis and presentation of crucial datasets pertaining to the calibration, validation and exploration of results of the model. The model is still in development and will include a larger spatial extent, sound calibration, validation and results analysis, and consideration of additional phenomena related to urban expansion and land cover change. Additionally, work is being done towards developing a plugin for GAMA, that will be dedicated to linking the model to a knowledge graph server. In doing so, by using the information contained inside this organised information database, agents may be able to dynamically adjust their behaviors and knowledge about their environment.

References

 L. An, V. Grimm, A. Sullivan, B. L. Turner II, N. Malleson, A. Heppenstall, C. Vincenot, D. Robinson, X. Ye, J. Liu, E. Lindkvist, and W. Tang. Challenges, tasks, and opportunities in modeling agent-based complex systems. *Ecological Modelling*, 457:109685, Oct. 2021.

- C. Cantergiani and M. Gómez Delgado. Urban growth simulation with AMEBA: Agentbased Model to residential occupation. *Boletín de la Asociación de Geógrafos Españoles*, 2020. Accepted: 2020-10-15T09:36:41Z Publisher: Asociacion Espanola de Geografia.
- A. Crooks, N. Malleson, E. Manley, and A. Heppenstall. Agent-Based Modelling and Geographical Information Systems: A Practical Primer. SAGE Publications Ltd, Thousand Oaks, CA, first edition edition, Jan. 2019.
- J. Groeneveld, B. Müller, C. M. Buchmann, G. Dressler, C. Guo, N. Hase, F. Hoffmann, F. John, C. Klassert, T. Lauf, V. Liebelt, H. Nolzen, N. Pannicke, J. Schulze, H. Weise, and N. Schwarz. Theoretical foundations of human decision-making in agent-based land use models – A review. *Environmental Modelling & Software*, 87:39–48, Jan. 2017.
- S. Z. Leao and C. Pettit. Mapping Bicycling Patterns with an Agent-Based Model, Census and Crowdsourced Data. In M.-R. Namazi-Rad, L. Padgham, P. Perez, K. Nagel, and A. Bazzan, editors, *Agent Based Modelling of Urban Systems*, Lecture Notes in Computer Science, pages 112–128, Cham, 2017. Springer International Publishing.
- N. Mozaffaree Pour and T. Oja. Urban Expansion Simulated by Integrated Cellular Automata and Agent-Based Models; An Example of Tallinn, Estonia. Urban Science, 5(4):85, Dec. 2021. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.
- C. M. Nwakaire, C. C. Onn, S. P. Yap, C. W. Yuen, and P. D. Onodagu. Urban Heat Island Studies with emphasis on urban pavements: A review. *Sustainable Cities and Society*, 63:102476, Dec. 2020.
- M. D. A. Rounsevell, D. T. Robinson, and D. Murray-Rust. From actors to agents in socio-ecological systems models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1586):259–269, Jan. 2012.
- K. Shi, Y. Chen, B. Yu, T. Xu, L. Li, C. Huang, R. Liu, Z. Chen, and J. Wu. Urban Expansion and Agricultural Land Loss in China: A Multiscale Perspective. *Sustain-ability*, 8(8):790, Aug. 2016. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.
- P. Taillandier, B. Gaudou, A. Grignard, Q.-N. Huynh, N. Marilleau, P. Caillou, D. Philippon, and A. Drogoul. Building, composing and experimenting complex spatial models with the GAMA platform. *GeoInformatica*, 23(2):299–322, Apr. 2019.
- L. Tang, X. Ke, Y. Chen, L. Wang, Q. Zhou, W. Zheng, and B. Xiao. Which impacts more seriously on natural habitat loss and degradation? Cropland expansion or urban expansion? Land Degradation & Development, 32(2):946–964, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ldr.3768.
- P. Tsagkis. Urban growth models and calibration methods: a case study of Athens, Greece. Bulletin of Geography. Socio-economic Series, (55):107–121, Mar. 2022. Number: 55.
- T. Xu, J. Gao, G. Coco, and S. Wang. Urban expansion in Auckland, New Zealand: a GIS simulation via an intelligent self-adapting multiscale agent-based model. *International Journal of Geographical Information Science*, 34(11):2136–2159, Nov. 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13658816.2020.1748192.



Radial analysis and scaling law of housing prices in French urban areas using DVF data

Gaëtan Laziou · Rémi Lemoy · Marion Le Texier

Abstract Using a nationwide dataset with millions of real estate transactions, this paper investigates the relationship between housing prices and city size through a radial (center-periphery) analysis. We find that housing price radial profiles scale in three dimensions with the power 1/5 of city population. Nonetheless, housing prices in the city center seem to be more sensitive to city population, raising the question of housing affordability in the center for low-income households.

Keywords Housing prices, DVF data, Radial analysis, Urban scaling

1 Introduction

Today, more than half of the world's population lives in urban settlements, and urbanization is still an ongoing trend [1]. Thus, it is of growing interest to understand how cities evolve to eventually make them more sustainable, because cities concentrate population, activities and innovation, but also pollution, as well as social inequalities. A growing number of studies look at cities as complex systems whose shape follows specific scaling laws [2]. The commonly used method consists in running the following model:

$$Y = \alpha N^{\beta} \tag{1}$$

where Y is an urban attribute, N the population of each city, α a normalization constant and β a scaling exponent [3,4].

However, aggregated quantities are assumed to be evenly distributed within cities, which is a questionable assumption. In line with well-known urban geography and economic literature [5-7], we would rather consider cities through their

Gaëtan Laziou

UMR CNRS IDEES 6266, University of Rouen E-mail: gaetan.laziou1@univ-rouen.fr

Rémi Lemoy UMR CNRS IDEES 6266, University of Rouen

Marion Le Texier LAGAM, University Paul Valéry, Montpellier 3 radial profile, and thus investigate their internal structures. Here, we look over the relationship between housing prices and city size, which has been little studied before (see e.g.[8]).

2 Materials and methods

Our main material is the *Demande de Valeurs Foncières* (DVF) dataset, produced by the central state tax services. This dataset provides data on millions of property transactions over the past five years.

Following the methodology of [9], a radial analysis is conducted on our sample of cities. We first define concentric rings of fixed width of 1km around the city hall, and then aggregate the price per square meter (obtained by regressing prices on dwelling size) of transactions within each ring at distance r to the city center. We express a volume scaling law for the radial function of housing prices p(r), following the equation:

$$p_{\beta}(r) = N^{\beta} f(r/N^{\beta}) \tag{2}$$

where N is the population of each city, β a scaling exponent and r the euclidian distance to the center in the Lambert 93 coordinates system.

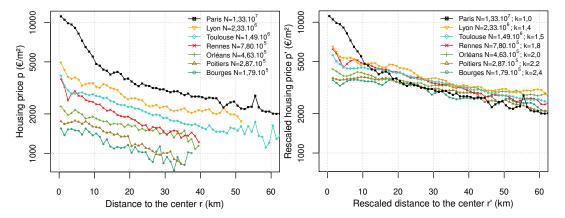


Fig. 1 Left panel: housing price as a function of the distance to the center in different French cities. Population N is given in the legend. Right panel: rescaled curves for the same cities. The rescaling factor k is given in the legend (r'=rk and p'=pk). Thus, r'=10km corresponds to r=10km in Paris, $r \approx 6.5$ km for Toulouse and $r \approx 4.2$ km for Bourges)

3 Results

Looking at the left panel of Figure 1, it can be seen that housing prices are decreasing with distance, and that the more populated a city is, the more expensive it is. We try different values of scaling exponents and find that the rescaling parameter given by the power 1/5 of city population, $N^{1/5}$, leads to the most homogeneous profiles between cities. Thus, the rescaled distance to the center is $r' = r \times (N_{Paris}/N)^{0.2}$ while the rescaled housing price is p'(r') =

 $p(r') \times (N_{Paris}/N)^{0.2}$. On the right panel, the rescaling seems to work satisfactorily with the sample of cities as it captures a common trend. Nonetheless, this is less the case for the first kilometers to the city center.

Thus, the power law relationship connecting housing prices and population in the city center was further investigated (housing price p_c within 0 < r < 1km from the center) on Figure 2. Overall, it appears that population can help predicting housing prices in city centers, as most of the variance can be explained by the models. However, we find a much higher exponent β than 1/5, which shows that real estate prices in the city center rise faster than the same prices in the periphery, with respect to population. Note that the α parameter (indicating the theorical housing price for N = 1 inhabitant) is close to zero.

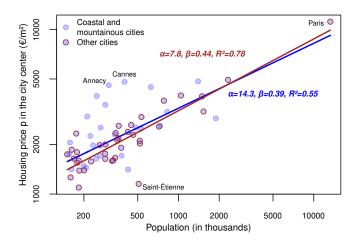


Fig. 2 Housing price in the city center of French cities as a function of city population. Non-coastal and non-mountainous cities are identified by a brown border. Note that we define coastal cities as cities whose boundaries coincide with the coastline, whether over a long or a short distance (latest delineations of functional urban areas provided by the French census bureau as a reference). Moreover, mountainous cities include the cities of Annecy, Chambéry and Grenoble, in the Alps. Lines show the best fit to a scaling relation $p_c(N) = \alpha N^{\beta}$ for all cities (blue line) and only non-coastal and non-mountainous cities (brown line).

4 Conclusion

Despite some fluctuations, explained by other locational amenities that remain to be studied, our empirical analysis tends to prove that the radial profiles of urban housing prices are homothetic (in the two euclidian dimensions of geographical space and in the vertical dimension of price). We also find that housing prices in the city center are more sensitive to city population, as we find a higher exponent. These results have strong implications for lower-income households, because prohibitive prices in the center may contribute to push them out into peripheral locations.

References

- 1. UNPD, World Urbanization Prospects: The 2018 Revision (2018)
- 2. M. Batty, The size, scale, and shape of cities, Science, **319**(5864), pp. 769-771 (2008)
- L. MA Bettencourt, The origins of scaling in cities, Science, **340**(6139), pp. 1438-1441 (2013)
- 4. L. MA Bettencourt, J. Lobo, Growth, innovation, scaling, and the pace of life in cities, Proceedings of the national academy of sciences, **104**(17), pp. 7301-7306 (2007)
- 5. W. Alonso, Location and land use. Harvard University Press, Cambridge, MA (1964)
- 6. M. Fujita, Urban economic theory. Cambridge University Press (1989)
- 7. J.H. Von Thünen, Der isolierte staat in beziehung auf landwirtschaft und nationalökonomie (1826)
- 8. P.P. Combes, G. Duranton, L. Gobillon, The costs of agglomeration: House and land prices in French cities, The Review of Economic Studies **86**(4), pp. 1556-1589 (2019)
- 9. R. Lemoy, G. Caruso, Evidence for the homothetic scaling of urban forms, Environment and Planning B: Urban Analytics and City Science, 47(5), pp. 870-888 (2020)



Towards a geographical theory different from that of the natural sciences: foundations for a relational complexity model

Olivier Bonin

Abstract We present in this paper a formal model that focuses on relational complexity. The model enables to connect concepts from the physical world and concepts from the realm of ideas both topologically and geometrically. The formal properties obtained in this model, such as identification, adjunction, copy and deletion are coherent with the phenomena that are modelled.

Keywords geography \cdot relational complexity \cdot theoretical model \cdot category theory

Introduction

Complexity is generally rooted in the desire to account for phenomena that seem to escape the reasonable criteria of system evolution, with mainly selforganisation and emergence. It is understandable that this problem interested geographers very early on, who quickly transposed the work of Prigogine [1], Haken [2] or Weidlich [3] to their field, to mention only three of the authors who developed an original approach of complexity originating from the field of dynamical systems. More recently, the development of agent-based modelling (ABM) has made it possible to model many social problems in a mimetic way, capturing complex behaviour.

For a complex model to be ontologically more than the sum of its parts, it seems logical to focus on the relationships between the elements of this system. This is the approach developed by the theoretical biologist Robert Rosen [4]. Rosen starts by defining simple models that he calls mechanisms. A simple model (1) belongs to a category of dynamical systems that are mathematical images of the system itself, and (2) among all the possible images of such a

O. Bonin

Univ Gustave Eiffel, Ecole des Ponts, LVMT Tel.: +33.1.81.66.88.68 E-mail: olivier.bonin@univ-eiffel.fr

system, there is one to which all the other images can be related (are homeomorphic): a maximal image. For Rosen, a complex system is a completely new mathematical object, whose mathematical image is not a dynamical system, and therefore does not have a maximal image. The memory evolutive systems [5] is an application of Rosen's idea to modelling memory and dementia.

We propose the foundations of a mathematical framework for developing new models of geographical phenomena, i.e., at the intersection of the natural sciences and the humanities and social sciences, building upon Rosen's idea and a ground-breaking paper by Pratt [6]. Generally, the natural world (described by physics) is best modelled as continuous, with motion, position, speed, etc., whereas the human world is best thought of as discrete: decisions, presence / absence, relations, membership, etc.. Interactions in the physical world are causal, follow the direction of time. Human decisions are logical, and logic swims upstream against time: the results of the actions anticipated before their implementation.

A hybrid geographical model could take the form of dynamic systems describing the physical world, coupled with an ABM. However, the formal coupling of dynamical systems and ABM in a way that makes them interact and respects the different movements of causality and intentionality with respect to time is, to our knowledge, an open problem. Although abstract and not directly implementable in a computer, the framework presented here is a formal way to do so. As the approach to complexity chosen here is relational, it is based on category theory, whose main purpose is to focus on the relationships between elements.

Category theory

In Category theory, the fundamental concept of homomorphism allows to compare two structured but very different sets, A and B. If we note \top the operation which structures A and \bot the operation which structures B, a homomorphism H of A in B is an application which to any element f of A associates an element H(f) of B such that : $H(f \top g) = H(f) \bot H(g)$, which amounts to saying that the homomorphism H translates A into B and the relation \top into \bot .

A category \mathcal{C} is a family of objects, with, for each pair of objects $A, B \in \mathcal{C}$, arrows connecting A to B, which we note $\mathcal{C}(A, B)$, and with a law statisfying : $\mathcal{C}(A, B) \times \mathcal{C}(B, C) \to \mathcal{C}(A, C)$.

Then, we introduce the concept of functor that allows to compare the morphology of two categories with homomorphisms. Let C_1 and C_2 be two categories. A functor F from C_1 into C_2 is a transformation which to any f of C_1 associates a f' = F(f) of C_2 such that : $F(f \circ g) = F(f) \circ F(g)$ and F transforms identical applications into identical applications.

Homomorphisms are often surjective, so they introduce simplifications. A functor will retain only some aspects of the starting category, which makes it a powerful tool for comparing structures.

The search for generalisation can lead to an interest in objects which, among the classes of objects, satisfy certain properties. In the language of category theory, the existence of these properties is expressed by arrows, i.e. by the existence of commutative diagrams between the objects of a category. These commutative diagrams constitute a kind of subcategory.

A diagram D in a category C is a directed graph whose nodes I are objects of the category and whose edges E are morphisms. We can also define a diagram D as a graph homomorphism between a graph of index I and the graph of category C.

A cone to D is an object c and a family of morphisms f_i such that :

$$\forall i, j \in I \ \forall e \in E \ f_e \in (C)[d_i, d_j] \Rightarrow f_e \circ f_i = f_j.$$

A limit for D is then a terminal object in the category of the set of cones to D. The concepts of cocones and colimites are defined in a dual way (the dual is defined by reversing all the arrows of the morphisms).

We will also need the concept of Chu Space. A Chu space over a set K consists of a set A of points, an antiset X of states, and an $X \times A$ matrix with entries drawn from K.

A formal model for coupling the natural world and the world of intentionality

We intend to describe with relevant mathematical concepts the possible interactions between the physical world and the behaviours of humans, communities and societies, which requires, from our point of view, to deal with interactions, both mental and physical, as well as to accommodate the downstream logic of causality and upstream logic of intentionality.

The approach we propose is to consider that we have two distinct worlds, the world of physics, and the world of decisions (from a human perspective), which must be connected. We consider all the class of all sets, each understood as a pure physical object (A, B, C, etc.), that we call **Set** and the class of all sets, each understood as a pure human construct (X, Y, Z, etc.), that we call **Setop**. Following Rosen's approach, these sets can be locally approximated by simple models (dynamical systems, ABM, etc.).

In **Set**, we construct the graph whose edges are the functions $f : A \to B$ connecting set A to set B, and transform this graph into a category (i.e. we add an identity function and a composition rule). Elements of **Setop** X transforms into Y with an antifunction, i.e. a binary relation whose converse is a function $Y \to X$. **Setop** can be turned into a category isomorphic to the dual to **Set**.

We now need to connect **Set** and **Setop**. If we worked in the realm of real numbers, we could place a first world at -1, the other one at +1 and connect them in two ways:

- algebraically: -(-1) = +1.

- geometrically: with the [-1, +1] interval.

Here, we connect the two sets in a similar way:

- algebraically with duality (the analogous to negation) that transforms Set into Setop
- geometrically: we introduce all Chu spaces and a graph which includes as subgraphs **Set** and **Setop**.

Properties of the model and interpretation

The properties of the model directly derive from its construction. They are consistent with the expectations related to the properties of the physical world and the world of ideas, and allow to link them in an elegant way.

Functions identify if they are not injective: f(a) = f(b) identifies a and b. It means that two different elements or models can be identified to represent the same phenomenon without being identical. Functions adjoin when they are not surjective: $f : A \to B$ transforms A onto f(A), and then adjoins Bf(A). Adjonction allows to associate a rich model to simpler one. Thus, functions are pertinent to describe operations of the physical world. In the physical world, there is no such thing as a strict copy, or two identical states of a system, but close enough states or models that can be identified.

Antifunctions make copies when their converses are not injective, and delete when their converses are not surjective. Antifunctions are pertinent to describe the mental world. Ideas can be strictly identical, exact copies of themselves, and can be totally suppressed.

This use of Chu spaces enables to describe formally several types of interactions. These interactions are directional, either causal (forward) or intentional (backward). This formalism can accommodate circular causalities, anticipation and finalism, as well as classical physical causalities.

Let us give a preliminary example of how such a model could be used. If we are interested in the residential choices in city, we can identify at least two spheres that interact: the sphere of economics, and the sphere of human decisions. The economic approach is causal, described for instance by Alonso's model [7]. The evolution of surfaces and prices follows the direction of time, as a consequence of the maximisation of the agents' utility and the determination of prices by an auction mechanism. The individual and collective decisions may rely on anticipations, beliefs, strategies. For instance, on can bet on the fact that a neighbourhood will be revitalised. This kind of reasoning flows against the direction of time. It can be efficiently modelled by ABM. A coupling of both approaches has been proposed for instance in [8]. However, this coupling consists in computing quantities from the ABM, introducing them into the economic model, which is also turned into a disequilibrium model. This limits the possibility to interpret the resulting model and to calibrate it. Moreover, it is not a complex model, in the sense that it cannot adapt to totally different logics, such as a financial crisis or changes in lifestyle that would require changes in the very structure of the models. What is needed is a model that can describe how the economic and the social approaches can interact, without hindering the natural behaviour of each domain.

The formal modelling framework presented here would enable to link a wider array of physical models (here, economical ones) and decision models, and to make explicit, via the functions and the antifunctions, the similarities and dissimilarities between all possible models, and the possible circulations from one model to another one. Notably, the construct with the Chu spaces enable an almost continuous coupling of completely different views on the same problem, based on the relational properties of the models.

Let use note however that this model is at this state purely formal and enables solely to make qualitative inferences. However, it enables to take account of many properties of geographical phenomena, or more generally of the coupling between the natural sciences and the social sciences and solves several apparent contradictions.

References

- I. Prigogine, G. Nicolis, The Journal of Chemical Physics 46(9), 3542 (1967). DOI 10.1063/1.1841255
- 2. H. Haken, Applied Physics A 57(2), 111 (1993)
- W. Weidlich, Chaos, Solitons & Fractals 18(3), 431 (2003). DOI 10.1016/S0960-0779(02)00666-5
- 4. R.J. Rosen, Theoretical Biology and Complexity: Three Essays on the Natural Philosophy of Complex Systems (Academic Press, 1985)
- A.C. Ehresmann, J.P. Vanbremeersch, Axiomathes 16(1-2), 137 (2006). DOI 10.1007/s10516-005-6001-0
- V. Pratt, in *TAPSOFT'95*, vol. 915 (Springer-Verlag, Aarhus, Denmark, 1995), vol. 915, pp. 108–122
- W. Alonso, Location and Land Use: Toward a General Theory of Land Rent (Harvard University Press, Cambridge (Mass.), Etats-Unis d'Amérique, 1964)
- O. Bonin, J.P. Hubert, Revue d'Économie Régionale & Urbaine (3), 471 (2014). DOI 10.3917/reru.143.0471

Structure & Dynamics



On the impact of introducing random modifications to the neighborhood of the abelian sandpile
Paulin Héleine, Juan Luis Jiménez Laredo, Frédéric Guinand and Damien Olivier
Asymptotic Dynamic Graph Order Evolution Analysis Vincent Bridonneau, Frédéric Guinand and Yoann Pigné
Who to Watch When? Strategic Observation in the Inverse Ising Problem Zhongqi Cai, Enrico Gerding and Markus Brede 285
The architecture of multifunctional ecological networks Mar Cuevas-Blanco, Sandra Hervías-Parejo, Victor Martínez Eguiluz, Lucas Lacasa, Isabel Donoso and Anna Traveset
Temporal Betweenness Centrality on Shortest Paths Mehdi Naima, Matthieu Latapy and Clémence Magnien



On the impact of introducing random modifications to the neighborhood of the abelian sandpile

Paulin Héleine · Juan Luis Jiménez Laredo · Frédéric Guinand · Damien Olivier

Abstract Proposed by Bak, Tang and Wiesenfeld in 1987, the Abelian sandpile was the first mathematical model to describe the phenomenon of Self-Organized Criticality (SOC). In its canonical form, the model is defined as a cellular automaton with cells arranged in a regular grid structure that can also be represented as a graph: nodes representing cells and edges the neighborhood of a cell. By introducing random modifications to the regular grid structure, this paper aims to explore the impact that other regular but non-grid topologies can have on the dynamics of the system. Starting with the canonical topology as a baseline, edges are rewired at random with the only constraint of keeping constant the indegree and outdegree of nodes. The different graphs studied start from a regular grid structures (when no edge is rewired), going through small-world graphs (for small rewiring coefficients), to end with random topology. The obtained results show that, among the wide range of used metrics, the signature of SOC is preserved despite different graph structures have an impact on the dynamics of the model.

 $\mathbf{Keywords}\ \mathrm{self-organised\ criticality} \cdot \mathrm{abelian\ sandpile} \cdot \mathrm{networks\ and\ complex\ systems}$

P. Héleine, F. Guinand, D. Olivier

LITIS Laboratory, Le Havre Normandie University E-mail: firstname.lastname@univ-lehavre.fr

Juan Luis Jiménez Laredo ICAR, University of Granada E-mail: juanlu@ugr.es

Juan Luis Jiménez would like to thank the Spanish project PID2020-115570GB-C22 (DemocratAI::UGR) funded by the Ministerio español de Economía y Competitividad.

1 Introduction

Understanding Self-Organized Criticality (SOC) is important because it is one of the signatures found in complex systems. SOC often explains how complex structures emerge and persist and why they exhibit similar scale-invariant properties. The underlying question is whether it is possible to formalize a mathematical system that describes these dynamics. Answering the question, in 1987, Per Bak, Chao Tank and Kurt Wiesenfeld defined the SOC concept using a cellular automaton modeling a sandpile [2] and showed the SOC was frequently showing up in our world [1].

In its simplest form, the sandpile model describes a system in which grains of sand are randomly thrown onto different cells of a cellular automaton, one grain at a time. When the number of grains in a cell exceeds a defined threshold, those grains in the cell are reassigned to the neighboring cells, triggering an avalanche that can produce minor or major disturbances. This system is opened at the borders of the cellular automaton so that grains disappear whenever they fall out of the automaton bounds.

The system dynamics present long periods of small avalanches punctuated by relatively short periods of catastrophic large avalanches. The system is said to display a SOC behaviour when the relation between the size and the frequency of the avalanches follows a power law, a behaviour that the sandpile model achieves in an emergent manner. The exponent of the power law is an indicator of the system behaviour and, in this paper, we show how this and other metrics can be altered by the neighborhood structure of the automaton cells.

Some previous works [6], [5] and [3] have already analysed the impact of the neighborhood structure on the dynamics of the model. In [3] in particular, authors analysed the evolution of the model when the grid is transformed into a small-world structure. This transformation is conducted by adding new links to randomly selected cells with a given probability. The higher the probability, the closer to a random network the system is. Since the number of connections in a cell can only increase with this method, to maintain a fair redistribution of grains between neighbors when an avalanche occurs, the critical threshold of the cell is adapted online to fit the number of neighbors. A consequence of those transformations is the impossibility for grains of sand to get out of the system since the automaton has no border anymore. Therefore, in order to preserve the opening of the system, the grains have a probability to disappear at each move during avalanches.

This paper proposes an approach for rewiring the neighborhood structure of the sandpile without altering any other component of the canonical model. As a result, the modified systems are as close as possible to the abelian model. The aim of this exploratory work is to determine the levers on which to play to control the avalanches within the model. A large panel of metrics is used for this purpose (history of the life of grains of sand, length and amplitude of the avalanches, etc.). The reminder of the paper is organized as follows: the experimental setup will firstly be exposed in the section 2 describing the rules and the implementation of the abelian sandpile, the random modifications of the neighborhood, and simulations parameters. Then, results will be presented and analysed in section 3. Finally, some conclusions and future lines of work will be presented in section 4.

2 Experimental setup

2.1 The abelian model

The sandpile model is typically represented using a grid lattice topology that uses a von Neumann neighborhood. For simulation purposes, we will model such structure using a graph in which nodes correspond to the cells of the grid and edges to the neighbors of a cell. By default, a cell has four neighbors except in the case that the cell is placed at the border, in which it will have three neighbors or at the corners with two neighbors.

The sandpile dynamics are governed by a set of rules. Every node of the graph has a critical threshold of four grains, which means that grains can accumulate on the cell until such critical value is met. When this threshold is reached or exceeded, the cell becomes unstable and the grains will topple to the neighboring cells. Since the threshold and the number of neighbors is four, one grain will be assigned to every neighboring cell. If the unstable cell is at the edge of the cellular automaton, grains will fall out of the system: the sandpile is an open system.

By following these simple rules, avalanches of different sizes may occur: as grains of sand are randomly thrown at the cellular automaton, some cells will eventually reach the critical threshold and grains will topple to the neighboring cells starting an avalanche. This may only have a local impact in the surrounding cells or start a chain reaction triggering a large avalanche. In the long term, the system dynamics evolves into a self-organized critical state in which the size and frequency of the avalanches follows a power law. When an avalanche occurs, the process of shedding grains is halted until the system recovers its stability (i.e. all cells are stable again). The duration and amplitude of the avalanches are therefore measured between two periods of stability of the system. This is how the first measurements were made in a real sand pile and later in the computer experiments conducted by Bak, Tang and Wiesenfeld.

Sandpile models implementing previous rules and mechanisms are said to be abelian since grains reassignments are commutative during topplings. That means the configuration of grains in the system at the end of an avalanche will be the same no matter the order in which grains are reassigned during the avalanche.

2.2 Random rewiring

As previously stated, the aim of this paper is to analyze the impact of random modifications to the neighborhood of the sandpile on the dynamics of the system. To that end, we propose an algorithm for rewiring the edges while keeping the indegree and outdegree of the nodes constant that uses the permutation principle of G. A. Croes [4]. The algorithm takes the graph and the number of edges that have to be rewired as an entry. This last parameter serves to gradually increase the degree of randomness in the graph that is initially a regular grid topology. Therefore, by increasing the number of edges rewired, we can start with a regular grid topology, go through a small-world topology [7] and end up with a completely random topology. As for the rest of the system, the exact same canonical rules are applied so that any change in the dynamics are necessarily due to the impact of the topology.

The algorithm will randomly pick two edges of the graph $\{u, v\}$ and $\{s, t\}$ defined by four nodes u, v, s, t. As exemplified in figure 1 a node in each edge is then swapped resulting in two new edges $\{u, t\}$ and $\{s, v\}$. This step is repeated $\frac{n}{2}$ times since the rewiring is achieved by pairs of edges, n being the number of edges to rewire. An edge can only be rewired once and the two edges selected must not have any node in common.

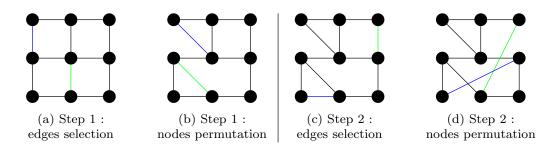


Fig. 1: Example of the rewiring of a lattice of size 3 with n = 4

2.3 Simulations parameters

In order to conduct an experimental analysis on the proposed algorithm, some decisions were taken and some parameters fixed to the following values. First of all, the simulated system is a grid lattice of size 32×32 . Simulations have a duration of 50,000 iterations, one iteration being a grain thrown followed by an avalanche if any cell becomes unstable as a result.

Experiments were conducted for rewiring rates going from 0% to 100% in steps of 10%. This allows to simulate the sandpile model on several graph structures going from the regular grid (no rewiring), through small-world graphs (for small rewiring coefficients), to random graph structures. The results of each rewiring coefficient are averaged over 25 simulations, each one being simulated with a different random seed.

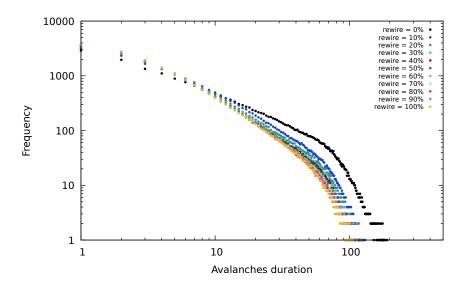


Fig. 2: Distribution of avalanches depending on their duration and frequency

3 Analysis of results

Through the different metrics used in this analysis, the results show that, in general, increasing the randomness of the topology increases the speed with which the grains that are deposited leave the system: avalanches become more effective in releasing the grains across the borders of the system. Nonetheless, the SOC signature remains intact, showing the same kind of power-law relationship between the frequency and duration of avalanches.

More in detail, the first metric employed measures the power-law relation between the duration and the frequency of avalanches as stated by Bak, Tang and Wiesenfeld. Figure 2 shows the results of simulations for the different rewiring coefficients. The shape of the plot shows that the power law is preserved throughout the process of rewiring with a decrease in the power law exponent for larger rewiring coefficients, which implies a decrease on the duration of large avalanches.

The second metric employed defines the number of grains released from the system with respect to the avalanche duration. Figure 3 depicts the different curves for the different rewiring coefficients under study. Once again, the figure shows the decrease of the avalanches duration as the rewiring coefficients increase. In addition to that, we can observe another phenomenon: as topologies are randomized, there is an increase on the number of grains released from the system. It can be argued that this change in the dynamics can be due to the decrease on the radius of the graph. As depicted in figure 4, the higher the rewiring coefficient, the smaller the radius. This translates into grains being able to find a shorter path to the borders in topologies with a smaller radius and, therefore, being released from the system more efficiently.

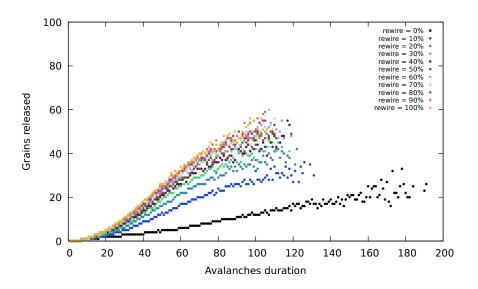


Fig. 3: Average number of grains released of the system by avalanches depending on their duration

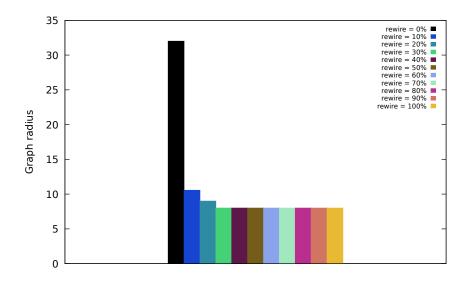


Fig. 4: Average radius of graphs for each rewiring coefficient

Finally, figure 5 shows the average grain density in the system normalized by the critical threshold. The results show that, independently of the rewiring coefficients, the dynamics of the system evolve to a constant density of roughly 50% of the threshold, acting as an attractor. In simple terms, the average density of the system stays at around 2 grains per cell. The fact that the rewiring does not have an impact on this metric, shows that the SOC signature of the system is preserved.

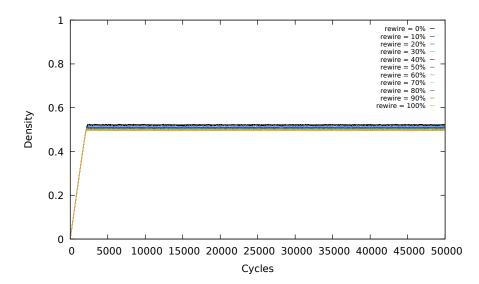


Fig. 5: Normalized density of grains in the system

4 Conclusions

The impact of introducing random modifications to the neighborhood of the abelian sandpile has been studied in this paper. The obtained results show that the SOC signature is kept despite the modifications of the graph structure. The rewiring process leads to a reduction of eccentricities of nodes which has several consequences. First, the path length from a cell to the border is, on average, reduced. This causes an increase in the number of grains that are released by the system during avalanches. Another consequence is that duration of avalanches is also reduced.

This study on the structure of the abelian sandpile aims at finding control levers to guide the self-organised criticality. The results give a first option: play on connections. Making the graph dynamic over some conditions on the state of the system might be a first approach. A second one could be to modify the transitivity of nodes making their critical threshold dynamic as well. Therefore, the progression of grains could be momentarily slowed down (increase of the threshold), or on the opposite accelerated (decrease of the threshold). However, for both options, the abelian aspect of the model would be lost since the neighbors to which reassign grains should be chosen.

The abelian sandpile is the simplest model that describes the SOC and because of its attractive properties, previous works have adapted this model to a sand sieve for addressing the distributed load balancing problem [6]. Our future works will focus on the detection and measure of the system overload as well as an online self-adaptation to it.

References

- 1. Bak, P.: How the nature works: the science of self-organized criticality. Copernicus (1996)
- Bak, P., Tang, C., Wiesenfeld, K.: Self-organized criticality. Physical Review A 38, 364–374 (1987)
- 3. Bhaumik, Η., Santra, S.B.: Critical properties of \mathbf{a} dissipative sandpile model onsmall-world networks. Physical Review Ε 88, 062,817 (2013).DOI 10.1103/PhysRevE.88.062817. URL https://link.aps.org/doi/10.1103/PhysRevE.88.062817
- Croes, G.A.: A method for solving traveling-salesman problems. Operations Research 6(6), 791–812 (1958). URL http://www.jstor.org/stable/167074
- de Arcangelis, L., Herrmann, H.: Self-organized criticality on small world networks. Physica A: Statistical Mechanics and its Applications **308**(1), 545–549 (2002). DOI https://doi.org/10.1016/S0378-4371(02)00549-6. URL https://www.sciencedirect.com/science/article/pii/S0378437102005496
- Jiménez Laredo, J.L., Bouvry, P., Guinand, F., Dorronsoro, B., Carlos, F.: The sandpile scheduler. Cluster Computing 17(2), 191–204 (2014). DOI 10.1007/s10586-013-0328-x. URL https://hal.archives-ouvertes.fr/hal-02501380
- Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (1998). DOI 10.1038/30918



Asymptotic Dynamic Graph Order Evolution Analysis

Vincent Bridonneau · Frédéric Guinand · Yoann Pigné

Abstract In this work, we investigate the analysis of generators for dynamic graphs, which are defined as graphs whose topology changes over time. We focus on generated graphs whose order (number of nodes) varies over time. We introduce a novel concept, called "sustainability," to qualify the long-term evolution of dynamic graphs. A dynamic graph is considered sustainable if its evolution does not result in a static, empty, or periodic graph. To illustrate how the analysis can be conducted, a parameterized generator, named D3G3 (Degree-Driven Dynamic Geometric Graphs Generator), that generates dynamic graph instances from an initial geometric graph, has been introduced in a recent work. The evolution of instances obtained from this model is driven by two rules that operate on the vertices based solely on their degree. In this work, we focus on particular sets of parameter leading the generator to produce graphs whose order increase or decrease exponentially or remain almost constant.

Keywords Dynamic Graphs \cdot Graph Generation \cdot Graph Properties \cdot Evolutionary models

1 Introduction

This work provides an analytical study for generated graphs obtained in the context of dynamic graph generators. A dynamic graph generator can be defined as a process that takes data as input (for instance, a seed graph) and produces a sequence of static snapshot graphs. More precisely, a generator will produce a snapshot graph G_{t+1} at a step t+1 considering t generated snapshot graphs $\{G_1, \ldots, G_t\}$ and the seed graph G_0 . The product of a dynamic graph generator is therefore a flow of static graphs ordered by a timestamp.

Corresponding author: Vincent Bridonneau

E-mail: vincent.bridonneau@univ-lehavre.fr

In that context, the present work focuses on the analysis of the evolution of the graph order (number of nodes) of dynamic graphs obtained by a specific generator. Many works have been dedicated to the generation of graphs. Most of them have been designed for a specific purpose [1-8] and the evolution of the order of generated graphs is known at each step. Most of the time, the order increases at each time step (growing networks) [1-5] or remains the same [6-8]. The purpose of this work is to address the question in the other way. Here the mechanism generating graphs is assumed to be known. Then, the problem is to find properties generated graphs satisfy. As a first study, this work deal with the evolution of graph order when the generator relies on rules enabling both the addition and the deletion of vertices. Example of real instances of graph of which the set of vertices is not constant may be find when looking at teams in team sports such as rugby or soccer. When seen as graphs, teams may be represented as dynamic graphs which order remains nearly constant through time. In such sports and during a single game, players may be replaced by others so that the set of vertices is not steady but its size does not change.

According to the generative mechanism, it may happen, after some time steps, that generated graphs become empty forever or periodic. A notion called "sustainability" is introduced to highlight this phenomenon. If there exists a time step t such that a generated graph becomes empty or periodic from that moment, then this graph is said to be "non sustainable". Otherwise, if no such time step exists, the graph is said "sustainable".

To better understand the purpose of this notion, a new version of the Degree Driven Dynamic Geometric Graph Generator (D3G3), introduced in [9], is considered. Graphs produced by D3G3 are geometric graphs. A geometric graph is defined by an euclidean space and a threshold d. For this study, without loss of generality we consider a 2D-unit-torus (i.e., a square $[0; 1]^2$ where the two opposite sides are connected). Each vertex is characterized by a set of coordinates, such that given two vertices u and v it is possible to compute their euclidean distance: dist(u, v). Given V the set of vertices, the set of edges E is defined in the following way: $E = \{(u, v) \in V^2 \mid dist(u, v) \leq d\}$.

Graphs generated by D3G3 are produced thanks to an evolution process. This mechanism is parameterized by an initial graph (the seed graph) and by two transition rules driving the evolution of the graph between two consecutive time steps. Apart from a random generator, no external decision or additional information is used by this mechanism. Rules are based on node degrees only and rely on a random generator for positioning new nodes in the 2D euclidean space. This leads to the name of the generator: *Degree-Driven Dynamic Geometric Graphs Generator* or D3G3.

Definition 1 Degree Driven Dynamic Geometric Graph Generator

An instance of D3G3 is defined by an initial graph, a set of parameters and two rules:

- $-G_0 \neq (\emptyset, \emptyset)$ the seed graph,
- parameters:

 - $(-d \in]0, \frac{\sqrt{2}}{2}[$ $(-S_s \text{ a set of non-negative integers})$
 - $-S_c$ a set of non-negative integers
- rules applied on G_t leading to G_{t+1} :
 - if $v \in V_t$, then $v \in V_{t+1}$ iff $\deg(v) \in S_S$ (conservation rule)
 - if $v \in V_t$ and if deg $(v) \in S_C$ then add a new vertex to V_{t+1} with a random position in the unit-torus (creation rule)

The order of the graph at each step is not set by any external process or as a parameter of the generator but rather emerges from the application of the rules on consecutive snapshot graphs. And the main issue addressed is to determine if, for a given parameter set, the generated graphs are sustainable or not.

For this new model, position of conserved nodes considered in D3G3 are changed so that their new position is independent from one step to the next one. The generator is parameterized by a set of non-negative integers S. The rule driving the evolution of the generator is the following: if a node at a given step t has its degree in S then it is conserved and it is at the origin of a new node at step t + 1. More precisely, considered values of set S are such that $S = \{sk + r \mid k \in \mathbb{N} + r \in A\}$, for a fixed positive integer s and a set $A \subset [0,s]$. For this very specific settings, we show that generated graphs have one of the three following behavior when their order is big enough: either their order increases exponentially, either it decreases exponentially or it is roughly constant. For each case, we provide criterion about sustainability. We show that proving whether a generated graph is sustainable for this three cases is not obvious and need to consider both small and big snapshot graphs.

2 Model and Concepts

The model we will discuss here is a variation of the model defined in the introduction and it is defined as follow:

Definition 2 The redistributed model

An instance of the Redistributed Degree Driven Dynamic Geometric Graph Generator (RD3G3) is defined by an initial graph, a set of integer and a rule:

 $-G_0 \neq (\emptyset, \emptyset)$ the initial graph,

- $-d \in (0, \frac{\sqrt{2}}{2})$ represents a connection threshold, S a set of non-negative integers
- rules applied on G_t leading to G_{t+1} :
 - for all $v \in V_t$, $v \in V_{t+1}$ (conserved node) and a new vertex is added to V_{t+1} with a random position in the unit-torus if and only if $\deg(v) \in S$ (duplication rule)

The term "redistributed" here comes from the new treatment of conserved nodes from D3G3. In the D3G3 model, if at a step t a node is conserved at step t+1 then its position does not change. Unlike this original version, at every time step t, the conserved nodes at step t+1 are uniformly redistributed over the torus so that new graphs are random geometric graphs whose order depends only on the number of conserved nodes. One can then find an estimation function $f_{S,d}$ of graph order at step t+1 knowing graph order at step t:

$$\forall n, f_{S,d}(n) = 2n \left(\sum_{k \in S} \binom{n-1}{k} p^k (1-p)^{n-1-k} \right) \tag{1}$$

where p is the probability for two nodes to be connected (for $d \in (0, \frac{1}{2})$, $p = \pi d^2$). Here n refers the order of the graph at step t. For the rest of the article, considered values of S will be restrained. These values are specified in the following section.

3 Asymptotic Graph Order Evolution

This section aims at presenting our work on the RD3G3 model for restrained values on the parameter S. Indeed, this work focuses on sets of the form $S = \{sk + r \mid r \in A, k \in \mathbb{N}\}$ for fixed $s \in \mathbb{Z}^+$ and $A \subset [0, s - 1]$. The main result of this paper is an equivalent of $f_{S,d}(n)$ for large values of n. This equivalent will also help understanding the behavior of generated graphs with high orders. It will also provide an answer to whether generated graphs are sustainable or not.

3.1 Intermediate Result

The result of this work relies on properties roots of unity satisfies. As a reminder, a *n*th root of unity for any positive integer n is defined as follow:

Definition 3 Let *n* be a positive integer. Then a *n*th root of unity is a complex number ω such that $\omega^n = 1$.

Such numbers satisfy several properties one may find in [10] at section 2.5. Most important ones for this article are gathered in the following lemma:

Lemma 1 Let n be a positive integer. Then the following holds:

- $\omega_n = \exp\left(\frac{2i\pi}{n}\right) \text{ is a nth root of unity;}$ $a complex number <math>\omega$ is a nth roots of unity if and only if there exist k such that $\omega = \omega_n^k$;
- if a complex number ω is a nth root of unity, then its modulus satisfies $|\omega| = 1;$
- sum of *j*th powers of *n*th root of unity, for any non-negative integer *j*, are such that:

$$\sum_{k=0}^{n-1} \left(\omega_n^k\right)^j = \begin{cases} n & \text{If } n \text{ divides } j \\ 0 & \text{Otherwise} \end{cases}$$

Such numbers are useful to prove the following result concerning infinite sums:

Lemma 2 Let $s \in \mathbb{Z}^+$, $n \in \mathbb{N}$ and $x \in \mathbb{R}$. Let $r \in [0, s - 1]$, then we get the following equality:

$$\sum_{k=0}^{+\infty} \binom{n}{sk+r} x^{sk+r} = \frac{1}{s} \sum_{j=0}^{s-1} \omega_s^{-jr} (1+\omega_s^j x)^n$$

where $\omega_s = \exp\left(\frac{2i\pi}{s}\right)$ is an sth root of unity.

Proof: Let $s \in \mathbb{Z}^+$, $n \in \mathbb{N}$ and $x \in \mathbb{R}$. Let $r \in [0, s - 1]$. Let $\omega_s = \exp(\frac{2i\pi}{s})$. The first thing to notice is that the infinite sum on the left side of the equality converges. For any values of k such that sk + r > n, the binomial $\binom{n}{sk+r} = 0$. Thus, the infinite sum contains only finitely many non-zero terms. Then, it is sufficient to notice that, according to properties roots of unity satisfy, the following holds:

$$\forall m, \frac{1}{s} \binom{n}{m} x^m \sum_{j=0}^{s-1} \omega_s^{j(m-r)} = \begin{cases} \binom{n}{m} x^m & \text{If there exists } k \text{ such that } m = sk+r \\ 0 & \text{Otherwise} \end{cases}$$

From this the following equations hold:

$$\sum_{k=0}^{+\infty} {n \choose sk+r} x^{sk+r} = \frac{1}{s} \sum_{m=0}^{+\infty} {n \choose m} x^m \sum_{j=0}^{s-1} (\omega_s^j)^{m-r}$$
$$= \frac{1}{s} \sum_{m=0}^{+\infty} \sum_{j=0}^{s-1} \omega_s^{-jr} {n \choose m} (\omega_s^j x)^m$$
$$= \frac{1}{s} \sum_{j=0}^{s-1} \omega_s^{-jr} \left(\sum_{m=0}^{+\infty} {n \choose m} (\omega_s^j x)^m \right)$$
$$= \frac{1}{s} \sum_{j=0}^{s-1} \omega_s^{-jr} \left(\sum_{m=0}^{n} {n \choose m} (\omega_s^j x)^m \right)$$
$$= \frac{1}{s} \sum_{j=0}^{s-1} \omega_s^{-jr} (1 + \omega_s^j x)^n$$

This ends the proof.

1 C

This lemma on roots of unity helps getting another expression of the function $f_{S,d}$:

Theorem 1 Let $s \in \mathbb{Z}^+$, $n \in \mathbb{N}$ and $A \subset [0, s - 1]$. Let S as defined above, then:

$$f_{S,d}(n) = \frac{2}{s} n \left(\sum_{r \in A} \left(\sum_{j=0}^{s-1} \omega_s^{-jr} \left(1 - p + \omega_s^j p \right)^{n-1} \right) \right)$$

Proof: Let $s \in \mathbb{Z}^+$, $n \in \mathbb{N}$ and $A \subset [0, s - 1]$. Rewriting $f_{S,d}(n)$ lead to the following expression

$$f_{S,d}(n) = 2n \sum_{r \in A} \left(\sum_{k=0}^{+\infty} \binom{n-1}{sk+r} p^{sk+r} (1-p)^{n-1-(sk+r)} \right)$$
$$= 2n(1-p)^{n-1} \sum_{r \in A} \left(\sum_{k=0}^{+\infty} \binom{n-1}{sk+r} \left(\frac{p}{1-p} \right)^{sk+r} \right)$$

Thus, applying result of lemma 2 provides:

$$f_{S,d}(n) = \frac{2}{s}n(1-p)^{n-1}\sum_{r\in A} \left(\sum_{j=0}^{s-1} \omega_s^{-jr} \left(1+\omega_s^{j}\frac{p}{1-p}\right)^{n-1}\right)$$
$$= \frac{2}{s}n\sum_{r\in A} \left(\sum_{j=0}^{s-1} \omega_s^{-jr} \left(1-p+\omega_s^{j}p\right)^{n-1}\right)$$

This ends the proof.

This theorem provides an exact formulae for the estimation function $f_{S,d}$. It is important to notice that this re-written formulae involves only finite sums. It is therefore easier to deal with its analysis which is the purpose of the following sub-section.

3.2 The Equivalent and First Interpretation

From result obtained in the last subsection, it is possible to get an equivalent for great values of n for $f_{S,d}$:

Theorem 2 Let $s \in \mathbb{Z}^+$, $n \in \mathbb{N}$ and $A \subset [0, s - 1]$. Let S as defined above, then for large values of n:

$$f_{S,d}(n) \sim \frac{2|A|}{s}n$$

Proof: This comes from theorem 1 and from properties on complex numbers. More precisely, for each value of $r \in A$ there is exactly one value of $j \in [0, s-1]$ such that $\omega_s^{-jr}(1-p+\omega_s^jp)=1$ (for j=0). For all other values of j, $\omega_s^{-jr}(1-p+\omega_s^jp)=1$ have a modulus lower than 1. The rest is computation of limits.

This result has an interpretation for graphs generated with the model. Indeed, for a given snapshot graph at step t of order n_t , the application of the rule will produce a graph with an expected order $\left(\frac{2|A|}{s}\right)n_t$ at step t+1. The next section goes further in the analysis of these three cases. It also highlights the difference between graph order evolution of big and small graphs: interpretation depends on different features of the parameter set S.

4 Generated Graphs Interpretation

This section aims at going further in the interpretation of previously stated results. More precisely, this section highlight three different asymptotic graph order evolution that occur from stated equivalent in 3.2. Moreover, interpretation for small graph order is given. This will help knowing whether generated graphs are likely to remain steady or not depending on the smallest values of the parameter S.

4.1 General Observations

Before dealing with each case, it is important to understand the meaning of theorem 2. This theorem states that for any given generated graph having n_t nodes at a step t and assuming n_t is big enough, then, at the next step, n_{t+1} is expected to be close to $\left(\frac{2|A|}{s}\right)n_t$. Therefore, starting with a seed graph of order N big enough would lead, after t steps, to a graph of order

$$n_t \simeq \left(\frac{2|A|}{s}\right)^t N$$

This is why graph order is said to grow exponentially. From this, three cases have to be observed:

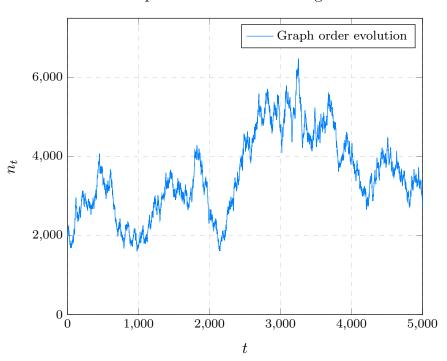
- The first case is $\frac{2|A|}{s} < 1$. This means generated graphs order are likely to decrease when their order is large.
- The second case is $\frac{2|A|}{s} > 1$. This means generated graphs order are likely to increase when their order is large.
- Finally, the third case is $\frac{2|A|}{s} = 1$.

4.2 Exponential Increasing

The first studied case is when s and A both satisfy $\frac{2|A|}{s} > 1$. For this case, as S is not bounded the order of generated graphs is likely to tend to infinity. Generated graphs are therefore likely to be sustainable.

4.3 Exponential Decreasing

The second studied case is when s and A both satisfy $\frac{2|A|}{s} < 1$. For this case, graph order of generated graphs is likely to decrease exponentially. However, it is not enough to conclude on the sustainability of generated graphs. Indeed, when graphs become small enough (close to 0), one may consider to take into account the smallest values of set S. This last case is further studied in section 4.5.



Graph order evolution through time.

Fig. 1 Simulation performed considering s = 4, A = [0, 1] and d = 0.05. The number of steps is 5000 and the initial seed graph is a random geometric graph of order 2000.

4.4 Quasi Constant Evolution

Two points must be noticed for the last case. First, this case happens if and only if s is even. Indeed, if s is odd, whatever the set A one may choose, the numerator will be even. Second, for a given time step t, application of the rule on a graph which order is n_t will produce a graph which order is expected to be $n_{t+1} = n_t$. It is however necessary to go further as $f_{S,d}$ only provides an expectation. The graph order will indeed change a little. An estimation for this change can be obtained with the standard deviation of a binomial law. Despite all these consideration simulations have been performed. They all show that graph order changes through time with little variations. These simulations are represented in figure 1. It is worth noticing graph order is not constant all along the simulation, but rather increasing or decreasing a little bit every time.

A further step to this study is to take into account the standard deviation $\sigma_{S,d}$ associated with graph order evolution. For a given $n_t \in \mathbb{Z}^+$ order of a graph at step $t, \sigma_{S,d}(n_t)$ tells how far away from n_{t+1} is $f_{S,d}(n_t)$, which in this case is roughly n_t . Thus, applying Chebishev's inequality [11], for instance, states that for any given real number k > 0:

$$\Pr[n_{t+1} \notin [n_t - k\sigma_{S,d}(n_t), n_t + k\sigma_{S,d}(n_t)]] \leqslant \frac{1}{k^2}$$

The computation of $\sigma_{S,d}(n)$ for large enough values of n lead to an equivalent which is the purpose of the following theorem:

Theorem 3 Let $s \in \mathbb{Z}^+$, $n \in \mathbb{N}$ and $A \subset [0, s - 1]$. Let S as defined above, then for large values of n:

$$\sigma_{S,d}(n) \sim \frac{1}{s} \sqrt{n|A|(s-|A|)}$$

Proof: The proof of this theorem relies on the same argument as for theorem 2 and on the definition of the standard deviation of binomial distributions. \Box

This theorem states that the standard deviation $\sigma_{S,d}(n)$ is proportional to \sqrt{n} for large values of n. This provides better information about the possible values n_{t+1} may have depending on n_t . Indeed, now above stated inequality can by rewritten as follow:

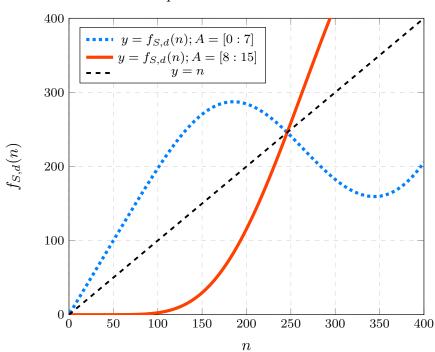
$$\Pr[n_{t+1} \notin [n_t - \frac{k}{2}\sqrt{n_t}, n_t + \frac{k}{2}\sqrt{n_t}]] \leqslant \frac{1}{k^2}$$

4.5 Sustainability of Small Generated Graphs

The question of whether a small generated graph is sustainable or not does not depend on the asymptotic variation of the graph order. The answer to this question relies on the smallest values that the parameter S contains.

Indeed, on the one hand, whatever the values of s one may consider, if $A \subset [k, s - 1)$ for any $k \ge \frac{s}{2}$, then graphs whose order does not exceed k do not have nodes with a degree greater than or equal to k. Therefore such graphs become empty because they do not have any node satisfying the duplication rule. A further step is to consider small values of parameter S. For instance, for d = 0.05, s = 16 and A = [8, 15], the full-lined curve of $f_{S,d}$ represented in figure 2 shows that for small values of n, $f_{S,d}(n) < n$. This means that graph order of small graph is expected to decrease between two consecutive steps and graphs are likely to become empty. Therefore generated graphs, for this configuration are likely not be sustainable.

On the other hand, whatever the values of s one may consider, if $A \subset [0, k+1]$ for any $k < \frac{s}{2}$, then graphs whose order does not exceed k have nodes with a degree lower than or equal to k. Therefore such graphs do not become empty because they have all their nodes satisfying the duplication rule. As for the first case, a further step is to consider small values of parameter S. For instance, for d = 0.05, s = 16 and A = [0,7], the dotted curve of $f_{S,d}$ represented in figure 2 shows that for small values of n, $f_{S,d}(n) \ge n$. This means that graph order of small graph is expected to increase between two consecutive steps. Therefore, as soon as graph order does not exceed a certain quantity, generated graphs are likely to conserve few nodes and therefore are likely to be sustainable.



Graphical representation of the function $f_{S,d}$ Considered parameters are s = 16 and d = 0.1.

Fig. 2 Theoretical graphical representation of $f_{S,d}$ for value of n from 0 to 400. The blue curve correspond to A = [0,7] and the red one correspond to A = [8,15]

Conclusion

This work is a new step in the study of dynamic graph generators. It studies a model based on the Degree Driven Dynamic Geometric Graph Generators (D3G3). It is a model parameterized by two quantities: a real positive number d and a set of non-negative integer S. In order to analyze this generator, a new property is introduced. This property is called "sustainability". A generated graphs is said to be sustainable if and only if it does not become empty nor periodic. This work aims at understanding the behavior of generated graphs for particular values of the set S. In these configurations, order of generated graphs is shown to have three different asymptotic evolution. Either it is exponentially increasing, exponentially decreasing or quasi constant. For the first case, generated graphs are sustainable with high probability. For the decreasing case, sustainability must be considered with respect to sustainability of small generated graphs. Indeed, graph order decreasing does not mean graphs become empty but small and, for that setting, it is important to consider smallest values of parameter S.

References

- 1. A.L. Barabási, R. Albert, Science **286**(5439), 509 (1999). DOI 10.1126/science.286.5439.509. URL https://doi.org/10.1126/science.286.5439.509
- D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, M. Boguñá, Physical Review E 82(3), 036106 (2010). DOI 10.1103/PhysRevE.82.036106. URL https://link.aps.org/doi/10.1103/PhysRevE.82.036106. Publisher: American Physical Society
- K. Zuev, M. Boguñá, G. Bianconi, D. Krioukov, Scientific Reports 5(1), 9421 (2015). DOI 10.1038/srep09421. URL http://www.nature.com/articles/srep09421
- F. Papadopoulos, M. Kitsak, M. Serrano, M. Boguñá, D. Krioukov, Nature 489(7417), 537 (2012). DOI 10.1038/nature11459. URL https://doi.org/10.1038/nature11459
- A. Muscoloni, C.V. Cannistraci, New Journal of Physics 20(5), 052002 (2018). DOI 10.1088/1367-2630/aac06f. URL https://iopscience.iop.org/article/10.1088/1367-2630/aac06f
- A.E.F. Clementi, C. Macci, A. Monti, F. Pasquale, R. Silvestri, SIAM Journal on Discrete Mathematics 24(4), 1694 (2010). DOI 10.1137/090756053
- 7. P. Erdős, A. Rényi,
- D.J. Watts, S.H. Strogatz, Nature **393**(6684), 440 (1998). DOI 10.1038/30918. URL https://www.nature.com/articles/30918. Number: 6684 Publisher: Nature Publishing Group
- V. Bridonneau, F. Guinand, Y. Pigné, Dynamic Graphs Generators Analysis : an Illustrative Case Study. Tech. rep., LITIS, Le Havre Normandie University (2023). URL https://hal.science/hal-03910386
- C.R. Hadlock, Field Theory and Its Classical Problems (Cambridge University Press, 2000). Google-Books-ID: 5s1p0CyafnEC
- 11. W. Feller, An Introduction to Probability Theory and Its Applications, Volume 2 (John Wiley & Sons, 1991). Google-Books-ID: rxadEAAAQBAJ



Who to Watch When? Strategic Observation in the Inverse Ising Problem

Zhongqi Cai · Enrico Gerding · Markus Brede

Abstract In this paper, we investigate the problem of inferring the network coupling strengths from partially observed time series data in an Ising model on scale-free networks. By assuming that only a certain fraction of observations for spin states are available, we study how an observer, who wants to maximise the accuracy of the network inference, should distribute a limited number of observations. Along with the benchmark case of randomly-chosen hidden nodes, we propose degree-dependent heuristics for observation allocations. We observe two regimes for the best observation strategies based on varying amounts of missing data. If only a small proportion of data cannot be observed, then one should focus on the observation of the states of periphery nodes. Otherwise, if a large number of states cannot be observed, allocating more observations to the high-degree nodes is preferable.

Keywords Network inference · Inverse Ising model · Complex networks

1 Introduction

The inverse statistical problem [1], which aims at inferring microscopic parameters of the underlying statistical model from observations, has gained significant attention in recent years [2]. One of the most typical and canonical settings for the inverse statistical problem is the inverse Ising model [3]. Specifically, for the inverse Ising model, one aims at reconstructing parameters of the Ising model such as coupling strengths between spins from data like spins' states or magnetization.

Due to the limitation of current data sampling techniques, it is common that the reconstruction can only be carried out on partial observations of complex systems. For instance, it is challenging to observe every neuron and track

Zhongqi Cai, Enrico Gerding, Markus Brede

School of Electronics and Computer Science, University of Southampton, UK

 $E\text{-mail: } Zhongqi. Cai@soton.ac.uk, \ eg@ecs.soton.ac.uk, \ Markus. Brede@soton.ac.uk \\$

complete spiking activities in the brain [4]. Therefore, a large part of the most recent literature on the inverse Ising problem focuses on inferring the modelling or structural parameters of the Ising model in the presence of missing data [5-11]. Of these, the works [5-9] assume the existence of hidden nodes whose states can not be observed for the whole Ising dynamics. Under this setting, the inference problem is addressed by utilizing the expectation-maximization (EM) algorithm. In addition to hidden nodes, the works of [10,11] consider a more practical scenario that even for the visible nodes, their states are not always trackable throughout the whole experiment. These two works investigate the inverse Ising problem with partially masked data by utilizing the meanfield approximation [10] and logistic regression [11]. Although much effort has been devoted to solving the inverse Ising problem with missing data from the algorithmic perspective, all of the above studies consider the dataset as given. In contrast, here, we study the problem of experimental design: How should a limited number of observations be allocated to generate data that allow for the most accurate possible inference? The latter question is of practical importance as it is always preferable to obtain lower inference errors using fewer observation resources, especially when experimental observations are costly.

Here, we propose heuristics for solving the inverse Ising problem on scalefree networks with missing data from the perspective of strategically allocating limited observation resources to reduce inference errors. We have made the following contributions. First, we propose a new framework for solving the inverse Ising model with regards to customizing the observed parts of a dataset for better inference. Second, our work sheds light on the role that network topology plays in network inference. Our results demonstrate that, depending on how many observations can be taken, there are two regimes for optimal allocations of observations. For a small proportion of missing data (i.e. large number of observations), it is better to predominantly focus the observation resources on periphery nodes. Otherwise, if only few observations are possible, it is preferable to concentrate observations on hub nodes. Additionally, we find that the accuracy of inference for in-linking and out-linking weights varies with node degree, with low-degree nodes allowing for more accuracy in inferring inlinking weights and high-degree nodes allowing for more accurate inference of out-weights.

2 Problem Formalization

Consider the stochastic dynamics of an Ising system with N binary spins. Each spin has two possible states at time t, denoted as $\sigma_i(t) = \pm 1$. Following the discrete-time and synchronously updated Glauber algorithm [12], the system evolves according to the conditional probability:

$$P(\sigma_i(t+1) \mid \{\sigma(t)\}_{i=1}^N) = \frac{e^{\sigma_i(t+1)H_i(\sigma(t))}}{2\cosh(H_i(\sigma(t)))},$$
(1)

for $i = 1, \dots, N, t \ge 1$ and $H_i(\sigma(t)) = \sum_j w_{ij}\sigma_j(t)$ where w_{ij} represents the coupling strength from spin j to spin i. Following the most commonly-used benchmark

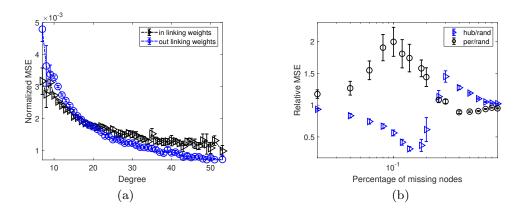


Fig. 1 (a) Dependence of the normalized mean square error (MSE) on nodes' degrees. The normalized MSE for in-linking weights is calculated as $k_i^{-1}(\sum_j \hat{w}_{ij} - w_{ij})^2$ for node *i* with degree k_i , and $k_i^{-1}(\sum_i \hat{w}_{ij} - w_{ij})^2$ for out-linking weights. (b) Dependence of relative MSE on percentage of unobserved nodes. The blue triangles represent the relative MSE between not observing the top x% of the highest degree nodes and not observing x% nodes chosen at random. The black circles are for the relative MSE between not observing x% of the lowest degree nodes and x% randomly-chosen nodes. The inference is performed based on observation length N^2 and repeated for 50 realizations.

in stochastic causality inference, here we use the Sherrington-Kirkpatrick (SK) model [13] to generate linking weights w_{ij} on top of scale-free networks. In more detail, all experiments in this work are performed on Barabási–Albert networks with network size N = 100 and average degree $\langle k \rangle = 20$, and the linking weights of existing connections are drawn from a normal distribution with mean 0 and standard deviation $4/\sqrt{N}$.

We are interested in designing observation schemes with the aim of reconstructing the coupling strengths $\{w_{ij}\}_{i,j=1}^N$ from partial observations with lower inference errors. More specifically, we consider the following context of two stages: 1) Assume only a fraction ρ of data points can be unmasked in a binary time-series dataset $\{\sigma(t)\}_{t=1}^{L}$ of length $L = N^2$. 2) Strategically choose a number of $LN\rho$ data points to be observed with the aim of maximally reducing the mean square inference error of w_{ij} , denoted as $N^{-1}\sum_{i,j}(\hat{w}_{ij}-w_{ij})^2$. Here, \hat{w}_{ij} represents the estimation for the true value w_{ij} obtained by applying the EM algorithm described in work [11]. As it is computationally demanding to perform direct optimization based on the iterative EM algorithm, here we focus on designing heuristics to select observed data points for better inference performance. Inspired by the works of network inference with hidden nodes [5,9], we propose a degree-dependent observation scheme. In more detail, we rank nodes by their degrees. Depending on the number of missing data, we either make the corresponding amounts of highest-degree nodes or lowest-degree nodes to be always un-observable throughout the time length L.

3 Results

In the following, we start by comparing the inference errors for nodes with different degrees. We then proceed by comparing the inference performance of the degree-dependent heuristics with the benchmark case of randomly choosing hidden nodes.

In Figure 1 (a), we show the dependence of inference errors of in(out)linking weights on nodes' degrees. For hub nodes, as they have strong impacts on network dynamics, it is more accurate to infer their out-linking weights compared to the in-linking weights. In contrast, for the periphery nodes, one can infer the in-linking weights with higher accuracy. Moreover, in Figure 1 (a), we observe that nodes are easier to predict the larger their degrees. To proceed, in Figure 1 (b), we compare the strategies of not observing the highest-degree nodes or lowest-degree nodes with the benchmark of not observing randomlychosen nodes. We clearly see two regimes in Figure 1 (b): if a small proportion of data is missing, then observing low-degree nodes will lead to high inference accuracy. However, the opposite holds if more data is missing. Inspired by the improvements of inference accuracy made by the degree-dependent heuristics, an interesting direction for future work is to look into details of the influence of spatial and temporal correlations of missing data on network inference.

References

- Arridge, S., Maass, P., Öktem, O., Schönlieb, C.-B.: Solving inverse problems using data-driven models. Acta Numer. 28, 1–174 (2019)
- Nguyen, H.C., Zecchina, R., Berg, J.: Inverse statistical problems: from the inverse ising problem to data science. Adv. Phys. 66(3), 197–261 (2017)
- Aurell, E., Ekeberg, M.: Inverse ising inference using all the data. Phys. Rev. Lett. 108(9), 090201 (2012)
- Soudry, D., Keshri, S., Stinson, P., Oh, M.-h., Iyengar, G., Paninski, L.: Efficient 'shotgun' inference of neural connectivity from highly sub-sampled activity data. PLoS Comput. Biol. 11(10), 1004464 (2015)
- 5. Dunn, B., Roudi, Y.: Learning and inference in a nonequilibrium ising model with hidden nodes. Phys. Rev. E 87(2), 022127 (2013)
- Tyrcha, J., Hertz, J.: Network inference with hidden units. Math Biosci Eng 11(1), 149–165 (2014)
- Bachschmid-Romano, L., Opper, M.: Inferring hidden states in a random kinetic ising model: replica analysis. J. Stat. Mech. Theory Exp. 2014(6), 06013 (2014)
- Battistin, C., Hertz, J., Tyrcha, J., Roudi, Y.: Belief propagation and replicas for inference and learning in a kinetic ising model with hidden spins. J. Stat. Mech. Theory Exp. 2015(5), 05021 (2015)
- Hoang, D.-T., Jo, J., Periwal, V.: Data-driven inference of hidden nodes in networks. Phys. Rev. E 99(4), 042114 (2019)
- 10. Campajola, C., Lillo, F., Tantari, D.: Inference of the kinetic ising model with heterogeneous missing data. Phys. Rev. E **99**(6), 062138 (2019)
- Lee, S., Periwal, V., Jo, J.: Inference of stochastic time series with missing data. Phys. Rev. E 104(2), 024119 (2021)
- Glauber, R.J.: Time-dependent statistics of the ising model. J. Math. Phys. 4(2), 294– 307 (1963)
- Sherrington, D., Kirkpatrick, S.: Solvable model of a spin-glass. Phys. Rev. Lett. 35(26), 1792 (1975)



The architecture of Multifunctional Ecological Networks

Mar Cuevas-Blanco · Sandra Hervías-Parejo · Victor M. Eguiluz · Lucas Lacasa · Isabel Donoso · Anna Traveset

Abstract We propose a novel approach to assess the functional importance of species in an ecosystem by considering their multiple ecological. The study is based on observations of 16 plant species and 675 animal/fungus species, interacting across six ecological functions. We formalize the relational dataset as a Resource-Consumer-Function tensor. By integrating out the consumer index, we construct a Multifunctional Ecological Network, which shows a stylized nested structure suggesting that certain functions and plant species can be classified as "generalists" or "specialists". We project MFEN into the function and plant class to quantify the heterogeneous roles and impacts plant species and ecological functions.

Mar Cuevas-Blanco IFISC (UIB-CSIC), Palma de Mallorca, Spain. E-mail: marcuevas@ifisc.uib-csic.es

Sandra Hervías-Parejo IMEDEA (UIB-CSIC), Esporles Illes Balears, Spain. E-mail: shervias@imedea.uib-csic.es

Victor M. Eguiluz IFISC (UIB-CSIC), Palma de Mallorca, Spain. E-mail: victor@ifisc.uib-csic.es

Lucas Lacasa IFISC (UIB-CSIC), Palma de Mallorca, Spain. E-mail: lucas@ifisc.uib-csic.es

Isabel Donoso IMEDEA (UIB-CSIC), Esporles Illes Balears, Spain. E-mail: isa.donoso@imedea.uib-csic.es

Anna Traveset IMEDEA (UIB-CSIC), Esporles Illes Balears, Spain. E-mail: atraveset@imedea.uib-csic.es

Lucas Lacasa IFISC (UIB-CSIC), Palma de Mallorca, Spain. E-mail: lucas@ifisc.uib-csic.es

Keywords Keystone Species \cdot Keystone Function \cdot Multilayer Networks \cdot Resilience \cdot Ecology.

1 Introduction

The identification and protection of keystone species are essential to maintain ecosystem stability and resilience [1]. To comprehensively assess species' functional importance, it is necessary to consider their multiple ecological roles e.g. multilayer networks [2]. To our knowledge only one study has empirically estimated the weight of edges across different ecological layers by quantifying the role of the same individual in two ecological processes [3].

2 Results

Our approach follows the consumer-resource paradigm, where plants are seen as "resources", and "consumers" encapsulate different types of animals or fungi. The precise span of dimensions is here driven by the extent of our data, based on observations recently collected in the islet Na Redona, which includes direct observation of 16 plant species, 675 animal/fungus species, interacting across six different ecological functions. Incorporating the functional dimension, the complete relational dataset is thus formalized in terms of a rank-3 tensor that we call the Resource-Consumer-Function tensor (RCF).

We interpret the architecture of this tensor as a weighted, multipartite, multilayer network, Figure 1. The network displays two types of nodes: resources (plant species) and consumers (animals and fungi), with interactions (links) taking place between groups but no direct intragroup links. Each layer of the network represents a specific function and the strength of each interaction is represented by a link weight. Consumers (animals, fungi) are often centered around a single plant species and thus form clusters. Interestingly, cross-cluster links are also present, thereby the ecosystem is entangled.

We first quantify the relationship between the resources of the ecosystem and the functions the system embodies. RCF encodes the architecture of MultiFunctional Ecological Networks (hereafter MFEN), decoded from the RCF by suitably contracting the consumer index, and yielding a resource-function matrix \mathbf{P} . The complex nested pattern observed suggests that certain ecological functions and plant species can be classified as "generalists" or "specialists". We formulate the concept of function keystonness, which focuses on the robustness of ecosystems with respect to perturbations to functions.

To further understand the multifunctional species keystone-ness and the role of plant species as ecosystem assemblers, we projected the MFEN into the function class and thus extract a function-function interaction network with weighted adjacency matrix $\mathbf{\Phi} = \mathbf{P}^{\top}\mathbf{P}$, that leverages how resources (plants) connect functions to assemble the ecosystem. We quantified the ecosystem robustness against perturbations (extinctions) of plant species by sequentially

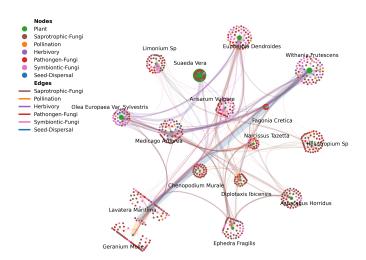


Fig. 1: Visualization of the Resource-Consumer-Function tensor (RCF) for the Na Redona dataset. Node colors account for plant species (green) and animal/fungus species according to the function they are involved in. The sizes of plant-nodes casptures their observed abundance. Edges represent function connections, their widths quantify the weight of interaction and the color denotes the functional interaction type. Species are clustered via Infomap.

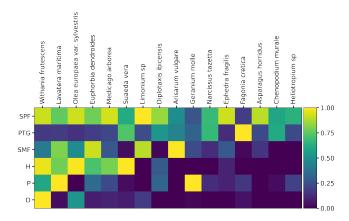


Fig. 2: Matrix of the Multifunctional Ecological Network for the Na Redona dataset. Rows represent functions and columns plant species.

pruning edges in Φ . Additionally, we ranked plant species, based on their multifunctional keystoneness, by conditioning Φ to each single plant species.

The dual concept of function keystonness can be addressed by following a similar manipulation: starting again from MFEN we project now on the plant class and thus construct a resource-resource interaction network withweighted adjacency matrix $\mathbf{\Pi} = \mathbf{P}\mathbf{P}^{\top}$, that leverages how functions broker the interaction between resources. We addressed two questions: (i) how robust is the ecosystem against perturbations of functions? (ii) how to quantify the heterogeneous roles and impacts of each different function in the ecosystem?

3 Discussion

Unfolding the different mathematical projections of the MFEN directly allows us to respond to (i) how robust is the ecosystem against perturbations of plant species or functions?, and (ii) how to quantify the heterogeneous roles and impacts not only of plant species, but also of functions, in the ecosystem? and quantify both (multifunction) species keystoness and (multispecies) function keystoness. We defined the novel and key concept of function keystoness. Just as keystone species encode, among other properties, robustness and resilience of an ecosystem [4], the response of the ecosystem to some disturbances may occur in the functional dimension, rather than in the resource (plant species) dimension. Defining and identifying key ecological functions and studying the robustness of functions to disturbances can be of great ecological importance to understand ecosystem functioning and resilience [5].

4 Methods

Estimation of the weights in RCF: The strength of an interaction between plant species *i*, and animal species *x* via a function α , was obtained using $f_{ix}^{\alpha} = \frac{m_{ix}^{\alpha}}{n_i}$. Where m_{ix}^{α} is the number of individuals of specie *i* where an animal species *x* was detected along α , and n_i is the number of individuals of specie *i* in the sample.

Resource-Consumer-Function tensor (RCF) visualization: By construction, f_{ix}^{α} is the probability of finding a resource *i* and a animal/fungus *x* interacting via a α . We thus obtain P_i^{α} , The probability that a plant species *i* participates in a function α , as $1 - \prod_x (1 - f_{ix}^{\alpha})$.

Acknowledgements We acknowledge support from the Spanish Agency of Research through Research and Development Projects Program (Grant PID2020-114324GB-C22 funded by MCIN/AEI/10.13039/501100011033). In particular, MCB thanks financial support Grant PID2020-114324GB-C22 funded by MCIN/AEI/10.1303/501100011033 and by "ESF Investing in your future".

- Galetti, M., Moleón, M., Jordano, P., Pires, M. M., Guimaraes Jr, P. R., Pape, T., et al , Ecological and evolutionary legacy of megafauna extinctions, Biological Reviews, 93(2), 845-862 (2018).
- 2. Bianconi, G., Multilayer networks: structure and function, Oxford university press (2018).
- Hervías-Parejo, S., Tur, C., Heleno, R., Nogales, M., Timóteo, S., and Traveset, A., Species functional traits and abundance as drivers of multiplex ecological networks: first empirical quantification of inter-layer edge weights. Proceedings of the Royal Society B, 287(1939), 20202127 (2020).
- 4. Mack, K. M., Eppinga, M. B., Bever, J. D., Plant-soil feedbacks promote coexistence and resilience in multi-species communities, Plos one, 14(2), e0211572 (2019).
- Yen, J. D., Cabral, R. B., Cantor, M., Hatton, I., Kortsch, S., Patrício, J., Yamamichi, M., Linking structure and function in food webs: maximization of different ecological functions generates distinct food web structures, Journal of Animal Ecology, 85(2), 537-547 (2016).



Temporal Betweenness Centrality on Shortest Paths

Mehdi Naima \cdot Matthieu Latapy \cdot Clémence Magnien

Abstract Betweenness centrality measure assesses the importance of nodes in a graph and has been used in a variety of contexts. Betweenness centrality has also been extended to temporal graphs. Temporal graphs have edges that bear labels according to the time of the interactions between the nodes. Betweenness centrality has been extended to the temporal graph settings, and the notion of paths has been extended to temporal paths. Recent results by Buß et al. [1] and Rymar et al. [2] showed that the betweenness centrality of all nodes in a temporal graph can be computed in $O(n^3 T^2)$ or $O(n^2 m T^2)$, where T is the number of time units, m the number of temporal edges and n the number of nodes. In this paper, we improve the running time analysis of these previous approaches to compute the betweenness centrality of all nodes in a temporal graph. We give an algorithm that runs in $O(n m T + n^2 T)$.

Keywords Graph mining \cdot Network Centrality \cdot Betwenness Centrality \cdot Temporal Graphs \cdot Temporal Paths \cdot Shortest Paths \cdot Graph algorithms.

1 Introduction

Betweenness centrality has been studied extensively in the literature and is a classical measure in network analysis and it is used in a number of different domains such that social networks [3], transports [4] and biology [5]. Moreover, betweenness centrality has been used as an efficient method for graph

M. Naima

M. Latapy and C. Magnien Sorbonne Université, CNRS, LIP6, F-75005 Paris, France E-mail: firstname.lastname@lip6.fr

This research is supported by the ANR projects 19-LCV1-0005 and FiT LabCom. The first author is supported by a DAAD Scholarship.

Department of Computer Science, RWTH Aachen University, 52074 Aachen, Germany E-mail: mehdi.naima@lip6.fr

partitioning and community detection [6]. Brandes in [7] used a method that allowed the computation of betweenness centrality for a whole graph in in $O(n m + n^2)$ which remains the fastest known algorithm.

More recently, betweenness centrality has been extended to dynamic graphs formalisms such that temporal graphs [8], and stream graphs [9]. The generalization of betweenness centrality to a temporal setting is not unique and many optimality criteria have been considered in the literature [1,2,10] such that shortest walks, fastest walks, foremost walks and shortest fastest walks. However, shortest paths are the most forward generalisation of the static case and in this paper we only focus on this criteria. It is then possible to define the betweenness centrality of a node v at time t by:

$$B(v,t) = \sum_{s \neq v \neq z \in V} \frac{\sigma_{sz}(v,t)}{\sigma_{sz}}$$

where $\frac{\sigma_{sz}(v,t)}{\sigma_{sz}}$ is the fraction of shortest temporal paths from s to z that pass through node v at time t. Recent results on temporal betweenness centrality tried with success to adapt Brandes algorithm to the temporal setting [1, 2]. For shortest paths their approach lead to time complexities of $O(n^3 T^2)$ and $O(n^2 m T^2)$ to compute the betweenness of a whole temporal graph. The authors did not apply Brandes algorithm to its full extent as we shall see. In this abstract we address this issue and show that for shortest paths we get an improved time complexity of $O(n m T + n^2 T)$ by using a suitable temporal BFS algorithm which improves the time complexity of previous approaches.

2 Formalism

We use a formalism close to the ones used in [1,2]. We define a directed temporal graph G as a triple $G = (V, \mathcal{E}, T)$ such that V is the set of vertices, $T \in \mathbb{N}$, is the maximal time step with $[T] := \{1, \ldots, T\}$ and $\mathcal{E} \subseteq V \times V \times [T]$ is the set of temporal arcs. We denote by n := |V| and $m := |\mathcal{E}|$. We call $V \times [T]$ the set of temporal nodes. Then $(v, w, t) \in \mathcal{E}$ represents a temporal arc from v to w at time t.

Definition 1 (Temporal walk) Given a temporal graph $G = (V, \mathcal{E}, T)$, a temporal walk W is a non-empty sequence of transitions $e \in \mathcal{E}^k$ where $e = (e_1, \ldots, e_k)$, with $e_i = (u_i, v_i, t_i)$ such that for each $i \in [k-1]$ of $v_i = u_{i+1}$ and $t_i \leq t_{i+1}$.

The length of a temporal walk W denoted len(W) is its number of arcs. We also denote by arr(W) the time of the last transition of W.

Definition 2 (Visited temporal nodes) For a temporal graph G, let W be a temporal walk of length k. We denote by $\mathcal{V}(W)$ the list of node appearances of W, it is defined as:

$$\mathcal{V}(W) = [(u_1, t_1)] + [(v_i, t_i) | 1 \le i \le k],$$

where + denotes list concatenation.

We can write a temporal walk $W = a \xrightarrow{1} b \xrightarrow{5} c$. Figure 1 represents a temporal graph having nodes $V = \{a, b, c, d\}$ and T = 6 with arrows representing the set \mathcal{E} . A temporal walk is called a *path* if each element in the visited nodes appears exactly once. In this paper all the walks considered are necessarily paths since we consider shortest walks and consequently these walks are all paths. Moreover, a temporal walk W is a **strict** temporal walk if for each transition time label is strictly larger than the previous one, that is

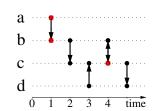


Fig. 1: The walk $W = a \xrightarrow{1} b \xrightarrow{4} c. \mathcal{V}(W) =$ [(a, 1), (b, 1), (c, 4)] are marked in red.

for $2 \leq i \leq k, t_i > t_{i-1}$ otherwise the temporal walk is a **non-strict** temporal walk. A non-empty walk W is an s - z walk if W starts in node s and ends in node z, we denote by W_{sz} the set of s - z walks. Then let $c_s(z) = \min_{W \in W_{sz}} len(w)$ if the minimum is not defined or the set of s - z walks is empty, then $c_s(v) = \infty$. Then we define the set of shortest paths as:

$$\mathcal{W} = \bigcup_{s,z \in V} \{ W | W \in W_{sz}, len(W) = c_s(z) \} \}.$$

Definition 3 Let $G = (V, \mathcal{E}, T)$ be a temporal graph. Let $s, v, z \in V$ and $t \in [T]$. We define the quantities,

- $-\sigma_{sz}$ is the number of s-z walks in \mathcal{W} and $\sigma_{vv}=0$.
- $-\sigma_{sz}(v,t)$ is the number of s-z walks in \mathcal{W} that pass through (v,t) in the sense of Definition 2.

We now define:

$$\delta_{sz}(v,t) = \begin{cases} 0 & \text{if } \sigma_{sz} = 0 ,\\ \frac{\sigma_{sz}(v,t)}{\sigma_{sz}} & \text{otherwise.} \end{cases} \qquad \delta_{s\bullet}(v,t) = \sum_{z \in V} \delta_{sz}(v,t).$$

Finally, our main quantity of interest is then defined by:

Definition 4 (Betweenness centrality)

$$B(v,t) = \sum_{s \neq v \neq z} \delta_{sz}(v,t), \quad B(v) = \sum_{t \in [T]} B(v,t).$$

Theorem 1 Let $G = (V, \mathcal{E}, T)$ be a temporal graph. Then the betweenness centrality of (strict and non-strict) shortest paths of all temporal nodes can be computed in $O(n m T + n^2 T)$ and if $m \ge n$ we obtain O(n m T).

The proof of Theorem 1 relies on several ingredients that we briefly describe in the following. For each node $s \in V$:

- 1. Build the Predecessor graph from node s where the predecessor graph is the temporal equivalent of the graph of shortest path in the static case.
- 2. The predecessor graph can be built using a Temporal BFS algorithm.
- 3. Extend Brandes recurrence (defined on $\delta_{s\bullet}$) to the temporal setting. This recurrence allows to compute the contributions of node s to the betweenness centrality of all other temporal node (v, t).

The authors of [1,2] used these steps and showed that the betweenness centrality can be computed in $O(n^3 T^2)$ and $O(n^2 m T^2)$ respectively. We improve the analysis of the algorithm used in [1] to construct the predecessor graph. Our improved analysis allows us then to conclude the proof of Theorem 1. In the introduction we mentioned using Brandes approach to its full extent, we see that the results of [1,2] when the temporal graph is static (i.e T = 1), lead to $O(n^3)$ and $O(n^2 m)$ while ours lead to $O(nm + n^2)$. Therefore our approach leads to the static optimal time algorithm if the temporal graph is static.

	PM	HS_2012	HS_2011	$hospital_ward$	WP_2013	ht09
temp/stat Kendall	0.81	0.80	0.83	0.78	0.81	0.80
temp/stat Top 20 $$	0.85	0.80	0.75	0.90	0.85	0.80

Table 1: Summary of experiments results on datasets of sociopatterns.org. PM refers to primary school, HS to highschool and WP to workplace. First row corresponds to the Kendall tau rank correlation coefficient between the temporal betweenness centrality (temp) and the static betweenness centrality on the aggregated graph (stat). Second row corresponds to the proportion of common nodes in the top 20 highest nodes between these measures.

Finally, we also provide an open-source implementation in $C++^{1}$ that we used to run the experiments summarized in Table 1.

- S. Buß, H. Molter, R. Niedermeier, M. Rymar, in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020), pp. 2084– 2092
- M. Rymar, H. Molter, A. Nichterlein, R. Niedermeier, in International Workshop on Graph-Theoretic Concepts in Computer Science (Springer, 2021), pp. 219–231
- 3. R.S. Burt, The new economic sociology: a reader pp. 325–348 (2004)
- R. Puzis, Y. Altshuler, Y. Elovici, S. Bekhor, Y. Shiftan, A. Pentland, Journal of Intelligent Transportation Systems 17(1), 91 (2013)
- S. Narayanan, The betweenness centrality of biological networks. Ph.D. thesis, Virginia Tech (2005)
- M. Girvan, M.E. Newman, Proceedings of the national academy of sciences 99(12), 7821 (2002)
- 7. U. Brandes, Journal of mathematical sociology **25**(2), 163 (2001)
- 8. V. Kostakos, Physica A: Statistical Mechanics and its Applications 388(6), 1007 (2009)
- 9. M. Latapy, T. Viard, C. Magnien, Social Network Analysis and Mining 8(1), 1 (2018)
- I. Tsalouchidou, R. Baeza-Yates, F. Bonchi, K. Liao, T. Sellis, International Journal of Data Science and Analytics 9, 257 (2020)

¹ github.com/busyweaver/code_temporal_betweenness_

Mobility



Delineation of city districts based on intraday commute patterns Yuri Bogomolov, Alexander Belyi, Ondrej Mikes and Stanislav Sobolevsky
Analysis of the German Commuter Network Christian Wolff, Markus Schaffert, Christophe Cruz and Hocine Cherifi
Academic Mobility as a Driver of Productivity: A Gender- centric Approach Mariana Macedo, Ana Maria Jaramillo and Ronaldo Menezes
Mobility networks as a predictor of socioeconomic status in urban systems Devashish Khulbe, Stanislav Sobolevsky, Alexander Belyi and Ondrej Mikes
Is Paris a good example for a X-minute city? Modeling city composition on POI data and X-minute statistics in Paris Sarah J Berkemer and Paola Tubaro
Impact of pedestrian flocking tactics on urban networks Guillaume Moinard and Matthieu Latapy 318
From CONSumers to PROSumers: spatially explicit agent-based model on achieving Positive Energy Districts Erkinai Derkenbaeva, Gert Jan Hofstede, Eveline van Leeuwen and Solmaria Halleck Vega



Delineation of city districts based on intraday commute patterns

Yuri Bogomolov¹ \cdot Alexander Belyi¹ \cdot Ondřej Mikeš² \cdot Stanislav Sobolevsky^{1,3,4}

Information and mobile technology have become essential in modern life, transforming the way we communicate, access information, entertain ourselves, and do business. Thereby we have the opportunity to access new datasets that did not exist for previous generations of scholars. Traditionally urban commute was studied based on census datasets, which represent the commute flow as a static number and get updated once in a few years. Over the last two decades, mobile phone datasets enabled new research avenues. In this paper, we utilize mobile phone mobility data to define a signature of urban districts in the city of Brno in the Czech Republic and leverage it for urban zoning.

Understanding intraday commute patterns is critical for city planners to optimize transportation infrastructure, public transport services, parking management, and mitigate transportation's environmental impact. The lack of hourly datasets limits the research of intraday commute patterns. Previous attempts covered only narrow mobility flows, like bicycle commuting in Melbourne [1] or taxi trips in New York City [2].

Defining city and district boundaries is known to be a challenge [3,4]. One approach to solve this problem is based on the connections graph using modularity maximization [5] or specialized algorithms to aggregate nodes [6], but they have limitations [7]. Another approach is based on the clustering of district signatures. Previous attempts to define district signatures were based on the municipal request datasets [8], financial activities [9], taxi trips [10, 11] and mobile call records [12,13]. However, all these events are generated only in case of particular activities and therefore have limited coverage. Also,

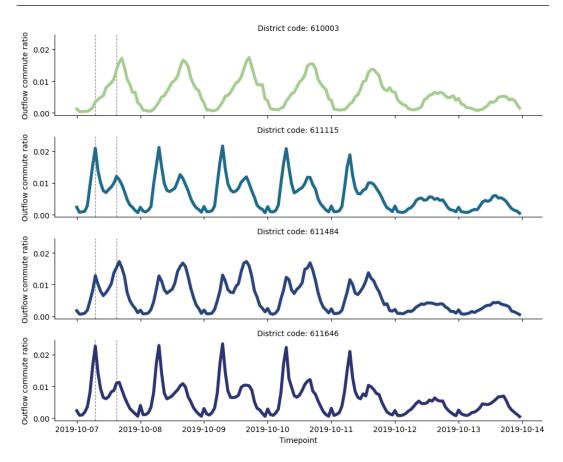
¹ Faculty of Science, Masaryk University, Brno 602 00, Czech Republic

² RECETOX, Masaryk University, Brno 625 00, Czech Republic

³ Center for Urban Science and Progress, New York University, Brooklyn, NY, USA

E-mail: ss9872@nyu.edu

⁴ Institute of Law and Technology, Masaryk University, Brno, 61180, Czech Republic



French Regional Conference on Complex Systems – Le Havre, May 31 June 2, 2023

Fig. 1 Hourly commute outflow ratios for four Brno districts. The vertical lines highlight commute spikes, and they correspond to 8 am and 4 pm on weekdays.

the resulting clusters did not reflect the mobility and transportation needs of inhabitants.

In this paper, we propose to use the mobile origin-destination commute data to build district mobility signatures and then use these signatures to group districts into areas with similar temporal patterns. Mobile flow data can be collected for any modern city, and it provides insights into mobility patterns and transportation needs.

Our research is based on the hourly aggregated origin-destination commute dataset. It covers 48 districts of Brno, the second-largest city in the Czech Republic. The commute flows are aggregated and do not contain any individual information.

In modern cities, people tend to carry their mobile phones for the whole day, and the dataset above provides unique insights into general mobility flows. We aggregated all outgoing trips for all districts for a given hour. As a result, every Brno district was represented by a vector with 168 (7 days \times 24 hours) outflow numbers. Due to the differences in population between districts, we normalized commute vectors by the total amount of outflow. Therefore every vector component represented the ratio of total weekly outflow that was observed in a given hour.

The normalized vectors had clear weekday and weekend patterns (see Fig. 1). We noticed a few repeating patterns across city districts. Districts that followed the same pattern had very similar outflow vectors (see the second and the fourth plot in Fig. 1), while patterns had significant differences. The initial investigation showed some correlation between commute pattern similarity and district location: e.g. the districts in the city center followed a similar pattern. To study the common patterns holistically we decided to reduce dimensionality and cluster the outflow vectors.

The outflow vector components are highly correlated due to the repeating patterns. We used PCA (Principal Component Analysis) to reduce the dimensionality of vectors and focus on the orthogonal components. Using just two components explains more than 73% of cumulative variance and we observe smaller incremental improvements for each additional component.

Dimensionality reduction allowed us to visualize all districts on a single two-dimensional plot. The visualization helped to identify patterns of city districts. The similarity of commute vectors for related districts gave us the idea to use commute vectors as a district signature. We used the K-means algorithm to cluster 2-component PCA feature vectors. Using the elbow method we decided to move forward with 3 clusters. The clustering output is presented in Fig. 2.

We identified strong patterns in the mobility outflow vectors. Clustering the signature vectors allows the delineation of city zones with similar mobility behavior. While the district signature does not contain any information about the spatial structure, we received a clear separation of the city center, residential area, and the mixed zone in between. The outflow profile of the residential cluster has a clear spike in the morning when people go to work. The city center outflow spikes in the evening. And mixed areas in between have traces of both patterns.

The general patterns matched our expectations, however, the clustering helped to identify some irregularities: e.g. the commute patterns in the southern outskirts of Brno have more similarity to the city center. We found a relationship between clusters and the ratio of municipal buildings, including libraries, schools, and city administration. While this fact is not sufficient to establish the causal relation we plan to conduct further research that will help us better understand the connection between the socioeconomic features of districts and their commute clusters.

To summarize, we proposed an approach leveraging mobility data from mobile phones to define unique signatures of urban locations, and apply them to urban zoning. The input dataset can be collected for any city. Relying only on the mobility data makes the signatures uniform for any country (in contrast to financial or municipal records), while wide dataset coverage enables capturing of general commute patterns (counter to taxi or biking datasets). Using the mobility signatures to delineate city districts provides great insights into mobility needs and can be directly used for transportation planning and land use classification. While the relationship between the municipal building and commute patterns contributes new information for city planning.

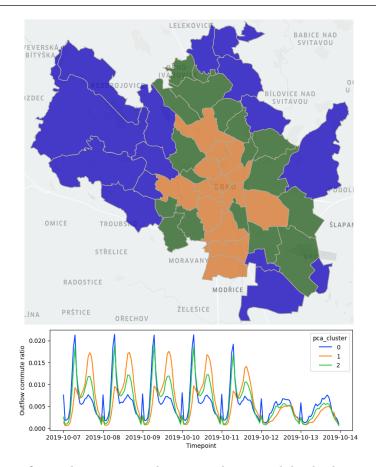


Fig. 2 The top figure demonstrates three city clusters, while the bottom figure shows aggregated hourly outflow commute profile for each cluster. The blue cluster corresponds to residential areas, the orange one covers the city center and business districts, while the green cluster has a combination of both.

- M.S. Smith, G. Kauermann, Transportation research part B: methodological 45(10), 1846 (2011)
- 2. N. Buchholz, Working paper, Tech. Rep. (2015)
- 3. L.M. Bettencourt, Science **340**(6139), 1438 (2013)
- 4. L.M. Bettencourt, J. Lobo, D. Strumsky, G.B. West, PloS one 5(11), e13541 (2010)
- 5. S. Sobolevsky, R. Campari, A. Belyi, C. Ratti, Physical Review E 90(1), 012811 (2014)
- 6. G. Duranton, The economics of interfirm networks pp. 107–133 (2015)
- L. Martínez-Bernabéu, J.M. Casado-Díaz, Papers in Regional Science 100(5), 1323 (2021)
- L. Wang, C. Qian, P. Kats, C. Kontokosta, S. Sobolevsky, PloS one **12**(10), e0186314 (2017)
- S. Sobolevsky, I. Sitko, R.T. Des Combes, B. Hawelka, J.M. Arias, C. Ratti, in 2014 IEEE international congress on big data (IEEE, 2014), pp. 136–143
- 10. X. Liu, L. Gong, Y. Gong, Y. Liu, Journal of transport Geography 43, 78 (2015)
- N.w. Kim, Y. Yoon, in Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising (2021), pp. 1–4
- C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, S.H. Strogatz, PloS one 5(12), e14248 (2010)
- T. Pei, S. Sobolevsky, C. Ratti, S.L. Shaw, T. Li, C. Zhou, International Journal of Geographical Information Science 28(9), 1988 (2014)



Analysis of the German Commuter Network

Period 2013 - 2021

Christian Wolff · Markus Schaffert · Christophe Cruz · Hocine Cherifi

Abstract Understanding the behavior of commuters is crucial as the number of commuters steadily rises, causing significant traffic congestion in many cities. Indeed, commuter behavior is vital in city and transport planning and policy-making[1–4]. Previous studies have investigated various factors that may impact commuting decisions. Still, these studies are often limited by the scale of data examined, including time duration, space, and the number of commuters. To address this gap, we gathered large-scale inter-city commuting data in Germany and analyzed the weighted commuting network from 2013 to 2021. This work relies on publicly available data so that the results can be reproduced.

Keywords Commuter · Complex Network Analysis · Spatial Data

Ch. Wolff

M. Schaffert Mainz University of Applied Sciences, i3mainz–Institute for Spatial Information and Surveying Technology, Germany E-mail: Markus.Schaffert@HS-Mainz.de

C. Cruz

Université de Bourgogne, Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB UMR CNRS 6303), France

E-mail: Christophe.Cruz @U-Bourgogne.fr

H. Cherifi Université de Bourgogne, Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB UMR CNRS 6303), France E-mail: Hocine.Cherifi@U-Bourgogne.fr

Mainz University of Applied Sciences, i3mainz–Institute for Spatial Information and Surveying Technology, Germany E-mail: Christian.Wolff@HS-Mainz.de

1 Introduction

The ongoing work presented here aims to analyze the Commuter flows in Germany from 2013 to 2021. Using all years of the period makes it possible to consider the short-term effects due to the COVID-19 pandemic. The network consists of the German districts (nodes) and the commuter flows (Einpendler) (edges) and incorporates weights based on the overall number of commuters. We exclude Commuter flows from outside German districts. The Federal Employment Agency, BA (Bundesagentur für Arbeit) (Dataset "Arbeitsmarkt in Zahlen, Sozialversicherungspflichtig Beschäftigte - Pendlernach Kreisen") provided us with the data representing the Commuter flows. One may notice different annotations in the dataset regarding the data quality of the years 2013/14, 2016/2017, and 2019 due to technical issues or local government reorganization. Coordinates were added to the dataset using the dataset Geographische Namen 1:250 000 (GN250), State geometries are added using the dataset Verwaltungsgebiete 1 : 250 000 (VG250-EW) (© GeoBasis-DE / BKG 2021) provided by the Federal Agency for Cartography and Geodesy (BKG).

2 Results

Preliminary results of the analysis confirm the findings of the BBSR regarding the increase in commuting activity. Indeed, the number of commuters in Germany has continuously increased in recent years. According to the Federal Office for Building and Regional Planning - BBSR, commuters rose by 6.6 percent in five years, from 18.4 million employees in 2016 to 19.6 million in 2021. This corresponds to 59.5% of all those subject to social security contributions. The group of commuters with a one-way commute of more than 50 kilometers experienced a sharp increase of 7.4%. It rose from 3.3 million in 2016 to 3.6 million in 2021. Among the major cities, Munich continues to lead the list with the most commuters [5].

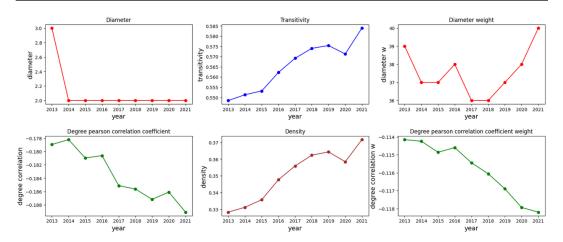
The analysis of the minimum Degree through time show that commuting increases. For example, the min degree increases from 30 in 2013 to over 40 in 2021. The median Degree rose from 115 in 2013 to more than 130 in 2021. The degree distribution shows that most districts have a low degree while a few districts have a high Degree.

Figure 1 shows the evolution throughout the global topological properties of the network. The diameter decreased from 3 in 2013 to 2 in 2014. Since then, it did not change. So, in terms of commuter routes, the German districts are pretty close. Taking into account the overall commuter counts as the weight adds a bit more dynamics to the diameter development as it decreases from 39 (2013) to 36 (2017/18) and increases again to 40 (2021). The density increases throughout the period (from 0.33 in 2013 to 0.37 in 2021) except for one drop from 2019 to 2020. It shows that commuting in Germany is increasing in numbers and connections between the districts. Although far from a complete graph, 0.37 seems pretty dense concerning the topic and the geographical extent of Germany. The development of transitivity follows the same trend, with the same drop between 2019 and 2020. It suggests that the increase also occurs between transitively connected nodes, for example, districts related to the same urban district. The Degree Correlation Coefficient shows a negative increase from -0.179 in 2013 to -0.188 in 2021. So, the network becomes more disassortative over time. It suggests that the rural-urban commuter relationships increased in the given period. The Federal Institute for Population Research (Bundesinstitut für Bevölkerungsforschung) reports that migration to urban municipalities has been steadily decreasing since 2012 and since 2014. They have even recorded migration losses. In contrast, rural districts have gained population for some years, indicating a renewed suburbanization phase in Germany [6]. One can assume that these people don't migrate to the rural areas because of a new job offer in these districts but due to other reasons, e.g., lower housing prices. Considering the weights, we observe the same trend but on a less disassortative level. The analysis of the community structure of the network also reveals interesting insights. Figure 2 shows the evolution of the community structure over the period uncovered by the Louvain Community Detection Algorithm. The communities fit the geographical areas quite well except for some large cities such as Berlin and Hamburg, which are in the southeastern cluster. We can also see that the community in central Germany disappeared in 2014, 2019, and 2021. The modularity exhibits low values around 0.2, characteristic of a weak community structure with many links between the communities. We observe a slightly decreasing trend except for a short increase between 2018 and 2020.

Using the overall commuter flows as weight. We can see that the community structure of the weighted network divides into smaller communities. The high modularity (0.7) shows that the communities are more cohesive. Their evolution in time follows the same trend as the unweighted community structure. The weighted community structure does not fit the state's borders well, except for some communities surrounding more prominent cities such as the Karlsruhe or the Mainz region. Furthermore, the number of communities varies between 10 and 13 (Fig.3). Some clusters disappear, e.g., in Lower Saxony / Eastern North Rhine-Westphalia, where the number of communities evolves between 1 and 3. It could be a result of the structure of the area consisting of rural areas and a few urban areas like Bremen, Hannover, Braunschweig, and Bielefeld.

We plan to extend the current analysis of static networks to include temporal networks, which can capture how the network structure and dynamics change over time[7]. It could involve investigating how different types of temporal dependencies affect the behavior of the network and the properties of its microscopic and mesoscopic structure. Additionally, incorporating information about the timing and duration of interactions between nodes could provide a more nuanced understanding of commuting patterns and dynamics.

Acknowledgements This work is part of the project Spatial Intelligence for the Integrated Care of Senior Citizens in Rural Neighbourhoods (Raumintelligenz für die integrierte



French Regional Conference on Complex Systems – Le Havre, May 31 June 2, 2023

Fig. 1 Global topological properties evolution from 2013 to 2021

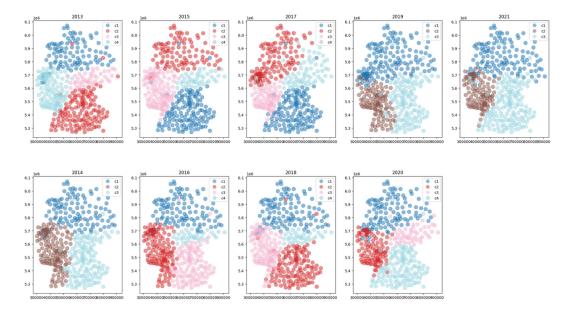


Fig. 2 Community structure evolution of the unweighted network uncovered by Louvain from 2013 to 202

Versorgung von Seniorinnen und Senioren in ländlichen Quartieren (RAFVINIERT)). It is funded by the Carl Zeiss Foundation in the program SocietyTransfer - Intelligent Solutions for an Ageing Society ("Transfer - Intelligente Lösungen für eine älter werdende").

- 1. I.M. Diop, C. Cherifi, C. Diallo, H. Cherifi, Applied Network Science 6, 1 (2021)
- 2. I. Moussa Diop, C. Cherifi, C. Diallo, H. Cherifi, arXiv e-prints pp. arXiv-2209 (2022)
- 3. I.M. Diop, C. Diallo, C. Cherifi, H. Cherifi, et al., Complexity 2022 (2022)
- M. Kumar, A. Singh, H. Cherifi, in Companion Proceedings of the The Web Conference 2018 (2018), pp. 1269–1275
- 5. BBSR. Länge der Arbeitswege unterscheidet sich regional erheblich (2022)
- M. Rosenbaum-Feldbrügge, N. Sander. Aktuelle Trends der Binnenwanderung in Deutschland (2020)
- 7. D. D. Mboup, D. Cherif, H. Cherifi, IEEE ACCESS 10, pp (2022)



Academic Mobility as a Driver of Productivity: A Gender-centric Approach

Mariana Macedo · Ana Maria Jaramillo · Ronaldo Menezes

1 Abstract

Academic mobility (the change of affiliations by academics) is usually positive in the career of researchers as it expands the collaboration networks and information flow between research groups [1,3,4,6,7]. However, under similar mobility circumstances, are the benefits equal amongst men and women? Do they exhibit similar impact factor via citations? Being a caregiver, for instance, may impact on the likelihood of moving, women being usually the one struggling to balance family and academic responsibilities, especially in senior positions [2,5]. Therefore, we highlight that even under similar mobility circumstances, gender norms play an important role.

We analysed the patterns in the career of researchers in the dataset of publications from ACM Digital Library (1980–2012). We have a total of 91,777 researches who 16% are women, and 809,397 publications in which around 22% women are at least one of the authors. Using network analyses, we found that the gender differences between the patterns found in the co-authorship networks tend to have similar characteristics across genders and career movements; the differences in the number of co-authors that men and women gain over their careers indicate that changing affiliations nationally and internationally benefits productivity. The increase in social ties can impact productivity, as writing papers collaboratively can speed up the process and lead to better quality work. However, we see that the small differences between the number of co-authors for women do not impact their productivity to make them more productive than men. Moreover, as men are the majority in our data, gender homophily benefits the high productivity levels more for men than for women.

Mariana Macedo

Center for Collective Learning, ANITI, University of Toulouse, FR E-mail: mmacedo@biocomplexlab.org

Ana Maria Jaramillo and Ronaldo Menezes

BioComplex Laboratory, Department of Computer Science, University of Exeter, UK

We plot the relationship between productivity (number of papers) and citations in Fig. 1; the distribution of women/men and their fraction in the four quadrants of the plots. The smallest fractions for both genders are for non-movers (top-right quadrant: 0.16% women and 0.23% men), and the largest fraction of both genders are also for non-movers (bottom-left quadrant: 98.52%women and 97.85% men). The highest difference between the movement categories is for researchers in the quadrant of high productivity-low citations, with national and international movers having, on average, 10 and 8 times more than non-movers. Regarding citations, women in both quadrants of high and low productivity get no differences when moving nationally (3.35%) or internationally (3.34%). In contrast, the fraction of men slightly increases when moving internationally (5.02%) compared to nationally (4.3%). We also found that the gender differences in productivity between **non-movers** researchers are smaller than for movers. Needless to say that our analyses come from assumptions and definitions limited by the data and methods we have available. Yet, our work sheds a light on where gender differences might be found in academic careers in Computer Science.

Changing affiliations might be a case of the rich-getting-richer or selection bias, potentially making men more prone to being hired in high-ranked institutions than women. There is a need to investigate the gender gap in women's representation in high-ranked institutions within and across countries. For instance, what is the relationship between moving from a developing nation to a developed one compared to moving across developed nations?

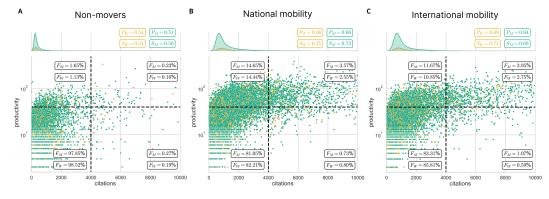


Fig. 1 Productivity versus citations across career movements: (A) Non-movers (B) National mobility (C) International mobility. Mobility has a role in how distributed women (yellow) and men (green) are in the plot, making the kurtosis smaller and increasing the number of productive and highly-cited researchers. The plot indicates the fraction of women (F_W) and men (F_M) for each quadrant, and it shows the Pearson $(P_{M|W})$ and Spearman correlations $(S_{M|W})$ between the productivity and citations for each gender.

References

 D. M. Jepsen, J. J. M. Sun, P. S. Budhwar, U. C. Klehe, A. Krausert, S. Raghuram, and M. Valcour, International academic careers: Personal reflections, International Journal of Human Resource Management (2014)

- 2. R. J. Leemann, Gender inequalities in transnational academic mobility and the ideal type of academic entrepreneur, Discourse (2010)
- 3. S. Marginson, What drives global science? The four competing narratives, Studies in Higher Education (2022)
- 4. A. M. Petersen, Multiscale impact of researcher mobility, Journal of the Royal Society Interface (2018)
- 5. M. Sautier. Move or perish? Sticky mobilities in the Swiss academic context. Higher Education (2021)
- G. Scellato, C. Franzoni, and P. Stephan. Migrant scientists and international networks. Research Policy, 44, 1, 108–120 (2015)
- W. Shen, X. Xu, and X. Wang, Reconceptualising international academic mobility in the global knowledge system: towards a new research agenda. Higher Education, 84, 6, 1317–1342 (2022)



Mobility networks as a predictor of socioeconomic status in urban systems

Devashish Khulbe · Alexander Belyi · Ondřej Mikeš · Stanislav Sobolevsky

Modeling socioeconomic dynamics has always been an area of focus for urban scientists and policymakers, who aim to better understand and predict the well-being of local neighborhoods. Such models can inform decision-makers early on about expected neighborhood performance under normal conditions, as well as in response to considered interventions before official statistical data is collected. While features such as population and job density, employment characteristics, and other neighborhood variables have been studied and evaluated extensively, research on using the underlying networks of human interactions and urban structures is less common in modeling techniques. We propose using the structure of the local urban mobility network (weighted by commute flows among a city's geographical units) as a signature of the neighborhood and as a source of features to model its socioeconomic quantities. The network structure is quantified through node embedding generated using a graph neural network representation learning model. In the proof-of-concept task of

D. Khulbe

A. Belyi Department of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno 61137, Czech Republic E-mail: bely@mail.muni.cz

O. Mikeš RECETOX, Masaryk University, Kamenice 753/5, Brno 62500, Czech Republic E-mail: ondrej.mikes@recetox.muni.cz

S. Sobolevsky

Institute of Law and Technology, Faculty of Law, Masaryk University, Brno 61180, Czech Republic

E-mail: sobolevsky@nyu.edu

Department of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno 61137, Czech Republic E-mail: dk3596@nyu.edu, khulbe@math.muni.cz

Center for Urban Science and Progress, New York University, Brooklyn 11201, NY, USA Department of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno 61137, Czech Republic

modeling the location's median income and housing profile in two different cities, such network structure features provide a noticeable performance advantage compared to using only the other available social features. This work can thus inform researchers and stakeholders about the utility of mobility network structure in a complex urban system for modeling various quantities of interest.

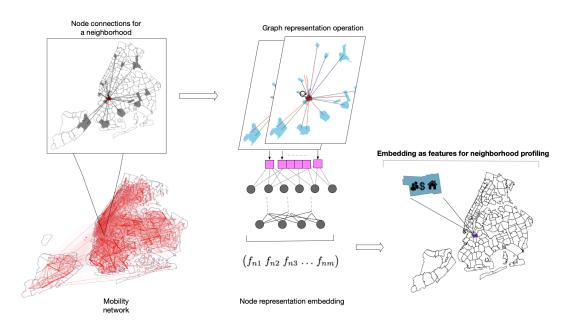


Fig.1 Structure of the locations' mobility network as a characteristic of its socioeconomic performance - an illustration for NYC

Large-scale mobility data has been used to study complex city dynamics while measuring the resilience of cities [1], although the network structure itself is not considered here to model socioeconomic performance indicators. Data-driven research has also emphasized modeling and studying dependencies among external features like 311 data [2], which proves importance of these in modeling. Using social media networks, metrics like centrality and average clustering coefficient have also been used as features to model economic growth in neighborhoods as a result of government investments [3]. These methods rely on combining external node variables with pre-defined network metrics as input features, rather than being able to learn the network representations of the whole urban area. Urban street networks using geo-located Twitter data have also been used to study public spaces and evaluate their success [4], indicating that granular networks are useful in analysis of sophisticated urban geographies. However, social media-based networks are not the best proxy for urban interactions, as they may exclude large populations based on age and location. A dynamic mobility data may help decision-makers assess the distribution and/or dynamics of the socioeconomic performance across the city in nearly real-time, without having to wait for years until Census and/or

	NYC			Brno		
	Node features	PPR	GNN	Node features	PPR	GNN
Lasso	0.22	0.25	0.5	0.03	0.15	0.51
\mathbf{RF}	0.22	0.45	0.48	0.04	0.4	0.59
VNN	0.22	0.58	0.54	0.04	0.06	0.43

 Table 1
 Out-of-sample R-squared values with supervised learning based on different model inputs

other official statistics reflecting such performance gets collected or updated. At the same time, mobility (represented by job flows or otherwise) is also an important metric to evaluate a neighborhood's economic profile. Urban researchers have established the connection between dynamic metrics like job density as indicators of a larger population standard of living [5]. More recently, researchers have also employed aggregate mobility indicators for socioeconomic modeling [6]. Within a larger urban mobility network, traditional machine learning modeling approaches are limited by the inability to properly define a feature space for a region. With the development and proven applicability of methods like Graph Neural Networks, there is a potential to leverage their power in urban contexts. Fig.1 illustrates the idea outlined with New York City's mobility among zip-code neighborhoods.

We consider the mobility network as a directed graph G(V, E), where the nodes V represent the geographical regions in a city and edges $E \subseteq V \times V$ are weighted by the mobility values between nodes. Two sets of techniques are considered to construct representation embedding of the network: 1. Personalized PageRank (PPR), which quantifies the significance of all nodes with respect to the given node. Formally, PPR embedding of a source node is the stationary probability distribution of a Markov chain that, with probability α , randomly transitions following the link structure of the network, and with probability $1 - \alpha$ teleports to a source node [7], and 2. Graph Neural Network (GNN) based embedding, which quantifies the node embedding as the series of network convolutions of certain initial embedding (e.g. one-hot encoding) of the connected nodes [8]. The goal is to learn a node function within the graph that preserves important structural properties of the network, such as local connectivity and symmetry. The formula for computing GNN embedding typically involves several iterations of message passing between nodes in the graph, which allows for the incorporation of local and global information. By repeating this message-passing process, GNN embedding can capture increasingly complex relationships and structures in the graph.

We experiment with mobility networks from two different cities - New York City (NYC), USA, and Brno, Czechia. NYC network is represented by the job flows among the zip codes job flows obtained from The Longitudinal Employer-Household Dynamics (LEHD), a U.S. census bureau program [9]. Whereas for Brno, we use aggregated hourly commute information from Tmobile cell phone-based location data. Two sets of socioeconomic target fea-

tures are considered - median income in 173 NYC zip codes and housing profile (% of single-family housing) in 48 Brno districts. Also, we consider external neighborhood features, such as zip code population and job densities for NYC, and employment rate in Brno. These are used as baseline features to compare against the utility of our embedding as input features. We experiment with three supervised learning approaches - Regularized Linear model (Lasso), Random Forests, and a Vanilla Neural Network (VNN). Table 1 presents out-ofsample results using various input features with the three models considered. The results show that mobility network embedding proves to have significant modeling capability, with GNN-based embedding proving to be particularly promising. Despite experimenting with data from two cities that differ in scale, geography, and economy, we demonstrate the ability of mobility networks to model socioeconomic characteristics. We also notice the results for NYC are much more stable, highlighting the differences in data and network sizes between the cities. Nevertheless, the better performance with embeddings for both cities shows the importance of mobility in socioeconomic profiling.

Our broader work aims to find the most efficient way of incorporating the network structure with other available neighborhood local contextual features to determine the maximum modeling capability. Overall, our findings can inform scientists and stakeholders to study and model complex urban systems from the perspective of a city's real-time networks while also leveraging the utility of local geographical contexts.

- Takahiro Yabe, P. Suresh C. Rao, Satish V. Ukkusuri, Resilience of Interdependent Urban Socio-Physical Systems using Large-Scale Mobility Data: Modeling Recovery Dynamics, Sustainable Cities and Society, Volume 75, 2021, 103237, ISSN 2210-6707, https://doi. org/10.1016/j.scs.2021.103237.
- Wang L, Qian C, Kats P, Kontokosta C, Sobolevsky S (2017) Structure of 311 service requests as a signature of urban location. PLoS ONE 12(10): e0186314. https://doi.org/ 10.1371/journal.pone.0186314.
- 3. Zhou X., Hristova D., Noulas A., Mascolo C. and Sklar M.:Cultural investment and urban socioeconomic development: a geosocial network approach. (2017) R. Soc. open sci.4170413170413 http://doi.org/10.1098/rsos.170413.
- Agryzkov, T., Martí, P., Nolasco-Cirugeda, A. et al. Analysing successful public spaces in an urban street network using data from the social networks Foursquare and Twitter. Appl Netw Sci 1, 12 (2016). https://doi.org/10.1007/s41109-016-0014-z.
- 5. Wen, M., Browning, C. R., & Cagney, K. A. (2003). Poverty, affluence, and income inequality: neighborhood economic structure and its implications for health. Social science & medicine, 57(5), 843-860.
- Xu, Y., Belyi, A., Bojic, I., & Ratti, C. (2018). Human mobility and socioeconomic status: Analysis of Singapore and Boston. Computers, Environment and Urban Systems, 72, 51-67.
- Gleich, D. F. (2015). PageRank Beyond the Web. SIAM Rev. 57, 3 (January 2015), (pp. 321–363). https://doi.org/10.1137/140976649
- Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Network (2017). International Conference on Learning Representations https: //openreview.net/forum?id=SJU4ayYgl.
- 9. United States Census Bureau (2021). Longitudinal Employer-Household Dynamics. https://lehd.ces.census.gov/. Accessed March 16, 2023.



Extended Abstract: Is Paris a good example for a X-minute city? Modeling city composition on POI data and X-minute statistics in Paris

Sarah J. Berkemer · Simon Genet · Léopold Maurice · Marie-Olive Thaury · Paola Tubaro

Keywords X-minute statistics \cdot City composition \cdot Urban modeling \cdot Paris \cdot OpenStreetMap

1 Motivation

Cities play an important role in various contexts such as society and economy but also sustainability and ecology. In many countries, urban areas host a large number of people working in the industrial sector, and also include main touristic sites as well as governments and administrations. City development has a high impact on the economics of a country as well as on the living quality of its inhabitants [7]. Therefore, cities can be considered as complex systems and summarizing all aspects of urban life is a challenging task [4]. Even though cities continuously change, many of them still show traces of historical events and keep historically grown aspects of composition and infrastructure. Thus, urban areas are not only a set of buildings and streets but they include complex dynamics of infrastructure and also their population [6]. Here, we do not only see daily movement in various transportation systems but also structural changes regarding the movement of subgroups of the population within the city regarding housing possibilities. Thus cities reflect socio-economic history [5] and at the same time show current societal and economic changes and their effects.

S. Berkemer LIX, Ecole Polytechnique E-mail: berkemer@lix.polytechnique.fr

P. Tubaro CREST CNRS-ENSAE E-mail: Paola.Tubaro@ensae.fr

M. Thaury, L.Maurice, S.Genet ENSAE

In recent years, urban development and design shifted towards a more pedestrian and cycling-friendly approach. One of the most popular models hereby is the 15-minute city concept [1] or its more generalized version, the X-minute city concept [2]. The Covid-19 pandemic triggered the emergence of a total rethinking of the functionality, composition and design of cities towards rather social, sustainable and resilient concepts. The aim is to provide access to a broad range of amenities within x minutes walking distance in order to recreate social links between the inhabitants and to avoid car journeys for public health and environmental reasons. X-minutes cities is an urban planning objective defined by Carlos Moreno [1] and made public by Anne Hidalgo during the 2020 municipal campaign and since then, Paris serves as an example city for many existing studies in the field. However, we did not find an extensive study showing to which extent Paris can be seen as a 15-minute city. In this project, we analyze the city of Paris, including its composition, accessibility to amenities and usage as an example for the 15- or X-minute city concept.

2 Data

OpenStreetMap (OSM, www.osm.org) data for Paris consisting of mapping data for streets and buildings annotated with names, numbers and points of interest (POI) such as shops, schools, restaurants, monuments, and parks as well as public transport annotations and bicycle paths. For many entries in OSM, the user can find very detailed information, e.g. names, categories of shops or food, or line numbers for public transport.

As a secondary source, we use a grid of squares of 200m x 200m on Paris with additional information on social and fiscal demographic properties. This grid - named *INSPIRE* - comes from the *FiLoSoFi* (https://www.insee.fr/fr/statistiques/4176290?sommaire=4176305) data set made by *INSEE* (www.insee.fr).

3 Methods and preliminary results

The purpose of our project is to develop a statistical model of city composition based on different types of points-of-interest (POIs). POIs are notable urban locations like schools or shops and are divided into several categories (public, health, leisure, catering, accommodation, shopping, money, tourism...), which are themselves subdivided. For example, OSM distinguishes bars from restaurants and fast food outlets. OSM also provides the ethnic character of the restaurants. These data thus make it possible to see the difference between a tourist, residential, commercial, wealthy or ethnic district.

First of all, we focus on different districts of Paris in order to compare their composition both quantitatively and qualitatively. In parallel, we are also looking to implement the x-minutes cities model to measure residents' accessibility to essential shops and amenities.

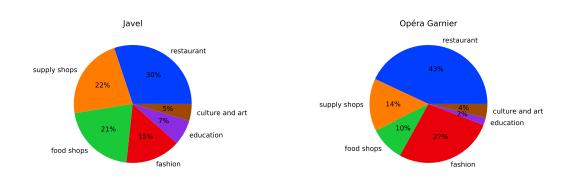


Fig. 1 Composition of two Parisian districts, Javel and Opéra Garnier regarding the amenity categories restaurants (blue), supply shops (orange), food shops (green), fashion (red), education (purple) and culture and arts (brown). We clearly see differences between a more touristic district such as Opéra Garnier and a more residential one like Javel.

Inspired by Knap et al [2] and Paris Mayor policy, we are concentrating on the composition of cities driven by services accessibility in a one kilometer radius (corresponding to a 15min walk at a 5km/h speed).

The categories for amenities that we are using during this study are the following: restaurants, supply shops, food shops, fashion, education, culture and art as it can be seen in Figure 1.

3.1 Composition of districts

As a first step towards the composition analysis of the complete city of Paris, we arbitrarily chose a dozen of places, such as monuments or metro stations, around which the social life of the neighborhood is organized. The neighborhood in question corresponds to a perimeter of one km of the pedestrian network around the chosen location. In each neighborhood we count the POIs for each of the categories defined above.

We find obvious disparities between neighborhoods in terms of the concentration of activities. In particular, the most outlying neighborhoods, such as Porte de la Chapelle or Eglise du Saint Esprit (Daumesnil district, 12th arrondissement), which are also relatively poor neighborhoods (particularly Porte de la Chapelle), are those containing the fewest amenities. On the contrary, the districts around Place des Vosges and Garnier Opera, which are more central and more affluent, have a much higher number of amenities.

The pie charts in Figure 1 enable us to distinguish a touristic district (Garnier Opera district, bottom) from a more residential district (Javel district, top). It is easy to see that more than 50% of the Javel district is composed of daily services (food shops, supply shops, education), compared to 26% in the Garnier Opera district. On the contrary, the Garnier Opera district has more than 70% of luxury services (restaurants, fashion), compared to 45% in the Javel district.

3.2 Accessibility statistics

Accessibility scores are an application of the two-step floating catchment area method (2SFCA) [3]. It has been first used to measure health services accessibility, but it can be applied to any sort of extensive variables like the number of amenities for instance. The idea behind 2SFCA is to measure for each service provider the surrounding demand. Then for each person (or place) asking for this service, we calculate the surrounding offer by considering that each service provider divides itself up on the previously calculated demand.

The FiLoSoFi data set also includes some socio-demographic variables like age pyramid or poverty rate. From those data, we calculate for each category the accessibility index using the 2SFCA method previously explained. The demand for a square is the weighted sum of the total population in the squares in 1 km radius. Weights are decreasing with the square of the distance (the weight of the center square is 1).

By using the accessibility index we consider that a square is influenced partially by its neighbors (in a 1 km radius), it is a proxy of the neighboring squares, while the total number is an "individual" measure. By combining the two, we can for instance separate a square which represents a commercial street (a lot of shops in the square) from a residential neighborhood or a commercial one.

Figure 2 clearly depicts in purple and yellow the Olympiades district in the South of Paris, historically known for gathering a large number of social housing units, as well as the low-income housing belt in the East of the city. You can also see the 16th, 7th, 8th and 6th arrondissements in the West and center of Paris (in red), which are known to be well-to-do districts.

4 Conclusion and Outlook

The aim of our analysis is twofold: (1) we plan to develop a model of the composition of a city's district and (2) we aim to analyze Paris regarding the X-minute city concept. For both parts, we took the first steps by defining data sets and formulating methods.

Next steps will include for (1) a more specific analysis of the city district composition by choosing a homogeneous set of district centers such as metro stations. For a more detailed analysis, we plan to include additional information provided by INSEE regarding socio-demographic variables. To refine the model, we plan to run clustering or regression methods. Regarding part (2) we plan to include further categories of amenities and run the analysis for the complete city of Paris. Hereby, we will be able to analyze how certain services are distributed among the city. A regression analysis of accessibility statistics will be conducted to identify the most important parameters for the accessibility measures inside a city as described by [2]. Both aspects of the project are planned to be extended to further (French) cities.

- Moreno, C., Allam, Z., Chabaud, D., Gall, C., Pratlong, F.: Introducing the 15-Minute City: Sustainability, Resilience and Place Identity in Future Post-Pandemic Cities. *MDPI* Smart Cities 4(1):93-111, 2021.
- Knap, E., Baran Ulak, M., Geurs, K.T., Mulders, A., van der Drift, S.. A composite Xminute city cycling accessibility metric and its role in assessing spatial and socioeconomic inequalities – A case study in Utrecht, the Netherlands. *Journal of Urban Mobility* 3, 2022.
- Wei Luo and Fahui Wang. Measures of Spatial Accessibility to Health Care in a GIS Environment: Synthesis and a Case Study in the Chicago Region. *Environment and Planning B: Planning and Design* 30.6:865–884, 2003.
- 4. Bettencourt, L. M. Cities as complex systems. *Modeling complex systems for public policies*:217-236, 2015.
- Almusaed, A., & Almssad, A. City phenomenon between urban structure and composition. Sustainability in Urban Planning and Design 3, 2019.
- Batty, M. Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies. Springer:1041-1071, 2009.
- 7. Abrahamson, M. Globalizing Cities: A Brief Introduction. Routledge 2019.

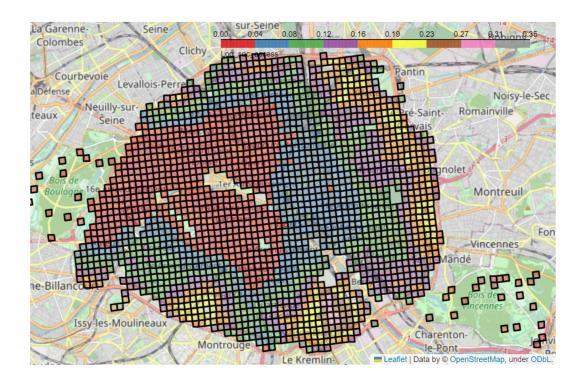


Fig. 2 Access to social housing in Paris where the red squares denote no access and the squares colored in yellow, brown or pink show a high access to social housing.



Impact of pedestrian flocking tactics on urban networks

Guillaume Moinard · Matthieu Latapy

Keywords Flocking \cdot Urban networks \cdot Protests \cdot Agent-based model \cdot OpenStreetMap \cdot Pedestrian dynamics \cdot Transport systems

1 Introduction

Urban networks are the foundation of modern cities, ensuring most mobilities, such as flows of goods and people. While most complex networks are vulnerable to attacks targeting most important nodes [1], urban networks, which possess an homogeneous node distribution, exhibit a greater robustness to nodes failure.

Still, demonstrations and protests can lead to massive perturbations. Many protesters gathering in a given place tend to overload its capacity. This impacts locally the traffic flow, but can also induce a cascade of congestion in the urban network. An important agent-based model literature inspired by Epstein's model [2] studied the microscopic interactions between protesters and counter forces. On the other side, some have modeled the macroscopic spreading dynamic of rioting behaviors from historical data [3].

However, to our knowledge, none have evaluated the direct impact of a collective action of protesters over an entire city. Unlike recent works on pedestrian dynamics [4], we do not pretend here to reproduce a realistic behavioral model for demonstrators. What we intend is to verify if an important disruption can emerge from a simple collective walking tactic on this network.

Guillaume Moinard

Sorbonne Université, CNRS, LIP6, F-75005 Paris, France E-mail: guillaume.moinard@lip6.fr

Matthieu Latapy Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

2 Models

2.1 Urban network

In our study urban networks are undirected graphs generated with Open-StreetMap data [5]. Intersections are nodes and edges are the pieces of streets linking those nodes. While this is a common representation for urban networks, we also chose to discretize our links by adding evenly spaced nodes on them, as shown on figure 1. We get a graph where walkers will make sufficiently small and equal moves at every step. This procedure has already been applied when studying transport systems on networks [6].



Fig. 1 The biggest connected component of Paris urban network. Dead ends have been removed and links have been discretized with a 50 meters step.

2.2 Pedestrian dynamic

What we call a tactic is a random walk which moves are biased by a set of rules. Instead of choosing its next node with equal probability among its neighbours, a walker might, for example, prefer to follow other agents. We only look for rules that are decentralized and comprehensive at a single agent level. This means our set of possible rules must be limited by the abilities of walkers, such as short range vision, bounded speed and finite energy. This forbids tactics too hard to compute for walkers or too slow, even if their convergence is proven [7]. Some specific studied tactics correspond to different already known dynamics, such as the simple inclusion process [8] or flocking on simple lattices [9]. We seek to identify, among those combination of rules, which ones are sufficient to generate gathering with disruptive effects.

3 Experiments

We first measure how agents following a common tactic perform at gathering on a given network. Walkers gathered on a same node form an aggregate. We consider that a node failure arises when an aggregate of walkers is bigger than a certain threshold C. In order to disrupt the network, walkers need to saturate many nodes. Therefore they seek to gather in numerous and big enough aggregates. For each tactic, and at every walking step, we measure the average size of aggregates and the number of aggregates bigger than C. We start with walkers uniformly distributed over the network. We also consider that agents won't walk more steps than the diameter of the network, so we run simulations only within that corresponding number of steps.

We observe that tactics biased by node's degree alone are not efficient, unlike those based on attraction towards walkers. Still, most are unable to form aggregates above a certain size. We find that only more refined tactics can, combining different rules.

However, an impacting tactic must not only be capable of forming big aggregates. They also have to move spontaneously in order to cover the network and disrupt as many nodes as possible. Therefore, to evaluate this, we compute the percent of nodes that have been saturated at least once during a simulation. Altogether, preliminary results show that the best tactic is a flocking similar to the Reynolds' model [10], combining rules based on alignment and cohesion with other walkers.

4 Conclusion

Our first results indicate large-scale impact is reachable with simple rules. Our on going simulations allow us to define explicitly what rules are sufficient for generating aggregates and flocking behaviors. Moreover we will use those results to rank tactics regarding their impact on urban networks.

- 1. Clémence Magnien, Matthieu Latapy, and Jean-Loup Guillaume. Impact of random failures and attacks on poisson and power-law random networks. 43(3):1–31.
- Joshua M Epstein. Modeling civil violence: An agent-based computational approach. Proceedings of the National Academy of Sciences, 99(suppl_3):7243-7250, 2002.
- 3. Laurent Bonnasse-Gahot, Henri Berestycki, Marie-Aude Depuiset, Mirta B Gordon, Sebastian Roché, Nancy Rodriguez, and Jean-Pierre Nadal. Epidemiological modelling of the 2005 french riots: a spreading wave and the role of contagion. *Scientific reports*, 8(1):107, 2018.
- Thibault Bonnemain, Matteo Butano, Théophile Bonnet, Iñaki Echeverría-Huarte, Antoine Seguin, Alexandre Nicolas, Cécile Appert-Rolland, and Denis Ullmo. Pedestrians in static crowds are not grains, but game players. *Physical Review E*, 107(2):024612, 2023.

- 5. Geoff Boeing. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. 65:126–139.
- 6. Izaak Neri, Norbert Kern, and Andrea Parmeggiani. Totally asymmetric simple exclusion process on networks. *Physical review letters*, 107(6):068702, 2011.
- 7. Subhash Bhagat and Andrzej Pelc. How to meet at a node of any connected graph. In 36th International Symposium on Distributed Computing (DISC 2022). Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.
- 8. Shlomi Reuveni, Iddo Eliazar, and Uri Yechiali. Asymmetric inclusion process. *Physical Review E*, 84(4):041101, 2011.
- 9. JR Raymond and MR Evans. Flocking regimes in a simple lattice model. *Physical Review E*, 73(3):036112, 2006.
- Craig W Reynolds. Flocks, herds and schools: A distributed behavioral model. In Proceedings of the 14th annual conference on Computer graphics and interactive techniques, pages 25–34, 1987.



From CONSumers to PROSumers: spatially explicit agent-based model on achieving Positive Energy Districts

Erkinai Derkenbaeva · Gert Jan Hofstede · Eveline van Leeuwen · Solmaria Halleck Vega

Introduction

The concept of Positive Energy Districts (PED) has emerged as a tool for urban energy transition toward climate neutrality through energy efficiency and net-zero energy balance. To achieve the PED, the urban energy system requires substantial changes in energy-efficient retrofitting (EER) and significant financial investments. This study focuses on developing an agent-based model called ENERGY Pro to explore households' decision-making on adopting EER, including double glazing, insulation of walls, roofs, floors, and the adoption of photovoltaic (PV) and heat pumps. Using the input survey to mimic the Dutch households, this model aims to understand households' contribution to urban energy transition in Amsterdam by 2030. The main contributions of this study are threefold. First, it offers an example of spatially-explicit empirically-driven energy models that are still scarce. In this study, the ENERGY Pro model explores the differences across the city districts, which leads to targeted policy implications at the local level that can also be useful for other places. Second, the generated synthetic population of Amsterdam using spatial microsimulation substantially extends the dataset and allocates households to places. This sheds light on more extensive information and helps thoroughly analyze the city districts. More, the created dataset will be made available and can be used for other studies on Amsterdam's households. Finally, the combination of agent-based modeling (ABM) and spatial microsimulation provides a deeper and more meaningful explorative analysis of the urban energy transition in Amsterdam by 2030, which is the first study of its kind.

E. Derkenbaeva

Wageningen University and Research

E-mail: erkinai.derkenbaeva@wur.nl

Method and data

In this study, an ABM approach is the main method for developing the EN-ERGY Pro model. Among different simulation approaches, ABM is the key approach to studying the behaviors of heterogeneous agents and their interactions over time in a quantitative manner [1]. Designed for bottom-up analysis, the ABM help capture individuals' emergent behavior and explain more complex macro behavior observed in the real world. The predominant advantage of using the ABM in researching the energy transition is its ability to account for complexity [2]. The model's conceptual framework is based on the Consumat meta-model [3] and incorporates micro and macro-driven factors. Micro-level factors are represented by individual households' sociodemographic (e.g., annual disposable income, age, education, social identity) and dwelling (e.g., dwelling type including apartment or non-apartment, construction year, annual energy consumption) characteristics. These characteristics affect households' satisfaction and uncertainty levels and, ultimately, the choice of a Consumat behavioral strategy. Behavioral options include different ways of adopting or not adopting EER. Macro-level factors are represented by the global variables (energy price and carbon emissions) influencing all households but to a different extent. The model includes two layers-spatial and social. The spatial layer represents residential buildings located in Amsterdam (i.e., agents' environment) and is informed by the BAG (Dutch automated system) data. The social layer denotes Amsterdam households informed by the WoON Dutch survey 2021 and Census data. Because of the limited number of respondents from Amsterdam (N=1630), we used a spatial microsimulation to expand the data. As a result, we created a usable dataset with 447 685 households assigned to a neighborhood based on their characteristics. The dataset is validated based on several goodness-of-fit measures. The temporal resolution of the model corresponds to one year (one step) covering a period of 10 years (2021-2030). The model is developed in NetLogo 6.3.0. Additionally, the model heavily relies on R programming language to perform more complex operations, such as spatial microsimulation, and time-expensive operations, such as sensitivity analysis. Finally, the model validation will rely mostly on expert validation, including consulting with urban energy transition policymakers in Amsterdam to examine whether patterns found in the model mimic reality.

Tentative results

The model demonstrates clear differences in households' contribution to energy transition across the districts of Amsterdam. First, we observe differences in adoption rate across the four EER measures investigated in this study. As such, the households in Amsterdam adopt double glazing and insulation, little fewer heat pumps, and much less PV. They tend to adopt fewer solar panels

because there are mostly multi-family houses in urban areas such as Amsterdam. The residents of this dwelling type need to make collective decisions on PV adoption as they share a common roof. Second, we observe differences in the impact of energy transition when only homeowners make adoption decisions compared to when both homeowners and tenants make these decisions. The energy transition in Amsterdam happens much faster when both groups of homeowners and tenants make adoption decisions. Third, the adoption of the EER differs across all districts. This can be explained by the differences in the context of each district, including households' and dwellings' characteristics. This study also investigates multiple scenarios, including the impact of a change in energy price on EER adoption and the energy transition in general. Additionally, we explore possible policy interventions that could inform policymakers in Amsterdam on observed energy-related behavioral patterns of homeowners and examined factors affecting these patterns. We use a One Factor At a Time (OFAT) approach of sensitivity analysis to explore the model's behavior. To validate the model, we will rely mostly on expert validation as there is no longitudinal data on EER uptake available in Amsterdam to validate the model against. Expert validation will include consulting with urban energy transition policymakers in Amsterdam to examine whether patterns in the model mimic reality.

Conclusion

Exploring households' decisions on EER adoption will allow an understanding of their contribution to urban energy transition and the pathways toward achieving PED goals. This study investigates how households' decisions vary across the districts in Amsterdam and why. Additionally, this study examines possible policy interventions and how the area-targeted policy approach can accelerate the energy transition. The findings of this study can also be useful for other cities. Future research can include other energy system elements, such as urban electric mobility, which is an important part of the energy transition.

- A. Ghorbani, L. Nascimento, T. Filatova, Energy Research & amp; Social Science 70, 101782 (2020). DOI 10.1016/j.erss.2020.101782. URL https: //doi.org/10.1016/j.erss.2020.101782
- P. Hansen, X. Liu, G.M. Morrison, Energy Research & amp; Social Science 49, 41 (2019). DOI 10.1016/j.erss.2018.10.021. URL https://doi.org/10. 1016/j.erss.2018.10.021
- W. Jager, M. Janssen, H.D. Vries, J.D. Greef, C. Vlek, Ecological Economics 35(3), 357 (2000). DOI 10.1016/s0921-8009(00)00220-2. URL https://doi.org/10.1016/s0921-8009(00)00220-2

Communities



Filtering the noise in consensual community detection Antoine Huchet, Jean-Loup Guillaume and Yacine
Ghamri-Doudane
Deep Learning Attention Model For Supervised and Unsuper- vised Network Community Detection Stanislav Sobolevsky
Quality certification of vertex cover heuristics on real-world net- works Fabrice Lecuyer
Identifying Influential Nodes: The Overlapping Modularity Vi- tality Framework Stephany Rajeh and Hocine Cherifi
Backbone Extraction of Weighted Modular Complex Networks based on their Component Structure Sanaa Hmaida, Hocine Cherifi and Mohammed El Hassouni
11a550um · · · · · · · · · · · · · · · · · · ·



Filtering the noise in consensual community detection

Antoine Huchet, Jean-Loup Guillaume · Yacine Ghamri-Doudane

Abstract Community detection allows understanding how networks are organised. Ranging from social, technological, information or biological networks, many real-world networks exhibit a community structure. *Consensual* community detection fixes some of the issues of classical community detection like non-determinism. This is often done through what is called a *consensus* matrix. We show that this consensus matrix is not filled with relevant information only, it is noisy. We then show how to filter out some of the noise and how it can benefit existing algorithms.

Keywords Real Graphs, Graph Algorithms, Consensual Communities, Noise

1 Introduction

Graphs representing real-world data exhibit particular features that make them far from regular. The distribution of edges is not homogeneous: parts of the graph are densely connected, while between such dense parts, there tend to be only a few edges. Such feature of real-world graphs is called *community structure*. Finding these densely connected parts is called *community detection*. In social graphs, community detection could help identify group of people such as families, friends or co-workers.

Many community detection algorithms exist like Walktrap, Infomap or Louvain [1]. Some algorithms are non-deterministic like the latter, where the communities produced are determined by the order in which the nodes are visited. Since the nodes may be visited in any order, such an algorithm may

This work has been partially funded by the ANR MITIK project, French National Research Agency (ANR), PRC AAPG2019.

A. Huchet · Jean-Loup Guillaume · Yacine Ghamri-Doudane

L3i La Rochelle Université, La Rochelle, France

E-mail: antoine.huchet@univ-lr.fr

produce different partitions of communities. To get a deterministic result, we combine the information of different partitions of communities into *consensual* communities [2].

Our contribution is twofold: we show that the information from the different partitions of communities is noisy. Then, when combining the partitions into consensual communities, we show that some of the noise can be avoided.

2 Consensual Communities

A graph G = (V, E) is made of a set of nodes V and a set of edges $E \subseteq V \times V$, where |V| = n and |E| = m. Communities form a partition of the nodes of the graph. That is, each node belongs to exactly one community. Lancichinetti, Fortunato and Radicchi (*LFR*) have proposed a model to generate synthetic graphs with a known community structure [3]. Those graphs are generated with a mixing parameter μ that defines the proportion of edges between communities. It ranges from 0 to 1, and the smaller μ , the easier it is to detect communities. The Modularity measures the quality of a partitions of a graph into communities. When the ground-truth communities are known, the *NMI* measures how close they are to a set of discovered communities. The Edge Clustering Coefficient (ECC(i, j)) is a similarity metric between two nodes *i* and *j* [5].

Since most community detection algorithms are non-deterministic, running n_p times an algorithm \mathscr{A} on a graph G may result in different partitions. We define a *consensus* matrix, C, where C_{ij} is the number of times that nodes i and j were put in the same community by \mathscr{A} across the n_p executions, called the *consensus* coefficient of i and j. Consensual communities can then be computed by building the *consensus* graph G_C , whose adjacency matrix is C. One way to derive the consensual communities is to set a threshold λ , and remove the edges of G_C whose weight is lower than λ . The resulting connected components would be the consensual communities [6]. It is also possible to execute again \mathscr{A} on G_C until convergence [2].

Note that the consensus matrix is an $n \times n$ matrix. Filling such a matrix is a lengthy process that requires a lot of memory. Different authors worked on improving this computation, as in Tandon's algorithm [7] or ECG [4], where only the entries C_{ij} that correspond to edges (i, j) of G are computed. This brings the number of entries from n^2 down to m, the number of edges in G.

3 Identifying and filtering the noise

We show that the consensus matrix C is noisy. To do so, we generate an LFR graph G (along with its ground-truth communities). We also use the ECC, which allows ordering our pairs of nodes. Next, we execute n_p times a community detection algorithm \mathscr{A} on G. Finally, we build a consensus matrix C, but we fill it one entry C_{ij} at a time in increasing order of ECC(i, j). In case

of a tie, we break it by selecting an arbitrary pair of nodes among the tie. After each addition in the incomplete consensus matrix C, we build the associated partial consensus graph G_C and execute \mathscr{A} on G_C . We then compute the NMI between the LFR ground-truth communities and the communities we just computed. We iterate this way until C is completely filled. This method allows us to study how the NMI varies based on the number of entries in C. Figure 1 shows the NMI as a function of the number of entries in C, for LFR graphs with 1 000 nodes, and 4 different values of μ . As we fill C, we observe that the NMI increases, reaches a maximum, then decreases. Notice that the maximum NMI corresponds to a consensus matrix that is far from filled.

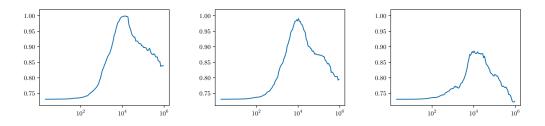


Fig. 1: NMI vs the number of entries of C, respectively for $\mu = .5, .6$ and .7

In real scenarios, we do not know the ground truth communities, so we cannot compute the NMI. We therefore need to decide beforehand how many entries of C need to be computed to get as close as possible to the maximum NMI. Our experiments show that the average modularity Q of the n_p executions of $\mathscr{A}(G)$ is strongly correlated to the threshold on the ECC τ below which entries should not be added to C. Thus, from Q, we can deduce when to stop filling C. We then feed the consensus graph G_C to existing algorithms. We call TANDON_FILTERED the heuristic that feeds the consensus graph G_C to TANDON, and ECG_FILTERED when we feed it to ECG. We call GENERIC_FILTER the approach that feeds G_C back to \mathscr{A} one last time. Note that when the algorithm that is fed with G_C supports weighted graphs, we weight the edges of G_C with their consensus coefficient.

To validate our approach, we generate LFR graphs with 10 000 nodes, with different values of μ . We choose $\mathscr{A} = \text{Louvain}$, and we execute our filtered and non-filtered algorithms on those graphs¹. Figure 2 shows the NMI and the running time as a function of μ . ECG_FILTERED provides a higher NMI than ECG, at the cost of a higher running time, due to the computation of the ECC. TANDON_FILTERED yields a comparable NMI, but takes longer than TANDON because of the computation of the ECC. The GENERIC_FILTER provides a higher NMI than ECG, but lower than TANDON, but allows working on bigger graphs than TANDON thanks to its better running time. We observe a high NMI, then a sharp decrease for our filtered approaches. This is because

¹ All the implementations are available on Software Heritage.

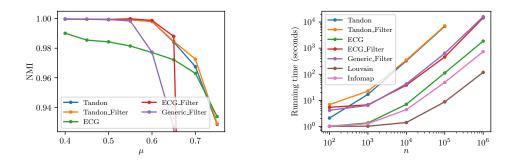


Fig. 2: NMI as a function of μ , LFR graphs with 10 000 nodes (left); Running time as a function of the size of the graph (number of nodes), for LFR graphs with $\mu = 0.6$ (right)

for high μ , the communities do not correspond to dense parts of the graphs anymore, so we tend to filter out intra-community edges.

We also applied our filters on real graphs and obtained similar results.

4 Conclusion and future work

We have shown that the information in the consensus matrix can be noisy but that it is possible to filter out some of the noise. We then used these observations to improve existing algorithms, and verified the effectiveness of our approach on synthetic and real graphs.

We believe that our noise filtering method would be useful for most algorithms that use a consensus matrix. Moreover, some algorithms perform community detection in several iterations, it could be worthwhile to study the effect of filtering the graph at each iterations.

- Blondel, V., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008(10), P10008 (2008)
- Lancichinetti, A., Fortunato, S.: Consensus clustering in complex networks. Scientific reports 2(1), 1–7 (2012)
- Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Physical review E 78(4), 046110 (2008)
- 4. Poulin, V., Théberge, F.: Ensemble clustering for graphs. In: International Conference on Complex Networks and their Applications, pp. 231–243. Springer (2018)
- Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proceedings of the national academy of sciences 101(9), 2658– 2663 (2004)
- Seifi, M., Guillaume, J.L.: Community cores in evolving networks. In: Proceedings of the 21st International Conference on World Wide Web, pp. 1173–1180 (2012)
- Tandon, A., Albeshri, A., Thayananthan, V., Alhalabi, W., Fortunato, S.: Fast consensus clustering in complex networks. Physical Review E 99(4), 042301 (2019)



Deep Learning Attention Model For Supervised and Unsupervised Network Community Detection

Stanislav Sobolevsky

Network community detection often relies on optimizing partition quality functions, like modularity or block-model likelihood. This appears to be a complex unsupervised learning problem (NP-hard in case of modularity optimization [1]) traditionally relying on heuristic algorithms, which often fail to reach the optimal partition, and, therefore, may require further fine-tuning. Over the last two decades, a large number of approaches and algorithms for community detection in complex networks have been suggested.

Some of them are straightforward heuristics such as hierarchical clustering [2] or the Girvan-Newman [3] algorithm, while the majority rely on optimization techniques for various objective functions. The most well-known partition quality function is modularity [4,5] assessing the relative strength of edges and quantifying the cumulative strength of the intra-community links. A large number of modularity optimization strategies have been suggested over the last two decades [6], [7], [4], [5], [8], [9], [10], [11], [12], [13], [14]. Comprehensive overviews are presented in [15], [16] and later surveys [17], [18].

However, the rise of deep learning and graph neural networks in particular, offer new opportunities. While graph neural networks have been widely used for supervised and unsupervised learning on networks, their application to unsupervised community detection optimization has been limited so far. This work proposes a suitable network augmentation with an additional layer of community meta-nodes, and a novel deep learning model over such a network for supervised and unsupervised community detection. The proposed model is efficient for unsupervised fine-tuning of the network community structure previously obtained by other algorithms with respect to a variety of objective functions, including modularity, blockmodel likelihood, description length etc.

Recently graph neural networks (GNNs) have become increasingly popular for supervised classifications and unsupervised embedding of the graph

Center for Urban Science and Progress, New York University, Brooklyn, NY, USA

Department of Mathematics and Statistics and Institute of Law and Technology, Masaryk University, Brno, Czech Republic

nodes with diverse applications in text classification, recommendation systems, traffic prediction, computer vision etc [19]. And GNNs were already attempted to be applied for community detection, including supervised learning of the ground-truth community structure [20] as well as some unsupervised learning of the node features enabling representation modeling of the network, including stochastic block-model [21] and other probabilistic models with overlapping communities [22] or more complex self-expressive representation [23]. However, existing GNN applications overlook unsupervised modularity optimization, which so far has been a major approach in community detection.

In our recent paper [24] we proposed a simple recurrent GNNinspired algorithm to serve as a proof-of-concept for unsupervised modularity optimization. The algorithm was tuning node community attachment through an iteration of the GNN-style transformations. It was capable of reaching the best-known partitions for some of the classical networks and provided a scheme for fine-tuning the network community structure with a flexible trade-off between performance and computational speed.

In order to further improve the efficiency of the approach, we propose a new model which consists of a two-layer bi-partite convolutional graph neural network, stacked with a fully connected attention vanilla neural network 1. The graph neural network generates embedding for both - the original network nodes as well as the

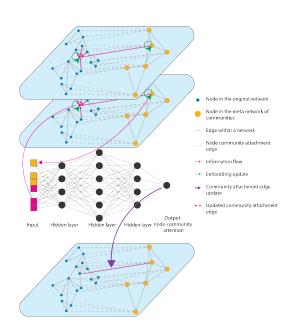


Fig. 1: A deep learning framework for network community detection.

community meta-nodes, while the vanilla neural network computes relative attention scores between each node of the original network and each community meta-node as a function of the stacked vectors of those node embedding, generated by the graph neural network.

The bi-partite network edges can be initially defined based on an initial community attachment to be fine-tuned. And the input node features for the graph neural network can represent any pre-defined network node embedding. The model can also be pre-trained through a supervised learning gradient descent optimization, aiming to reconstruct a previously known community structure. Then the model can be further tuned in an unsupervised manner in order to optimize a given objective function, such as network modularity.

 Table 1: Performance on the proposed deep learning algorithm improving partition modularity from Louvain algorithm for some classic network examples

Network	Louvain	Improvement
Word adjacency network in David Copperfield [25] Amazon co-purchases of political books, orgnet.com	$\begin{array}{c} 0.305052 \\ 0.527082 \end{array}$	$\begin{array}{c} 0.309279 \\ 0.527236 \end{array}$

Table 2: Out-of-sample performance on the proposed deep learning algorithm in supervised learning of the best-known or given partition for some network examples (community reconstruction accuracy for the 40% randomly masked nodes)

Network	Accuracy
Co-appeareance of characters in Les Miserables [26]	85.18%
Amazon.com co-purchases of political books, www.orgnet.com	89.47%
Dolphins' Social Network [27]	90.00%
Network of Jazz Musicians [28]	89.41%
Neural network of C. Elegans [29]	86.96%
LFR synthetic network with 500 nodes [30]	96.83%
LFR synthetic network with 1000 nodes [30]	98.70%

The approach turned out to be efficient in fine-tuning the results of other algorithms, e.g. a popular Louvain algorithm [9]. The table 1 provides examples of such improvement reached by the Python 3.7 implementation of the proposed model. It can also be applied to other quality functions, such as a block-model likelihood or description length.

Furthermore, the model can perform supervised community detection, extrapolating the community structure provided for a certain part of the network nodes to the rest of them. The out-of-sample reconstruction accuracy for the best-known partition typically ranges within 85-99% for a number of classic and Lancichinetti-Fortunato-Radicchi (LFR) synthetic networks (table 2). For comparison, the out-of-sample accuracy of 93.51% of the simple label propagation baseline algorithm (nodes with unknown community attachment get attached according to the majority of their neighbors) for the largest LFR network with 1000 nodes falls noticeably short of the 98.70% accuracy achieved by the proposed deep learning approach. Those cases represent initial proofof-concept results, while fine-tuning of the model's configuration could further help improve the performance. Also, evaluation of the approach on a broader range of examples and comparison against known state-of-the-art/baseline supervised community detection approaches remains the subject of future work.

Finally, as the deep learning model configuration does not depend on the dimensionality of the network or the number of network communities, but only on the selected dimensionality of the node embedding, it makes it possible to consider transferring the pre-trained model architectures and parameters between the networks. And similarly to [24], iterating an ensemble of partition fine-tuning models (pre-trained over select sample networks) over the target network partition may provide the best practical results.

While the presented results serve as a proof of concept of the proposed deep learning model's utility for supervised and unsupervised community detection, its further fine-tuning and extensive evaluation, and exploring the potential of transfer learning between the networks, is the subject of our future research.

- U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, D. Wagner, arXiv preprint physics/0608255 (2006)
- 2. T. Hastie, The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations (Springer, New York, 2001)
- 3. M. Girvan, M. Newman, Proc. Natl. Acad. Sci. USA 99 (12), 7821 (2002)
- 4. M. Newman, M. Girvan, Phys. Rev. E 69 (2), 026113 (2004)
- 5. M. Newman, Proceedings of the National Academy of Sciences 103(23), 8577 (2006)
- M.E.J. Newman, Phys. Rev. E 69, 066133 (2004). DOI 10.1103/PhysRevE.69.066133. URL http://link.aps.org/doi/10.1103/PhysRevE.69.066133
- A. Clauset, M.E.J. Newman, C. Moore, Phys. Rev. E 70, 066111 (2004). DOI 10.1103/ PhysRevE.70.066111. URL http://link.aps.org/doi/10.1103/PhysRevE.70.066111
- Y. Sun, B. Danila, K. Josić, K.E. Bassler, EPL (Europhysics Letters) 86(2), 28004 (2009). URL http://stacks.iop.org/0295-5075/86/i=2/a=28004
- V.D. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefebvre, Journal of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (2008)
- 10. R. Guimera, M. Sales-Pardo, L.A.N. Amaral, Physical Review E 70(2), 025101 (2004)
- B.H. Good, Y.A. de Montjoye, A. Clauset, Phys. Rev. E 81, 046106 (2010). DOI 10.1103/PhysRevE.81.046106. URL http://link.aps.org/doi/10.1103/PhysRevE.81. 046106
- J. Duch, A. Arenas, Phys. Rev. E 72, 027104 (2005). DOI 10.1103/PhysRevE.72.027104. URL http://link.aps.org/doi/10.1103/PhysRevE.72.027104
- J. Lee, S.P. Gross, J. Lee, Phys. Rev. E 85, 056702 (2012). DOI 10.1103/PhysRevE.85. 056702. URL http://link.aps.org/doi/10.1103/PhysRevE.85.056702
- 14. S. Sobolevsky, R. Campari, A. Belyi, C. Ratti, Physical Review E 90(1), 012811 (2014)
- 15. S. Fortunato, Physics Report **486**, 75 (2010)
- 16. S. Fortunato, D. Hric, Physics reports **659**, 1 (2016)
- 17. B.S. Khan, M.A. Niazi, arXiv preprint arXiv:1708.00977 (2017)
- M.A. Javed, M.S. Younis, S. Latif, J. Qadir, A. Baig, Journal of Network and Computer Applications 108, 87 (2018)
- 19. Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S.Y. Philip, IEEE transactions on neural networks and learning systems (2020)
- 20. Z. Chen, X. Li, J. Bruna, arXiv preprint arXiv:1705.08415 (2017)
- 21. J. Bruna, X. Li, stat **1050**, 27 (2017)
- 22. O. Shchur, S. Günnemann, arXiv preprint arXiv:1909.12201 (2019)
- 23. S. Bandyopadhyay, V. Peter, arXiv preprint arXiv:2011.14078 (2020)
- 24. S. Sobolevsky, A. Belyi, Applied Network Science 7(1), 1 (2022)
- 25. M.E. Newman, Physical review E 74(3), 036104 (2006)
- 26. D.E. Knuth, The Stanford GraphBase: a platform for combinatorial computing (Addison-Wesley, 1993). URL http://www-cs-staff.stanford.edu/~{}uno/sgb.html
- D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, Behavioral Ecology and Sociobiology 54(4), 396 (2003). DOI 10.1007/s00265-003-0651-y. URL http://dx.doi.org/10.1007/s00265-003-0651-y
- 28. P.M. Gleiser, L. Danon, Advances in Complex Systems 06(04), 565 (2003). DOI 10. 1142/S0219525903001067. URL http://www.worldscientific.com/doi/abs/10.1142/ S0219525903001067
- J.G. White, E. Southgate, J.N. Thomson, S. Brenner, Philosophical Transactions of the Royal Society of London. B, Biological Sciences **314**(1165), 1 (1986). DOI 10.1098/ rstb.1986.0056. URL http://rstb.royalsocietypublishing.org/content/314/1165/ 1.abstract
- 30. A. Lancichinetti, S. Fortunato, F. Radicchi, Phys. Rev. E 78 (4), 046110 (2008)



Quality certification of vertex cover heuristics on real-world networks

Fabrice Lécuyer

Keywords Complex networks \cdot Vertex cover \cdot Quality certification \cdot Heuristics

1 Addressing the vertex cover problem on real-world networks

The vertex cover problem consists in finding a subset of nodes that touch all the edges of a graph. Numerous real-world applications are equivalent to finding a vertex cover with minimum number of nodes: vaccinating as few people as possible to prevent any transmission of a virus, destroying as few routers as possible to shutdown a network, using as little energy as possible to provide wireless connectivity to a given area... Yet, finding the exact result is generally not feasible under time constraints: no polynomial algorithm is known [1], even when the graph is planar or when nodes have small degree.

When facing this problem on datasets that represent large complex networks, engineers have two options. The first one consists in using a preprecessing method that reduces the graph with specific rules before applying an exact exponential-time algorithm. In 2019, a programming challenge fostered efforts in this direction and rewarded the implementation of [3]. If the network has suitable properties, it can lead to an exact minimum cover in reasonable time; but in general, there is no guarantee that the execution terminates in the next hours, days or years.

The alternative option is to use fast and intuitive algorithms that have no theoretical guarantees, called heuristics. Several of them are combined in the linear-time implementation of [4]. Their execution time can be predicted and the quality of their result can be excellent, but it is not mathematically guaranteed: there is no indication on how far the heuristic result is from optimum.

This work is funded by the French National Agency of Research through ANR FiT LabCom.

Fabrice Lécuyer, E-mail: fabrice.lecuyer@lip6.fr

Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

To bridge the gap, we propose a method that certifies the quality of a heuristic on a given network. The quality certification takes a network and gives both an approximate result and a certificate of its quality, defined as the ratio between the heuristic result and a bound on the optimum value. For example, the shortest path between two cities is lower-bounded by the distance as the crow flies; the certified quality of a path is then given by the ratio between its length and the lower-bound. To obtain lower-bounds for vertex cover, we design the two strategies presented in Section 2. The first one uses the well-known dual problem of maximum matching. The second uses a greedy clique partition that particularly fits complex networks.

The experiments of Section 3 test the method on 114 real-world networks with up to three billion edges: we certify that the results of state-of-the-art heuristics for the minimum vertex cover problem are within 1% of the optimum value on two thirds of the networks. This work shows that valuable quality certificates can be given for existing heuristics on specific networks without loosing on scalability: both the heuristic and the certification take linear time. It outlines the best practice of providing certifications for heuristics in general. As it may generalise to other algorithmic problems, it opens a door for further research and for deployment in real-world applications.

2 Designing lower-bounds to certify the quality of a vertex cover

Given a graph of n nodes, finding the size c^* of a minimum vertex cover is NP-hard. Various heuristics can be used to obtain a *small* but not *minimum* cover of size $c \ge c^*$. Besides, vertex cover has a well-known 2-approximation algorithm that implies a lower-bound on c^* : for any maximal set of eindependent edges, a vertex cover needs at least one node to cover each of these e edges, and at most two nodes to cover all the graph. This gives: $e \le c^* \le 2e$.

To obtain a useful lower-bound, the aim is to compute the maximum value of e^* , also called the maximum matching number. This problem can be solved in polynomial-time by the blossom algorithm [5]. Faster heuristics can find a close but smaller value e, in particular a linear greedy algorithm: as long as independent edges are in the graph, it selects the one that has the less neighbouring edges.

Definition 1 (Certification by matching) Given a vertex cover of c nodes and a matching of e edges, the quality certification method guarantees that the vertex cover is within factor μ of the minimum cover, with the quality ratio μ defined as:

$$\mu = \frac{c}{e}$$
 then $c \le \mu c^*$

However, as we will see in Section 3, this method does not succeed on denser graphs. In particular, when there is a clique of 2k nodes, a matching can have at most k edges while a vertex cover has at least 2k-1 nodes: for high k, the quality ratio tends towards 2, which is not better than the mathematical

guarantee in any graph. Still, the certification by matching is an important building block for the next bounding heuristic that we propose below.

With their strong community structure, complex networks are known to contain high numbers of cliques [2]. From this observation, we propose a new lower-bounding technique based on a partition of the graph into x independent cliques. To cover all the edges of a clique, a vertex cover needs to contain at least all but one of the nodes of the clique. Summing over the x cliques, we obtain $c^* \ge n - x$.

Finding a small value x^* is called the clique cover problem and is NP-hard; but a greedy linear method exists: it grows a clique progressively by selecting adjacent nodes of small degree. When the clique cannot grow, it is added to the partition. Note that a matching is a particular instance of a clique partition: it partitions the nodes into independent edges (cliques of two nodes) and isolated nodes (cliques of one node). Thus, we can always obtain x such that $n-x \ge e$. In the end, we have the following inequalities:

$$e \le e^* \le n - x \le n - x^* \le c^* \le c \le 2e$$

Definition 2 (Certification by cliques) Given a vertex cover of c nodes and a partition of the nodes into x cliques, the quality certification method guarantees that the vertex cover is within factor γ of the minimum cover, with the quality ratio γ defined as:

$$\gamma = \frac{c}{n-x}$$
 then $c \le \gamma c^*$

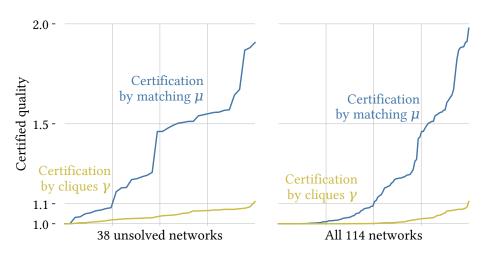


Fig. 1 Ratio of certified quality for vertex cover using matching (μ) or cliques (γ) . Left: 38 unsolved networks (exact solution unknown). Right: all 114 networks. Networks are ranked by their certified quality. Horizontal grids indicate 1.1 and 1.5 thresholds and vertical grids cut networks in four even groups. Observe that half of the networks obtain a certified quality within the 1.5 line with matching. With cliques, 92 networks including 18 unsolved have a ratio below 1.03 with cliques: the lower-bound certifies that the smallest cover found by the heuristic is at most 3% larger than minimum.

3 Certifying results on real-world networks

To assess the quality certification method using these bounds, we gather 114 real-world networks (web graphs, social networks, biological interactions...) and we apply the following algorithms on each of them:

- the exponential algorithm of [3] that is able to compute the size c^* of a minimum vertex cover for 76 of the 114 networks in less than six hours;
- the greedy heuristic implemented in [4] to find a small cover of size c;
- our implementation of a linear-time greedy matching algorithm that obtains e close to the e^* of the blossom algorithm (which is not linear and has an average relative improvement under 0.3%);
- our implementation of a linear-time greedy algorithm to partition n nodes into x cliques.

The certification ratio μ obtained with a matching is shown in blue on Figure 1. Unsolved networks are those where the exact algorithm could not compute a minimum vertex cover in six hours; the matching guarantees that, for half of them, the approximate cover is at most 1.5 times as big as the minimum. Among solved networks, the quality ratio is even lower, and the execution of these greedy algorithms can be much faster than the exact exponential algorithm.

As a matching is a special case of clique partition, we know that the clique method gives better results. Indeed, the yellow lines of the figure show that the ratio γ is always under 1.11 even for unsolved networks. Strikingly, it is under 1.01 for 76/114 networks: the minimum value is not known but the certification guarantees that the found cover is at most 1% larger than optimum.

Conclusion

Altogether, the results show that the certification can be used as an efficient shortcut for hard problems: in linear time, we obtain a proof that an approximate result is close to the unknown optimum. Further research needs to translate this principle into applications and to extend it to other algorithmic problem. The hope is that heuristics will always go along with a quality certification method, thus bridging the gap between predictable execution time and guarantees on the results.

- 1. R.M. Karp, *Reducibility among Combinatorial Problems*, Complexity of Computer Computations, 1972. https://doi.org/10.1007/978-1-4684-2001-2_9
- 2. A. Baudin et. al, Clique percolation method: memory efficient almost exact communities, ADMA, 2021. http://arxiv.org/abs/2110.01213
- 3. D. Hespe et. al, WeGotYouCovered: the winning solver from the PACE 2019 Challenge, Workshop CSC, 2020. https://doi.org/10.1137/1.9781611976229.1
- S. Cai et. al, Finding a small vertex cover in massive sparse graphs, Journal of Artificial Intelligence Research, 2017. https://doi.org/10.1613/jair.5443
- 5. J. Edmonds, Paths, trees, and flowers, Canadian Journal of Mathematics, 1965.



Identifying Influential Nodes: The Overlapping Modularity Vitality Framework

Stephany Rajeh · Hocine Cherifi

Abstract This paper proposes an Overlapping Modularity Vitality framework for identifying influential nodes in networks with overlapping community structures. The framework uses a generalized modularity equation and the concept of vitality to calculate the centrality of a node. We investigate three definitions of overlapping modularity and three ranking strategies prioritizing hubs, bridges, or both types of nodes. Experimental investigations involving real-world networks show that the proposed framework demonstrates the benefit of incorporating overlapping community structure information to identify critical nodes in a network.

Keywords Influential nodes \cdot Modularity Vitality \cdot Overlapping Community structure

1 Introduction

Real-world networks often have a community structure [1,2]. Recent research has focused on designing "community-aware" centrality measures that take advantage of this structure [3]. However, these measures typically assume that a node belongs to a single community, while they may often belong to multiple communities. This overlapping community structure is a significant feature in many real-world networks, including social networks and protein networks [4].

While there are many existing works on detecting overlapping network communities, few researchers have leveraged it to identify critical nodes [5–12]. Moreover, the existing methods have some drawbacks. "Membership" is a straightforward approach quantifying the number of communities shared by

S. Rajeh

Laboratoire d'Informatique de Bourgogne - University of Burgundy - France E-mail: stephany.rajeh@u-bourgogne.fr

H. Cherifi ICB UMR 6303 CNRS - University of Burgundy - France

a node [5]. It cannot differentiate between nodes with the same community affiliations. "OverlapNeighborhood" focuses on identifying hubs located in the vicinity of overlapping nodes without considering the nature of their connections [7]. In contrast, "Random Walk Overlap Selection" prioritizes only highly connected overlapping nodes [8]. The Overlapping Modular Centrality measure combines the communities of overlapping nodes to calculate their local importance, which leads to the loss of important information about the overlap [9]. These methods fail to consider the node's role within its shared communities. For example, a node with connections with several communities based on "Membership" may be more influential than another with the same number of links because it belongs to large communities. Moreover, a node can pertain to one community more than another community. In addition, all these methods suffer from incomplete and fluctuating information. For instance, overlapping information may be available for a set of nodes, not all of them, and it may be available in a different format for these sets of nodes.

2 Overlapping Modularity Vitality

To solve these issues, we propose a flexible framework for identifying influential nodes in networks with an overlapping community structure called Overlapping Modularity Vitality [13]. It is based on the vitality concept [14]. Vitality assesses the influence of a node by looking at the variation of a given quality function when removing it from the network. Building on previous work [15], we extend Modularity Vitality to deal with various overlapping community structures. The proposed framework uses a generalized modularity quality function to model multiple overlapping community typologies, including crisp and fuzzy overlapping information. In crisp communities, a node has full membership in communities, while it can have partial membership in a fuzzy community structure.

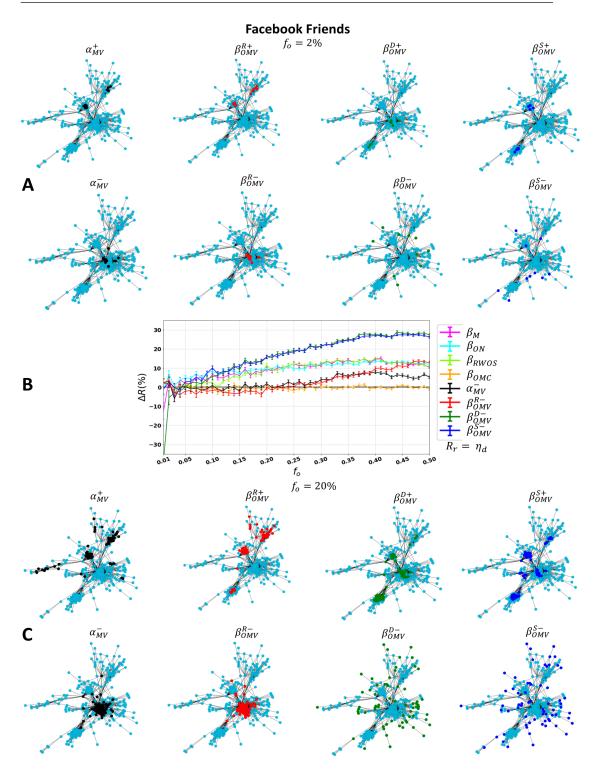
3 Experimental results

We investigate three definitions of overlapping modularity with different membership strengths, including reciprocity membership, degree membership, and node similarity. The measures differ in the information they use about the overlaps of communities and their ability to discriminate between nodes. Reciprocity membership is less discriminating since nodes may share the same number of communities, while node similarity and degree membership encode more nuanced differences. Additionally, since the framework is based on vitality, the Overlapping Modularity Vitality framework can have three ranking strategies for prioritizing hubs, bridges, or both. The effectiveness of the proposed framework is evaluated on 21 real-world networks using the SIR epidemic spreading scenario. Results show that the framework's performance depends on the available resources (f_o) , with different versions being more effective depending on a budget of nodes. Overall, results suggest that the distance between initially infected nodes is a good indicator of when enough resources are available. Furthermore, the framework that uses degree membership and node similarity with a bridges-first ranking scheme is most effective in this scenario, as seen in the Facebook friends network in Figure 1.

4 Conclusion

These findings demonstrate the benefit of incorporating overlapping community structure information to identify critical nodes in a network. The strength of the proposed framework lies in its capacity to integrate various types of structural information by using customized overlapping modularity measures. Indeed, as no one-size-fits-all definition applies to all real-world networks, a flexible approach is necessary to address different scenarios.

- 1. S. Fortunato, D. Hric, Physics reports 659, 1 (2016)
- 2. H. Cherifi, G. Palla, B.K. Szymanski, X. Lu, Applied Network Science 4(1), 1 (2019)
- 3. S. Rajeh, M. Savonnet, E. Leclercq, H. Cherifi, Quality & Quantity pp. 1–30 (2022)
- F. Reid, A. McDaid, N. Hurley, in *Mining Social Networks and Security Informatics* (Springer, Dordrecht, 2013), pp. 79–105
- 5. L. Hébert-Dufresne, A. Allard, J.G. Young, L.J. Dubé, Scientific reports 3(1), 1 (2013)
- D. Chakraborty, A. Singh, H. Cherifi, in Computational Social Networks: 5th International Conference, CSoNet 2016, Ho Chi Minh City, Vietnam, August 2-4, 2016, Proceedings 5 (Springer International Publishing, 2016), pp. 62–73
- M. Kumar, A. Singh, H. Cherifi, in Companion Proceedings of the The Web Conference 2018 (2018), pp. 1269–1275
- F. Taghavian, M. Salehi, M. Teimouri, Physica A: Statistical Mechanics and its Applications 467, 148 (2017)
- 9. Z. Ghalmane, C. Cherifi, H. Cherifi, M. El Hassouni, Scientific reports 9(1), 1 (2019)
- 10. S. Rajeh, M. Savonnet, E. Leclercq, H. Cherifi, IEEE Access 8, 129717 (2020)
- 11. N. Gupta, A. Singh, H. Cherifi, in 2015 7th international conference on communication systems and networks (COMSNETS) (IEEE, 2015), pp. 1–6
- 12. S. Rajeh, M. Savonnet, E. Leclercq, H. Cherifi, Scientific reports 11(1), 10088 (2021)
- 13. S. Rajeh, H. Cherifi, Social Network Analysis and Mining 13(1), 37 (2023)
- 14. D. Koschützki, K.A. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl, O. Zlotowski, in *Network analysis* (Springer, Berlin, Heidelberg, 2005), pp. 16–61
- T. Magelinski, M. Bartulovic, K. M. Carley, IEEE Transactions on Network Science and Engineering 8(1), 707 (2021). DOI 10.1109/TNSE.2020.3049068



French Regional Conference on Complex Systems – Le Havre, May 31 June 2, 2023

Fig. 1 The Facebook Friends network (A and C) with the nodes chosen to be initially infected by the vitality measures: Modularity Vitality (β_{MV}), and the three different versions of Overlapping Modularity Vitality, namely: reciprocity membership (β_{OMV}^R), degree membership (β_{OMV}^D), and node similarity (β_{OMV}^S). The measures use a hubs-first ranking scheme denoted by a "+" sign and a bridges-first ranking scheme marked by a "-" sign. The top two rows and the bottom two rows represent 2% and 20% of the fraction of initially infected nodes, respectively. In the middle (B), the figure represents the relative difference of the outbreak size (ΔR) as a function of the fraction of initially infected nodes (f_o) Facebook Friends. The reference centrality (R_r) is degree centrality (β_D). The centrality measures under test are: Membership (β_M), OverlapNeighborhood (β_{ON}), Random Walk Overlap Selection (β_{RWOS}), Overlapping M§44 lar Centrality (β_{OMC}), Modularity Vitality (β_{MV}^-), and the three different versions of Overlapping Modularity Vitality, namely: reciprocity membership (β_{OMV}^{R-}), degree membership (β_{OMV}^{D-}), and node similarity (β_{OMV}^{S-}).



Backbone Extraction of Weighted Modular Complex Networks based on their Component Structure

Sanaa Hmaida
1 $\,\cdot\,$ Hocine Cherifi
2 $\,\cdot\,$ Mohammed El Hassouni 1

Abstract This work introduces a generic backbone extraction framework exploiting the mesoscopic network structure. Indeed, numerous real-world networks are made of dense groups of nodes called communities, multi-core or local components. To deal with these groups' heterogeneity, we propose to extract the backbones independently from their various components and fuse them. Experimental investigations on real-world networks demonstrate the effectiveness of the proposed approach compared to the classical techniques' agnostic of the mesoscopic structure of real-world networks.

Keywords Community structure \cdot Component Structure \cdot Backbone extraction \cdot Multi-Core Structure.

1 Introduction

As data production grows, network analysis is becoming more challenging. Therefore, preserving relevant information while minimizing the network size is crucial. Backbones simplify the network's underlying structure, making identification easier for patterns, communities, and other essential network features. They have many applications in various fields, such as social, transportation, biological, and telecommunication networks. Researchers are developing new techniques to make networks smaller while maintaining their essential structure. The backbone extraction process identifies critical edges and nodes while removing irrelevant data [1,2,3,4,5,6,7,8,9]. One can distinguish two main approaches in backbones extraction techniques. Structural methods tend to remove nodes or links while preserving critical topological properties of the network [10]. In contrast, statistical techniques eliminate noisy edges or nodes based on statistical significance [11]. Here, we introduce an approach based on the mesoscopic properties of networks.

¹ FLSH, FSR, Mohammed V University in Rabat, Morocco E-mail: sanaa.hmaida.sh@gmail.com

 2 ICB UMR 6303 CNRS University of Burgundy, Dijon, France E-mail: hocine.cherifi@u-bourgogne.fr

E-mail: mohamed.elhassouni@flsh.um5.ac.ma

These node subgroups within the network are commonly apprehended through the community or the core-periphery structure. This work builds on the component structure recently introduced due to its adaptability and flexibility [12]. However, using alternative mesoscopic representations is straightforward. We propose a generic framework for backbone extraction that can use any backbone extraction technique developed for weighted networks. Experimental investigations with the influential threshold and Disparity Filter methods on real-world networks from various origins demonstrate the effectiveness of exploiting the component structure to extract the backbone.

2 Multiscale Backbone extraction framework

The backbone extraction framework is based on simple ideas. It is well-recognized that real-world networks contain various groups of nodes with different densities of interactions. They can be called communities or multi-core groups. A backbone technique ignoring this mesoscopic structure treats equally groups that can be quite different. Indeed, for example, the weights of the links in the various groups may vary over many orders of magnitude, making extracting the truly relevant connections challenging. To overcome this drawback, we propose tailoring the backbone extraction technique to the group of nodes rather than the entire network. The component structure is a very flexible representation for this purpose. Indeed, it splits the networks into two types of components: local and global components: The local components are the dense parts of the network. The intercommunity links and the nodes to which these links are connected form the global component. So local components are the networks' communities or cores, while the global components include the nodes and links joining the core or communities. The proposed framework contains three steps. First, one needs to build the network component structure. Then, one applies a backbone extraction technique to each component rather than the original network. The third step consists in merging the backbones of the global and local components to uncover the overall network backbone. This strategy allows the backbone extraction technique to adapt to the component's topology heterogeneity.

3 Experimental results

We perform an extensive experimental evaluation. It includes real-world networks selected from different domains with numbers of nodes and edges, ranging from hundreds to thousands, to cover a variety of situations. To evaluate the backbone extractor's effectiveness, we use classical measures such as average weighted degree, average link weight, average betweenness, weighted modularity, and qualitative comparisons. The component structure uncovering process is similar to the one proposed in the literature. 1) The Louvain algorithm allows the detection of the network community structure. 2) The communities form the local components; the global components contain all the nodes and links joining the communities. 3) The backbones of the components are extracted. We use the influential Disparity Filter and the classical global thresholding method. Indeed, our goal is to evaluate the

	Average Weighted	Average Link	Average Between-	Weighted Modularity
	Degree	Weight	ness	
Original Network	56.09	3.59	0.03	0.37
Classical Global Threshold	44.05	8.50	0.07	0.43
Multiscale Global Threshold	47.14	8.17	0.06	0.38
Classical Disparity Filter	6.94	1.24	0.04	0.15
Multiscale Disparity Filter	6.60	1.31	0.04	0.31

Table 1 The basic properties values calculated of Wind Surfers network backbones.

framework rather than the backbone extraction technique. 4) Finally, the union of the various backbones forms the overall network backbone.

In the experiments, we set the backbone size to 30% of the edges of the original network. Figure 1 illustrates the process with the Wind Surfers network. Firstly, we visualize at the top the original network and its community structure uncovered by the Louvain algorithm. The figures in the middle represent its component structure. It contains two local components corresponding to the two communities colored in pink and green, respectively, and one global component with nodes belonging to each community and the links joining these nodes. The figures at the bottom represent the classical and the multiscale backbone extracted using the Global Threshold algorithm. We observe that the multiscale backbone preserves all the components. In contrast, the classical backbone eliminates the global component. Indeed, only one link between node 1 and 7 is preserved, which connect the two local components. All the other links of the global component are filtered out because of their low weights. This figure illustrates that the Multiscale framework acts as a multi-threshold technique tuned to each component weight distribution rather than the overall network weight distribution.

Table 1 reports topological properties of the original network and the various backbones for the Wind Surfers network. It concerns the backbones extracted with the classical global threshold, the disparity filter algorithms, and their multiscale counterparts. The global threshold backbone is good at keeping high weights links. Results show that the Multiscale global threshold technique compares favorably with it for these related properties. Furthermore, its weighted Modularity is almost identical to the original network, demonstrating its ability to preserve the community structure of the original network. The classical Disparity filter preserves links with low weights better than the global threshold. Results show that this property is enhanced in its multiscale version. Once again, the community structure is better preserved. Indeed, whatever filtering technique is used, the global component is more impacted in the classical methods.

To summarize, the proposed framework effectively exploits real-world network heterogeneity, enhancing the advantages of the classical techniques. Future work will explore its association with more sophisticated backbone extraction techniques. Additionally, we plan to investigate the impact of alternative component structure detection algorithms.

Acknowledgements This research is carried out as a part of the project Plateforme logicielle d'intégration de stratégies d'immunisation contre la pandémie COVID-19 funded by the grant of the Hassan II Academy of Sciences and Technology of Morocco.

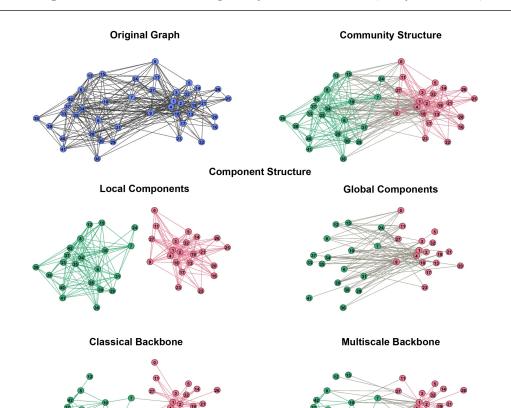


Fig. 1 The backbone extraction of 'Wind Surfers' network using Global Threshold algorithm.

- C.H. Gomes Ferreira, F. Murai, A.P. Silva, M. Trevisan, L. Vassio, I. Drago, M. Mellia, J.M. Almeida, Plos one 17(9), e0274218 (2022)
- 2. Z. Ghalmane, C. Cherifi, H. Cherifi, M. El Hassouni, Scientific Reports 10(1), 1 (2020)
- 3. Z. Ghalmane, C. Cherifi, H. Cherifi, M. El Hassouni, in 9th International Conference on Complex Networks and Their Applications (2020), pp. p-3
- 4. V. Gemmetto, A. Cardillo, D. Garlaschelli, arXiv preprint arXiv:1706.00230 (2017)
- 5. Z. Ghalmane, C. Cherifi, H. Cherifi, M. El Hassouni, Information Sciences 576, 454 (2021)
- S. Rajeh, M. Savonnet, E. Leclercq, H. Cherifi, in Network Science: 7th International Winter Conference, NetSci-X 2022, Porto, Portugal, February 8-11, 2022, Proceedings (Springer International Publishing Cham, 2022), pp. 67-79
- A. Yassin, H. Cherifi, H. Seba, O. Togni, in 2022 IEEE Workshop on Complexity in Engineering (COMPENG) (IEEE, 2022), pp. 1–8
- A. Yassin, H. Cherifi, H. Seba, O. Togni, in Complex Networks and Their Applications XI: Proceedings of The Eleventh International Conference on Complex Networks and their Applications: COMPLEX NETWORKS 2022—Volume 2 (Springer International Publishing Cham, 2023), pp. 551–564
- 9. L. Dai, B. Derudder, X. Liu, Journal of Transport Geography 69, 271 (2018)
- S. Rajeh, M. Savonnet, E. Leclercq, H. Cherifi, in *Network Science*, ed. by P. Ribeiro, F. Silva, J.F. Mendes, R. Laureano (Springer International Publishing, Cham, 2022), pp. 67–79
- A. Yassin, H. Cherifi, H. Seba, O. Togni, in *Complex Networks and Their Applications XI*, ed. by H. Cherifi, R.N. Mantegna, L.M. Rocha, C. Cherifi, S. Micciche (Springer International Publishing, Cham, 2023), pp. 551–564
- I.M. Diop, C. Cherifi, C. Diallo, H. Cherifi, Applied Network Science 6(1), 92 (2021). DOI 10.1007/s41109-021-00430-2. URL https://hal.science/hal-03560111

Innovation Diffusion



The impact of heterachical ties on information diffusion Sasha Piccione and Marco Tolotti
A pattern of diffusion of artificial intelligence in science: the development of an AI scientific specialty in neuroscience Sylvain Fontaine, Floriana Gargiulo, Michel Dubois and Paola Tubaro
Junk science bubbles and the abnormal growth of giants Floriana Gargiulo, Tommaso Venturini and Antoine Houssard
Unpacking popularity: volume, longevity, connectivity and glob- ality Mariana Macedo, Melanie Oyarzun and Cesar A. Hidalgo



The impact of heterarchical ties on information diffusion

Sasha Piccione · Marco Tolotti

Abstract This paper positions itself in the stream of literature that complements the classical theories of knowledge transfer, acquisition, and absorption with the studies of networks and their properties. The novelty that this paper brings is twofold. On the one hand, by relying on the concept of Simmelian ties, we further the discussion regarding the role that the strength of a tie plays in the transfer of knowledge and information across a network. In particular, by considering the organizational structure of a company and the personal relationships of the employees of the company, we want to study the transfer of information between subjects belonging to two different hierarchical levels and the role that Simmelian ties play. On the other hand, we enrich the classical innovation diffusion and opinion dynamics models with a characterization that is linked with the structural characteristics of the network of the population analyzed. Eventually, we propose a model that accounts for the structural characteristics under discussion. Simmelian ties opinion dynamics innovation diffusion knowledge transfer multilayer networks complex networks

Keywords Simmelian ties \cdot Opinion dynamics \cdot Innovation diffusion \cdot Knowledge transfer \cdot Multilayer networks \cdot Complex networks

1 Extended abstract

In the last decades, there has been a long debate about the role of agent's characteristics and structural properties on the type and quality of communi-

Sasha Piccione

Department of Management, Ca' Foscari University of Venice, Cannaregio 873, 30121 Venice, Italy

E-mail: sasha.piccione@unive.it

Marco Tolotti Department of Management, Ca' Foscari University of Venice, Cannaregio 873, 30121 Venice, Italy E-mail: tolotti@unive.it

cation among actors in complex organizations. More specifically, a segment of the literature has been trying to complement the classical theories of knowledge transfer with the study of networks and their structural properties ([1]; [2]). Scholars have tackled this issue under several angles, mostly focusing on that may facilitate or hinder the transfer of knowledge within an organization, such as sparseness, strength of tie and local topology.

For instance, the literature that studies the factors facilitating/impeding the transfer of knowledge within an organization has focused the efforts on different levels of analysis, i.e., on the characteristics of the ego subject [3], on the dyadic level [4], on microstructural aspects [5] and on the characteristics of the whole network [6]. Moreover, it has been recognized that outputs of communication effort may depend on whether the transferred knowledge is uncodified [7], implicit [8] or diversified [9]

Grounding on this literature, our aim is to dig deeper on the role of trust/distrust between two social parties within a company (*i.e.*, superior and subordinates). To this aim, in this study we focus our attention on the "verticality" and "heterarchicity" of communication. These traits are clearly related to the characterization of communication (and knowledge transfer) between one employee and its superior (or subordinate).

A second specific miscrostructural conformation on which we put attention is the "simmelianity" of ties [10]. A Simmelian tie is a tie between two subjects who have a third relationship in common, i.e. a tie within a triad [11]. Simmelian ties are considered to be strong ties due to the reciprocity, the symmetry and transparency that characterize such ties [10] and the consequent social forces that comes into play [5]. Simmelian ties are deemed to be interesting for knowledge and informaton transfer studies as they favor the development of a "common language and shared understanding among the parties involved" [2]regardless the distance between subjects [12]. Having to deal with the impact of informal ties on knowledge transfer across a formal organization, we intend to tackle such issue by adopting a multilayer approach.

From a methodological viewpoint, we start from the pioneering works by [13] in the context of opinion dynamics and its subsequent adaptations to innovation diffusion [14] and propose a mathematical model that not only takes into consideration the relational differences for what regards the trust between subjects, but also for what regards the willingness and the likelihood of an interaction between two subjects.

Let us consider a group of individuals, represented by the presence of a formal and informal network. Let $G(\mathcal{N}, \mathcal{R})$ be the network representing the informal relationships and let $G^F(\mathcal{N}, \mathcal{E})$ be the formal network where the set of members is the same for G, whereas $R \neq E$. Moreover, we assume the formal network to be a directed tree graph; therefore, E = N - 1. As far as R is concerned, it is composed by actual ties [11]. Consequently, G is an undirected graph. Let us now assume that such group has diverse opinions regarding a specific topic represented by a vector of opinions that vary in time $\mathbf{p}(t) = (p_0(t), ... p_n(t), ... p_N(t))$, where $p_n(t) \in [0, 1]$ and $\mathbf{p}(0)$ is the vector that represents the set of initial opinions. Each subject updates her opinion according to the trust she poses on her neighbours opinion $T_{i,j}$. The row stochastic matrix **T**, represents the overall set of "trust" ties within the sample.

The beliefs of the sample are updated at each time period so that:

$$\mathbf{p}(t) = \mathbf{T}p(t-1) \tag{1}$$

Eventually, if the network that is analyzed is strongly connected and is aperiodic [13], a consensus (p^*) is reached. Trust is not only relevant for opinion dynamics, but also for the transfer of knowledge [15]. More recently, [14] has tried to model the diffusion of innovation by adapting the DeGroot model. The author, in fact, poses as an adoption condition an idiosyncratic threshold level γ_i , characterizing the willingness to adopt of actor *i*:

agent *i* adopts by time
$$t \iff p_i(t) \ge \gamma_i$$
. (2)

The DeGroot model as described in (1) does not tell us anything whether subject i decided to adopt the good or the idea. Interestingly, [14] proposes a solution by introducing a threshold and the condition expressed in (2). Similarly to studies of adoption threshold theory, in this research we assume that a subject will influence the others regarding the value of a good if the value she attributes is sufficiently high (Valente, 1996). Additionally, we want to go over the base assumption that affirms that a subject will influence all its neighbours at each generic time step t. We, in fact, assume that the willingness of a subject to interact with one of her neighbours at a generic time step t determines the probability of an *information episode* [16] and depends on the specific relationship she has with that person. In order to account for such assumption, we must introduce a new condition, *i.e.*, that the transfer of information between i and j at a generic time step t depends on a specific probability $\xi_{i,j}$. The overall interactions that occur at the time step t are summarized by the matrix $\mathbf{B}(t) = (\beta_{i,j}(t))$, where $\beta_{i,j}(t)$ expresses whether the subjects i and j have interacted at the time step t. Additionally, similarly to [14], we assume that subject i will talk about a given topic or information if the relevance p_i regarding a such topic is higher than a idiosyncratic threshold γ_i . Therefore, $\beta_{i,j}(t)$ will depend on the probability $\xi_{i,j}$ and on whether p_i is above γ_i

More concisely the matrix $\mathbf{B}(t)$ is specified as follows for each $(i, j) \in \mathcal{R}$ and $t \ge 0$:

$$\beta_{i,j}(t) = \begin{cases} 1 & \text{with probability } \xi_{i,j} \\ 0 & \text{with probability } 1 - \xi_{i,j} \end{cases}$$
(3)

 $\xi_{i,j}$ depends on the characteristics of the tie between *i* and *j* and that $\beta_{i,i} = 1$ at any time step. Our assumption is that subjects from different hierarchical level are less prone to interact and to trust each other's belief ([15]). By including $\beta_{i,j}$, we are taking into consideration the fact that *i* and *j* are connected it does not necessarily mean that they are always communicating and that, in particular, the first will receive the information by the second.

The introduction of a stochastic element forces us to apply some adjustments to the Trust matrix so that the rows of resulting matrix sum to 1 at any time step. We, therefore, introduce a new matrix $\tilde{\mathbf{T}}(\mathbf{t})$ that represents the redistributed weights that the subjects pose on the persons they talked with at time t. Specifically, $\tilde{T}_{i,j}(t)$ is defined as follows:

$$\tilde{T}_{i,j}(t) = \frac{T_{i,j}\beta_{i,j}(t)}{\sum_{j=1}^{N} T_{i,j}\beta_{i,j}(t)}$$
(4)

Some remarks on $\tilde{\mathbf{T}}(\mathbf{t})$ are due. First of all, in (4)the numerator can either be equal to 0 or to $T_{i,j}$ while the denominator acts as a normalization factor that takes into consideration all the interactions that subject *i* had at time *t*. For instance, if *i* does not interact with any of her neighbour at time *t*, the trust she poses on her own opinion $(T_{i,i})$ will be equal to 1. On the contrary, if *i* interacts with all her neighbours at time *t*, $\tilde{T}_{i,j}(t) = T_{i,j}$. Secondly, $\tilde{\mathbf{T}}(\mathbf{t})$ is still a row stochastic matrix. Note that

$$\tilde{T}_{i,i} = \frac{T_{i,i}}{\sum_{j=1}^{N} T_{i,j}\beta_{i,j}}$$
(5)

and if all $\beta_{i,j} = 0$, $\tilde{T}_{i,i} = 1$.

Consequently, each subject will update her beliefs as follows:

$$p_i(t) = \sum_{j=1}^{N} \tilde{T}_{i,j}(t) p_j(t-1)$$
(6)

Eventually, the beliefs of the sample are updated at each time period so that:

$$\mathbf{p}(t) = \tilde{\mathbf{T}}(t)\mathbf{p}(t-1) \tag{7}$$

Differently from the classical DeGroot model, it is not possible to find \mathbf{p}^* as solution of the fixed point problem $\mathbf{p} = \tilde{\mathbf{T}}\mathbf{p}$, since $\tilde{\mathbf{T}}$ is not constant in time. Nevertheless, by definition:

$$\mathbf{p}(t) = \tilde{\mathbf{T}}(t) \cdot \tilde{\mathbf{T}}(t-1) \cdot \dots \cdot \tilde{\mathbf{T}}(1) \cdot \mathbf{p}(0)$$
(8)

Furthermore, if the system converges, we can approx \mathbf{p}^* as:

$$\mathbf{p}^* = \tilde{\mathbf{T}}(t^*) \cdot \tilde{\mathbf{T}}(t^* - 1) \cdot \dots \cdot \tilde{\mathbf{T}}(1) \cdot \mathbf{p}(0)$$
(9)

where (t^*) is such that

$$|\tilde{\mathbf{T}}(t^* - 1) - \tilde{\mathbf{T}}(t^*)| < \varepsilon$$
(10)

where ε is small enough.

In this sense, $\tilde{\mathbf{T}}(t^*)$ approximates $\tilde{\mathbf{T}}^{\infty}$ so then $\mathbf{p}(t^*)$ approximates \mathbf{p}^{∞} .

Therefore, we identify a generic row $\tilde{\mathbf{s}}(t^*)$ of the matrix $\mathbf{T}(t^*)$ as the social influence vector. This vector represents the overall impact that each subject had in the formation of the consensus. The $\tilde{\mathbf{s}}(t^*)$ vector helps answering the question 'who is the most influential?'. Focusing on the "equilibrium" social influence vector can be of interest for studies regarding centrality that aim

	n	ANOVA Model 1: Interaction probability 0.1	ANOVA Model 2: Interaction probability 0.3	ANOVA Model 4: Interaction probability 0.7	ANOVA Model 5: Interaction probability 0.9
NonSimmelian $_{deg4}$	80 000	0.0275	0.0270	0.0257	0.0255
N C:	45000	(0.0035) 0.0279	(0.0140) 0.0282	(0.0245) 0.0299	(0.0258) 0.0300
NonSimmelian $_{deg5}$	45 000	(0.0279) (0.0034)	(0.0282)	(0.0299)	(0.0305)
$Simmelian_{deg4}$	15000	0.0276	0.0269	0.0249	0.0247
G : 1:	10.000	(0.0036)	(0.0143)	(0.0239)	(0.0253)
Simmelian $_{deg5}$	40 000	$0.0280 \\ (0.0038)$	$0.0289 \\ (0.0164)$	$0.0305 \\ (0.0296)$	$0.0307 \\ (0.0313)$

Table 1: ANOVA Analysis on lattice network

at understanding what are the most relevant positions within a network. In particular, as we will see, when stochasticity is introduced, some actors can play an interestingly relevant role even though they would not in the case stochasticity is ignored.

In particular, we intend to focus on analyzing who, by virtue of their position within the network and the topological characteristics of the neighbourhood, appears to be more influential. Answering such question can be of particular relevance at furthering the discussion regarding of seeding [18] and company communication. In particular, considering the relevance that Simmelian ties have on the transfer of information, we want to explore whether such relevance is caused by the topological characteristics or whether it is the result of social forces. Also, we want to explore whether and the role of Simmelian ties changes adding social context, *i.e.* differences in terms of trust and willingness to share information depending on hierarchical positions

The analysis is carried out on computer simulations utilizing *ad hoc* networks like the Kite network [11] and lattice networks. The results as summarized in Table 1.

In Table 1 we can see the statistics regarding the simulations on a modified lattice network. For instance, in this specific experiment we wanted to see whether *ceteris paribus* there were specific topological conformations that would allow subjects to be systematically more influential than the others. Also, we can see the results of the ANOVA analysis carried out on the entire sample and also on samples grouped by the degree. As it can be seen, the degree of the node is, clearly, a driving force. For instance, there are some changes in the means and medians of influence in each group. The subjects with a higher degree are characterized by a positive increase in terms of means while the others by a decrease. Interestingly, such changes are more marked in the case of subjects who have, at least, one Simmelian tie. For instance, as the probability of interaction grows, the difference between groups becomes more net. Differently from what we expected, the "Simmelianity" does not imply that all subjects have an higher influence in the consensus formation. In

fact, it appears that subjects who have, at least, one Simmelian tie and has a lower degree than the neighbours "suffer" from such a condition. In particular, being a subject that has one connection with a Simmelian tie appears to be more beneficial for the subjects who have a higher degree. Nevertheless, it is unclear the magnitude of the role played by the degree of the nodes and by the Simmelianity of the ties. On this point, further investigation is necessary. With our model we aim at overcoming the assumption according to which all the subject talk with their neighbours at each time step by introducing the $\mathbf{B}(t)$ matrix which represents the interactions that actually occur at time t. By overcoming such assumption, we highlight how deciding the proper seeding point can be of crucial relevance if the aim is to achieve the widest spread of information or knowledge. By introducing stochasticity, we allow the representation of a wider spectrum of possible outcomes. Also, the scope of the paper is to explore the understanding of which topological characteristics of the ego subject and her neighbour can be more or less affected by the introduction of stochasticity.

- Hansen M. T.: The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunit. Administrative Science Quarterly, Vol. 44: 82-111 (1999)
- Tortoriello M. and Krackhardt D.: Activating Cross-boundary Knowledge: the role of Simmelian Ties in the Generation of Innovations. Academy of Management Journal, Vol. 53(1): 167-181 (2010).
- 3. Haas A.: Crowding at the frontier: boundary spanners, gatekeepers and knowledge brokers. Journal of Knowledge Management, Vol. 19(5): 1029-1047 (2015).
- Friedkin N. E.: Information Flow Through Strong and Weak Ties in Intraorganizational Social Networks. Social Networks, Vol. 3: 273-285 (1982).
- Solorzano M. G., Tortoriello M. and Soda G.: Instrumental and affective ties within the laboratory: The impact of informal cliques on innovative productivity. Strategic Management Journal, Vol. 40: 1593-1609 (2015).
- 6. Alguezaui S. and Filieri R.: Investigating the role of social capital in innovation: sparse versus dense network. Journal of Knowledge Management, Vol. 14(6):891-909 (2010).
- 7. Reagans R. and McEvily B.: Network Structure and Knowledge Transfer: effects of Cohesion and Range. Administrative Science Quarterly, Vol. 48(2): 240-26 (2003).
- 8. Fritsch M. and Kauffeld-Monz M.:The impact of network structure on knowledge transfer: an application of social network analysis in the context of regional innovation networks. Annual Regional Science, Vol. 44: 21-38 (2010).
- Gargiulo M., Ertug G. and Galunic C.: The Two Faces of Control: Network Closure and Individual Performance among Knowledge Workers. Administrative Science Quarterly, Vol. 54: 299-333 (2009).
- Simmel G.: The Sociology of Georg Simmel. Translated and Edited by Wolff K. H., Collier McMillan, Ontario (Canada) 1950.
- Krackhardt D.: Simmelian Tie: Super Strong and Sticky. In Roderick Kramer and Margaret Neale (eds.). Power and Influence in Organizations. Thousand Oaks, CA: Sage, pp. 21-38 (1998)..
- Tortoriello M., McEvily B. and Krackhardt D.: Being a Catalyst of Innovation: The Role of Knowledge Diversity and Network Closure. Organization Science, Vol. 26(2): 423-438 (2015).
- DeGroot M.H., Reaching a Consensus. Journal of the American Statistical Association, Vol. 69(345):118–121 (1974).

- 14. Assenova V. A.: Modelling the diffusion of complex innovations as a process of opinion formation through social networks. PLoS ONE, Vol. 13(5): 1-18 (2018).
- 15. Rutten W., Blaas-Franken J. and Martin H.: The impact of (low) trust on knowledge sharing. Journal of Knowledge Management, Vol. 20(2): 199-214 (2016).
- Yale L. J. and Gilly M. C.: Dyadic Perceptions in Personal Source Information Search. Journal of Business Research, Vol. 32: 225-237 (1995).
- 17. Tasselli S. and Caimo A.: Does it take three to dance the Tango? Organizational design, triadic structures and boundary spanning across subunits. Social Networks, Vol. 59: 10-22 (2019).
- Peres R., Muller E. and Mahajan V.: Innovation diffusion and new product growth models: A critical review and research directions. International Journal of Research in Marketing, Vol. 27(2): 91-106 (2010).



A pattern of diffusion of artificial intelligence in science: the development of an AI scientific specialty in neuroscience

Sylvain Fontaine¹ · Floriana Gargiulo¹ · Michel Dubois¹ · Paola Tubaro²

Abstract Artificial intelligence (AI) commonly refers to both a research program and, more generally, a set of complex computer-based programs which aim to mimic human mind processes with high reckoning power. These algorithms are amenable to applications in a variety of disciplines which use them for scientific advancements and which sometimes improve them for their conceptual and methodological needs. From this assertion we formulate the hypothesis that AI is an "adjacent possible" in Kauffman's sense [1], that is, AI first emerged in a specific scientific and technological context thanks to a combination of existing knowledge or innovations, and it is now expanding by blending with other novelties from specific disciplines and by reshaping some disciplinary practices and knowledge structures.

This work intends to address the process of diffusion of AI in neuroscience in a scientometric manner within this framework, namely how some AI-related knowledge have been brought from different scientific disciplines or specialties into neuroscience, and then how they are used. The corresponding underlying dynamics can be captured through both the disciplinary ecosystem around the whole neuroscience and the structure of collaborations among the scientists involved in this field of research.

So we first build a neuroscience bibliometric corpus extracted from the Microsoft Academic Knowledge Graph database [2], in which we separate AI-related articles from non-AI ones with a dedicated keywords' filter applied to the titles and abstracts, that is provided by Gargiulo et al. [3]. This corpus includes scientific articles published between 1970 and 2019 in peer-reviewed journals that are referenced by the Web Of Science. Then we explore specifically the development of a dedicated AI specialty in neuroscience with its own scientific community by conducting analyses of both the egocentric citation

 $^{^1}$ GEMASS, CNRS-Sorbonne Université, 59-61 rue Pouchet, 75017 Paris

 $^{^2}$ CREST, CNRS-IPP, 91120 Palaiseau

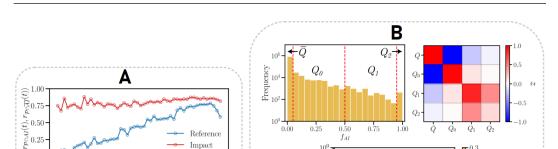
Corresponding email: sylvain.fontaine@cnrs.fr

network around these neuroscience articles and the associated co-authorship network.

From the citation network we first propose an aggregated cartography of the main fields of research cited by these articles, as well as those that are most impacted by them just one year after publication. It illustrates a particular evolution of both AI-related references and generated citations that is not borrowed by the core of the neuroscience literature, namely a progressive specialization of references toward computer science, mathematics and engineering, while its impact is more broadly distributed across the entirety of the neuroscience field (see Fig.1A), mainly into engineering, radiology and neuroimaging technologies for clinical research and medicine. This preliminary analysis especially indicates a growth of such references and generated citation a around the 1990's just after the second AI winter.

To reinforce these findings, the time-aggregated co-authorship network, including the main collaborations since 1970, first exhibits a small set of researchers (around 1.3% of them) that together have authored the most publications in AI research in peer-reviewed journals oriented toward computational neuroscience, mostly based on neural networks, and that does not maintain links with the rest of neuroscience community which does not publish AI research (see z-scores heat-map on Fig.1B). By considering the disciplinary backgrounds of these 'AI specialists', we divide this subset in two groups, namely a first one including authors who are trained in the main disciplines that are shaping neuroscience since the 1940's and who have a low AI activity (Q_0 in Fig.1B), and another smaller one including authors who are trained in other disciplines that are not represented in the first group, such as computer science and engineering $(Q_1 \text{ and } Q_2 \text{ in Fig.1B})$. More precisely, the second group is made up of authors who are not fully involved in the neuroscience field and who are keeping to publish within their original disciplines. These results have led us to qualify the AI scientists as 'outsiders' of neuroscience [4].

According to the state of the art of the formation of a scientific specialty encountered in science studies [5], this work thus shows through AI in neuroscience a pattern of diffusion of knowledge in a scientific field of research, namely a formation and a transformation of a special AI-research alongside neuroscience, with its own evolving scientific community and bibliographic references, and which seems though to contribute to the main challenges of this field of research. It aims to open some research directions within science studies. A first work in progress is pushing further the idea of the development of an AI specialty in neuroscience by mobilizing other indicators which would assess the common cognitive structure of its members (vocabulary used in the articles, specific patterns of co-citations) and the institutionalization of such a specialty in academia or industry (main research centers, conferences and scientific societies around the world, the associated university training tracks and the accumulated grants in this domain). Another worth study would address the issues about interdisciplinarity and the potential diffusion of AI-related knowledge and technologies outside neuroscience. In particular, the impact of AI-related works in conference proceedings, which are not considered here but



0.00

Reference Impact

2020

2010

0.25

10

10 f_{AI} 10

0.50

0.75

 10^{-1} a_{AI}

1.00

Fig. 1 A: Temporal Jaccard index between AI and non-AI rankings of disciplines appearing in references (blue) and generated citations (red) in the neuroscience corpus (P). It shows a progressive uniformization of references in both corpora while the generated citations keep to be similar in the studied period. B: Top left: Distribution of the share of AI-related publications per author (left), separated in quartiles. Top right: Z-score matrix of the number of edges shared between the quartiles compared to randomized collaboration networks based on the same set of authors. Bottom: Distribution of the studied neuroscience journals in the two-dimensional space defined by their activity in AI a_{AI} and the average AI activity f_{AI} of the authors who published in it up 2019. One point represent one journal. The plot indicates a quasi-linear relation between the two parameters in the highest values of a_{AI} , therefore a concentration of the most active authors in AI in mostly AI-oriented journals as well.

are widely favored in computer science-related disciplines, would be useful in measuring this aspect.

This preliminary work is thus a part of a larger study of the history of neuroscience through AI. It is also a first step towards a road map to investigate the diffusion of AI into other fields of research that are receptive to it, in order to observe more broadly the construction of AI in science.

Keywords Science of science · scientometrics · Artificial Intelligence · knowledge diffusion \cdot technological specialization

Acknowledgements S.F. is founded by the a PhD fellowship from the CNRS-MiTi program EpiAI. This work is partially founded by the ANR project ScientIA.

References

1970

1980

1990

2000

- 1. Stuart A. Kauffman, Investigations. Oxford University Press (2000)
- 2. Michael Färber, The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data. In: C. Ghidini et al. (Eds.), The Semantic Web - ISWC 2019 (Vol. 11779, p. 113-129). Springer International Publishing (2019)
- 3. Floriana Gargiulo, Sylvain Fontaine, Michel Dubois, Paola Tubaro, A meso-scale cartography of the AI ecosystem. Available in : http://arxiv.org/abs/2212.12263 (2022)
- 4. Diana Crane, Social Structure in a Group of Scientists: A Test of the "Invisible College" Hypothesis, American Sociological Review, 34-3, 335-362 (1969)
- 5. Daryl E. Chubin, The Conceptualization of Scientific Specialties, The Sociological Quarterly, 17-4, 448-476 (1976)



Junk science bubbles and the abnormal growth of giants

Floriana Gargiulo $\,\cdot\,$ Antoine Houssard $\,\cdot\,$ Tommaso Venturini

Abstract As we showed in [1], the typical property of online recommendation algorithms of reinforcing the attention on the latest trends can be responsible for the ever faster rise and fall of collective attention around online objects (i.e. YouTube videos), an effect that we call "junk news bubbles".

In an accelerating society, it is not only the consumption of online information that is becoming more ephemeral: a similar phenomenon has been observed for the citation cycles of scientific production [2] In the case of scientific production, however this acceleration may depend less on specific recommendation algorithms than on the practices of researchers for linking their work to the existing literature, the "shoulders of giants" on which the scientific innovation is supposed to stand.

In this context, indeed, other phenomena concerning the temporality of research has been observed: the consolidations of canons and the slowing down of disruptiveness (the low turnover of central ideas, the consolidation of scientific visibility produced by Matthew effects) [3,4].

We start to analyze these two apparently opposite phenomena, the acceleration of attention cycles and crystallization of canons, applying a varied set of statistical measures to several corpora, at different scales of granularity (from precise thematic domains, like i.e., astrophysics or cognitive science or artificial intelligence, to large domains identified in the Web of Science (WOS), i.e., life science or technology).

For collecting these corpora we use different techniques to query the OpenAlex platform: the papers in the large fields of study are selected searching

Floriana Gargiulo GEMASS-CNRS E-mail: floriana.gargiulo@ecnrs.fr

Antoine Houssard CIS-CNRS

Tommaso Venturini CIS-CNRS the papers published in the journals that the WOS classify in that area. The papers in the thematic domains are collected using the concepts provided by OpenAlex.

We observe that in all the different corpora some trends are highly universal: first of all the citation inequality among the papers , measured by the Gini index (and confirmed by entropy measures), increases in time (In Fig1 this measure is displayed for the macro-areas of the WOS.). This is a first signal of the fact that the most cited papers have an even higher probability to be cited than it was in the past. We reinforce this measure calculating, year by year, the difference, in terms of citations, between the top cited papers and a random sample of the same size (normalized by the total number of papers published in the year). This measure provide another indication of the fact that the top papers acquired in time en even higher attractiveness.

At the same time we analyzed the "fertility" patterns of publications, analyzing the capacity of each publication to give rise to other papers. If we display, for each year, the number of different cited papers as a function of the total number of published papers, also including another well known fact that is the increasing number of references in the papers, we observe (Right plot of Fig.1 that, for all the disciplines, the growing trends slow down starting from the nineties.

In the left lower plot of Fig.1, using a modified jaccard index taking into account the position of the elements, we show another interesting effect connected to crystallization of canons: the top-k ranking for each domain remains more and more stable over time.

Finally we analyze the attention cycles of papers, showing a general tendency for early discovery (and the gradual disappearing of the sleeping beauty phenomenon) and early reach of the moment of maximum citations for random samples. Top papers are also discovered earlier but the time to the max increases in time (in agreement with what we observed before).

These effects that we observed in our corpora suggest the existence of a supercumulative advantage process promoting again and again the same classic papers and penalizing most of the novelties, that become more ephemeral. We argue that this effect can be directly reconducted to the exponential growth of the scientific literature.

We propose therefore a simple model to reproduce the citation patterns inspired, as [1], by the competition of objects in an attention arena. We show that an elevated growing trends of the number of objects make the less popular ones less competitive in the attention market and leave a higher dominance to the more popular ones.

Keywords Science of science, attention cycles, citation networks

^{1.} Castaldo, M., Venturini, T., Frasca, P., Gargiulo, F. Junk news bubbles modelling the rise and fall of attention in online arenas. new media & society, 24(9), 2027-2045 (2022)

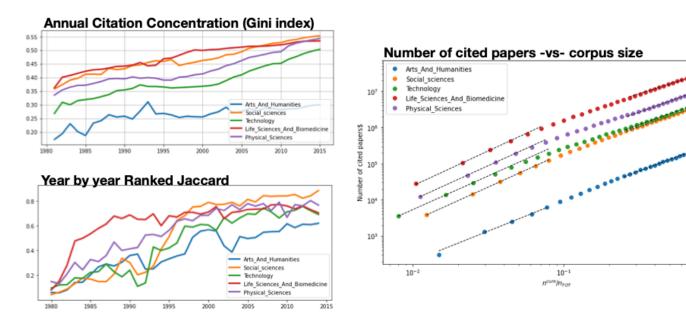


Fig. 1 Please write your figure caption here

- 2. Lorenz-Spreen, Philipp, et al. Accelerating dynamics of collective attention. Nature communications 10.1:1-9 (2019)
- 3. Chu, Johan SG, and James A. Evans. Slowed canonical progress in large fields of science. Proceedings of the National Academy of Sciences 118.41 (2021)
- 4. Park, Michael, Erin Leahey, and Russell J. Funk. Papers and patents are becoming less disruptive over time. Nature613.7942: 138-144 (2023)



Unpacking popularity: volume, longevity, connectivity and globality

Mariana Macedo $\,\cdot\,$ Melanie Oyarzun $\,\cdot\,$ Cesar A. Hidalgo

1 Abstract

Measures of citations, fame, and popularity are used frequently as proxies for the quality of scientific and cultural work to study the dynamics of success, attention, and collective memory [1-3]. The advent of the internet has made it possible to estimate popularity online by measuring the overall quantity of attention through searches, video reproductions, and page views [4–6]. Yet, raw measures of attention often conflate multiple forms of popularity. To address this issue, we unpack popularity in a framework focused on four dimensions: (i) volume, (ii) longevity, (iii) connectivity, and (iv) globality. These dimensions are computed from the page views and demographic information on Wikipedia of the 100k most famous people collected by the Pantheon database in 2020 [7]. Volume refers to the overall quantity of attention measured by page views. Longevity captures the temporal extent of attention [3]. Connectivity captures how people are related to other people's fame, measured by the number of times that a person was mentioned on other pages of Wikipedia [8]. Lastly, globality captures how attention varies across space, measured by the diversity and concentration of attention across languages [9, 10].

We find that globality, connectivity, and longevity explain 66%, 20%, and 3% of the variance in volume, respectively. Taken together, the globality, connectivity, and longevity dimensions explain 76% of volume, after controlling for individual characteristics such as occupation, gender, age and nationality. Our results are also robust while considering the creation date of the wikipedia pages.

Melanie Oyarzun

Mariana Macedo and Cesar A. Hidalgo

Center for Collective Learning, ANITI, University of Toulouse, FR E-mail: mmacedo@biocomplexlab.org, cesifoti@gmail.com

Centro de Investigación en Complejidad Social (CICS), Facultad de Gobierno, Universidad del Desarrollo, Chile. E-mail: moyarzunw@udd.cl

Figure1 illustrates our decomposition for four famous actors and soccer players. For instance, the actress Meryl Streep has the highest globality, connectivity, and volume, and Henry Irving has the highest longevity. Although Irving, an English stage actor born in 1838, is still famous in our sample, he is not as famous as Meryl Streep. For the soccer players, Ahmed El Shenawy is a young Egyptian player who has only played professionally in his country scores low on globality and longevity, but relatively high on volume indicating he is a local famous player. Thus, each dimension contributes to understanding the heterogeneities of popularity across occupations and generations. Our findings take one step further in mapping the historical geography of fame and provide a more nuanced and comprehensive framework for measuring popularity.

Finally, we state that our model and findings are limited to the collected data, and, the four unpacked dimensions, even though normalized, their scales are not comparable, and the area of a polygon varies across occupations (being a famous soccer player is not the same as a famous actor).

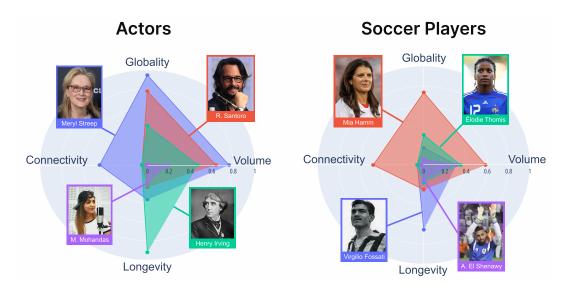


Fig. 1 Dimensions of popularity: volume, longevity, connectivity and globality.

References

- S. Fortunato, C.T. Bergstrom, K. Börner, J.A. Evans, D. Helbing, S. Milojević, A.M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, A.L. Barabási, Science **359**(6379), eaao0185 (2018). DOI 10.1126/science.aao0185. URL https://www.science.org/doi/abs/10.1126/science.aao0185
- F. Wu, B.A. Huberman, Proceedings of the National Academy of Sciences 104(45), 17599 (2007)
- C. Candia, C. Jara-Figueroa, C. Rodriguez-Sickert, A.L. Barabási, C.A. Hidalgo, Nature human behaviour 3(1), 82 (2019)
- 4. W. Lewoniewski, K. Wecel, W. Abramowicz, Inf. (2020). DOI 10.20944/preprints202003.0460.v1

- K. Abbas, M. Shang, A. Abbasi, X. Luo, J.J. Xu, Y.X. Zhang, Scientific reports 8(1), 1 (2018)
- 6. F. Ogushi, J. Kertész, K. Kaski, T. Shimada, Scientific Reports $\mathbf{11}(1),\,18371$ (2021)
- 7. A.Z. Yu, S. Ronen, K. Hu, T. Lu, C.A. Hidalgo, Scientific data **3**(1), 1 (2016)
- P. Beytía, J. Schobin, Research Data Journal for the Humanities and Social Sciences 5(1), 50 (2020)
- Y. Gandica, R. Lambiotte, T. Carletti, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 10 (2016), vol. 10, pp. 43–46
- D. Jemielniak, M. Wilamowski, Journal of the association for information science and technology 68(10), 2460 (2017)

Socio-Technical Systems



Integrated bi-objective model for berth scheduling and quay crane assignment with transshipment operations
Marwa Samrout, Adnan Yassine and Abdelkader
Sbihi
Network structures of a centralized and a decentralized market. A direct comparison.
Sylvain Mignot and Annick Vignes
Alignment of Multinational Firms along Global Value Chains: A network-based perspective Charlie Joyez
Is Conservation Agriculture the Future of Farming in France? Damien Calais
Let's Tweet about Soccer? A Gender-centric Question
Akrati Saxena and Mariana Macedo
From geographic data to spatial knowledge in agent-based mod- eling applied to land use simulation Severin Vianey Kakeu Tuekam, Eric Fotsing and
Marcellin Julius Antonio Nkenlifack

Integrated bi-objective nmodel for berth scheduling and quay crane assignment with transshipment operations

Marwa Samrout · Adnan Yassin · Abdelkader Sbihi

Received: date / Accepted: date

Abstract A container terminal is a complex system where different service level conditions are required for different vessels and customers. Thus, simulating terminal processes is often a complex task requiring a modeling approach that facilitates decision analysis. In this study, we first propose to model the berth allocation (BAP) integrated with the quay crane assignment (QCAP). A new mixed integer bi-objective linear program is proposed to reduce the dwell time of each ship, the penalty by late ships, and to find the optimal number of quay cranes (QCs) needed per ship. Then, we develop a resolution procedure based on the non-dominated genetic sorting algorithm (NSGA-III). We also use statistical analysis to identify the control parameters of (NSGA-III). Then, we implement the algorithm calibrated with the obtained control parameters. We conduct a computational experiment on a set of large randomly generated instances to highlight the benefits and suitability of the proposed approach. The numerical results show the efficiency of the approach.

Keywords Berth Allocation · Quay Crane Assignment · Container Terminal · Bi-Objective Optimization · Transshipment · Scheduling

1 Introduction

Container terminals form by their nature focal nodes at the heart of the dense freight flow, being a crucial part of the supply chain network. They tend to be the most complex environments within the sector of international transportation. They are classified into four main subsystems in the literature (see Fig 1): (1) port operations (Ship-to-Shore subsystem), (2) transport/transfer operations (Horizontal-Transport/ Transfert subsystem), (3) storage operations (Storage subsystem) and (4) Delivery-receipt subsystem operations. Basically, there are two routing options for unloaded shipping containers ([15]): (i) to transport them to a storage area while awaiting the arrival of their exit modes of transport (indirect transshipment), and (ii) to carry out a direct transfer towards a mode of transport without passing by the storage zone (direct transshipment). This form of transfer was once widespread but is now rarely used, because synchronizing the arrival of hinterland modes with ship unloading operations is very complex. The unloading and loading tasks are guided by an operations plan drawn up before the ships' arrival. It contains a sequence of containers to be handled in order of priority, i.e., unloading precedes loading, unloading the deck before the hold and loading the reverse. The first step in the unloading process is to pick up the container by a crane. Once the unloading ends, the ship is allowed to leave the terminal.

2 Literature Review

The integrated Berth Allocation and Crane Assignment Problem (BACAP) has lured the interest of many investigators in the domain over the last decade. [8] were the pioneers. They developed an MIP for a static-continuous variant of this problem and used a Lagrangian relaxation for the solution. A discrete quay with static vessel arrivals aspect were considered by (were the pioneers. They developed an MIP for a static-continuous variant of this

Marwa Samrout

Adnan Yassin

Abdelkader Sbihi

Normandie Univ, UNIHAVRE, LMAH, FR CNRS 3335, ISCN, 76600 Le Havre, France E-mail: marwa.m.samrout@gmail.com

Normandie Univ, UNIHAVRE, Institut Supérieur d'Etudes Logistiques, 76600 Le Havre, France.

University of South-Eastern Norway, Campus Kongsberg, Kongsberg, Norway

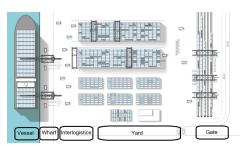


Fig. 1 Description of the operations involved to move containers in and out of the container terminal

problemand used a Lagrangian relaxation for the solution. A discrete quay with static vessel arrivals aspect were considered by ([6],[7]) as both a mono and multi-objective approach. In both cases, they used a GA solve the problem. as both [10] gave an enhanced MILP of continuous quay with dynamic arrivals of ships to minimize: (i) the total charging/discharging time, (ii) the penalty term due to an undesired location, and (iii) the number of cranes needed to handle vessels. [5] adopted the concept of "profiles" while assigning cranes to ships. The authors developed a Biased Random Key GA. [9] built a new model of the Discrete and Dynamic BAP (DDBAP). They proposed a Bee Colony Optimization (BCO). [3] devised three distinct formulations for the integrated continuous Berth Allocation and Quay Crane Assignment Problem (BACAP). A self-adaptive meta- heuristic algorithm to find solutions for a large-scale problem. The priority of the entering ships was determined by a powerful composite dispatching rule through three objectives: maximizing the use of the (QCs) and minimizing the total service time and charges. Comparative studies were performed for large-scale problems based on the literature. Recently, [11] have proposed a new MILP to optimize the berthing schedule and build a transshipment connections planning between feeder and mother vessels to minimize vessels dwell times, the late penalty and decide the mode of transshipment needed

3 Problem description

The interactions between several entities of varied nature evolving in a lower level in a CT result in the dynamics of complex behaviors. Therefore, to better understand the functioning of the system it necessary to understand the role and the phenomena resulting from the actions produced by these entities. Our research work combines two seaside operations namely the BAP and the QCAP considering the tasks of direct and indirect transshipment. one Our aim consists to assign a berthing position and a berthing time to all the incoming vessels. Another purpose of this study is to make connections between transshipment vessels, which mean to choose the best possible transshipment mode between two vessels. The objectives are: (1) to minimize the dwell time for each vessel and the penalty of tardiness and, (2) to minimize the number of (QCs) used. The paper expands the relative position formulation for a (BAP) ([11]). We integrate the (QCAP) with ship-to-ship transshipment consideration to simulate the effective interaction between these operations. The issue lies in finding the right balance between these objectives that is, on one hand, the higher the number of cranes, the longer the time of ship processing is reduced, on the other hand, as we have a limited number of cranes, the optimized resource management must be applied in CT for a high performance. To formulate the problem, we consider a quay divided into B equal-sized sections. Only one vessel can be assigned to a specific berth at a specific moment. Segments of 4 hours form the time horizon. The originality of our model lies in adding new constraints that extends the model of [2] as follows:

- A couple of coming vessels breed only one kind of transshipment operation including direct or indirect transshipment.
- Only the couple of vessels arriving at the same time can transfer directly their containers.
- Every bay is served by exactly one QC.
- A safe distance is kept between the (QCs) from each other when functioning simultaneously.
- The middle-indexed cranes cannot be positioned at end bays.
- A lower-indexed (a higher-indexed) crane will not be deposited to the right of a higher-indexed crane (the left of a lower-indexed crane) ([12]).

The transshipment and the QCs assignment operations were modeled using the following decision variables:

$$D_{ij} = \begin{cases} 1 & \text{if there is a direct transshipment between vessel } i \text{ and vessel } j \\ 0 & \text{otherwise;} \end{cases}$$

$$I_{ij} = \begin{cases} 1 & \text{if there is an indirect transshipment between vessel } i \text{ and vessel } j \\ 0 & \text{otherwise;} \end{cases}$$

$$C_{ug} = \begin{cases} 1 & \text{if crane } g \text{ is assigned to bay } u \\ 0 & \text{otherwise;} \end{cases}$$

4 Non-dominated Sorting Genetic Algorithm III (NSGA-III)

The basic structure of our NSGA-III is similar to the original NSGA-II [13] with substantial modifications in its selection operator. Unlike the NSGA-II, the diversity amid population members in NSGA- III is ensured by adaptively upgrading a number of well-distributed reference points. The selected reference points can either be preordained in a structured way or supplied by the user. In this paper, we use [14]'s approach that sets points on a normalized hyper-plane, an (M-1)-dimensional unit simplex which is evenly inclined to the two objective axes and cross each axis at one intercept. The initial population of the implemented NSGA is filled with random solutions. The crossover operators employed to solve the proposed (BQCAP) are: the Whole Arithmetic Recombination (WAR), the Single Point Crossover (SPC), the Double Point Crossover (DPC) and the Uniform Crossover (UC). In this work, we have two fitness functions, each is composed by an objective function and a penalty function. To enhance the NSGA performance, we have calibrated parameters values. These include the population size, the crossover and mutation probabilities. To tune the best parameters settings, we computed the mean position of results obtained by each parameters combination, then making an assessment where the bigger average position is 2 better. Several parameter settings were examined (Table 1). We run the algorithm 30 times for each parameter with 81 combinations and 150 iterations. Table 2 shows the tested sets that lead to the best computational results. The best population size is obtained for 100. Also, the most effective crossover and mutation probabilities are obtained for 0.7 and 0.0001 respectively. We notice that all the mutation rates can produce satisfactory results. Typically, we use 3 levels for each of the 4 tested factors. With these values, the algorithm showed quite promising results. Obviously, as the number of levels increases, the number of possible combinations of parameters in a factorial experiment increases rapidly; for instance, four factors each at four levels would demand 256 trials. This rapidly becomes impracticable from the viewpoint of time and other resources.

Population size (nPop):	100,	200,	350
Crossover percentage (pc):	0.7,	0.8,	0.9
Mutation percentage (pm):	0.0001,	0.2,	0.5
Mutation rate (mu):	0.1,	0.01,	0.001

Table 1 Evolutionary algorithm's parameters values tested

Combination N°	pc	pm	mu	nPop	Avg position
3	0.7	0.0001	0.001	100	56.4661
2	0.7	0.0001	0.01	100	56.6284
1	0.7	0.0001	0.1	100	57.0413
20	0.9	0.0001	0.01	100	59.9840
11	0.8	0.0001	0.01	100	63.7026
10	0.8	0.0001	0.1	100	64.3363
21	0.9	0.0001	0.001	200	46.3886
12	0.8	0.0001	0.001	100	66.6152
5	0.7	0.2	0.01	100	68.5585
19	0.9	0.0001	0.1	100	68.7928

Table 2 Best parameters combinations among all datasets in terms of CPU time.

5 ANOVA single factor

The Analysis of Variance (ANOVA) test is used to check if the real average of a NSGA with parameters setting differs from the real average of the algorithm with other parameters settings. We accept or reject the hypothesis by comparing this probability with a significance fixed level α . If the *p*-value obtained for each combination of NSGA variations is smaller than α , then the variances vary such that a significant difference is observed between the algorithms. We use MS Excel to carry out the One -Way ANOVA test where α is equal to 0.05 (Fig ??). For each type of crossover, the average of fitness function and computational time are determined for 30 runs. Four different instances (TP2, TP3 and TP4). Figure ?? shows that all the instances have a p-value smaller than α except TP3 where the probability is about 0.316. This means that the time required for NSGA to carry out all the required generations is not the same for TP2 and TP4 when using different crossover operators (Fig. 2). Obviously, WAR is the slowest operator in TP4 while it is the fastest in TP2 and TP3.

 Data Stor V Salavies
 TP2
 Produbility
 Critical velocity of program

 Source of vorsitions
 Sone of source of vorsitions
 Sone of vorsitions
 Produbility
 Critical velocity of productions

 Between groups
 302.2566103
 3
 100.352034
 41.9341772
 30.687-60
 4.066180551

 Total
 321.7967071
 11
 Produbility
 Critical velocity of productions
 Address of source of vorsitions
 Critical velocity of vorsitions
 Critital vorsitions
 Critital velocity of vo

Fig. 2 ANOVA of the crossover performance

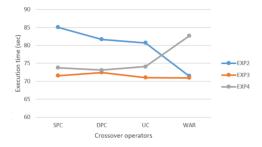


Fig. 3 Average execution time for 150 generations

6 Experimental design

All the numerical experiments conducted in this paper are generated randomly. The model was solved optimally for several sizes by Cplex V.12.7.0. We present here small and medium- size generated instances, where B = 20, the number of vessels ranges from N = 2 to 16 ships, the total number of quay cranes (QCs), q = 2, 3 and 4 and the total number of bays ranges from 103 to 500. Larger instances where N = 20, 30 and 32 vessels are examinate. All the applied simulations extract the parameter values ranges from [2]. The collected results are compared to those of the NSGA-III to demonstrate the efficiency of this algorithm. For each large instance, the exploited software was run during an overall time limit of 3600s.

7 Results and Discussion

Figures 5 & 6 exhibit the objective values achieved by the exact method and the metaheuristic in the columns Cplex and NSGA-III under the heading ObjVal. These two tables also report the CPU times needed for solving the model under the heading "CPU(s)". Clearly, all the instances can be resolved in less than two minutes. According to these results, we can affirm the validity of our mathematical model proposed as well as the used approach for all the considered instances. A convex combination of objectives is used in the weighted sum approach. In this method, specified weights w_1 and w_2 are assigned to the objective functions. We note that the sum of the weights is constant and negative weights are not authorized ($w_1 + w_2 = 1$ and $w_1, w_2 \ge 0$). Equation (1) shows the formula for this method in our problem.: Min $w_1 * f_1 + w_2 * f_2$ (1) As this combination is linear, it can be represented as a straight line possessing the weights as slope in the objective function space. When modifying the weight combinations, different lines can be obtained. For convex Pareto fronts, there is enough room to quantify such solutions with various weights. Nevertheless, for non-convex cases, as seen in Figure 4, there exist points in the non-convex zone of the Pareto front that cannot be attained for any combinations of the weight. This is accomplished by archiving and analyzing all of the results obtained during the algorithm's execution: the weight vector, the corresponding solutions, and the objective function values. To determine the bounds and the region where the true Pareto front lies, we use the weight vectors $w = [w_1 = 0, w_2 = 1]$ and $w = [w_1 = 0, w_2 = 1]$. Then, the method chooses the weights depending on the region where no solution is found yet. A uniform mesh with points embodied he Pareto front and a "weight" vector $w = [\alpha, 1 - \alpha]$ is used where α is going from 0.1 to 0.9 in steps of 0.1.

8 Conclusions and future work

This paper proposes a new MIP formulation for a dynamic BQCAP. A non-dominated sorting genetic algorithm (NSGA-III) for solving large problems is proposed. The comparison between the results obtained by CPLEX and the ones obtained by NSGA-III exhibit high quality solutions in relatively short computation times for small, medium and large-sized problems. A statistical analysis demonstrates that the execution time can be reduced in a significant way depending on which crossover operator is used. Moreover, the optimal processing parameters of NSGA-III were statistically identified in section 4. For further work, we suggest to try to investigate the scheduling interventions which inherit a lot of uncertainties in real ports as the fluctuations of freight transportation demand and unforeseen events. These fluctuations can give rise to an uncertain number of loading/unloading containers when a vessel moors and to uncertain arrival times at the port. These uncertainties further confuse the tactical berth and transshipment tasks, on which the deterministic model has already been very complex.

References

1. Y.-M. Park, K. H. Kim, A scheduling method for berth and quay cranes, in: Container terminals and automated transport systems, Springer, 2005, pp. 159–181.

2. A. Ak, Berth and quay crane scheduling: problems, models and solution methods, Georgia Inst. of Tech. 2008.

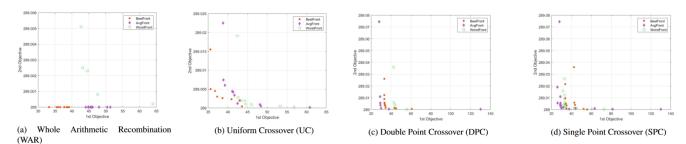


Fig. 4 Pareto front of NSGA-III on instance "EXP10" using 4 different operators

Instance		CPLEX	NSGA		
	Obj. value	CPU(sec)	Solution status	Obj. value	CPU(s)
TP1	59.0000	0.23	Optimal	64.1704875	77.002260
TP2	99.8000	0.34	Optimal	100.7946033333	89.413214
TP3	44.3999	0.05	Optimal	47.2936	71.707091
TP4	38.3000	0.05	Optimal	41.487715	95.051326
TP5	74.5999	0.28	Optimal	79.2445	86.230071
TP6	61.2999	0.69	Optimal	68.92008	79.723702

Instance		CPLEX	NSGA			
-	Obj. value	CPU(sec)	Solution status	Obj. value	CPU(s)	
1LS	101.2999	3600	Optimal	108.33465	93.990599	
2LS	105.7999	3600	Optimal	106.6599	94.303513	
3LS	108.4999	3600	Optimal	114.345368	93.539775	
5LS	154.4000	3600	Optimal	157.23639	94.084486	

5

Fig. 5 Computational results for small and mediumsized tested problem instances

- 3. T. El Boghdadly, Evolutionary optimization approach for the single and multiple-port berth allocations and Quay Crane Assignment Problem, Ph.D. thesis, University of Portsmouth, 2018.
- 4. J. Dubreuil, (2008). La logistique des terminaux portuaires de conteneurs. Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport, 2008.
- E. Lalla-Ruiz, C. Exposito-Izquierdo, J. De Armas, B. Melian-Batista, J. M. Moreno-Vega, Migrating birds' optimization for the seaside problems at maritime container terminals, Journal of Applied Mathematics (2015).
- C. Liang, L. Lin, J. Jo, Multiobjective hybrid genetic algorithm for quay crane scheduling in berth allocation planning, International Journal of Manufacturing Technology and Management 16 (2009a) 127–146.
- Liang, Y. Huang, Y. Yang, A quay crane dynamic scheduling problem by hybrid evolutionary algorithm for berth allocation planning, Computers & Industrial Engineering 56 (2009b) 1021–1028.
- Y.-M. Park, K. H. Kim, A scheduling method for berth and quay cranes, in: Container terminals and automated transport systems, Springer, 2005, pp. 159–181.
- 9. L. P. Prencipe, M. Marinelli, A novel mathematical formulation for solving the dynamic and discrete berth allocation problem by using the bee colony optimization algorithm, Applied Intelligence 51 (2021) 4127–4142.
- B. Raa, W. Dullaert, R. Van Schaeren, An enriched model for the integrated berth allocation and quay crane assignment problem, Expert Systems with Applications 38 (2011) 14136–14147.
- 11. M. Samrout, A. Sbihi, A. Yassine (2023), A genetic algorithm for the berth scheduling with ship-to-ship transshipment operations integrated model, submitted to Computers & Operations Research (under review)
- 12. E. Theodorou, A. Diabat, A joint quay crane assignment and scheduling problem: formulation, solution algorithm and computational results, Optimization Letters 9 (2015) 799-817.
- 13. K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan, "A fast and elitist multi-objective genetic algorithm: NSGA- II," IEEE Trans. Evol. Comput., vol. 6, no. 2, pp. 182–197, Apr. 2002
- I. Das and J. Dennis, "Normal-boundary intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems," SIAM J. Optimization, vol. 8, no. 3, pp. 631–657, 1998.
- 15. J. Dubreuil, (2008). La logistique des terminaux portuaires de conteneurs. Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport, 2008.

Fig. 6 Computational results for large-scale tested problem instances



Trust somebody but choose carefully : an empirical analysis of social relationships on an exchange market

Sylvain Mignot · Annick Vignes

Abstract This article analyses the influence of trust on the functioning of a market for perishable goods, where there exists no label or quality signal and quantities can be scarce. This daily market exhibits a specific bi-organization. Every morning, agents can choose between bidding or exchanging through bilateral transactions. Surprisingly, this organization is a stable one and it sounds like an economic paradox. Our hypothesis is that agents behave in a different way according to their level of trust or reputation. It is well accepted in economy that trust plays an important role in transactions but its definition and measurement stay, as far as we know, very elusive. This article provide a measurement of loyalty, based on the dynamics of agents' encounters. It brings into the light that, when the transaction links on the auction market reflects the economic constraints of the partners, the relationships on the bilateral market depends on something more. Clearly, the prices of the bilateral transactions are the consequences of economics and non economics determinants. Our results help to understand the distinctive characteristics and functioning of each sub-market. This discussion contributes to the debate about the efficiency of market structures.

Keywords market design \cdot trust \cdot social networks

1 Introduction

A fundamental assumption in economics is that rational individuals act in their own self interest. One implication is that, when trading, buyers are supposed to seek for the lowest price and sellers for the highest one and social interactions

Annick Vignes

Sylvain Mignot

Lille Catholic University, 60 Boulevard Vauban, 59800 Lille, France

ENPC, LISIS-INRAE and CAMS-EHESS, UMR 8557, 54 boulevard Raspail, 75006, Paris, France

are not considered. It is now largely accepted that social relationships affect the efficiency of a market structure (centralized or decentralized) [1–3].

The objectives of the current study is to examine the network structures of a very specific market : the Boulogne-sur-mer fish market. On this market two market structures coexist, each beeing used by the same buyers and sellers, exchanging similar goods. The two submarkets are a centralized one (Auctions) and a decentralized one (over-the-counter market). For each sub-market we examine (1) the global network structure, (2) the local network structure, and (3) we identify the traders characteristics that best explain the network structures. by comparing the results, we can compare the role of trust (bilateral market) and reputation (auction market) in the individual choices of trading partners.

Structural measures are used to characterize networks structures. Exponential random graph models are used to evaluate how trader characteristics explain purchasing patterns, and how the influence of these characteristics vary with the market mechanism.

We bring into the light that, when the transaction links on the auction market reflects the economic constraints of the partners, the relationships on the bilateral market depends on something more. Clearly, the prices of the bilateral transactions are the consequences of economics and non economics determinants. At first glance, the stable co-existence of two market structures looks like a paradox. Our results help to understand the distinctive characteristics and functioning of each sub-market. This discussion contributes to the debate about the efficiency of market structures.

2 The main market features and the data

We present here some particular features of the Boulogne s/mer fish market, through the analysis of a detailed database, consisting of 300 000 daily transactions on the period 2006-2007.

The market: The Boulogne s/mer fish market is located in the North of France near Belgium. It is considered as the most important fish market in France and one of the most important in Europe, in terms of quantity. On this market, the catch becomes scarce: this is due to the rarefaction of fish and a policy of quotas decided by the Common Fisheries Policy of the European Union, to protect the resource. This market uses a double mechanism where both auction and bilateral sub-markets coexist.

This market is a daily one, open 6 days a week. Transactions begin early in the morning. Agents are heterogeneous on both sides of the market. They are or sellers or buyers. There is no possibility of arbitrage. Buyers form an heterogeneous population, facing different budget and time constraints. They can freely buy on both sub-markets. Each day, sellers have the possibility to choose how to sell their fish (auctions or pairwise exchanges). Once the sub-market chosen, they cannot change their strategy until the next day for practical reasons (costs of bringing the merchandise from one part of the market to the other are very high). [4] show the existence of two behaviors: some agents purchase most of the time on the same sub-market, when others switch regularly. Loyal sellers, the ones who change rarely, are mainly present on the bilateral market.

On the auction sub-market sellers can't choose their buyers: the buyers are not supposed to interact with the auctioneer, apart from the prices formation mechanism. But, of course, they can decide not to bid when the catch of certain boats is being sold. Indirect trust can play a role, but not direct trust. The time constraint is high while all the transactions take place in a very short time. Important volumes of fish are traded and transactions occur at a fast rate.

On the bilateral market, the prices are not displayed and emerge from a bargaining process. Buyers, who are retailers are looking for specific species, that correspond to their expected demand. Here agents have different source of private information, depending on their past history, their ability to bargain and transact and the special links they can have with agents of the other type (buyers or sellers), here direct trust can exist and influence market outcome.

The data: 200 boats are registered in this market and designated as "sellers" in what follows. 100 buyers purchase regularly, most of them on both sub-markets. The database we use covers a year and a half (2006-2007) where both sub-markets coexist. For each transaction, the date, the species, the characteristics of the traded fish (size, presentation, quality), buyer's and seller's identities, the type of trade mechanism (auction or negotiated), the quantity exchanged and the transaction price are known. The analysis of the database tells a story of heterogeneity. First statistical results exhibit heterogeneous behaviors in terms of quality and quantities exchanged, on the both sides of the market. On the period studied, the two sub-markets (auctions and negotiated) are of equal importance (45%) of volume for the auctions market, 55% for the bilateral one): the same agents transact on the two "sub-markets" and the same types of fish are sold through both mechanisms (80 different species of fish are traded). Between 37% and 54% of each of the four main fish species (in term of quantities) are sold on the auction market which suggests an equivalent distribution of the production between the two market mechanisms.

3 Methodology and preliminary results

The first observation we can make is that the prices are higher on the negotiated market (average and median) and that the prices distributions behave differently on the two markets. The auction distribution, even if not following a normal law, is less asymmetric than the pairwise one (skewness of 0.87 vs. 3.00 and kurtosis of 1.71 vs. 16.74 on the bilateral market) and then exhibits relatively few high values. Clearly, pairwise exchanges are more risky and this result is in line with the literature.

When looking at the buyers strategy, we observe a propensity to exchange with a higher number of sellers on the negotiated market than on the auction one. We guess here that the trade network is more dense on the negotiated market that on the auction one. In the same way a simple correlation between the number of time a couple is present at the same time on a market and the number of times they transact together is higher on the negotiated market. The matching between buyers and sellers seems to follow different rules on both sub market. We will analyse the behaviors of buyers and sellers by doing a network analysis of the trading network of both submarket.

We first analyze the structure of the two networks (centralized and decentralized), looking at the difference in density, clustering and centralisation. Our preliminary results show similar number of nodes (same traders go on both markets), relatively comparable densities but very different clustering coefficients. Clustering is much higher on the auction market than on the pairwise one. In the same way, distributions of centralities on the projected networks are quite different. The two networks are structurally different, even if buyers, sellers, and goods exchanged are the same on both submarkets.

We then turn to ERGM to evaluate which of our measures are associated with a tie between a buyer and a seller, in order to estimate the nature of linking on the two sub-markets. do pairs of people exchange because a kind of informal contract (we talk about trust) or do they exchange because of an economic specialization? We can then compare the influence of these parameters on linking, allowing us to compare the effect of reputation (indirect trust on an auction market) and direct trust (on a pairwise market).

References

- A. Babus, P. Kondor, et al., Manuscript (R&R Econometrica), London School of Economics (2013)
- 2. C. Opp, V. Glode, in 2016 Meeting Papers (Society for Economic Dynamics, 2016), 1591
- 3. V. Glode, C. Opp, Available at SSRN: https://ssrn.com/abstract=2697281 (2017)
- S. Mignot, G. Tedeschi, A. Vignes, Journal of Artificial Societies and Social Simulation 15 (2) 3 15 (2), 3 (2012)

Alignment of Multinational Firms along Global Value Chains: A network-based perspective

Charlie Joyez *

February 10, 2023

Abstract

Multinationals' expansion over the last decades is undoubtedly linked to the rise of Global Value Chains (GVCs), but the extent and evolution of this link is still little documented. This paper studies this relationship by assessing the influence of the GVC network on the network of French multinational's foreign affiliates, and show that firms have increasingly settled their foreign affiliates along the developing Global Value Chains (GVC) since the late 1990s. More specifically, we compare the co-evolution of the two networks and use a quadratic assignment procedures to reveal the increasing influence of GVCs on multinationals' network structure. Standard econometric panel regressions also support this result. We then show that this alignment is more driven by a move toward upstream locations on the GVC network.

 $Keywords\colon$ Global Value Chains; Multinational Firms; Location Choices; Weighted Directed Networks.

 $JEL \ codes : F02; F23; F60; C45$

*Université Cote d'Azur, GREDEG, 06560 Valbonne, FRANCE Contact: charlie.joyez@unice.fr

This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'avenir" program (reference : ANR-10-EQPX-17 – Centre d'accès sécurisé aux données – CASD) .

1 Introduction

"Networks within networks". This is how the OECD described the interplay of Multinational Enterprises (MNEs) over Global Value Chains (Cadestin et al., 2018b).^{*} Indeed, over the last decades, Global Value Chains (GVCs) have largely expanded and diversified, reshuffling the global organisation of production. This production is increasingly provided and coordinated by multinational firms, whose number, size and activity developed until roughly accounting for one third of global output and half of global exports (UNCTAD, 2013; Cadestin et al., 2018a). These two phenomenons are intrinsically linked, with MNEs being for part a subset of Global Value Chains that is organized through explicit coordination and control within firms boundaries. Yet the matching between this global ownership network and the international flows of value added is largely imperfect, for two main reasons. First, some of the affiliates of global groups are not designed to internalize the value chain, but undertake local production and sale activities as are horizontal Foreign Direct Investments (FDIs) in the traditional horizontal/vertical FDI distinction (Markusen, 1984). Conversely vertical FDIs are those who inscribed the more into the GVCs, but existing estimates seem to show that vertical FDIs are a minority (Blonigen, 2005). Second, a large part of GVCs is still organized through arm's length trade - either by MNEs or not -. This is even more likely to be true with the uncertainty that follows GVCs' geography changes as we know multinationals prefer to start to trade with unaffiliated parties in a new country before opening an affiliate in it (Gazaniol, 2014).

Although the literature describing the international fragmentation of the production processes and the increasing interdependencies it implies is abundant (e.g. Baldwin, 2006), little attention has been placed on the firm-level coverage of GVCs. Yet, these complex production networks have also large repercussions on the organizational structure of firms that face the key decision of whether or not to expand geographically beyond their domestic borders to control the foreign segments of their production process. It is worth noting that over the same period, the worldwide number of multinational enterprises (MNEs) gradually increased[†] and the average size of multinationals also inflated. Yet, the role of global value chains in this dynamic remains

^{* &}quot;In the global economy of today MNEs increasingly function as networks within the international production networks of GVCs", Cadestin et al. (2018b, p.9)

[†]Until its 2007 edition, the UNCTAD World Investment Report reported an estimate number of MNEs, growing from 38,000 in 1993 to 80,000 in 2006 (UNCTAD, 1995, 2007). In 2018, the Financial Times reported the estimation of 100,000 multinational firms worldwide (fDI market database, 23 Feb 2018).

to be examined.

This first contribution of this paper is to show how the multinational enterprises' geographical expansion followed the development of GVCs since the mid 1990s. It is somehow related with the literature studying the optimal location of production stages (Costinot et al., 2013; Antràs and De Gortari, 2017; Antràs and Chor, 2018). But, rather to focus on the countries' position into global value chains, it focuses on the complex system made of the firms' multiple locations and their accordance with the geography of global value chains. It is therefore closer to the contributions of Antràs and Chor (2013) and Alfaro et al. (Forthcoming) that focused on the extent of firms' boundaries in a context of internationally sequential production. Indeed, a second contribution of this paper is also close to these two works, as we also examine the direction - upstream or downstream the GVCs - of the expansion of firms. To overcome the data limitation on firm-level global intra-firm trade flows that would precisely describe the firms' inclusion in GVCs, these previous works rely on detailed industrial affiliation of the firms or of each of their plants, to reconstruct their assumed trade pattern using Input-Output tables. They deduce an industry or firm-level propensity to integrate upstream or downstream stages on the GVCs.

This paper develops an alternative approach of companies' integration into GVCs, using firmlevel information on their international expansion, to examine how their network of foreign plants fits with the network structure of global value chains. Such network approach were underlying in many studies of value chains, as notably shows the description of "spiders" and "snakes" supply chains from Baldwin and Venables (2013), but have often been overlooked. Yet, the explicit use of complex networks is increasingly popular in various fields of economics to study the structure of connections, the strategic interactions, and the interdependence between actors (see Jackson (2014) for a comprehensive review). Economic geography is particularly sensitive to such tools, that fits well to study locally clustered structures, and for a long time used in geographic studies (Glückler, 2007; Ter Wal and Boschma, 2009). Other pioneers of network analysis in economics are obviously to be found in econophysics, with notably the seminal elaboration of the "product space", a network depicting the relatedness of traded goods by Hidalgo et al. (2007), and their subsequent works up to the economic complexity index (Hidalgo and Hausmann, 2009; Hausmann et al., 2014).

Therefore, this work extends the few ones that have already mapped the GVCs' network, and examined the topology their trade flows, benefiting from the increasing availability and coverage of international input-output tables (Cerina et al., 2015; Amador and Cabral, 2017; Criscuolo and Timmis, 2018). Yet, our topic is wider than these developments, as we also study the geographical network of multinational firms, to focus on the complex mechanism that links two networks' structures. In this second perspective, this paper follows existing works studying multinational firms' network of locations at the city level (Ducruet et al., 2011; Rozenblat, 2015; Wall and van der Knaap, 2011). At the host-country scale though, only a scarce number of paper detailed the network structure of Foreign Direct Investments (FDIs) (De Masi et al., 2013; De Masi and Ricchiuti, 2018a; Joyez, 2017).

This paper compares the co-evolution of the geographical coverage of French MNEs with the network of global value chains, and reports how multinationals' plants are increasingly located along GVCs between 1996 and 2011. We provide some evidence of an international reorganization of French firms following the emerging GVCs. Specifically, the development of GVCs led French firms to shift their strategies and move up the value chains as the MNEs' alignment on GVCs turned to be more driven by upstream expansion of firms (toward prior stages of production) than by downstream one (toward subsequent stages). This representation of the global economy as a multiplex system, provides new insights on the firms' sensitivity to the risk of dismantling GVCs that policy maker and firms' managers themselves should be aware of, especially since the attraction of GVCs increased after 2008, once they started to stabilize.

The rest of the paper is organized as follows. Section 2 reviews the literature that introduced a network approach of global valued chains and multinational firms. Section 3 presents the methodology used to reconstruct both GVCs' and MNEs' networks and the database used. In section 4, the structural evolution of each network is detailed. Section 5 presents the econometric estimations of the networks' alignment. Finally section 6 concludes and delivers some policy recommendations.

2 Networks analysis of Global Value Chains and Multinational Firms

2.1 GVCs' networks

Since the early 2000s, the overall world trade flows have repeatedly been studied as a network, spreading out of its initial field of econophysics (Serrano and Boguñá, 2003; Garlaschelli and Loffredo, 2005) to more conventional empirical literature on trade flows (Kali and Reyes, 2007; Fagiolo et al., 2010; De Benedictis and Tajoli, 2011; De Benedictis et al., 2014). These papers agree on the hierarchical structure of the World Trade Web (WTW), characterized by several network metrics, such as high clustering coefficient, a disassortative pattern, and a right-skew distribution of networks' connectivity indexes revealing a core-periphery structure.

Global value chains differ from global trade flows though. Since Hummels et al. (2001) the participation in GVCs is assessed by breaking down gross trade flows along original sources and final destination of value added. A first attempt of retracing GVCs' networks is therefore by focusing only in trade in intermediates. Using flows classified as trade of parts and components from BACI dataset, Ferrarini (2013) has been the first use network visualization tools to map "vertical trade" - an explicit reference to Hummels et al. (2001) concept of vertical specialization. Yet, an alternative methodology was rapidly adopted. Instead of using international classifications, further works relied on the increasing accessibility and coverage of international Input-Output (I-O) tables to capture more precisely trade in intermediate according to their final destination. Both Cerina et al. (2015) with the World Input-Output Database (WIOD) and Criscuolo and Timmis (2018) with the OECD Inter Country Input Output (ICIO) tables, developed this methodology at the country-sector level given in the I-O tables.

Progressively Daudin et al. (2011), Johnson and Noguera (2012) or Koopman et al. (2014) provided methodologies to decompose these I-O tables, and identify the foreign valued added content of a country's exports, and its detailed source. Because these flows corresponds even better to the intrinsic idea of GVCs, they recently have been studied through a network approach. Using the WIOD data, Xiao et al. (2017) divide the WTW into several flows following Koopman et al. (2014) decomposition, and details notably the network of Foreign Value Added (FVA). Also with the WIOD dataset, Amador and Cabral (2017) detail their own computation of bilateral

Foreign Valued Added in exports and describe the decentralization trend from a binary network perspective from 1995 to 2011. The weighted analysis of their FVA network is detailed in Amador et al. (2018).

2.2 Multinationals networks

Despite their obvious web structure made of several plants in various countries, the networks of multinationals firms may be harder to conceive, because of the absence of a unique relational variable drawing the links between the networks' nodes. A large variety of flows could be used to define and weight the networks' edges, such as financial flows, technologies, trade in goods or services, or even people transiting in the network. But choosing one of these flows to define the networks' connection pattern brings us back to the limited availability of such data. Then, we simply link all hosting localization of a same firm, without weighting this firm-level network, to build a map of each firms' geographic coverage, as previously done by geographers at the city scale (Hussain et al., 2018; Ducruet et al., 2011; Rozenblat, 2015). The global network is made by superimposition of all firm-level network, leading to weighting edges proportionally to their frequency of use. At the country level (taking the host countries as the network nodes), it formally corresponds to the projection into the "country space" of the initial bipartite network made of each multinational firms and their Foreign Direct Investments destinations (see De Masi and Ricchiuti (2018b) for a detailed explanation of projection from bipartite graphs).

The examination of this network of the firms' FDI destination extends the vast literature on multinationals' location choices (see Alfaro and Chen (2017) for a recent survey), but allows to consider simultaneously all locations of a firm instead of focusing on isolated choices. This is far from being trivial because location choices of firms are largely interdependent decisions: the set of existing foreign location of the firm necessarily influences the new ones, and all plants play a role in a complex global design of the firm that includes the multiple locations of the firms. To the contrary, traditional empirical estimations of locations as discrete choices assume FDIs to be independent from each other, and occasionally use firm-level fixed effects that control for the firms' previous locations (Chen and Moore, 2010). Although they can allows to reach unbiased estimates for FDI drivers, they act as black boxes, preventing precisely to study the pattern of interconnections among host countries. It is this pattern of pairs of countries jointly invested by multinationals that the multinationals' network reveals. De Masi et al. (2013) examined this network structure from Italian MNEs, and Joyez (2017) reveals the role of firms' heterogeneity into the French firms' network topology. De Masi and Ricchiuti (2018a) offer an extension to all UE28 countries, focusing on the propagation of risks created by the geographic diversification and intensification of such networks. All these works described the MNEs' network as a small-world network, highly clustered and hierarchical. Although this is a common point with the world trade web and GVCs, the truth is that most of real-life networks exhibits such topology from the world wide web, to the airport hubs. This common point is insufficient to draw immediate conclusions. To look further at the networks structures' similarities, we reconstruct the two networks as detailed in the next section.

3 Data and network reconstruction

3.1 Global Value Chains networks

Following the recent developments in the GVCs' network literature reviewed in the previous section, and to fit closely to the initial idea of Hummels et al. (2001), we define the GVCs' network as the network of foreign valued added content of exports. These flows are particularly adapted to picture GVCs because unlike row intermediate trade flows, they capture the value added flows that cross at leas two borders. Yet, contrary to most of the existing works, we rely on the UNCTAD-Eora GVCs database. Built on Eora multi regions I-O (MRIO) tables, the UNCTAD-Eora GVCs provides a square matrix of 190 countries, in which it reports the value added from country i embedded in country j's exports, expressed in thousands of current U.S. dollars from 1990 to 2015. Its geographic coverage is far greater than the ones offered by the WIOD or the OECD-ICIO datasets, respectively gathering 40 and 63 countries[‡] This Eora-MRIO data have already been used to study centrality in GVCs, notably by Antràs and De Gortari (2017). One particularity of all IO-derived measures, consists in the computation of both direct and indirect foreign valued added. Two countries that are not direct trade partners can have some of each others' value added embedded in their exports, from their own imports. Because of this recursive computation, virtually all pairs countries are linked in these data. Yet, to clear the network picture we only focus on substantial GVCs' linkages. Marginal links below

[‡]For more information on the construction of Eora dataset and the necessary estimations in it, see Lenzen et al. (2012, 2013). The UNCTAD also provides comparison with OECD ICIO data at http://worldmrio.com/unctadgvc/

100,000 current USD are removed[§]

Formally, for each year t we define the a weighted directed GVCs' network as $GVC_t = (C, E_t)$, formed by the set of nodes (countries) $C = \{i; i = 1, 2, ..., N\}$ and by the set of links E_t . It is fully characterized by its binary and weighted adjacency matrices of dimension NxN, defined as follows. The binary adjacency matrix $A_t^G = [a_{ij,t}^G]$ with

$$a_{ij,t}^{G} = \begin{cases} 0 & \text{if } FVA_{ij} < 100 \text{ at year } t \\ 1 & \text{otherwise} \end{cases}$$

where FVA_{ij} is the foreign valued added from *i* into *j*'s exports as reported in Eora.

The weighted adjacency matrix is $W_t^G = [w_{ij}]$ where

$$w_{ij,t}^{G} = \begin{cases} 0 & \text{if } a_{ij}^{G} = 0 \\ FVA_{ij} & \text{otherwise} \end{cases}$$

While A_t^G is symmetric by construction, the weighted adjacency matrix W_t^G isn't so, and account fors the direction of value added flows. The bottom layer of figure 1 is an illustration of this network structure.

3.2 Multinationals' network of French foreign affiliates

The yearly LiFi survey from the French national statistic institute (*INSEE*) reports the financial linkages between domestic firms and French or foreign entities through shares ownership, and benefits from an very good coverage of multinational firms.[¶] Specifically it lists upstream linkages, identifying one company being the "head of group", and downstream linkages of all entities it posses. To keep on with the traditional conception of ownership, we only focused at majority owned entities. The sample of French firms is made of all French "head of groups" or independent firms that hold at least one foreign affiliate. Foreign-owned french companies are therefore removed from the sample, because the global location of the whole group isn't known, and

[§]The value of the threshold -from 0 to 10 millions of dollars actually doesn't substantially change the findings. Results from alternative thresholds are available on demand.

[¶]All firms previously identified as "head of groups" or whose a participation in other firms above 1.2 million euros, or with more than 500 employees, or with a turnover superior to 60 millions euros foreign affiliates are surveyed. Such thresholds are low, especially for multinational enterprises and guaranties a large coverage of French MNEs

their location choices can be biased by their foreign owners. Also, all firms without any foreign affiliates were removed from the sample. The sample size varies each years but grows from 1122 firms and 7,491 foreign affiliates in 129 countries in 1996 to 2996 firms owning 19,872 affiliates in 167 countries as for 2011. As explained in the previous section, we first draw the individual network of foreign location choices for each firm, and then compile them all to obtain the yearly network of foreign host countries of French multinationals, henceforth labeled the MNEs' network for simplicity. This is a weighted, non-directed network, in which the nodes are the host-countries of foreign affiliates and obviously France. The edges' weight indicate the number of French firms realizing the country pairing.

Formally, each year t network is defined as $MNE_t = (C_t, F_t)$ where $C_t = \{i; i = 1, 2, ..., N\}$ is the set of host countries at time t (i.e. the networks' nodes), and F_t describes their dyadic connections. These connections are given by the binary and weighted adjacency matrices, the first one being $A_t^M = [a_{ij,t}^M]$ with

$$a_{ij,t}^{M} = \begin{cases} 0 & \text{if not any French MNEs have simultaneously an affiliate in i and j at time t} \\ 1 & \text{otherwise} \end{cases}$$

The weighted adjacency matrix $W_t^M = [w_{ij,t}^M]$ reflects the number of French firms using any country pairing $\{i; j\}$ at time t. The network being undirected, $\forall \{i, j\}, w_{ij,t} = w_{ji,t}$. The top layer of figure 1 illustrates this network.

3.3 Harmonizing networks

Each of the GVCs' and MNEs' networks are populated by the same type of nodes: countries. The former network details their tights in terms of trade in value added, and the latter details the geographic coverage of French firms' international expansion. Therefore, these two networks can be seen as two layers of a "multiplex" or "multilayer" network (Kivelä et al., 2014), as depicted in the figure 1 below. This multiplex network is actually a representation of the international economy where all kind of international economic relationships are can be pictured in a different layer. Our aim is to examine whether the superimposition of these two layers fits well, or if the change in the structure in the GVCs' layer generates changes in MNE's one.

Though, to investigate the co-evolution of these two layers, they have to be made of the same

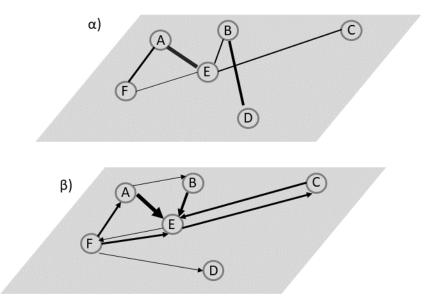


Figure 1: MNEs' and GVCs' networks as a multilayer network Note: Nodes A to E represents the set of countries. The layer α) is undirected and illustrate the MNEs' network, the edges' size is proportional to the number of French firms located simultaneously in both countries. The layer β) is directed and illustrates the GVCs' network. The arrows represent the origin and the amount of foreign value added in exports.

set of countries. These leads us to drop the "rest of the world" line initially in Eora database, that we cannot find in the MNEs' network. With a constant number of 189 countries, the resulting GVCs' network outpaces the set of countries in MNEs' network. The missing nodes in the firms' network can be added as isolated nodes, which allow us to preserve virtually all countries. Yet, our estimation strategy involves to add control variables, and notably dyadic variables that could also be described as layers of international networks. Specifically, we control for the bilateral distances between any countries (from CEPII gravity dataset), and bilateral trade in million of current U.S. dollars (from Correlates of War project). The initial coverage of these two datasets was respectively of 224 and 195 countries, but the common subset between these two dataset and GVCs' dataset is of 171 countries, which will be the final sample for our estimations.^{**} The removed nations are necessarily marginal countries in MNEs' network, because they all are isolated nodes. Their weight on the world economy is also relative as the final sample represents 99.05% of the world's GDP in 2011, according to the CEPII gravity dataset.

A last change is required to harmonize the networks, as two of them are undirected (MNEs'

^{||}Barbieri, Katherine and Omar M. G. Omar Keshk. 2016. Correlates of War Project Trade Data Set Codebook, Version 4.0. Online: http://correlatesofwar.org.

^{**}The complete country list available in Appendix B.

coverage and bilateral distance), while the other two are directed (trade and VA flows). Since the initial aim is to focus on the determinants of an undirected network, the direction of the GVCs' and trade flows are irrelevant. We therefore erase the direction information by summing the two bilateral arrows. Section 5.2 details a directed approach, and the reconstruction of a directed (sequential) MNEs' network to study the upstream or downstream direction of the alignment of the two networks.

4 Structural Evolution of the two networks

Before to focus on the co-evolution of the networks, it is instructive to see that the two networks' structures have changed. This can be done by studying the evolution of the aggregate network metrics on connectivity and centralization. In line with previous works using WIOD data, we report a decentralization trend of the GVCs' network and increasing cross-country connectivity (Cerina et al., 2015; Amador and Cabral, 2017; Amador et al., 2018). Interestingly, a similar pattern emerges from the MNEs' network evolution. This shift in the global location of French MNEs is crucial to prove that the alignment on GVCs is not only due to the emergence of GVCs across a constant location pattern of firms.

The most immediate metrics of connectivity is the network *density*, which reports the share of (non-null) edges over the possible number of edges. Let d_t^n be the density of the network n = G, M at time t. Formally, $d_t^n = \frac{\sum_{i \neq j} \sum_{j \neq i} a_{ij,t}^n}{N(N-1)}$ where the $a_{ij,t}^n$ are the elements of the binary adjacency matrices of network n, and N its number of nodes.

The figure 2 reports the evolution of the density in the two networks. The GVCs' network shows a net increase in density especially since the early 2000s, implying that an increasing pairs of countries are linked through substantial value added linkages in GVCs. The drop in 2009 shows how the great depression that followed the financial crisis diminished these linkages, but was rapidly recovered, although a ceiling seems reached since 2010, but this is possibly due to the high level of density that cannot increase much more. During the same period, the density of the MNEs' network grew even more than the GVCs' one, implying that a larger combination of countries have been done by French firms, denoting a diversification of internationalization pattern. The drop in 1998 is intriguing, and we have no explanation of it at this stage![†]

^{††}More generally, the year 1998 is an outlier in several of our results concerning the MNEs' network. Although identifying why is a continuing question, these intriguing results are not sufficient to shift the global conclusions

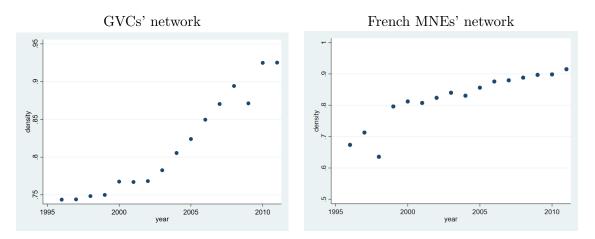


Figure 2: Density of networks 1996-2011

Yet, this index only partially reflects the evolution of the network connectivity because of its binary dimension. A weighted approach of the connectivity is accessible through the weighted overall clustering coefficient. The clustering coefficient reflects the tendency that two neighbors of one node are themselves directly connected, forming triangles. It is defined for each node *i* as $C_i = 2t_i/k_i(k_i - 1)$, where t_i is the number of triangles attached to the node, and k_i , the node degree, that is its number of neighbors $k_i = \sum_{j \neq i} a_{ij}$.

The weighted generalization of this basic index accounts for the breakdown of weights between the three edges as suggested by Onnela et al. (2005). Their index derives from the unweighted version, but replaces the number of triangles t_i with the sum of the triangles intensity.

$$\tilde{C}_i = \frac{2}{k_i(k_i - 1)} \sum_{j,k} \left(\tilde{w}_{ij} \tilde{w}_{jk} \tilde{w}_{ki} \right)^{1/3}$$

Where $\tilde{w}_{ij} = w_{ij}/max(w_{ij})$ is a relative measure of the edges' weight. By construction this index reaches lower values than the unweighted version since $\tilde{C}_i \to C_i$ when the network becomes binary. The index equals one when all possible triangles are closed and equally weighted.

The figure 3 reports the evolution of this weighted clustering coefficient, which increased in the two types of networks especially since 2002. This increase in the clustering structure of the networks means that the edges of any triangle of nodes in the network are increasingly balanced, implying a better connectivity, and a lower centralization of the networks.

of this work.

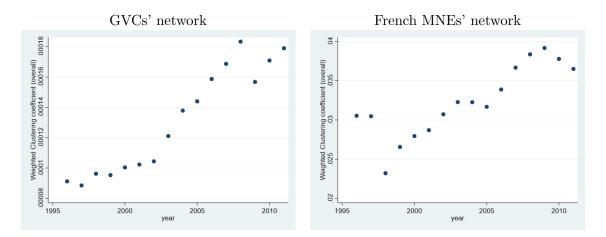


Figure 3: Weighted clustering coefficient of networks 1996-2011

The centralization of a network refers to the relative importance of few central nodes (cores) on the network. Traditionally, the centralization of a network is studied through the degree or the *strength* distribution. The strength is the weighted generalization of the degree index for each node, and is equal to the sum of the weights of all its edges. Formally, $s_i = \sum_j w_{ij}$. The more concentrated the degree or strength distribution, the more the network is centralized around a few number of central entities.

The figure 4 reports the inverse cumulative distribution functions (CDF) of the nodes' strength for each network in 1996 and 2011. The network of global value chains displays a largely unequal distribution of the nodes' strength, with a few numbers of nodes at the top of the distribution, i.e. largely linked with many others countries in terms of value added trade. The very long tail is made of all other countries that are far less connected to others. This is consistent with a core-periphery structure already reported in the foreign value added network (Amador et al., 2018), and more generally in the world trade web (Fagiolo et al., 2009). The pattern is less stressed in the case of the MNEs' network, but the convexity of the CDF also suggests a similar structure.

One could notice a slightly smoother CDF for both networks' strength distribution in 2011 compared to 1996. To investigate more precisely the evolution of the strength's distribution, we report the kurtosis of the distribution from 1996 to 2011 for each network in figure 5. The Kurtosis of strength distribution is always positive in both networks, implying a right-skewed distribution, with a lot of nodes with low strength and few nodes with large strength. Yet, the two networks reflect an unambiguous trend toward a decrease in the strength kurtosis, confirming a

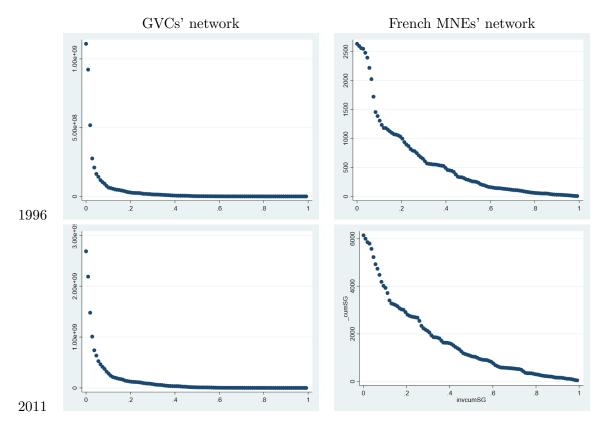


Figure 4: Strength Cumulative distribution function

re-balance of the nodes' strength, although the core-periphery structure is maintained, especially in the GVCs' network.

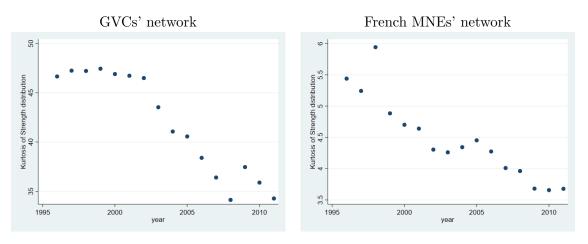


Figure 5: Kurtosis of Strength distribution of networks 1996-2011

This deconcentration of the strength distribution reveals a decentralization trend in the two networks structure, which is confirmed by a last index: the strength centralization index, a weighted generalization of Freeman (1979) centralization index, and formally defined as:

$$F_{s} = \frac{\sum_{i}^{N} [s_{i^{*}} - s_{i}]}{max \sum_{i}^{N} [s_{i^{*}} - s_{i}]}$$

Where $s_{i^*} = max(s_i)$, and $max \sum_{i}^{N} [s_{i^*} - s_i] = (\sum_{i} \sum_{j} w_{ij} - min(s_i))(n-1)$ is the maximum possible sum of differences in strength for a network of same dimensions (number of nodes and total weight). This centralization index have two valuable properties: it is bounded between zero and one, the higher implying the greater strength centralization. Also, this index equals the standard degree centralization index when the graph is binary. Its evolution (figure 6), is very similar to the drop in the kurtosis, and confirms a decreasing centralization of the valued network.

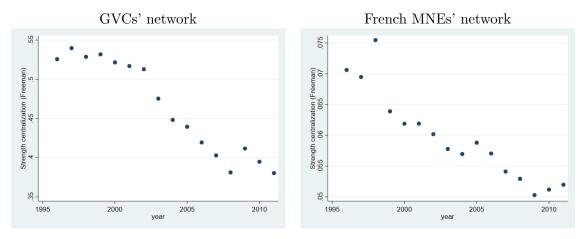


Figure 6: Strength centralization of networks 1996-2011

The several metrics describing the networks topology agree on revealing a structural change in the two networks structures, toward less centralized networks made of better and more equally connected nodes. Of course, a hierarchical structure still exists, with some nodes being more important than others. Yet, there is a clear trend in both networks toward more horizontal structure, where previously peripheral countries play a greater role in the world's production and contribution to global value chains. This actually isn't surprising but shows how the network analysis is designed to capture the "great convergence" that accompanied recent globalization (Baldwin, 2016). Yet, these countries are also increasingly considered by French firms in organizing their global activities, and associated with others in a more diverse way. This co-movement of the two network structure is too striking not to see there some dynamics and influence from one structural change to another. The next section assesses such hypotheses.

5 Firm's alignment on GVCs

5.1 Quadratic Assignment Procedure

A popular methodology in complex networks analysis to run regressions when both dependent and independent variables are relational matrices of same size, is the multiple regression Quadratic Assignment Procedure (MRQAP or simply QAP) as details Krackhardt (1988). The very concept of network structure is to assume that the nodes are not independent from each other, which violates the i.i.d assumption required for most traditional estimates. The MRQAP provides an alternative that preserves the network structure and compares the estimates of the model to the distribution of such statistics resulting from large numbers of simultaneous rows and columns permutation of the considered variables. This permutation method allows us to include multiple matrices in one analysis while accounting for inherent structural autocorrelation. The MRQAP model used in this paper consider relational variables and their structural interdependencies when assessing the coefficients' statistical relevance. Specifically, if the initial (positive) coefficient is superior to 99% of the coefficients estimated from the randomly permutated samples, it represents a significance level of 0.01.

We begin with a very naive model defined in equation (1) where $MNEnet_t = W_t^M$ corresponds to the 171x171 adjacency matrix of the MNEs' network in year t, and $GVCnet_{t-1} = W_{t-1}^G$ the network of Foreign VA in exports across the same 171 countries the year before that. Both variables are log-transformed to express the coefficient as an elasticity.

$$ln(MNEnet_t) = \beta_O + \beta_1 * ln(GVCnet_{t-1}) + \epsilon$$
(1)

The one-year lag of the independent variable has a double purpose: first it aims at acknowledging the necessary time to build foreign plants for firms that want to adapt to a changing environment. Second, it is designed to control for simultaneity bias. Although there is an obvious simultaneous development of the MNEs and GVCs, the change in current MNEs position can hardly explain the past flows of value added. More about the simultaneity threat: let's remind that our sample only accounts for French multinationals' plants, that are unlikely to shape the global value added flows as they represent 3 to 4% of the worldwide number of MNEs^{‡‡}

 $^{^{\}ddagger\ddagger} \mathrm{Comparing}$ the number of French MNEs in our sample (from 1122 to 2996) with the UNCTAD reported number of MNEs

The threat could still exist though if this small sample of global MNEs were reflecting accurately the pattern of international position of all multinationals across the globe. But this is also improbable, as the FDI location choices largely depend on the origin country trade patterns and geographic position. Therefore, using only the French sample of MNEs and the lag values of global value chains flows removes the threat of endogeneity bias from the simultaneity of GVCs' and MNEs' developments.

Table 1: Naive QAP estimates

	'97	'98	'99	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11
$\operatorname{GVCnet}_{t-1}$	0.106	0.107	0.123	0.127	0.133	0.139	0.140	0.143	0.142	0.150	0.154	0.158	0.171	0.173	0.194
	$(109.34)^{**}$	$(106.93)^{**}$	$(117.03)^{**}$	$(117.98)^{**}$	$(120.03)^{**}$	$(122.86)^{**}$	$(120.74)^{**}$	$(121.78)^{**}$	$(120.00)^{**}$	$(122.39)^{**}$	$(124.98)^{**}$	$(127.20)^{**}$	$(131.51)^{**}$	$(133.42)^{**}$	(135.50)**
_cons	-0.127	-0.125	-0.127	-0.121	-0.134	-0.139	-0.126	-0.141	-0.158	-0.192	-0.233	-0.277	-0.362	-0.270	-0.515
	$(20.72)^{**}$	$(19.55)^{**}$	$(18.79)^{**}$	(17.41)**	$(18.41)^{**}$	$(18.56)^{**}$	$(16.37)^{**}$	$(17.75)^{**}$	$(18.98)^{**}$	$(21.69)^{**}$	(25.48)**	(29.13)**	(35.29)**	(27.46)**	$(44.59)^{**}$
R2	0.29	0.28	0.32	0.32	0.33	0.34	0.33	0.34	0.33	0.34	0.35	0.36	0.37	0.38	0.39
Ν	29,070	29,070	29,070	29,070	29,070	29,070	29,070	29,070	29,070	29,070	29,070	29,070	29,070	29,070	29,070

The model in (1) was repeated from 1997 to 2011, and the results in table 1 show a positive, highly significant, and quite steadily increasing influence of GVCs on building MNEs' network dyads over time, with a coefficient experiencing a 83% increase in 15 years. This shows the growing correlation between the two networks' structures, and an increase of the explanatory power of GVCs about the MNEs' geographic linkages through the ten point increase of the R-square. This first result is purely descriptive, but in line with hypothesis of the two networks' synchronization. However, the obvious omitted variables bias cast a doubt on the estimates' validity, and call for preciser estimations.

The next model developed in equation (2) includes the two other country-level networks that likely determine MNEs' one: bilateral distances and bilateral trade.

$$ln(MNEnet_t) = \beta_O + \beta_1 * ln(GVCnet_{t-1}) + \beta_3 * ln(Tradenet_{t-1}) + \beta_1 * ln(Distnet) + \epsilon$$
(2)

The expected role of bilateral distance is not obvious: following gravity-like models, the geographic distance to previous locations should reduce probability of new linkages. But foreign affiliates can also act as regional branch, where an affiliate in a country would reduce the probability to invest in its neighbor, and increase the probability to target a remote region. Including bilateral trade is also important to make sure that the GVCs' coefficient doesn't only captures the fact that host countries are trade partner. One could worry about the potential

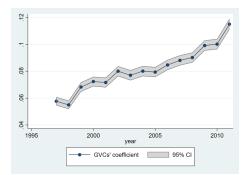


Figure 7: GVCs' coefficients from MRQAP

multicollinearity bias from GVCs' and Trade networks. Yet, the GVCs' network is not a subsample of the global trade network as would be the network of trade in intermediates, because GVCs only capture the value that crosses two borders, and reconstructs indirect value added linkages absent from the trade network. In addition, the bias involved by multicolinearity would be less severe than the one from omitted variable, and finally, the large level of significance of both variables displayed in table ?? rejects the fear of a severe multicollinearity.

The addition of the control variables reduce slightly the GVCs' coefficient, but it remains positive, significant, and increasing, although more volatile in the late 1990s. The bilateral trade control displays the expected results, increasing the MNEs' linkages. Interestingly, the bilateral distance is also found to be positive, which validates the assumption that existing regional presence deters neighboring investments. It is worth noting though that the GVCs highlight higher absolute coefficients than both trade and distance, especially at the end of the period studied. It means that the importance of value added flows between countries is stronger than the one of raw trade flows in determining the multiple location choices of MNEs than the one of row trade flows. Although the elasticity might seem low - between 10 and 20% -, keep in mind that opening of foreign affiliate is a long-term choice, and that only a few foreign affiliates are created each years compared to the existing stocks.

To offer a clear picture of the increase in GVCs' coefficients, I reported them and their 95% confidence interval in figure 7. The upward trend is clear and even reveals an acceleration since 2008, after a stagnation in the early 2000s. The last point deserves particular attention, as we know the GVCs to have stabilized after the 2009 crisis (as reveals notably the smoother changes in GVCs' topology reported in section 4, and already discussed by Los et al. (2015)or Timmer et al. (2016)). Yet, their consequence, at least in shaping location choices, are still increasing.

	1997	1998	1999	2000	2001	2002	2003	2004
$\operatorname{GVCnet}_{t-1}$	0.058	0.055	0.068	0.072	0.072	0.080	0.077	0.080
	(35.93)**	$(34.15)^{**}$	(38.34)**	(39.62)**	(38.02)**	$(41.72)^{**}$	(39.72)**	$(41.26)^{**}$
$\operatorname{Tradenet}_{t-1}$	0.056	0.062	0.063	0.063	0.071	0.069	0.074	0.073
	$(37.49)^{**}$	(41.55)**	(38.04)**	$(36.98)^{**}$	$(40.01)^{**}$	(38.31)**	$(40.46)^{**}$	$(40.43)^{**}$
Distnet	0.056	0.052	0.054	0.054	0.062	0.063	0.076	0.071
	$(10.21)^{**}$	$(9.25)^{**}$	$(9.10)^{**}$	(8.87)**	$(9.75)^{**}$	$(9.72)^{**}$	$(11.59)^{**}$	$(10.54)^{**}$
_cons	-0.651	-0.622	-0.637	-0.634	-0.712	-0.733	-0.845	-0.807
	$(13.29)^{**}$	$(12.37)^{**}$	$(11.98)^{**}$	$(11.58)^{**}$	$(12.62)^{**}$	$(12.65)^{**}$	$(14.31)^{**}$	$(13.41)^{**}$
R^2	0.32	0.32	0.35	0.35	0.37	0.37	0.37	0.37
Ν	$29,\!070$	29,070	29,070	29,070	$29,\!070$	$29,\!070$	$29,\!070$	29,070
	2005	2006	2007	2008	2009	2010	2011	
$\operatorname{GVCnet}_{t-1}$	2005 0.079	2006 0.085	2007 0.088	2008 0.090	2009 0.099	2010 0.100	2011 0.115	_
$\operatorname{GVCnet}_{t-1}$								-
$\operatorname{GVCnet}_{t-1}$ Tradenet _{t-1}	0.079	0.085	0.088	0.090	0.099	0.100	0.115	_
	0.079 (41.50)**	0.085 (43.76)**	0.088 (45.75)**	0.090 (47.12)**	0.099 (50.45)**	0.100 (51.09)**	0.115 (54.79)**	_
	0.079 (41.50)** 0.073	0.085 (43.76)** 0.074	0.088 (45.75)** 0.074	$0.090 \\ (47.12)^{**} \\ 0.075$	0.099 (50.45)** 0.078	$0.100 \\ (51.09)^{**} \\ 0.081$	$0.115 (54.79)^{**} 0.082$	-
$\operatorname{Tradenet}_{t-1}$	$\begin{array}{c} 0.079 \\ (41.50)^{**} \\ 0.073 \\ (41.30)^{**} \end{array}$	$\begin{array}{c} 0.085 \\ (43.76)^{**} \\ 0.074 \\ (42.65)^{**} \end{array}$	0.088 (45.75)** 0.074 (43.54)**	$\begin{array}{c} 0.090 \\ (47.12)^{**} \\ 0.075 \\ (45.44)^{**} \end{array}$	$\begin{array}{c} 0.099 \\ (50.45)^{**} \\ 0.078 \\ (47.72)^{**} \end{array}$	0.100 (51.09)** 0.081 (47.82)**	$\begin{array}{c} 0.115 \\ (54.79)^{**} \\ 0.082 \\ (49.98)^{**} \end{array}$	_
$\operatorname{Tradenet}_{t-1}$	0.079 (41.50)** 0.073 (41.30)** 0.053	0.085 (43.76)** 0.074 (42.65)** 0.050	0.088 (45.75)** 0.074 (43.54)** 0.039	0.090 (47.12)** 0.075 (45.44)** 0.038	0.099 (50.45)** 0.078 (47.72)** 0.053	0.100 (51.09)** 0.081 (47.82)** 0.042	0.115 (54.79)** 0.082 (49.98)** 0.046	_
$\operatorname{Tradenet}_{t-1}$ lDistnet	$\begin{array}{c} 0.079 \\ (41.50)^{**} \\ 0.073 \\ (41.30)^{**} \\ 0.053 \\ (7.85)^{**} \end{array}$	$\begin{array}{c} 0.085 \\ (43.76)^{**} \\ 0.074 \\ (42.65)^{**} \\ 0.050 \\ (7.28)^{**} \end{array}$	$\begin{array}{c} 0.088 \\ (45.75)^{**} \\ 0.074 \\ (43.54)^{**} \\ 0.039 \\ (5.67)^{**} \end{array}$	$\begin{array}{c} 0.090 \\ (47.12)^{**} \\ 0.075 \\ (45.44)^{**} \\ 0.038 \\ (5.48)^{**} \end{array}$	0.099 (50.45)** 0.078 (47.72)** 0.053 (7.59)**	0.100 (51.09)** 0.081 (47.82)** 0.042 (5.87)**	$\begin{array}{c} 0.115 \\ (54.79)^{**} \\ 0.082 \\ (49.98)^{**} \\ 0.046 \\ (6.37)^{**} \end{array}$	_
$\operatorname{Tradenet}_{t-1}$ lDistnet	$\begin{array}{c} 0.079 \\ (41.50)^{**} \\ 0.073 \\ (41.30)^{**} \\ 0.053 \\ (7.85)^{**} \\ -0.656 \end{array}$	0.085 (43.76)** 0.074 (42.65)** 0.050 (7.28)** -0.653	$\begin{array}{c} 0.088 \\ (45.75)^{**} \\ 0.074 \\ (43.54)^{**} \\ 0.039 \\ (5.67)^{**} \\ -0.581 \end{array}$	$\begin{array}{c} 0.090 \\ (47.12)^{**} \\ 0.075 \\ (45.44)^{**} \\ 0.038 \\ (5.48)^{**} \\ -0.595 \end{array}$	0.099 (50.45)** 0.078 (47.72)** 0.053 (7.59)** -0.797	0.100 (51.09)** 0.081 (47.82)** 0.042 (5.87)** -0.636	$\begin{array}{c} 0.115 \\ (54.79)^{**} \\ 0.082 \\ (49.98)^{**} \\ 0.046 \\ (6.37)^{**} \\ -0.826 \end{array}$	_

Table 2: MRQAP estimations

5.2 Upstream and Downstream alignment

After having documented this increasing alignment of MNEs along global value chains, an immediate question arises. Do the firms create new plants upstream of existing plants (closer to source of the VA) as they would do if they aimed at controlling initial production steps, or downstream of them (closer to the final use of the good) to control subsequent production and distribution steps? To answer this question a slightly different network analysis is required, using a directed network of multinationals, in which the direction of the edge from node i to node j indicate the sequence of the firms internationalization from country i to country j. Because the precision of the data is only at yearly intervals, if two countries are targeted the same year by a given firm, the edge is set as reciprocate. Otherwise, the arrow indicates the sequence of internationalization. As underlined in section 2, the GVCs' network is initially directed because the foreign content of exports isn't symmetric. This directed network is therefore used in this section. Remind that each of its edges w_{ij}^G is equal to the value added from i embedded in country j exports, therefore an alignment of the sequential MNEs' network on the GVCs' one would characterize downstream reorganization of firms, as going from i to j they would go toward the next destination of the value added. To capture the upstream alignment of firms, we repeat the analysis using the transposed adjacency matrix of GVCs, that reflects the reverse flow. Formally, each edge on the transposed GVCs' matrix can be expressed as $w_{ij,t}^G$ ' = { $w_{ji,t}^G$ }, and reflects the VA from j into i's exports. Henceforth, an alignment between the MNEs' network and this transposed network, would imply that firms go back toward the source of the value added. Similarly we use the transposed matrix of Trade network to capture whether firms move up or down trade flows. The new model to estimate becomes equation 3:

$$ln(MNEnet_t) = \beta_0 + \beta_1 * ln(upGVCnet_{t-1}) + \beta_2 * ln(downGVCnet_{t-1}) + \beta_3 * ln(Exportsnet_{t-1}) + \beta_4 * ln(Importsnet_{t-1}) + \beta_5 * ln(Distnet) + \epsilon$$
(3)

Where $UpGVCnet_{t-1}$ corresponds to the value added from j embedded in i's exports and captures the upstream alignment on GVCs of firms settling in j after i. Symmetrically, $DownGVCnet_{t-1}$ captures the downstream alignment on GVCs. The results of this directed analysis are in line with the general conclusions in describing an increasing influence of GVCs on the geographical coverage of French MNEs. From a sensibly similar upward and downward alignment patterns, the upward trend increases more, ending up with a marginal effect higher of 18% than the downward synchronization. The evolution of the two coefficients and their confidence interval is reported in figure ??

5.3 Dyadic panel econometric estimations

The MRQAP though suffers limitation to account for dependent variables that aren't matrices. To include additional control variables as robustness checks of the preliminary results achieved by the QAP estimations we reshaped the network data into a dyadic panel, in which the unit of observation is any pairs of countries from the multilayer network of 105 countries defined above, with a time dimension being the year of observations. We then include country-level determinants of multiple location choices derived from gravity-like model that appear to hold in FDI location choices (Alfaro and Chen, 2017). Specifically we add each countries' GDP. Their distance to France would complete this gravity-like model, but is captured by the origin and destination-specific fixed effects added in the model, to correct from the networks' non i.i.d observations. Yet, besides of capturing all country-specific variable (such as the distance to

	1997	1998	1999	2000	2001	2002	2003	2004
$GVCs_{t-1}$ (down)	0.029	0.025	0.032	0.033	0.031	0.035	0.034	0.032
	(18.35)**	(16.27)**	(18.17)**	(18.22)**	$(16.63)^{**}$	(18.54)**	$(17.65)^{**}$	(16.31)**
$GVCs_{t-1}$ (up)	0.027	0.026	0.034	0.036	0.035	0.038	0.037	0.037
	(17.25)**	$(16.79)^{**}$	(19.48)**	(20.12)**	(18.86)**	(20.09)**	$(18.93)^{**}$	(18.55)**
$\operatorname{Trade}_{t-1}$ (exports)	0.015	0.020	0.017	0.017	0.021	0.020	0.022	0.024
	$(10.72)^{**}$	$(14.21)^{**}$	(11.29)**	$(10.90)^{**}$	$(13.19)^{**}$	(12.22)**	$(12.86)^{**}$	$(14.36)^{**}$
$\operatorname{Trade}_{t-1}$ (imports)	0.013	0.016	0.016	0.017	0.019	0.019	0.022	0.022
	$(9.51)^{**}$	(11.55)**	$(10.31)^{**}$	(10.76)**	(11.80)**	(11.51)**	$(12.95)^{**}$	(13.13)**
Distnet	0.031	0.038	0.042	0.045	0.050	0.057	0.066	0.065
	(7.23)**	(8.68)**	$(8.99)^{**}$	$(9.22)^{**}$	(9.96)**	(11.06)**	$(12.51)^{**}$	(12.13)**
_cons	-0.394	-0.466	-0.511	-0.532	-0.600	-0.667	-0.743	-0.748
	$(10.39)^{**}$	$(11.97)^{**}$	$(12.17)^{**}$	(12.31)**	(13.32)**	(14.39)**	(15.74)**	$(15.53)^{**}$
R^2	0.28	0.28	0.31	0.31	0.32	0.33	0.32	0.32
Ν	29,070	29,070	29,070	29,070	29,070	29,070	29,070	$29,\!070$
	2005	2006	2007	2008	2009	2010	2011	
$GVCs_{t-1}$ (down)	0.031	0.033	0.034	0.032	0.035	0.037	0.039	-
	(16.05)**	$(16.64)^{**}$	(16.90)**	(15.71)**	$(16.78)^{**}$	(17.32)**	(17.83)**	
GVCs_{t-1} (up)	0.035	0.036	0.037	0.038	0.039	0.042	0.046	
(1)	(17.74)**	(17.93)**	$(18.66)^{**}$	(18.95)**	(18.75)**	(19.71)**	(21.04)**	
$\operatorname{Trade}_{t-1}$ (exports)	0.025	0.026	0.026	0.027	0.028	0.030	0.031	
	(15.35)**	(15.73)**	(15.94)**	(16.95)**	(17.16)**	(17.63)**	(18.62)**	
$\operatorname{Trade}_{t-1}$ (imports)	0.022	0.022	0.022	0.022	0.025	0.022	0.024	
	(13.58)**	(13.59)**	(13.35)**	(13.72)**	(15.24)**	(13.42)**	$(14.19)^{**}$	
Distnet				· /	· /	· /	. ,	
Distinct	0.057	0.056	0.054	0.054	0.066	0.062	0.064	
Distilet	0.057 (10.60)**	0.056 (9.98)**	0.054 (9.50)**	0.054 (9.51)**	0.066 (11.19)**	0.062 (10.29)**	0.064 (10.49)**	
_cons								
	$(10.60)^{**}$	$(9.98)^{**}$	$(9.50)^{**}$	$(9.51)^{**}$	$(11.19)^{**}$	$(10.29)^{**}$	$(10.49)^{**}$	
	(10.60)** -0.690	(9.98)** -0.696	(9.50)** -0.700	$(9.51)^{**}$ -0.722	(11.19)** -0.874	(10.29)** -0.818	(10.49)** -0.934	_

Table 3: Directed QAP estimations

France) that can no longer be directly interpreted, this solution offers less flexibility than the QAP, because only invariant characteristics are included in the fixed effects, and didn't adapt for new plants within the whole period.

Formally, we are estimating the model of equation (4) using an OLS, firstly without and then with origin and destination fixed effects. All variables are log-transformed.

$$MNE_{ij,t} = \beta_0 + \beta_1 * upGVC_{ij,t-1} + \beta_2 * downGVC_{ji,t-1} + \beta_3 * Exports_{ij,t-1} + \beta_4 * Imports_{ij,t-1} + \beta_5 * Distance_{ij} + \beta_6 * GDP_{i,t-1} + \beta_6 * GDP_{j,t-1} + \beta_7 * Dist_{FRA,i} + \beta_8 * Dist_{FRA,j} + \epsilon$$

$$(4)$$

The result of this estimation is reported in table ??, and each variable shows high significance

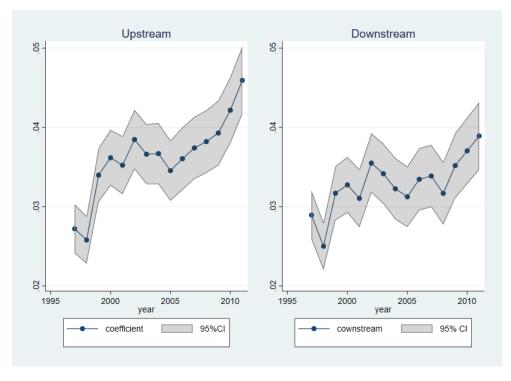


Figure 8: Directed QAP coefficients

level and expected signs, close to the QAP results.

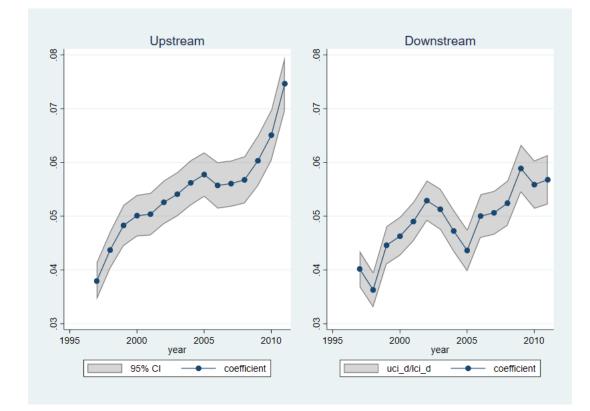


Figure 9: Directed OLS coefficients

	1997	1998	1999	2000	2001	2002	2003	2004
LlUGVCup	0.038	0.044	0.048	0.050	0.050	0.053	0.054	0.056
	$(21.43)^{**}$	$(24.76)^{**}$	(24.84)**	(25.51)**	$(25.12)^{**}$	(25.53)**	$(25.94)^{**}$	(26.30)**
LlUGVCdown	0.040	0.036	0.045	0.046	0.049	0.053	0.051	0.047
	(23.39)**	(21.64)**	$(24.67)^{**}$	(25.27)**	(26.36)**	$(27.69)^{**}$	(26.37)**	(23.70)**
L.lTrade	-0.000	0.000	-0.001	-0.002	-0.001	-0.001	-0.002	-0.002
	(0.24)	(0.03)	(0.39)	(1.25)	(0.56)	(0.46)	(1.22)	(1.34)
L.linvTrade	-0.004	-0.006	-0.006	-0.006	-0.005	-0.005	-0.005	-0.003
	$(2.70)^{**}$	$(3.55)^{**}$	$(3.32)^{**}$	$(3.34)^{**}$	$(3.05)^{**}$	$(2.92)^{**}$	$(2.87)^{**}$	(1.83)
lbildist	0.003	0.005	0.015	0.018	0.016	0.025	0.028	0.022
	(0.59)	(0.98)	$(2.88)^{**}$	$(3.46)^{**}$	(3.03)**	$(4.56)^{**}$	$(4.99)^{**}$	$(3.91)^{**}$
L.lgdp_0	0.034	0.041	0.040	0.037	0.039	0.040	0.044	0.047
01	$(4.06)^{**}$	$(4.81)^{**}$	$(4.14)^{**}$	(3.84)**	$(3.92)^{**}$	$(3.82)^{**}$	$(4.11)^{**}$	$(4.33)^{**}$
L.lgdp_d	0.003	-0.001	-0.000	0.000	0.000	-0.001	-0.003	-0.005
0	(1.60)	(0.40)	(0.25)	(0.03)	(0.03)	(0.60)	(1.61)	$(2.27)^{*}$
_cons	-0.963	-1.056	-1.129	-1.095	-1.149	-1.217	-1.279	-1.285
	$(4.64)^{**}$	$(4.98)^{**}$	$(4.78)^{**}$	$(4.67)^{**}$	$(4.69)^{**}$	$(4.73)^{**}$	$(4.84)^{**}$	$(4.72)^{**}$
R^2 (overall)	0.29	0.29	0.32	0.33	.33	0.34	0.34	0.33
N	22,916	$22,\!916$	23,220	$23,\!672$	$23,\!981$	$23,\!981$	$23,\!981$	$24,\!291$
	2005	2006	2007	2008	2009	2010	2011	
	2005	2006	2007	2008	2009	2010	2011	-
LUCYC	2005	2006	2007	2008	2009	2010	2011	
LlUGVCup	$2005 \\ 0.058$	2006 0.056	2007 0.056	2008 0.057	2009 0.060	2010 0.065	2011 0.075	
-	2005 0.058 (27.54)**	2006 0.056 (25.43)**	2007 0.056 (25.70)**	2008 0.057 (25.45)**	2009 0.060 (25.16)**	2010 0.065 (26.82)**	2011 0.075 (28.97)**	
LlUGVCup LlUGVCdown	2005 0.058 (27.54)** 0.044	2006 0.056 (25.43)** 0.050	2007 0.056 (25.70)** 0.051	2008 0.057 (25.45)** 0.052	2009 0.060 (25.16)** 0.059	2010 0.065 (26.82)** 0.056	2011 0.075 (28.97)** 0.057	
LlUGVCdown	2005 0.058 (27.54)** 0.044 (22.02)**	2006 0.056 (25.43)** 0.050 (24.28)**	2007 0.056 (25.70)** 0.051 (24.48)**	2008 0.057 (25.45)** 0.052 (24.61)**	2009 0.060 (25.16)** 0.059 (25.97)**	2010 0.065 (26.82)** 0.056 (24.47)**	2011 0.075 (28.97)** 0.057 (24.32)**	-
LlUGVCdown	2005 0.058 (27.54)** 0.044 (22.02)** -0.001	2006 0.056 (25.43)** 0.050 (24.28)** 0.001	2007 0.056 (25.70)** 0.051 (24.48)** -0.001	2008 0.057 (25.45)** 0.052 (24.61)** -0.001	2009 0.060 (25.16)** 0.059 (25.97)** 0.000	2010 0.065 (26.82)** 0.056 (24.47)** 0.002	2011 0.075 (28.97)** 0.057 (24.32)** 0.001	-
LlUGVCdown L.lTrade	2005 0.058 (27.54)** 0.044 (22.02)** -0.001 (0.29)	2006 0.056 (25.43)** 0.050 (24.28)** 0.001 (0.36)	2007 0.056 (25.70)** 0.051 (24.48)** -0.001 (0.29)	$\begin{array}{c} 2008 \\ 0.057 \\ (25.45)^{**} \\ 0.052 \\ (24.61)^{**} \\ -0.001 \\ (0.48) \end{array}$	2009 0.060 (25.16)** 0.059 (25.97)** 0.000 (0.08)	2010 0.065 (26.82)** 0.056 (24.47)** 0.002 (0.78)	2011 0.075 (28.97)** 0.057 (24.32)** 0.001 (0.57)	-
-	2005 0.058 (27.54)** 0.044 (22.02)** -0.001 (0.29) -0.004	2006 0.056 (25.43)** 0.050 (24.28)** 0.001 (0.36) -0.003	2007 0.056 (25.70)** 0.051 (24.48)** -0.001 (0.29) -0.005	$\begin{array}{c} 2008 \\ 0.057 \\ (25.45)^{**} \\ 0.052 \\ (24.61)^{**} \\ -0.001 \\ (0.48) \\ -0.004 \end{array}$	2009 0.060 (25.16)** 0.059 (25.97)** 0.000 (0.08) -0.005	2010 0.065 (26.82)** 0.056 (24.47)** 0.002 (0.78) -0.005	2011 0.075 (28.97)** 0.057 (24.32)** 0.001 (0.57) -0.004	-
LlUGVCdown L.lTrade L.linvTrade	2005 0.058 (27.54)** 0.044 (22.02)** -0.001 (0.29) -0.004 (2.16)*	$\begin{array}{c} 2006 \\ 0.056 \\ (25.43)^{**} \\ 0.050 \\ (24.28)^{**} \\ 0.001 \\ (0.36) \\ -0.003 \\ (1.74) \end{array}$	$\begin{array}{c} 2007\\ 0.056\\ (25.70)^{**}\\ 0.051\\ (24.48)^{**}\\ -0.001\\ (0.29)\\ -0.005\\ (2.53)^{*} \end{array}$	$\begin{array}{c} 2008 \\ 0.057 \\ (25.45)^{**} \\ 0.052 \\ (24.61)^{**} \\ -0.001 \\ (0.48) \\ -0.004 \\ (2.29)^{*} \end{array}$	$\begin{array}{c} 2009 \\ 0.060 \\ (25.16)^{**} \\ 0.059 \\ (25.97)^{**} \\ 0.000 \\ (0.08) \\ -0.005 \\ (2.45)^{*} \end{array}$	$\begin{array}{c} 2010 \\ 0.065 \\ (26.82)^{**} \\ 0.056 \\ (24.47)^{**} \\ 0.002 \\ (0.78) \\ -0.005 \\ (2.41)^{*} \end{array}$	$\begin{array}{c} 2011\\ 0.075\\ (28.97)^{**}\\ 0.057\\ (24.32)^{**}\\ 0.001\\ (0.57)\\ -0.004\\ (2.24)^{*} \end{array}$	-
LlUGVCdown L.lTrade	2005 0.058 (27.54)** 0.044 (22.02)** -0.001 (0.29) -0.004 (2.16)* 0.008	$\begin{array}{c} 2006 \\ 0.056 \\ (25.43)^{**} \\ 0.050 \\ (24.28)^{**} \\ 0.001 \\ (0.36) \\ -0.003 \\ (1.74) \\ 0.004 \end{array}$	$\begin{array}{c} 2007\\ 0.056\\ (25.70)^{**}\\ 0.051\\ (24.48)^{**}\\ -0.001\\ (0.29)\\ -0.005\\ (2.53)^{*}\\ 0.002 \end{array}$	$\begin{array}{c} 2008 \\ 0.057 \\ (25.45)^{**} \\ 0.052 \\ (24.61)^{**} \\ -0.001 \\ (0.48) \\ -0.004 \\ (2.29)^{*} \\ -0.000 \end{array}$	$\begin{array}{c} 2009\\ 0.060\\ (25.16)^{**}\\ 0.059\\ (25.97)^{**}\\ 0.000\\ (0.08)\\ -0.005\\ (2.45)^{*}\\ 0.011\end{array}$	$\begin{array}{c} 2010\\ 0.065\\ (26.82)^{**}\\ 0.056\\ (24.47)^{**}\\ 0.002\\ (0.78)\\ -0.005\\ (2.41)^{*}\\ 0.006 \end{array}$	2011 0.075 (28.97)** 0.057 (24.32)** 0.001 (0.57) -0.004 (2.24)* 0.009	-
LlUGVCdown L.lTrade L.linvTrade lbildist	$\begin{array}{c} 2005 \\ 0.058 \\ (27.54)^{**} \\ 0.044 \\ (22.02)^{**} \\ -0.001 \\ (0.29) \\ -0.004 \\ (2.16)^{*} \\ 0.008 \\ (1.34) \end{array}$	$\begin{array}{c} 2006 \\ 0.056 \\ (25.43)^{**} \\ 0.050 \\ (24.28)^{**} \\ 0.001 \\ (0.36) \\ -0.003 \\ (1.74) \\ 0.004 \\ (0.62) \end{array}$	$\begin{array}{c} 2007\\ 0.056\\ (25.70)^{**}\\ 0.051\\ (24.48)^{**}\\ -0.001\\ (0.29)\\ -0.005\\ (2.53)^{*}\\ 0.002\\ (0.33) \end{array}$	$\begin{array}{c} 2008 \\ 0.057 \\ (25.45)^{**} \\ 0.052 \\ (24.61)^{**} \\ -0.001 \\ (0.48) \\ -0.004 \\ (2.29)^{*} \\ -0.000 \\ (0.04) \end{array}$	$\begin{array}{c} 2009\\ 0.060\\ (25.16)^{**}\\ 0.059\\ (25.97)^{**}\\ 0.000\\ (0.08)\\ -0.005\\ (2.45)^{*}\\ 0.011\\ (1.67)\end{array}$	$\begin{array}{c} 2010\\ 0.065\\ (26.82)^{**}\\ 0.056\\ (24.47)^{**}\\ 0.002\\ (0.78)\\ -0.005\\ (2.41)^{*}\\ 0.006\\ (0.89) \end{array}$	$\begin{array}{c} 2011\\ 0.075\\ (28.97)^{**}\\ 0.057\\ (24.32)^{**}\\ 0.001\\ (0.57)\\ -0.004\\ (2.24)^{*}\\ 0.009\\ (1.30) \end{array}$	-
LlUGVCdown L.lTrade L.linvTrade	$\begin{array}{c} 2005 \\ 0.058 \\ (27.54)^{**} \\ 0.044 \\ (22.02)^{**} \\ -0.001 \\ (0.29) \\ -0.004 \\ (2.16)^{*} \\ 0.008 \\ (1.34) \\ 0.047 \end{array}$	$\begin{array}{c} 2006 \\ 0.056 \\ (25.43)^{**} \\ 0.050 \\ (24.28)^{**} \\ 0.001 \\ (0.36) \\ -0.003 \\ (1.74) \\ 0.004 \\ (0.62) \\ 0.047 \end{array}$	$\begin{array}{c} 2007\\ 0.056\\ (25.70)^{**}\\ 0.051\\ (24.48)^{**}\\ -0.001\\ (0.29)\\ -0.005\\ (2.53)^{*}\\ 0.002\\ (0.33)\\ 0.051 \end{array}$	$\begin{array}{c} 2008 \\ 0.057 \\ (25.45)^{**} \\ 0.052 \\ (24.61)^{**} \\ -0.001 \\ (0.48) \\ -0.004 \\ (2.29)^{*} \\ -0.000 \\ (0.04) \\ 0.048 \end{array}$	$\begin{array}{c} 2009\\ 0.060\\ (25.16)^{**}\\ 0.059\\ (25.97)^{**}\\ 0.000\\ (0.08)\\ -0.005\\ (2.45)^{*}\\ 0.011\\ (1.67)\\ 0.052 \end{array}$	$\begin{array}{c} 2010\\ 0.065\\ (26.82)^{**}\\ 0.056\\ (24.47)^{**}\\ 0.002\\ (0.78)\\ -0.005\\ (2.41)^{*}\\ 0.006\\ (0.89)\\ 0.053 \end{array}$	$\begin{array}{c} 2011\\ 0.075\\ (28.97)^{**}\\ 0.057\\ (24.32)^{**}\\ 0.001\\ (0.57)\\ -0.004\\ (2.24)^{*}\\ 0.009\\ (1.30)\\ 0.061 \end{array}$	-
LlUGVCdown L.lTrade L.linvTrade lbildist L.lgdp_0	$\begin{array}{c} 2005 \\ 0.058 \\ (27.54)^{**} \\ 0.044 \\ (22.02)^{**} \\ -0.001 \\ (0.29) \\ -0.004 \\ (2.16)^{*} \\ 0.008 \\ (1.34) \\ 0.047 \\ (4.22)^{**} \end{array}$	$\begin{array}{c} 2006 \\ 0.056 \\ (25.43)^{**} \\ 0.050 \\ (24.28)^{**} \\ 0.001 \\ (0.36) \\ -0.003 \\ (1.74) \\ 0.004 \\ (0.62) \\ 0.047 \\ (4.05)^{**} \end{array}$	$\begin{array}{c} 2007\\ 0.056\\ (25.70)^{**}\\ 0.051\\ (24.48)^{**}\\ -0.001\\ (0.29)\\ -0.005\\ (2.53)^{*}\\ 0.002\\ (0.33)\\ 0.051\\ (4.34)^{**} \end{array}$	$\begin{array}{c} 2008 \\ 0.057 \\ (25.45)^{**} \\ 0.052 \\ (24.61)^{**} \\ -0.001 \\ (0.48) \\ -0.004 \\ (2.29)^{*} \\ -0.000 \\ (0.04) \\ 0.048 \\ (4.06)^{**} \end{array}$	$\begin{array}{c} 2009\\ 0.060\\ (25.16)^{**}\\ 0.059\\ (25.97)^{**}\\ 0.000\\ (0.08)\\ -0.005\\ (2.45)^{*}\\ 0.011\\ (1.67)\\ 0.052\\ (4.09)^{**} \end{array}$	$\begin{array}{c} 2010\\ 0.065\\ (26.82)^{**}\\ 0.056\\ (24.47)^{**}\\ 0.002\\ (0.78)\\ -0.005\\ (2.41)^{*}\\ 0.006\\ (0.89)\\ 0.053\\ (4.22)^{**} \end{array}$	$\begin{array}{c} 2011\\ 0.075\\ (28.97)^{**}\\ 0.057\\ (24.32)^{**}\\ 0.001\\ (0.57)\\ -0.004\\ (2.24)^{*}\\ 0.009\\ (1.30)\\ 0.061\\ (4.66)^{**} \end{array}$	-
LlUGVCdown L.lTrade L.linvTrade lbildist	$\begin{array}{c} 2005 \\ 0.058 \\ (27.54)^{**} \\ 0.044 \\ (22.02)^{**} \\ -0.001 \\ (0.29) \\ -0.004 \\ (2.16)^{*} \\ 0.008 \\ (1.34) \\ 0.047 \\ (4.22)^{**} \\ -0.007 \end{array}$	$\begin{array}{c} 2006 \\ 0.056 \\ (25.43)^{**} \\ 0.050 \\ (24.28)^{**} \\ 0.001 \\ (0.36) \\ -0.003 \\ (1.74) \\ 0.004 \\ (0.62) \\ 0.047 \\ (4.05)^{**} \\ -0.006 \end{array}$	$\begin{array}{c} 2007\\ 0.056\\ (25.70)^{**}\\ 0.051\\ (24.48)^{**}\\ -0.001\\ (0.29)\\ -0.005\\ (2.53)^{*}\\ 0.002\\ (0.33)\\ 0.051\\ (4.34)^{**}\\ -0.006 \end{array}$	$\begin{array}{c} 2008 \\ 0.057 \\ (25.45)^{**} \\ 0.052 \\ (24.61)^{**} \\ -0.001 \\ (0.48) \\ -0.004 \\ (2.29)^{*} \\ -0.000 \\ (0.04) \\ 0.048 \\ (4.06)^{**} \\ -0.007 \end{array}$	$\begin{array}{c} 2009\\ 0.060\\ (25.16)^{**}\\ 0.059\\ (25.97)^{**}\\ 0.000\\ (0.08)\\ -0.005\\ (2.45)^{*}\\ 0.011\\ (1.67)\\ 0.052\\ (4.09)^{**}\\ -0.005 \end{array}$	$\begin{array}{c} 2010\\ 0.065\\ (26.82)^{**}\\ 0.056\\ (24.47)^{**}\\ 0.002\\ (0.78)\\ -0.005\\ (2.41)^{*}\\ 0.006\\ (0.89)\\ 0.053\\ (4.22)^{**}\\ -0.010\\ \end{array}$	$\begin{array}{c} 2011\\ 0.075\\ (28.97)^{**}\\ 0.057\\ (24.32)^{**}\\ 0.001\\ (0.57)\\ -0.004\\ (2.24)^{*}\\ 0.009\\ (1.30)\\ 0.061\\ (4.66)^{**}\\ -0.011 \end{array}$	-
LlUGVCdown L.lTrade L.linvTrade lbildist L.lgdp_0 L.lgdp_d	$\begin{array}{c} 2005\\ 0.058\\ (27.54)^{**}\\ 0.044\\ (22.02)^{**}\\ -0.001\\ (0.29)\\ -0.004\\ (2.16)^{*}\\ 0.008\\ (1.34)\\ 0.047\\ (4.22)^{**}\\ -0.007\\ (3.36)^{**} \end{array}$	$\begin{array}{c} 2006 \\ 0.056 \\ (25.43)^{**} \\ 0.050 \\ (24.28)^{**} \\ 0.001 \\ (0.36) \\ -0.003 \\ (1.74) \\ 0.004 \\ (0.62) \\ 0.047 \\ (4.05)^{**} \\ -0.006 \\ (2.54)^{*} \end{array}$	$\begin{array}{c} 2007\\ 0.056\\ (25.70)^{**}\\ 0.051\\ (24.48)^{**}\\ -0.001\\ (0.29)\\ -0.005\\ (2.53)^{*}\\ 0.002\\ (0.33)\\ 0.051\\ (4.34)^{**}\\ -0.006\\ (2.72)^{**} \end{array}$	$\begin{array}{c} 2008 \\ 0.057 \\ (25.45)^{**} \\ 0.052 \\ (24.61)^{**} \\ -0.001 \\ (0.48) \\ -0.004 \\ (2.29)^{*} \\ -0.000 \\ (0.04) \\ 0.048 \\ (4.06)^{**} \\ -0.007 \\ (2.95)^{**} \end{array}$	$\begin{array}{c} 2009\\ 0.060\\ (25.16)^{**}\\ 0.059\\ (25.97)^{**}\\ 0.000\\ (0.08)\\ -0.005\\ (2.45)^{*}\\ 0.011\\ (1.67)\\ 0.052\\ (4.09)^{**}\\ -0.005\\ (1.95)\end{array}$	$\begin{array}{c} 2010\\ 0.065\\ (26.82)^{**}\\ 0.056\\ (24.47)^{**}\\ 0.002\\ (0.78)\\ -0.005\\ (2.41)^{*}\\ 0.006\\ (0.89)\\ 0.053\\ (4.22)^{**}\\ -0.010\\ (3.93)^{**} \end{array}$	$\begin{array}{c} 2011\\ 0.075\\ (28.97)^{**}\\ 0.057\\ (24.32)^{**}\\ 0.001\\ (0.57)\\ -0.004\\ (2.24)^{*}\\ 0.009\\ (1.30)\\ 0.061\\ (4.66)^{**}\\ -0.011\\ (3.99)^{**} \end{array}$	-
LlUGVCdown L.lTrade L.linvTrade lbildist L.lgdp_0	$\begin{array}{c} 2005 \\ 0.058 \\ (27.54)^{**} \\ 0.044 \\ (22.02)^{**} \\ -0.001 \\ (0.29) \\ -0.004 \\ (2.16)^{*} \\ 0.008 \\ (1.34) \\ 0.047 \\ (4.22)^{**} \\ -0.007 \\ (3.36)^{**} \\ -1.111 \end{array}$	$\begin{array}{c} 2006 \\ 0.056 \\ (25.43)^{**} \\ 0.050 \\ (24.28)^{**} \\ 0.001 \\ (0.36) \\ -0.003 \\ (1.74) \\ 0.004 \\ (0.62) \\ 0.047 \\ (4.05)^{**} \\ -0.006 \\ (2.54)^{*} \\ -1.141 \end{array}$	$\begin{array}{c} 2007\\ 0.056\\ (25.70)^{**}\\ 0.051\\ (24.48)^{**}\\ -0.001\\ (0.29)\\ -0.005\\ (2.53)^{*}\\ 0.002\\ (0.33)\\ 0.051\\ (4.34)^{**}\\ -0.006\\ (2.72)^{**}\\ -1.229\end{array}$	$\begin{array}{c} 2008 \\ 0.057 \\ (25.45)^{**} \\ 0.052 \\ (24.61)^{**} \\ -0.001 \\ (0.48) \\ -0.004 \\ (2.29)^{*} \\ -0.000 \\ (0.04) \\ 0.048 \\ (4.06)^{**} \\ -0.007 \\ (2.95)^{**} \\ -1.168 \end{array}$	$\begin{array}{c} 2009\\ 0.060\\ (25.16)^{**}\\ 0.059\\ (25.97)^{**}\\ 0.000\\ (0.08)\\ -0.005\\ (2.45)^{*}\\ 0.011\\ (1.67)\\ 0.052\\ (4.09)^{**}\\ -0.005\\ (1.95)\\ -1.470\\ \end{array}$	$\begin{array}{c} 2010\\ 0.065\\ (26.82)^{**}\\ 0.056\\ (24.47)^{**}\\ 0.002\\ (0.78)\\ -0.005\\ (2.41)^{*}\\ 0.006\\ (0.89)\\ 0.053\\ (4.22)^{**}\\ -0.010\\ (3.93)^{**}\\ -1.290 \end{array}$	$\begin{array}{c} 2011\\ 0.075\\ (28.97)^{**}\\ 0.057\\ (24.32)^{**}\\ 0.001\\ (0.57)\\ -0.004\\ (2.24)^{*}\\ 0.009\\ (1.30)\\ 0.061\\ (4.66)^{**}\\ -0.011\\ (3.99)^{**}\\ -1.617\end{array}$	-
LlUGVCdown L.lTrade L.linvTrade lbildist L.lgdp_0 L.lgdp_d _cons	$\begin{array}{c} 2005 \\ 0.058 \\ (27.54)^{**} \\ 0.044 \\ (22.02)^{**} \\ -0.001 \\ (0.29) \\ -0.004 \\ (2.16)^{*} \\ 0.008 \\ (1.34) \\ 0.047 \\ (4.22)^{**} \\ -0.007 \\ (3.36)^{**} \\ -1.111 \\ (3.97)^{**} \end{array}$	$\begin{array}{c} 2006 \\ 0.056 \\ (25.43)^{**} \\ 0.050 \\ (24.28)^{**} \\ 0.001 \\ (0.36) \\ -0.003 \\ (1.74) \\ 0.004 \\ (0.62) \\ 0.047 \\ (4.05)^{**} \\ -0.006 \\ (2.54)^{*} \\ -1.141 \\ (3.90)^{**} \end{array}$	$\begin{array}{c} 2007\\ 0.056\\ (25.70)^{**}\\ 0.051\\ (24.48)^{**}\\ -0.001\\ (0.29)\\ -0.005\\ (2.53)^{*}\\ 0.002\\ (0.33)\\ 0.051\\ (4.34)^{**}\\ -0.006\\ (2.72)^{**}\\ -1.229\\ (4.15)^{**} \end{array}$	$\begin{array}{c} 2008 \\ 0.057 \\ (25.45)^{**} \\ 0.052 \\ (24.61)^{**} \\ -0.001 \\ (0.48) \\ -0.004 \\ (2.29)^{*} \\ -0.000 \\ (0.04) \\ 0.048 \\ (4.06)^{**} \\ -0.007 \\ (2.95)^{**} \\ -1.168 \\ (3.89)^{**} \end{array}$	$\begin{array}{c} 2009\\ 0.060\\ (25.16)^{**}\\ 0.059\\ (25.97)^{**}\\ 0.000\\ (0.08)\\ -0.005\\ (2.45)^{*}\\ 0.011\\ (1.67)\\ 0.052\\ (4.09)^{**}\\ -0.005\\ (1.95)\\ -1.470\\ (4.54)^{**} \end{array}$	$\begin{array}{c} 2010\\ 0.065\\ (26.82)^{**}\\ 0.056\\ (24.47)^{**}\\ 0.002\\ (0.78)\\ -0.005\\ (2.41)^{*}\\ 0.006\\ (0.89)\\ 0.053\\ (4.22)^{**}\\ -0.010\\ (3.93)^{**}\\ -1.290\\ (4.00)^{**} \end{array}$	$\begin{array}{c} 2011\\ 0.075\\ (28.97)^{**}\\ 0.057\\ (24.32)^{**}\\ 0.001\\ (0.57)\\ -0.004\\ (2.24)^{*}\\ 0.009\\ (1.30)\\ 0.061\\ (4.66)^{**}\\ -0.011\\ (3.99)^{**}\\ -1.617\\ (4.86)^{**} \end{array}$	-
LlUGVCdown L.lTrade L.linvTrade lbildist L.lgdp_0 L.lgdp_d _cons R^2 (overall)	$\begin{array}{c} 2005\\ 0.058\\ (27.54)^{**}\\ 0.044\\ (22.02)^{**}\\ -0.001\\ (0.29)\\ -0.004\\ (2.16)^{*}\\ 0.008\\ (1.34)\\ 0.047\\ (4.22)^{**}\\ -0.007\\ (3.36)^{**}\\ -1.111\\ (3.97)^{**}\\ \hline 0.32 \end{array}$	$\begin{array}{c} 2006\\ 0.056\\ (25.43)^{**}\\ 0.050\\ (24.28)^{**}\\ 0.001\\ (0.36)\\ -0.003\\ (1.74)\\ 0.004\\ (0.62)\\ 0.047\\ (4.05)^{**}\\ -0.006\\ (2.54)^{*}\\ -1.141\\ (3.90)^{**}\\ \hline 0.33 \end{array}$	$\begin{array}{c} 2007\\ 0.056\\ (25.70)^{**}\\ 0.051\\ (24.48)^{**}\\ -0.001\\ (0.29)\\ -0.005\\ (2.53)^{*}\\ 0.002\\ (0.33)\\ 0.051\\ (4.34)^{**}\\ -0.006\\ (2.72)^{**}\\ -1.229\\ (4.15)^{**}\\ \hline 0.33 \end{array}$	$\begin{array}{c} 2008 \\ 0.057 \\ (25.45)^{**} \\ 0.052 \\ (24.61)^{**} \\ -0.001 \\ (0.48) \\ -0.004 \\ (2.29)^{*} \\ -0.000 \\ (0.04) \\ 0.048 \\ (4.06)^{**} \\ -0.007 \\ (2.95)^{**} \\ -1.168 \\ (3.89)^{**} \\ \hline 0.33 \end{array}$	$\begin{array}{c} 2009\\ 0.060\\ (25.16)^{**}\\ 0.059\\ (25.97)^{**}\\ 0.000\\ (0.08)\\ -0.005\\ (2.45)^{*}\\ 0.011\\ (1.67)\\ 0.052\\ (4.09)^{**}\\ -0.005\\ (1.95)\\ -1.470\\ (4.54)^{**}\\ \hline 0.34 \end{array}$	$\begin{array}{c} 2010\\ 0.065\\ (26.82)^{**}\\ 0.056\\ (24.47)^{**}\\ 0.002\\ (0.78)\\ -0.005\\ (2.41)^{*}\\ 0.006\\ (0.89)\\ 0.053\\ (4.22)^{**}\\ -0.010\\ (3.93)^{**}\\ -1.290\\ (4.00)^{**}\\ 0.35\end{array}$	$\begin{array}{c} 2011\\ 0.075\\ (28.97)^{**}\\ 0.057\\ (24.32)^{**}\\ 0.001\\ (0.57)\\ -0.004\\ (2.24)^{*}\\ 0.009\\ (1.30)\\ 0.061\\ (4.66)^{**}\\ -0.011\\ (3.99)^{**}\\ -1.617\\ (4.86)^{**}\\ 0.35 \end{array}$	-
LlUGVCdown L.lTrade L.linvTrade lbildist L.lgdp_0 L.lgdp_d _cons	$\begin{array}{c} 2005 \\ 0.058 \\ (27.54)^{**} \\ 0.044 \\ (22.02)^{**} \\ -0.001 \\ (0.29) \\ -0.004 \\ (2.16)^{*} \\ 0.008 \\ (1.34) \\ 0.047 \\ (4.22)^{**} \\ -0.007 \\ (3.36)^{**} \\ -1.111 \\ (3.97)^{**} \end{array}$	$\begin{array}{c} 2006\\ 0.056\\ (25.43)^{**}\\ 0.050\\ (24.28)^{**}\\ 0.001\\ (0.36)\\ -0.003\\ (1.74)\\ 0.004\\ (0.62)\\ 0.047\\ (4.05)^{**}\\ -0.006\\ (2.54)^{*}\\ -1.141\\ (3.90)^{**} \end{array}$	$\begin{array}{c} 2007\\ 0.056\\ (25.70)^{**}\\ 0.051\\ (24.48)^{**}\\ -0.001\\ (0.29)\\ -0.005\\ (2.53)^{*}\\ 0.002\\ (0.33)\\ 0.051\\ (4.34)^{**}\\ -0.006\\ (2.72)^{**}\\ -1.229\\ (4.15)^{**} \end{array}$	$\begin{array}{c} 2008 \\ 0.057 \\ (25.45)^{**} \\ 0.052 \\ (24.61)^{**} \\ -0.001 \\ (0.48) \\ -0.004 \\ (2.29)^{*} \\ -0.000 \\ (0.04) \\ 0.048 \\ (4.06)^{**} \\ -0.007 \\ (2.95)^{**} \\ -1.168 \\ (3.89)^{**} \end{array}$	$\begin{array}{c} 2009\\ 0.060\\ (25.16)^{**}\\ 0.059\\ (25.97)^{**}\\ 0.000\\ (0.08)\\ -0.005\\ (2.45)^{*}\\ 0.011\\ (1.67)\\ 0.052\\ (4.09)^{**}\\ -0.005\\ (1.95)\\ -1.470\\ (4.54)^{**} \end{array}$	$\begin{array}{c} 2010\\ 0.065\\ (26.82)^{**}\\ 0.056\\ (24.47)^{**}\\ 0.002\\ (0.78)\\ -0.005\\ (2.41)^{*}\\ 0.006\\ (0.89)\\ 0.053\\ (4.22)^{**}\\ -0.010\\ (3.93)^{**}\\ -1.290\\ (4.00)^{**} \end{array}$	$\begin{array}{c} 2011\\ 0.075\\ (28.97)^{**}\\ 0.057\\ (24.32)^{**}\\ 0.001\\ (0.57)\\ -0.004\\ (2.24)^{*}\\ 0.009\\ (1.30)\\ 0.061\\ (4.66)^{**}\\ -0.011\\ (3.99)^{**}\\ -1.617\\ (4.86)^{**} \end{array}$	-

Table 4: Directed OLS estimations

The results of the interacted model show an increasing influence of GVCs, pictured in figure 9 for clarity purposes. An interesting result is the acceleration of the increasing trend since 2009, in line with the finding of the QAP model, of increased GVCs' influence over location choice when they were stabilizing.

The results are reported in table ??, and the evolution of the GVCs' coefficient is pictured in figure ??.

The estimations on the whole panel without and with fixed effects confirm the co-existence of both upstream and downstream alignment channels, with an overall similar effect of each, notably when controlling for country pairs fixed effects. Yet, the separate regressions on the two sub-samples show an opposite trend of each direction. The downward trend, that was less important in the initial period increases and becomes dominant in the most recent period. To the contrary, upward alignment of multinationals on GVCs decreases, although remaining positive and significant. This result is particularly interesting because it contradicts a common perception that links global value chains and international fragmentation of the production to outsourcing only. This result suggest that French firms increasingly go toward countries making subsequent stages of production in GVCs compared to their previous locations. To confirm this particularly interesting result, we run a similar interacted model than done with the undirected flows, to enable to see a yearly coefficient of both Upward and Downward alignment, to be interpreted as a yearly marginal effect compared to baseline level (of 1997). The table **??** reports the estimates, and figure **??** displays the coefficients of upstream and downstream alignment, and their 95% confidence interval. Each of these coefficients are significant at the 1% level

These results confirm the conclusion drawn from sub-sample analysis. with a stronger initial effect of upstream alignment in 1997 with respect to downstream trend. Yet, the upstream trend decreases each year, despite a small rebound after the 2008 crisis while the downstream channel increases at a stable pace.

6 Concluding remarks

In this paper, we develop a network based approach of Multinationals' geographical expansion using micro data from a French firm-level sample. The network analysis allows to unveil the complexity of multiple locations decision, to study some determinants of the international coverage of firms and its structural changes. Specifically, this approach reveals a decentralization trend of global location choices toward a less hierarchical structure. We further link this trend with an alignment on the emerging Global Value Chains and the slicing up of production processes. Following a flourishing literature, the GVCs' network is estimated through the foreign value added embedded from any country to any one else exports from international Input-Output tables, and reflects the true international production networks. Focusing on the co-evolution of several layers from the complex network that is the international economy through a Multiple Regression Quadratic Assignment Procedure shows that the MNEs follow more and more these cross-countries value added linkages when settling new subsidiaries abroad. This result holds when adding gravity-like control variable of the location choices such as host countries GDP, bilateral distances, as well as years and countries fixed effects in a traditional OLS model. Then, using a directed network which reveals the sequence of internationalization steps of French firms, I investigated whether this global reorganization of firms' affiliates all along the GVCs is mainly drawn toward upstream or downstream stages of production. It appears that a shift occurred between these two strategies. While the downstream and upstream alignment were sensitively equal in the late 1990s, the upstream alignment grew more rapidly and became the main driver of the synchronization of MNEs' and GVCs' networks since the mid-2000s. French multinationals' are increasingly locating in countries whose value added are largely represented in the exports of their previous host countries.

Despite these robust results, the limit of this approach is the absence of direct evidence about MNEs going in these countries to control an internationally fragmented production. Such proof would require detailed intra-firm flows between all foreign plants, to compare this within-trade network to the actual GVCs. But such data isn't currently available for researchers. Therefore, we only report that the MNEs' geographical coverage is getting closer to the GVCs flows, which suggests an international fragmentation of their production, and notably by offshoring early production stages, or integrating their foreign suppliers.

Nevertheless, French firms have increasingly opened new affiliates all along the roads value added follows, and whatever are the exact intra-firm flows, this increases their sensitivity to the GVCs fluctuations. Any dismantling of global value chains would negatively impact firms that have established their foreign plants all along them. Even for those who do not directly participate to them, they are part of an economic environment that rely on GVCs, which will deteriorate in case of dismantlement of GVCs. In addition, the threats to GVCs would impact all kind of transactions between these affiliates, and not only the specific trade in value added. Therefore, the inherent vulnerability of multinational firms to the sustainability of global value chains, is strengthened by the alignment of their plants on global value chains I report. Policymakers should therefore be aware than although being the most efficient firms, and benefiting from the globalization, the multinational firms are actually increasingly tied to the future of global value chains, and their brutal dismantling would affect them directly.

International extension of this research would be meaningful, not only to know which are the countries whose firms are the most (or the least) geographically aligned on GVCs, but also to examine whether the upward-driven alignment depends on the domestic country's position on GVCs. Further extensions should also include firm heterogeneity to identify the most sensible firms to this worldwide setup of their production units.

References

- Alfaro, L. and Chen, M. X. Transportation Cost and the Geography of Foreign Investment. Harvard Business School Working Papers 17-061, Harvard Business School, January 2017.
- Alfaro, L., Antràs, P., Chor, D., and Conconi, P. Internalizing global value chains: A firm-level analysis. *Journal of Political Economy*, 2016 Forthcoming.
- Amador, J. and Cabral, S. Networks of value-added trade. *The World Economy*, 40(7):1291–1313, 2017.
- Amador, J., Cabral, S., Mastrandrea, R., and Ruzzenenti, F. Who's who in global value chains? a weighted network approach. Open Economies Review, 29(5):1039–1059, Nov 2018.
- Antràs, P. and Chor, D. Organizing the global value chain. *Econometrica*, 81(6):2127–2204, 2013.
- Antràs, P. and Chor, D. On the measurement of upstreamness and downstreamness in global value chains. Working paper, National Bureau of Economic Research, 2018.
- Antràs, P. and De Gortari, A. On the geography of global value chains. Working paper, National Bureau of Economic Research, 2017.
- Baldwin, R. Globalisation: the great unbundling (s). Economic Council of Finland, 20, 2006.
- Baldwin, R. The great convergence. Harvard University Press, 2016.
- Baldwin, R. and Venables, A. J. Spiders and snakes: Offshoring and agglomeration in the global economy. *Journal of International Economics*, March 2013.
- Blonigen, B. A. A review of the empirical literature on FDI determinants. Atlantic Economic Journal, 33(4):383–403, 2005.
- Cadestin, C., De Backer, K., Desnoyers-James, I., Miroudot, S., Rigo, D., and Ye, M. Multinational enterprises and global value chains: the oecd analytical amne database. 2018a.
- Cadestin, C., De Backer, K., Desnoyers-James, I., Miroudot, S., Ye, M., and Rigo, D. Multinational enterprises and global value chains: New insights on the trade-investment nexus. 2018b.
- Cerina, F., Zhu, Z., Chessa, A., and Riccaboni, M. World input-output network. *PloS one*, 10 (7):e0134025, 2015.
- Chen, M. X. and Moore, M. O. Location decision of heterogeneous multinational firms. *Journal* of International Economics, 80(2):188–199, 2010.
- Costinot, A., Vogel, J., and Wang, S. An elementary theory of global supply chains. *The Review* of Economic Studies, 80(1):109–144, 2013.
- Criscuolo, C. and Timmis, J. GVCs and centrality: Mapping key hubs, spokes and the periphery. Working paper, OECD Publishing, 2018.
- Daudin, G., Rifflart, C., and Schweisguth, D. Who produces for whom in the world economy? Canadian Journal of Economics/Revue canadienne d'économique, 44(4):1403–1437, 2011.
- De Benedictis, L. and Tajoli, L. The world trade network. The World Economy, 34(8):1417–1454, 2011.
- De Benedictis, L., Nenci, S., Santoni, G., Tajoli, L., and Vicarelli, C. Network analysis of world trade using the baci-cepii dataset. *Global Economy Journal*, 14(3-4):287–343, 2014.
- De Masi, G., Giovannetti, G., and Ricchiuti, G. Network analysis to detect common strategies in italian foreign direct investment. *Physica A: Statistical Mechanics and its Applications*, 392 (5):1202 – 1214, 2013.
- De Masi, G. and Ricchiuti, G. Eu fdi network and systemic risks. Working paper, DISEI, Università degli Studi di Firenze, 2018a. 27/2018.

- De Masi, G. and Ricchiuti, G. The network of european outward foreign direct investments. In Gorgoni, S., Amighini, A., and Smith, M., editors, *Networks of International Trade and Investment*. Vernon Press, March 2018b.
- Ducruet, C., Ietri, D., and Rozenblat, C. Cities in worldwide air and sea flows: A multiple networks analysis. *Cybergeo: European Journal of Geography*, 2011.
- Fagiolo, G., Reyes, J., and Schiavo, S. World-trade web: Topological properties, dynamics, and evolution. *Physical Review E*, 79(3):036115, 2009.
- Fagiolo, G., Reyes, J., and Schiavo, S. The evolution of the world trade web: a weighted-network analysis. *Journal of Evolutionary Economics*, 20(4):479–514, 2010.
- Ferrarini, B. Vertical trade maps. Asian Economic Journal, 27(2):105–123, 2013.
- Freeman, L. C. Centrality in social networks conceptual clarification. *Social networks*, 1(3): 215–239, 1979.
- Garlaschelli, D. and Loffredo, M. I. Structure and evolution of the world trade network. *Physica* A: Statistical Mechanics and its Applications, 355(1):138–144, 2005.
- Gazaniol, A. The location choices of multinational firms: The role of internationalisation experience and group affiliation. *The World Economy*, 2014.
- Glückler, J. Economic geography and the evolution of networks. *Journal of Economic Geography*, 7(5):619–634, 2007.
- Hausmann, R., Hidalgo, C. A., Bustos, S., Coscia, M., Simoes, A., and Yildirim, M. A. The atlas of economic complexity: Mapping paths to prosperity. Mit Press, 2014.
- Hidalgo, C. A., Klinger, B., Barabasi, A.-L., and Hausmann, R. The product space conditions the development of nations. Paper 0708.2090, arXiv.org, 2007. 00468.
- Hidalgo, C. A. and Hausmann, R. The building blocks of economic complexity. proceedings of the national academy of sciences, 106(26):10570–10575, 2009.
- Hummels, D., Ishii, J., and Yi, K.-M. The nature and growth of vertical specialization in world trade. *Journal of International Economics*, 54(1):75–96, 2001.
- Hussain, O. A., Zaidi, F., and Rozenblat, C. Analyzing diversity, strength and centrality of cities using networks of multinational firms. *Networks and Spatial Economics*, pages 1–27, 2018.
- Jackson, M. O. Networks in the understanding of economic behaviors. Journal of Economic Perspectives, 28(4):3–22, 2014.
- Johnson, R. C. and Noguera, G. Accounting for intermediates: Production sharing and trade in value added. *Journal of international Economics*, 86(2):224–236, 2012.
- Joyez, C. On the topological structure of multinationals network. *Physica A: Statistical Mechanics and its Applications*, 473:578 588, 2017.
- Kali, R. and Reyes, J. The architecture of globalization: a network approach to international economic integration. *Journal of International Business Studies*, 38(4):595–620, 2007.
- Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.
- Koopman, R., Wang, Z., and Wei, S.-J. Tracing value-added and double counting in gross exports. *American Economic Review*, 104(2):459–94, 2014.
- Krackhardt, D. Predicting with networks: Nonparametric multiple regression analysis of dyadic data. Social networks, 10(4):359–381, 1988.
- Lenzen, M., Kanemoto, K., Moran, D., and Geschke, A. Mapping the structure of the world economy. *Environmental science & technology*, 46(15):8374–8381, 2012.

- Lenzen, M., Moran, D., Kanemoto, K., and Geschke, A. Building eora: A global multi-region input-output database at high country and sector resolution. *Economic Systems Research*, 25 (1):20–49, 2013.
- Los, B., Timmer, M. P., and de Vries, G. J. How global are global value chains? a new approach to measure international fragmentation. *Journal of Regional Science*, 55(1):66–92, 2015.
- Markusen, J. R. Multinationals, multi-plant economies, and the gains from trade. *Journal of International Economics*, 16(3–4):205–226, 1984.
- Onnela, J.-P., Saramäki, J., Kertész, J., and Kaski, K. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6):065103, 2005.
- Rozenblat, C. Inter-cities' multinational firm networks and gravitation model (international forum). Annals of the Association of Economic Geographers, 61(3):219–237, 2015.
- Serrano, M. Á. and Boguñá, M. Topology of the world trade web. *Physical Review E*, 68(1): 015101, 2003.
- Ter Wal, A. L. and Boschma, R. A. Applying social network analysis in economic geography: framing some key analytic issues. *The Annals of Regional Science*, 43(3):739–756, 2009.
- Timmer, M., Los, B., Stehrer, R., and de Vries, G. An anatomy of the global trade slowdown based on the wiod 2016 release. Working paper, Groningen Growth and Development Centre, University of Groningen, 2016.
- UNCTAD. World Investment Report: Transnational Corporations and Competitiveness. United Nations, 1995.
- UNCTAD. World Investment Report: Transnational Corporations, Extractive Industries and development. United Nations, 2007.
- UNCTAD. World Investment Report: Global value chains: Investment and trade for development. United Nations, 2013.
- Wall, R. S. and van der Knaap, G. A. Sectoral differentiation and network structure within contemporary worldwide corporate networks. *Economic Geography*, 87(3):267–308, 2011.
- Xiao, H., Sun, T., Meng, B., and Cheng, L. Complex network analysis for characterizing global value chains in equipment manufacturing. *PloS one*, 12(1):e0169549, 2017.

A ICIO data

Using the inter-country trade in intermediate from ICIO-OECD database to map the GVCs leads to a drastic reduction of the size of the networks considered because of the smaller geographic coverage (54 countries). The following tables report the results from the undirected and directed QAP regression. The last figure reports the evolution of the upstream and downstream forces of reorganization.

MNEs net (log) 1997	MNEs net (log) 2011
5.212	
$(33.68)^{**}$	
	10.778
	(39.84)**
-9.443	-26.062
$(24.95)^{**}$	$(30.42)^{**}$
0.28	0.36
2,862	2,862
	$5.212 \\ (33.68)^{**} \\ -9.443 \\ (24.95)^{**} \\ 0.28$

Table 5: QAP estimations - Undirected

* p < 0.05; ** p < 0.01

These results confirm the synchronization of the two networks, as the more two countries trade in intermediate, the higher the probability that French firms goes into these two countries via FDIs. Moreover, this determinant increases over time, and offers a better prediction of multinationals' multiple location choices.

B List of countries included in the estimations.

countries marked with a O are also included in the ICIO OECD data.

Albania, Algeria, Andorra, Angola, Argentina^O, Australia^O, Austria^O, Bahamas, Bahrain, Bangladesh, Barbados, Belgium^O, Benin, Brazil^O, Bulgaria^O, Burkina Faso, Cambodia^O, Cameroon, Canada^O, Cayman Islands, Central African Republic, Chad, Chile^O, China^O, Colombia^O, Congo, Costa Rica^O, Cote d'Ivoire, Cyprus^O, Czech Republic^O, Denmark^O, Djibouti, Ecuador, Egypt, Estonia^O, Ethiopia, Fiji, Finland^O, Gabon, Gambia, Germany^O, Ghana, Greece^O, Guatemala, Guinea, Hungary^O, India^O, Indonesia, Iran, Ireland^O, Israel^O, Italy^O, Japan^O, Kenya, Korea (rep)^O, Kuwait, Lebanon, Liberia, Liechtenstein, Lithuania^O, Luxembourg^O, Madagascar, Malaysia^O, Mali, Mauritania, Mauritius, Mexico^O, Morocco^O, Mozambique, Myanmar, Netherlands^O, New Zealand^O, Niger, Nigeria, Norway^O, Pakistan, Panama, Paraguay, Peru^O, Philippines^O, Poland^O, Portugal^O, Russia^O, Saudi Arabia^O, Senegal, Serbia, Singapore^O, Slovenia^O, South Africa^O, Spain^O, Sweden^O, United States of America^O, Uganda, Ukraine, Uruguay, Vanuatu, Venezuela, Viet Nam^O, Zambia, Zimbabwe, 3rd French Regional Conference on Complex Systems (FRCCS)

May 31 – June 02, 2023 / Le Havre, France

Submission of a Work-in-Progress Abstract: Camera-Ready Version (March 2023)

Author: Damien Calais, Ph.D., AXEMA, 19 rue Jacques Bingen, 75 016 Paris.



For the purpose of Open Access, a CC-BY public copyright licence has been applied by the author to the present document and will be applied to all subsequent versions arising from this submission.

Title: Is Conservation Agriculture the Future of Farming in France?

<u>Abstract</u>: French and European Union institutions as well as agrifood companies are calling for an agroecological transition in which soil conservation play a major role. Low-till farming accounts at present for 35 pct of the field crops area in France but conservation agriculture (CA) covers only 4 pct (vs. over 50 pct in some of the Americas countries). How can CA become the new French conventional agriculture by 2040? An ongoing strategic foresight study is using the Multi-Level Perspective (MLP) to assess the capability of French agriculture as a social-ecological-technical system (SETS) to achieve such a goal. The feasible ways to reach the targets of the called-for transition will be detailed. The value the stakeholders can derive from the shift to CA will be assessed along with the extra cost they may be asked to bear. Strategies will be designed to anticipate and manage possible conflicts. The discussion will appraise to what extent the research protocol of this study can help to solve issues of contention in transforming agricultural systems elsewhere.

<u>Keywords</u>: conservation agriculture – social-ecological-technical system – multi-level perspective – foresight – France

Introduction:

First designed to maintain the viability of cultivating lands degraded by severe erosion (Scopel, 2022), conservation agriculture (CA) is "a farming system that promotes minimum soil disturbance (i.e. no tillage), maintenance of a permanent soil cover, and diversification of plant species" (FAO, 2023). France and the European Union address CA to tackle threats on soil fertility, water resources, biodiversity, and move towards carbon neutrality while adapting agriculture to climate change (European Commission, 2022, 2020a, 2020b; Ministère de l'Agriculture, 2022; Ministère de la Transition écologique, 2020).

Up to now, the greening of agricultural policies and industry-led initiatives for regenerative agriculture have been irrelevant to achieve significant progress in disseminating CA in France and Europe, or even in enhancing the environmental impact of agriculture (European Environment Agency, 2019). Farmers' training and technical expertise are challenges to be taken forward. CA should be accepted by society: little known by the general public, its implementation has nevertheless given rise to controversy over the use of glyphosate and the proliferation of biogas digesters. New evidence should be collected to firm up knowledge of what can be expected from CA in environmental matters. Those issues are arising in a context of global uncertainty where the problem of food sovereignty is coming to the fore.

Applying the multi-level perspective (Geels, 2002; Geels, Schot, 2007), CA in France can be considered as a niche within a changing social-ecological-technical system where conventional and organic agriculture are the dominant regimes. The system encompasses farms, agriculture supply chains, and a socio-economic and political landscape open to the world. What are the ways to take bold action toward the goal of making CA a new regime in France, beside or in place of the existing ones?

Field crops and mixed farming will constitute the scope of the study. Permanent crops are excluded as tillage intensity is low by nature; in addition, orchard or vineyard grassing is not a cover crop in the sense of an off-season plant between two cash crops.

494

Methods:

The expectations placed on CA will be quantified through public policy documents, press releases from food industries and restaurant chains, websites dedicated to CA, and reports from associations that promote it. The ability of CA to meet the goals and to do so better than conventional or organic farming will be assessed. Possible trade-offs will be described when it appears that CA cannot reach the targets due to internal contradictions or obstacles from the legal and regulatory, social, cultural or economic environment. The feasibility of different ways of disseminating CA under various pedoclimates will be explored, whether CA emerges as a third path alongside conventional agriculture and organic farming, replaces them, or combines with them.

Ongoing semi directive interviews are conducted with agri-equipment manufacturers (company managers, marketing departments, technical sales staff), agri-food industry professionals, service providers related to CA (insurance companies that insure the risk taken by farmers who change their cropping systems to move towards agro-ecology, companies that perform and interpret soil analyses, companies that certify carbon credits), farmers, and research engineers.

To ensure readability, not more than twenty actors will be selected for an in-depth analysis. An actor is not understood as a natural or legal person but as a social or economic group with means of action, organization, and a strategy to achieve its goals (Bassaler, 2004). The profile of each actor will include the purpose of its actions, prioritized objectives, projects, motivations, constraints, resources, past strategies, and levers for influencing the others. To deal with those features, the MACTOR method will be applied, aiming at a typology of actors according to their influence on the system, their dependence on other actors, and a valuation of their capacity to set power relations favorable to them (Godet, 2001).

4Q5

<u>Results</u>:

The feasible ways to reach the targets of the called-for transition will be detailed. The value the stakeholders can derive from the shift to CA will be assessed along with the extra cost they may be asked to bear. Strategies will be designed to anticipate and manage possible conflicts between them, according to tables of their relative power, conflict potentials, and commitments on the various objectives.

What are the conditions under which CA can become the future of conventional agriculture in France? The preliminary findings of the study highlight that economic hindrances are obvious: direct seeding (without tilling the soil beforehand) or strip-tilling require expensive equipment; the risk of yield loss is real during the first years of transition. Banks and insurance companies can play a role in moderating this risk. However, the keys to a massive diffusion of CA seem to lie in the training of farmers. Requiring a keen sense of observation and a permanent ability to adapt the cropping system to the hazards that arise, this highly technical agriculture will be even more so if it must be conducted without glyphosate. Breaking with the habits of previous generations, it will only take off if cultural and psychosocial barriers are lifted. Far from the model of a technical itinerary defined by crop and by plot, CA seems today to be poorly supported by policies that condition public aid on an obligation of means rather than results. The growing integration of digital technology could be an opportunity for CA, as it offers a flexibility that mechanical equipment does not have to adapt machines and farming operations to local specificities.

At the heart of CA innovation, technical progress in seeding equipment is continuing and could lead to collaboration between manufacturers and seed companies. Intercropping (several crops at the same time in the same field) also raise challenges for harvesting and crop protection equipment.

426

Discussion:

The discussion will appraise to what extent the research protocol of this study can help to solve issues of contention in transforming agricultural systems elsewhere. The case of the Netherlands, where agriculture is currently being sacrificed to the objectives of reducing nitrogen pollution, will be especially addressed.

References:

Bassaler N (2004) Le jeu des acteurs de l'information géographique : un cas appliqué de la méthode MACTOR. Cahiers du LIPSOR 17:1-64. laprospective.fr/dyn/francais/memoire/jeudesacteurs.pdf European Commission (2022) Proposal for a Regulation of the European Parliament and of the Council on the sustainable use of plant protection products and amending Regulation (EU) 2021/2115. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12413-Pesticides-sustainable-use-updated-EU-rules- en

European Commission (2020a) A Farm to Fork strategy for a fair, healthy and environmentallyfriendly food system. https://food.ec.europa.eu/horizontal-topics/farm-fork-strategy_en

European Commission (2020b) EU 2030 Biodiversity Strategy: Bringing nature back into our lives. https://environment.ec.europa.eu/strategy/biodiversity-strategy-2030_en

European Environment Agency (2019) Climate change adaptation in the agriculture sector in Europe. Publications Office of the European Union, Luxembourg, 108 p. doi:10.2800/537176

Geels FW (2002) Technological transitions as evolutionary reconfiguration processes: a multi-level perspective and a case-study. Research Policy, 31, 8-9:1257–1274. doi:10.1016/s0048-7333(02)00062-8

4**Q**7

Geels FW, Schot J (2007) Typology of sociotechnical transition pathways. Research Policy, 36, 3:399-417. sciencedirect.com/science/article/abs/pii/S0048733307000248

Godet M (2001) Manuel de prospective stratégique, 2e éd.. Dunod, Paris, 2 t.

Ministère de l'Agriculture (2022) Plan stratégique national PAC 2023-2027. agriculture.gouv.fr/pac-

2023-2027-le-plan-strategique-national

Ministère de la Transition écologique (2020) Stratégie nationale bas-carbone.

https://www.ecologie.gouv.fr/strategie-nationale-bas-carbone-snbc

Scopel É (2022) Les modalités de développement de l'ACS dans le monde : l'évolution et l'adaptation d'un concept. Paper presented at Académie d'Agriculture de France, Paris, France, June 2022. www.youtube.com/watch?v=0puGU4Na_-M&t=5s



Let's Tweet about Soccer? A Gender-centric Question

Mariana Macedo · Akrati Saxena

Fans, sports organizations, as well as players use social networks like Twitter to build and maintain their identity and sense of community [8,9]. Breaking news about soccer sometimes comes first on Twitter than traditional media channels and provides an excellent means to access instantaneous information from official and unofficial sources [9]. Soccer has more than 3.5 billion fans worldwide, and it is estimated that 1.3 billion of them are females (around 38%) [4]. Our work examines whether, in a male-dominated environment, women and men differ in communication patterns. Women soccer fans tend to experience biases and prejudices, and our question relies on how this is translated to online spaces [7,1]. Through our study, we look into the patterns of interaction (tweets, retweets and replies) between men and women, and how communication evolves over 3 months (March 7 to June 6, 2022) for English and Portuguese languages. To identify the gender, we extract the first name of the user and use the GenderGuesser API [5] to assign a gender. We considered all the names for which gender was not classified as in conflict, unknown, or unisex.

After our data preprocessing, we have 7,676,624 tweets in English (6,365,239 are from males and 1,313,731 are from females) and 2,958,443 tweets in Portuguese (2,312,415 are from males and 648,395 from females). The highest women rate (when analysing temporally) in our data reaches 28% female users, i.e., close to the overall Twitter ratio (29.6%).

We build directed networks from the tweets in Portuguese and English where the nodes are the users, and the links are the connections based on retweets and replies. Portuguese networks have 193,574 nodes and 1,946,932

A. Saxena

M. Macedo

Center for Collective Learning, ANITI, University of Toulouse, FR, E-mail: mmacedo@biocomplexlab.org

LIACS, Leiden University, The Netherlands E-mail: a.saxena@liacs.leidenuniv.nl



French Regional Conference on Complex Systems – Le Havre, May 31 June 2, 2023

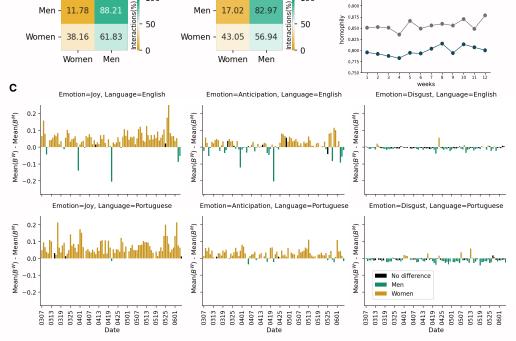


Fig. 1: Gender differences of the communication patterns on Twitter. **A.** Percentage of interactions between Women and Men. In Portuguese, the interactions between women and men are more gender-neutral than in English. **B.** Homophily values per week. English network shows higher homophily rates than in Portuguese regardless of the week. **C.** Gender differences between the emotion levels (Joy, Anticipation and Digust). Women have higher levels of joy and anticipation (lower levels of disgust) than men.

edges, and English network has 340,840 nodes and 1,158,344 edges. We then first study both the networks and observe that the Portuguese network has a higher women ratio in interactions (Figure 1.A) and a lower homophily (Figure 1.B) than the English network. One of the reasons might be that in Portuguese, people belong to specific places, mostly from Brazil and Portugal, where soccer is more popular. The network structure from women's tweets tends to be denser, with higher average clustering and lower assortativity than men's one, aligned to previous research in other types of communication such as research collaborations [3].

We further carried out a text analysis of whether men and women exhibit emotions differently in soccer. The emotions are computed using NRC Emotion Intensity Lexicon, VADER, and Google's Perspective API [2,6]. We plot, in Figure 1.C, the gender differences of the emotions extracted from the tweets over time. We found that, in soccer, women tend to express higher levels of joy and anticipation than men in both languages, and disgust tends to be more gender-neutral with slightly higher levels for males. Interestingly, we did not find any qualitative difference in relation to the gender differences in emotion between the English and Portuguese collected tweets. We thus found that the emotional response across genders seems independent of the overall network structure.

Our work takes a step further to understand gender differences in communication patterns exposing possible misleading perceptions of free speech in social media. Regardless of soccer being more popular for Portuguese speakers, the overall communication patterns vary across languages, but still, gender differences seem to be focused on women being more positive and less active than men. In the future, we plan to investigate further the explainability of patterns extracted from the networks and their communities, and compare them with a female-dominated sport.

References

- E. S. Cavalier and K. E. Newhall. 'stick to soccer:'fan reaction and inclusion rhetoric on social media. Sport in Society, 21(7):1078–1095, 2018.
- C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1):216–225, May 2014.
- A. M. Jaramillo, M. Macedo, and R. Menezes. Reaching to the top: The gender effect in highly-ranked academics in computer science. Advances in Complex Systems, 24(03n04):2150008, 2021.
- 4. D. Lange. Soccer fans by gender 2019, Nov 2020.
- Lead-ratings. Gender Guesser API. https://github.com/lead-ratings/gender-guesser, 2015. [Online; accessed 2-Aug-2022].
- R. Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.
- K. Toffoletti. Sexy women sports fans: Femininity, sexuality, and the global sport spectacle. *Feminist Media Studies*, 17(3):457–472, 2017.
- J. Williams, S. J. Chinn, and J. Suleiman. The value of twitter for sports fans. Journal of Direct, Data and Digital Marketing Practice, 16(1):36–50, Jul 2014.
- H. M. Wold., L. Vikre., J. A. Gulla., Özlem Özgöbek., and X. Su. Twitter topic modeling for breaking news detection. In *Proceedings of the 12th International Conference on Web Information Systems and Technologies - Volume 2: WEBIST*, pages 211–218. INSTICC, SciTePress, 2016.



From geographic data to spatial knowledge in agent-based modeling applied to land use simulation Toward the integration of spatial dimension in human agent

Severin Vianey Kakeu Tuekam · Eric Fotsing · Marcellin Julius Antonio Nkenlifack

Abstract The last decades have been marked by an increasing number of scientific papers dedicated to agent-based modeling applied in various domains and especially the integration or coupling with GIS data for more realism. This dynamic approach of modeling involves autonomous agents cooperating to solve a complex problem underlined in a domain such as a land use-cover system where human agents consequently modify their environment to live. However, the spatial aspect of intelligence led by the representation and reasoning of an environment is not explicitly captured during the modeling. Nowadays, the rise of its technologies offers new research opportunities to think and work on the analysis, formalized representation, and the use of the physical space of agents using spatial data. This naturally brings the agent more information, increases its spatial knowledge, and thus strengthens its decision-making process. This paper proposes a conceptual model for spatial knowledge representation from spatial data in view to improve the cognitive dimension of agents in land-use simulations. To illustrate the design of spatial knowledge semantic net, some examples have been presented. Finally, some ways have also been investigated about knowledge acquisition by cognitive agents.

Severin Vianey Kakeu Tuekam

Department of Mathematics and Computer Science, University of Dschang P.O. BOX 67, Dschang-Cameroon

E-mail: severinkakeu @gmail.com

Eric Fotsing

Marcellin Julius Antonio Nkenlifack

Department of Mathematics and Computer Science, University of Dschang P.O. BOX 67, Dschang-Cameroon E-mail: marcellin.nkenlifack@gmail.com

Department of Computer Engineering, Fotso Victor University Institute of Technology P.O. BOX 134, Bandjoun-Cameroon E-mail: efotsing@gmail.com

Keywords Agent-based modeling \cdot Multi-agent system \cdot Geographic data \cdot Spatial knowledge representation \cdot Semantic networks \cdot Spatial predicate and reasoning \cdot Social modelling and simulation

1 Introduction

The last few decades have been marked by an increasing number of scientific papers dedicated to Agent-Based Modeling (ABM) applied in various domains (geography, ecology, sociology, robotics, etc.). Moreover, the integration or coupling of Geographic Information Systems (GIS) and agent-based models (ABMs) to produce more realistic multi-agent systems is becoming a big challenge in which we are working respectively in micro (agent) and macro (multi-agent) levels. This dynamic approach of modeling involves autonomous agents (process, robot, human being, animal, government, etc.) cooperating in an environment defined to solve a complex problem. In a domain such as a Land Use-Cover System (LUCS) where human agents should consequently modify their environment to live. The land use concept represents the process in which human beings employ the resources of land (e.g. Forests, Mountains, Water, Agricultural Land, etc.). The human being is one of the main actors that causes the land use-cover changes and the disappearance of some biophysical entities at the land surface in which an abnormal discrepancy is observed in the use of resources [1]. So, there is a big stake in the land use-cover modeling and simulations on the resource management research with the growth of population in the world [2]. The environmental sciences study the facts produced by individual actors in their space at the land surface (climate, waters, soils, plant cover and animal space, deforestation, etc.) using sociology and psychology concepts to model actors [3,4] and then implementing them in computer science with ABM as in [5–7]. These works study the ecosystems of natural resources management including the analysis of behavior and relationships between the individual entities through simulations, where methods are also proposed to build ABM from actor [8–10] generally based on the old methodologies cited in [11].

However, the spatial aspect of intelligence-led by cognitive representation, reasoning, and management of the environment knowledge is not explicitly captured during the ABM. The spatial intelligence based on these forms of knowledge is one of several others cited in the theory of multiple intelligences that allows us to keep our space in mind [12]. Moreover, if knowledge management represents a major economic challenge for the future, it is also important in a virtual system representing society such as multi-agent systems (MAS). The spatial knowledge system of the particular human agent can thus be seen as information that takes on a certain meaning in a given context of its environment at a given time such as defined by [13] about general knowledge system. It has been studied in knowledge engineering for over a decade. The rise of its technologies offers new research opportunities to think and work on the analysis, formalized representation, and the use of the physical space of agents using spatial data. This naturally brings more information to the agent, increases its spatial knowledge (SK), and thus strengthens its decision-making process. Moreover, agents need to take into account not only changing situations and interactions with other agents in the world but possibly also their thinking about their physical environment present in the system. Consequently, agents should be able to learn to behave optimally from their interaction with other agents, but also from their spatial background embedding more information about the physical world load in the system. This paper proposes a conceptual model for spatial knowledge representation from spatial data in view to improve the cognitive dimension of agents in land-use simulations. To illustrate the design of spatial knowledge semantic net, some examples have been presented. Finally, some ways have also been investigated about knowledge acquisition by cognitive agents.

This manuscript is organized into five sections structured as follows. The first section presents the literature on knowledge representation models. The second focuses on the spatial knowledge representation insisting on the geographic data types and region connection calculus (RCC) theory. The third section is dedicated to our proposed model for representing spatial knowledge based on a semantic network and a sample case study. In section four, we propose three ways for spatial knowledge acquisition use to build the cognitive agents in a MAS while the last section is devoted to the conclusion and prospects.

2 State-of-the-art on the Knowledge Representation Approaches

Knowledge engineering is an artificial intelligence domain that appeared about fifteen years ago where the problem of formalizing knowledge was posed to be able to create computerized systems. fundamentally in 2002, [14] presented the knowledge representation as a surrogate, a substitute for the thing itself, that is used to enable an entity to determine consequences by thinking rather than acting, that is, by reasoning about the world rather than taking action in it. Also as a set of ontological commitments and a fragmentary theory of intelligent reasoning expressed in terms of several components (representation and a set of inferences). This section presents a state of the art on knowledge representation approaches or models used in AI illustrated in Figure 1.

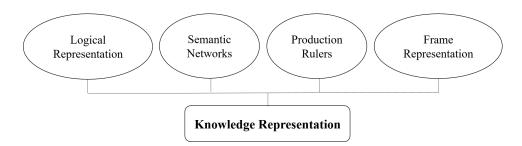


Fig. 1 Knowledge representation approaches

2.1 Logical Representation

Logical representation (LR) is a language with some concrete rules which deal with propositions and has no ambiguity in representation. LR means drawing a conclusion based on different conditions. This representation lays down some important communication rules. Categorized in two logics (propositional and predicate logics), it defines the syntax and semantics which supports sound inference. Each sentence can be translated into logic using syntax and semantics where the first concept represents the set of rules which decide how we can construct legal sentences in the logic, which symbol is used, and how to write those symbols. The semantics are the set of rules by which we can interpret the sentence in the logic (using P and Q preposition and their negation $\neg P$, $\neg Q$, and some symbols/operators like \Rightarrow , =, AND, OR, XOR, etc.). It allows to an assignment of a meaning to each sentence. For example, *Road exists in Region and Road is crossed by a railway* are propositions that can be returned true or false.

2.2 Semantic Network

It is a good alternative to predicate logic where we can represent our knowledge in the form of graphical networks [15]. This network consists of nodes representing objects and arcs which describe the relationship between those objects. Semantic networks (SN) can categorize the object in different forms and can also link those objects with two types of relations (*IS-A for Inheritance and Kind-of*, [16]). Semantic networks are easy to understand and can be easily extended (e.g. Figure 2).

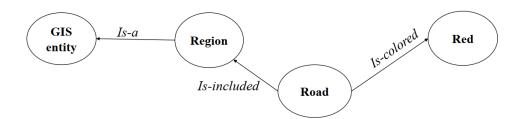


Fig. 2 An illustration of SN for describing an area

One of the drawbacks of SN comes from the fact that it takes more computational time at runtime. So, we need more time to traverse the complete network tree to answer some questions. It might be possible in the worst-case scenario that after traversing the entire tree, we find that the solution does not exist in this network.

2.3 Production Rules

This technique consists of a pair of *condition* and *action* (which means, *If condition then action*). It has mainly three parts: A set of production rules, a working memory, and a recognize-act-cycle.

- IF (River is included in Region AND period of Hunting started) THEN action (hunting in the Region)
- IF (River crosses Road AND (Road AND River in Forest)) THEN action (set trap)

In these production rules, the agent checks for the condition and if the condition exists then the production rule is set up and the corresponding action is carried out. The condition part of the rule determines which rule may be applied to the situation for optimal decision-making. Then, the action part carries out the associated problem-solving steps. This complete process is called a recognize-act cycle. These production rules are generally used in agent-based simulation platform programs as Netlogo¹, GAMA², MadKid³, Cormas⁴, etc.

2.4 Frame Representation

It represents a record-like structure that consists of a collection of attributes and their values to describe an entity in the world. Frames are the AI data structure that divides knowledge into substructures by representing stereotyped situations [17]. It consists of a collection of slots and slot values. These slots may be of any type and size. Each slot has a name and value. The various aspects of a slot are known as Facets. Those are features of frames that enable us to put constraints on the frames. Example: IF-CONDITION facts are called when data of any particular slot is needed. A frame may consist of any number of slots, and a slot may include any number of facets and facets may have any number of values. Frames derive from SN and later evolved into our modern-day classes and objects. In the frame, knowledge about a spatial object or event can be stored together in the knowledge base. The frame is a type of technology that is widely used in various applications including Natural language processing and Computer vision.

3 Spatial knowledge representation

In this section, we focus on knowledge acquisition and representation from spatial data.

¹ https://ccl.northwestern.edu/netlogo/

² https://gama-platform.org/wiki

³ http://www.madkit.net/madkit/madkit.php

⁴ http://cormas.cirad.fr/fr/demarch/sma.htm

3.1 Spatial data for the virtual and physical environment in MAS

Spatial data comprise the geographic information about the object on the earth and also its features [18]. A single piece of data is constituted by latitude, longitude, and altitude coordinates defining a specific location on earth. GIS experts classify the spatial data into two types according to the storing technique, namely, raster data and vector data model. These are used in a GIS environment to represent the real-world observations (objects or events that can be designed in 2D or 3D) by spatial entities as illustrated in Figure 3 for roads. We note that an objective view of the world treats entities as discrete objects. Point locations of cities or villages would be an example of an object.

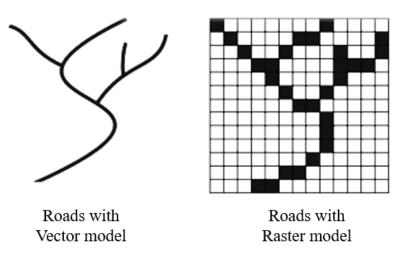


Fig. 3 Representation of vector and raster models in GIS

The vector model or features is decomposed into three different geometric primitives called points, lines/polylines, and polygons (Figure 4), while the raster data model uses an array of cells representing pixels, to represent realworld objects. In GIS applications, Raster datasets are commonly used for representing and managing imagery, surface temperatures, digital elevation models, and numerous other entities.

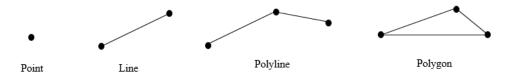


Fig. 4 The various geometric primitives for Vector representation

The representation of a real-world entity in GIS tools or spatially explicit agent-based simulation platforms is in large part dictated by the scale of the analysis. The scale represents the ratio of the distance on the map to that in the real world. So a large-scale map implies a relatively large ratio and thus a small extent. This is counter to the laypersons interpretation of large scale which focuses on the scope or extent of a study; so a large scale analysis would imply one that covers a large area [18]. A region of a country can be represented at a small scale (e.g. 1:10,000,000), Thus, the cities may be best represented as points. At a large scale (e.g. 1:25,000), They may be best represented as a polygon. Those strategies are used to design the virtual and physical world in a multi-agent platform during land use-cover modeling and can be exploited for spatial thinking based on spatial knowledge extracted.

3.2 Definition and role of spatial knowledge

Generally, knowledge is seen as understood information, i.e. assimilated and used, leading to action [13]. Nonaka and Takeuchi, two experts in knowledge management, differentiate between two forms of knowledge: tacit and explicit knowledge, which is much more detailed in [19]. Much of our spatial knowledge comes from our current experience within the space in which we live but is also acquired in a variety of other ways such as from maps, verbal descriptions of other experiments, virtual realities, etc. Map experience led to more accurate estimation of the straight line or Euclidean distances between locations characteristic of knowledge, whereas actual navigation led to more accurate route distance estimation and more accurate judgments of the actual direction of locations from station points within the building. As an illustration, a human agent can use a vector map to estimate more precisely the length of a straight line (roads, rivers, etc.), the Euclidean distances between places, to determine, for example, the nearest buildings, the neighboring roads, the areas crossed by rivers.

Indeed, the SK is often spatially fuzzy but important in our decisionmaking process. For describing fuzzy knowledge in a situation described, we need some membership degree as illustrated by [20] in Figure 5.

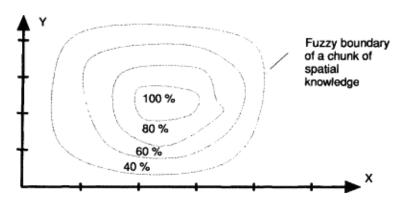


Fig. 5 An illustration of the fuzzy boundary of a spatial knowledge

As an illustration of the location of city blocks and a central Bank, we can say that some city blocks belong 100 percent to this Bank, another one only at 95 percent, some in the middle at 50 percent, and so on. The most important is that we have the information that the city and Bank are very close.

Spatial knowledge includes the semantics of the spatial relationships between geographic objects of an environment in the individual's wayfinding process and the knowledge about how to move in this physical (virtual) environment [21]. In daily life, human beings can acquire spatial knowledge about their environment in three distinct stages: landmark knowledge is acquired first, then route knowledge, and finally survey knowledge. How that acquisition is explained or practiced in agent-based modeling for land use simulation is the main issue of our investigation. First of all, we formalize the comprehension of the world based on spatial data using mathematical concepts and then, how to get a computerized version of the spatial thinking model will be the next steps of work.

3.3 Background on Region Connection Calculus theory

RCC is one of the most widely referenced systems of qualitative spatial representation and reasoning proposed by [22] and used in various research works in literature including [23,24]. RCC assumes a continuous representation of space also called Mereological theory. Among these works, [25] developed a theory of discrete space using this theory but the remaining question is, can we use the same practice to propose the SK representation model using continuous spaces described by vectors layer? Assuming that each entity of a vector layer (or entire layer) can be a shape. Thus, the theory is a primitive binary relation of parthood relating parts to wholes (of which they are part of). The region in this scope noted represents any geographic object designed from *point, line, polygon.* If x and y are two regions (*e.g. city and quarter layers*) of the space denoted by S, several relations can be defined between them as follows:

- 1. A Part of (\mathbf{P})
 - Reflexivity: $\forall x [P(x x)]$
 - Anti-symmetry: $\forall \ x,y \ [P(x, \ y) \ \land \ P(y, \ x) \ \rightarrow \ x=y]$
 - Transitivity: $\forall x, y [P(x, y) \land P(y, z) \rightarrow P(x, z)]$
- 2. Proper Part (PP)

 $PP(\mathbf{x},\mathbf{y}) \equiv_{def} P(\mathbf{x},\mathbf{y}) \land \neg P(\mathbf{y},\mathbf{x})$

3. Overlap (O)

It is the next most important relation after P. Further relations and operations such as union (or fusion) and intersection are also defined in terms of P or a combination of both.

Then, $O(\mathbf{x}, \mathbf{y}) \equiv_{def} \exists \mathbf{z} (P(\mathbf{z}, \mathbf{x}) \land P(\mathbf{z}, \mathbf{y}))$

Also, $\forall x, y \ [\neg P(y,x) \rightarrow \exists z \ (P(z,y) \land \neg O(z,x))]$

It means that PP is rendered extensional, meaning that two distinct entities differ in at least one part. Thus,

 $\forall x, y \; [\; \forall z, \; (O(z,x) \longleftrightarrow O(z,y)) \rightarrow x{=}y]$

- Disjoint Connection (DC) x is disconnected from y means that x and y are not connected
 Externally Connection (EC)
 - x is externally connected to y means that x and y are connected but do not overlap
- 6. Equal (EQ) x is equal to y means, each point of x and y is part of the other.
- 7. Tangential Proper Part (TPP) x is a TPP of y (covers) means that x is a proper part of y and some region is EC to both.
- 8. Non-Tangential Proper Part (NTPP)

x(or y) is a non-tangential proper of y(or x) but not a TPP part of y (contains), so x(or y) is a proper part of y(or x) but not a TPP.

The formal description of the representation of these relations between regions can be defined expressed as follows with more details available in [22]:

- Each region is characterized by its interior and boundary.
- The description of the topological relationship between two regions is a 2×2 matrix (\mathcal{M}) that represents the 4 intersections. \mathcal{M} indicates the intersection between the interior (*int*) and boundary (*bdy*) of one with both the interior and boundary of other or indicates an emptiness. Let X, Y $\in \mathcal{S}$, the intersections (I) between region X and Y is represented as follows:

$$\mathcal{M}(X,Y) = \begin{bmatrix} I_{bdy-bdy}(X,Y) & I_{bdy-int}(X,Y) \\ I_{int-bdy}(X,Y) & I_{int-int}(X,Y) \end{bmatrix}$$

In this matrix, each variable can take a value of 0 or 1 depending on the intersection (empty or not) and become information for the agent. Thus, in the RCC8 configuration, we count eight possible matrices that correspond to the 4-intersection configuration representing the topological information (Table 1).

Table 1 The 4-intersection formalism with RCC [25]

RCC8	DC	EC	РО	$\mathbf{E}\mathbf{Q}$	\mathbf{TPP}	NTPP	\mathbf{TPP}^{-1}	$NTPP^{-1}$
\mathcal{M}	[0 0]	$\begin{bmatrix} 1 & 0 \end{bmatrix}$	[1 1]	$\begin{bmatrix} 1 & 0 \end{bmatrix}$	[1 1]	01	[1 0]	[0 0]
	0 0	0 0	1 1	0 1	$\begin{bmatrix} 0 & 1 \end{bmatrix}$	0 1	1 1	1 1

A major utility of a cognitive map built from the observation of the land use is remarked in the moving plan of humans where they find out a shortcut which minimizes for example a qualitative or quantitative traveling cost such as energy consumption and risks. Therefore, it is desirable to minimize navigation even when it is necessary to have more spatial information in the mind of the agent about a region (village) for a hunting activity. This observation motivates us to introduce another representation taking into account the exterior of the regions rather than using the 3×3 -matrix in 9-Intersections designed in Egenhofer configuration [22] to get more spatial knowledge.

$$\mathcal{M}(X,Y) = \begin{bmatrix} \partial X \bigcap \partial Y \ \partial X \bigcap Y^{\circ} \ \partial X \bigcap Y^{-} \\ X^{\circ} \bigcap \partial Y \ X^{\circ} \bigcap Y^{\circ} \ X^{\circ} \bigcap Y^{-} \\ X^{-} \bigcap \partial Y \ X^{-} \bigcap Y^{\circ} \ Y^{-} \bigcap Y^{-} \end{bmatrix}$$

Where: the *interior* of a region noted R is the union of all open sets in R and denoted by R° , the *exterior* of R is the set of all points available in R^2 and denoted by R^- . *boundary* of R is the intersection of her closure and closure of her exterior of R and denoted ∂R .

This is important in spatially explicit agent-based simulations. For example, the relationship between a linear region L (road) and an area A (forest or protected area) can be represented as follows (Figure 6).

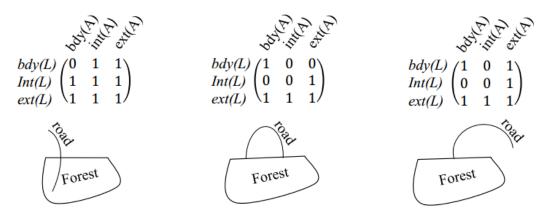


Fig. 6 Representation of three relationships between roads (lines) and forests (areas)

In sum, all of these configurations used as spatial knowledge matrices allow us to make the spatial knowledge representation in a qualitative context. Then, it constitutes a base of the conceptualization of a simple cognitive map with data structures and logic predicates.

4 Formal definition of the spatial knowledge

This section is focused on the conceptualization of spatial knowledge and how to concretely use it in agent-based modeling.

4.1 Defining the SK as a simple Predicate to evaluate

In agent theory, an agent can use the information captured in his environment and others received from fellow creatures to construct an internal information base. So, the belief's elements come from the evaluation of that base (relational or spatial base, ontological base, etc.). Concerning the spatial side of intelligence, we investigate two ways: spatial knowledge can be extracted from that base to get a whole knowledge base or can be directly designed by users and accessible to the agent. The second way is defined as follows:

Topological information between two geographical entities is represented by 2×2 Matrix where each element allows to deduce of spatial knowledge. Thus, this one can be represented by a predicate named Spatial Predicate(SP) defined as

 $\mathcal{P}_{i\ (r,w)}(\mathcal{F}(\mathbf{r}_j,\,\mathbf{w}_k))$ Where,

- $-\mathcal{P}$ is a name that identifies the predicate,
- -i represents the agent getting the spatial information,
- $-r_j$ and w_k represent the geographical entity (call Regions in RCC theory) concerned by computation,
- F is a function (e.g. C, EQ, PO, DC, TPP, etc.) returning several values (4 in RCC8 or 9 if the 9-Intersection approach is used) stored in predicate P.

SP mays change depending on the simulation context. In some scenarios, we can have one instance of road, river, forest, protected area, or another spatial entity. Consequently, the general predicate becomes $\mathcal{P}_{i(r,w)}(\mathcal{F}(\mathbf{r}, w))$ in the case that we only have one instance for each entity (e.g. a road, railway or river comes from GIS file). If we want to represent the fact that an agent called *toto* knows that a river(\mathcal{R}) crosses village (\mathcal{V}), *sp_riverExistsInVillage* represents the associated predicate: *sp_riverExistsInVillagetoto*(\mathcal{R}, \mathcal{V})(*PO*(\mathcal{R}, \mathcal{V})).

In this expression, the agent is familiarized with the fact that \mathcal{R} is partially overlapped to \mathcal{V} , So a river exists in the village. That knowledge can be shared with other agents. For example, if *toto* is a hunter, then he knows that it is possible to catch the small antelopes in that village along the river.

4.2 Defining the mental map as a computing variable

We suppose that each agent has a Spatial Mental Map (SMM) that represents its thoughts on the world and contains several SK as already explained in [26]. The mental state of Belief-Desire-Intension architecture is driven by belief, desire, and intention databases. Conceptually, we can define the global mental state as the sum of each knowledge:

$$SMM_a = \sum_{0}^{n-1} (Pr, Val, Po, LT)_i$$

where,

- $-SK_a$: the name of spatial knowledge getting by agent a,
- *Pr*: the name or identifier of a predicate,
- Val : the value of Pr,
- Po: the importance of knowledge for an agent (this value is situated between 0 and 1),

-LT a Life Time value that indicates how long this knowledge can stay in the agent memory (Defined by modeler like Po).

5 The proposed SK model based on the semantic network

As described in section 2.2, we are going to use the SN as a support of spatial knowledge storing and sharing in agent society because it is a knowledge representation technique used for propositional information, so a mathematical tool (graph) adapt to our kind of information managed using the propositional sentences come from RCC formula. This mathematical structure can help agents to save and share more spatial knowledge in their environment. This can help agents (AI systems) to better understand their physical environment and reason about it using strong decision-making support for complex problems solving.

5.1 The semantic network model proposed

Our spatial knowledge semantic network called SKSN is defined by a set of binary relations on a set of nodes (regions or geographic entities came from GIS data and specify by the modeler):

- A set of nodes (N) representing all the regions considered in the simulation.
- A set of arcs (A) appear as arrows to express the relationships between regions, and link labels specifying the relations (Figure 7).

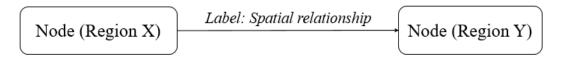
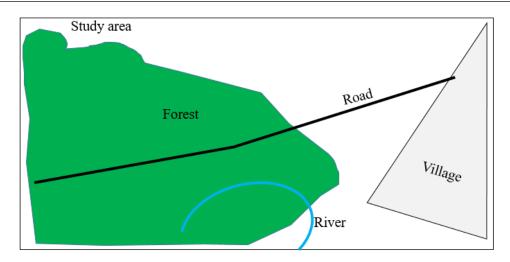


Fig. 7 Model of SKSN

5.2 An example of SK representation for agent-based modeling

Generally, the land use models deal with describing activities of land-consuming actors and their competition for land in an urban or rural setting. These actors are households, firms, and retail establishments, each with particular requirements for space and access to jobs (industrial houses), schools, forests, markets, etc. The study area is generally designed using a GIS tool and imported into an agent-based simulation tool such as GAMA or Netlogo for more analysis and comprehension. In this example, to illustrate the use of the previous SN model (SKSN) and build a knowledge base ready for a computerize process, we suppose a sample map representing a study area (Figure 8):

The following instance of SKSN can be designed during the modeling process (Figure 9):



French Regional Conference on Complex Systems - Le Havre, May 31 June 2, 2023

Fig. 8 An illustration of a sample map for a study area

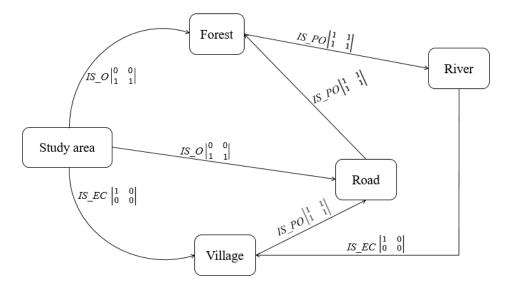


Fig. 9 A sample SKSN deduced from the previous map

6 Spatial knowledge acquisitions techniques use for cognitive agents

For an agent simulation scenario, we propose different kinds and several steps for the spatial knowledge acquisition process. It can be done manually (modeler by coding/specification in simulation platform) or automatically (using an AI process to build an SKSN from GIS data). The spatial knowledge acquisition (SKA) b agents are one of the most difficult tasks in this work. For example, in society, a human being spends many years attending schools, colleges, and universities, home, for the sole purpose of learning knowledge and skills. In general, there are three ways for a human being to acquire knowledge. As in AI, we investigate three ways of acquisition to use in a MAS.

 The first way to help the agent to learn about his physical environment is the supervised-learning, in which agents can interact with a super-agent that share his SK base with them (Figure 10)

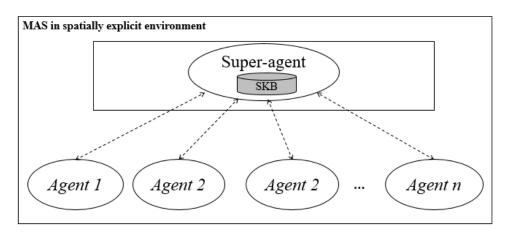


Fig. 10 MAS with a super agent (coordinator)

 The second way is a self-study, in which the agent could gain SK from reading texts, SKSN, and doing experiments without too many interventions of any super-agent (Figure 11).

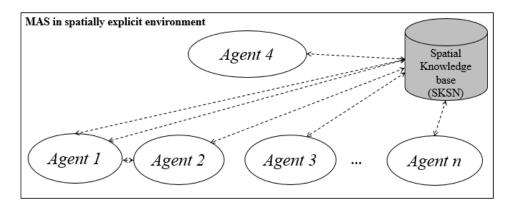


Fig. 11 MAS with self agents only and a common SK base

- The third way is autonomous perception or discovery, building SKSN in which agents' minds transform received signals into cognitive states of knowing the meanings (i.e. spatial knowledge) behind these signals (Figure 12). Now, it is the mental capability behind the transformations from signals to meanings that lays the foundation of an agent's self-intelligence. Without such capability, an entity will not be able to acquire knowledge at all. Due to the huge amount of spatial knowledge in the physical environment, it is not possible to cover the entire set of spatial knowledge acquisition in a MAS.

All these propositions are good to experiment but the third is more adapted to the agent paradigm where the agent should be autonomous during its life cycle in simulation.

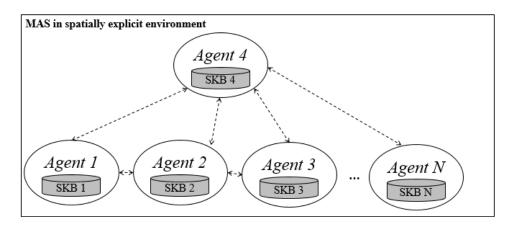


Fig. 12 MAS with autonomous agents only and their own SK base

7 Discussion and Conclusion

This paper represents the first effort to investigate and provide conceptual models as a result of how cognitive agents should reason spatially in a multiagent system. This result is directly applied to be applied in the domain of agent-based modeling where a system is modeled as a collection of autonomous decision-making agents, one method to synthesize prior knowledge of society (actor/people, activities, and environment). However, the challenge in ABM or MAS research is not to simply model and simulate an ecosystem, but to integrate a maximum of intelligence in these systems in view to increase the validity and reality. Thus, we have presented some ways to conceptualize the knowledge and then we are only focused on the following tool: RCC, semantic networks, and logical predicate. Using these we have presented a graph model of a spatial knowledge base called SKSN that agents can be used to perform their decision-making capabilities (spatial intelligence about the physical environment in which they move) in complex problem-solving. Our proposed model based on a semantic network has been instantiated by a sample case of knowledge representation.

A direct prospect in this work is the experimentation of all MAS models presented respecting the learning approach mentioned in each case. Theoretically, the third way of cognitive model design presented seems good for the MAS paradigm although it requires many computer resources to manage the knowledge base of each agent. The supervised learning approach by each agent is also a good alternative to make the best cognitive system representing the land use model or other.

Acknowledgements This work is done conjointly in a land use-cover project framework called COMECA⁵ in an interdisciplinary research team conducted by cooperation between IRAD of Cameroon and Kyoto University of Japan. Thus, we would like to express our deep gratitude to the socio-anthropologist and ecological researchers for their relevant information

 $^{^5}$ Co-creation of innovative forest resources management combining ecological methods and indigenous knowledge

about land use-cover activities, their patient guidance during the fieldwork in East region and their enthusiastic encouragement for this research work.

References

- 1. H. Briassoulis, Regional Research Institute, West Virginia University (2020). URL https://researchrepository.wvu.edu/cgi/viewcontent.cgi? article=1000&context=rriweb-book.
- 2. S. Preston, Popul Res Policy Rev (1996). URL https://doi.org/10.1007/BF00126129.
- 3. W. Jager, Journal of Artificial Societies and Social Simulation (2017)
- 4. W. de Groot, Environmental Science Theory. Concepts and methods in a one-world, Problem oriented paradigm (1992)
- K. Stanilov, Space in Agent-Based Models (Springer Netherlands, Dordrecht, 2012), pp. 253–269
- Q.C. Truong, T.H. Nguyen, K. Tatsumi, V.T. Pham, V.P.D. Tri, Land **11**(2) (2022). URL https://www.mdpi.com/2073-445X/11/2/297. 10.3390/land11020297
- F. Bousquet, C.L. Page, I. Bakam, A. Takforyan, Ecological Modelling 138(1), 331 (2001). Doi.org/10.1016/S0304-3800(00)00412-9
- E.D. Kameni, T.P. van der Weide, W.T. de Groot, Complex Systems Informatics and Modeling Quarterly 12, 86 (2017). 10.7250/csimq.2017-12.05
- 9. R. Ullah, pp. 11–18 (2017). DOI https://doi.org/10.1016/B978-0-12-805451-2.00002-8. URL https://www.sciencedirect.com/science/article/pii/B9780128054512000028
- T. Jarraya, Z. Guessoum, pp. 122–136 (2007). URL https://hal.archives-ouvertes.fr/hal-01311646
- M. Wooldridge, N. Jennings, D. Kinny, Autonomous Agents and Multi-Agent Systems 3, 285 (2000). DOI 10.1023/A:1010071910869
- 12. K. Davis, J. Christodoulou, S. Seider, H. Gardner, pp. 485–503 (2011)
- 13. J.L. Ermine, La gestion des connaissances (Hermès sciences publications, 2003)
- 14. R. Davis, H. Shrobe, P. Szolovits, AI Magazine 14 (2002)
- 15. R. Brachman, Readings in Knowledge Representation pp. 191–216
- 16. H.B. de Barros Pereira, M. Grilo, I. de Sousa Fadigas, C.T. de Souza Junior, M. do Vale Cunha, R.S.F.D. Barreto, J.C. Andrade, T. Henrique, Expert Systems with Applications **210**, 118455 (2022). DOI https://doi.org/10.1016/j.eswa.2022.118455. URL https://www.sciencedirect.com/science/article/pii/S0957417422015500
- D. Hommen, Synthese **196**(10), 4155 (2019). DOI 10.1007/s11229-017-1649-8. URL https://doi.org/10.1007/s11229-017-1649-8
- 18. F.M. Howari, H. Ghrefat, in Pollution Assessment for Sustainable Pracin Applied Sciences and Engineering, ed. by A.M.O. ticesMohamed, E.K. F.M. Howari (Butterworth-Heinemann, Paleologos, 2021),pp. 165 https://doi.org/10.1016/B978-0-12-809582-9.00004-9. 198. DOI URL https://www.sciencedirect.com/science/article/pii/B9780128095829000049
- 19. H.T. Ikujiro Nonaka, The knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. (Oxford University Press, 1995)
- 20. H.J. Zimmerman, Fuzzy Sets Theory and Its Application (Kluwer Academic, 1985)
- P. Sorrentino, A. Lardone, M. Pesoli, M. Liparoti, S. Montuori, G. Curcio, G. Sorrentino, L. Mandolesi, F. Foti, Frontiers in psychology 10, 728 (2019)
- M.J. Egenhofer, K.K. Al-Taha, in *Theories and Methods of Spatio-Temporal Reasoning* in *Geographic Space*, ed. by A.U. Frank, I. Campari, U. Formentini (Springer Berlin Heidelberg, Berlin, Heidelberg, 1992), pp. 196–219
- 23. B. Nebel, C. Freksa, Inform. Spektrum (2011). Doi:10.1007/s00287-011-0555-6
- 24. S. Du, Q. Qin, Q. Wang, H. Ma, International Journal of Approximate Reasoning 47(2), 219 (2008). DOI https://doi.org/10.1016/j.ijar.2007.05.002. URL https://www.sciencedirect.com/science/article/pii/S0888613X07000497
- 25. A. Galton, Earth Sci Inform 2 (2009). DOI DOI 10.1007/s12145-009-0027-6
- S.V. Kakeu Tuekam, E. Fotsing, M.J. Nkenlifack, An agent architecture embedding spatial reasoning for actors design in land use modelling (2022). URL https://hal.inria.fr/hal-03706614. CARI 2022

Blockchain



Deep Reinforcement Learning for Selfish Nodes Detection in a Blockchain
Md Muhidul Islam Khan
Blockchain for the maritime: A modular proposition Rim Abdallah, Cyrille Bertelle, Jerome Besan- cenot, Claude Duvallet and Frederic Gilletta 438
Automation and fluidity of logistics transactions through blockchain technologies Maxence Lambard, Cyrille Bertelle and Claude Duvallet
 Privileging permissioned blockchain deployment for the mar- itime sector Rim Abdallah, Cyrille Bertelle, Jérôme Besan- cenot, Claude Duvallet and Frederic Gilletta 446
Secure access control to data in off-chain storage on blockchain- based consent systems by cryptography Mongetro Goint, Cyrille Bertelle and Claude Du- vallet



Deep Reinforcement Learning for Selfish Nodes Detection in a Blockchain

Muhidul Islam Khan

Abstract Blockchain-based secure resource allocation offers a highly secure and transparent resource distribution method. Blockchain systems ensure that transactions are tamper-proof and can be validated by decentralized nodes, which removes the need for a centralized authority, reducing the risk of fraud or corruption. Additionally, smart contracts can automate resource allocation, ensuring the process is fair, transparent, and efficient. In a private blockchain, the number of nodes is often smaller than in public blockchains, making the network more vulnerable to attacks by selfish nodes. This can lead to a higher concentration of power among a smaller group of actors, making it easier for selfish nodes to manipulate the system for their benefit. Deep reinforcement learning helps to detect selfish nodes in a blockchain adaptively. We propose a deep reinforcement learning-based method for detecting selfish nodes in a blockchain. Simulation results show that our proposed method outperforms the reference solution.

Keywords Deep reinforcement learning \cdot Network slicing \cdot Blockchain \cdot Selfish nodes.

1 Introduction

Blockchain provides immutable transactions in a distributed setting. After performing a successful transaction, a block has been added with the hash. Blockchain requires participants to provide extensive resources for computing purposes and making transactions correct [1]. Most blockchain-based technologies are therefore based on incentive-based mechanisms, where participants get incentives in a consensus mechanism. The incentive mechanism is key to most

Muhidul Islam Khan

Dept. of Electrical Engineering and Computer Science Stavanger, Norway E-mail: md.m.khan@uis.no second address

permissionless blockchain-based technology. In Bitcoin, some miners get rewarded for mining a block in the system or performing a transaction [2].

Among the miners, there are two types: honest and selfish miners [3]. Our task is to identify these selfish miners, which hamper the overall system.

The contributions of the paper are as follows:

- We propose a deep reinforcement learning method for selfish miner detection.
- We propose a consensus mechanism for learning about the selfish miners.
- Overall, the proposed method helps to increase the incentive/reward, which means that the system is in the right direction.
- Our proposed selfish miner detection is helpful for the Blockchain-based system based on our simulation results for the higher and less selfish rewards. Finally, this method will be beneficial for the secure systems providing privacy, security and dependability.

The organization of the paper is as follows. Section 2 mentions the related works. Section 3 defines the problem. Section 4 provides the solution to the problem. Section 5 summarises the results, and Section 6 concludes with future works.

2 Related Works

In [4], the authors focus on the impact of a selfish miner using a Markov chain that models selfish behavior both from mining and block distribution time. They find that blocks mined by honest miners undergo a longer distribution time compared to blocks mined by selfish miners. This delay results in intentional forking, and the resulting network inconsistency provides more opportunities for selfish miners to gain unfair revenue. However, their method is probabilistic based, and there is no adaptive learning of adversaries/selfish miners. Our proposed method is an adaptive learning mechanism for finding the impact of selfish miners on the environment. We detect the selfish miners and also do neutralize the impact on the reward gained by the miners using the consensus part in our learning method.

3 Problem Definition

We design a model that formulates the Bitcoin mining behavior. After that, we detail the selfish mining algorithm.

Selfish miners achieve its goal by revealing its own mined blocks to postpone the normal workflow with the honest miners' work. Selfish miners keep its mined blocks private. On the other hand, the honest miners continue mining for the public branch. By revealing the privately owned blocks, selfish miners do engage the honest miners for achieving reward.

In Figure 1, the scenario of a miner acts as honest and selfish. The orangecolored miners denote the honest branch, H_O , which is public, and when it

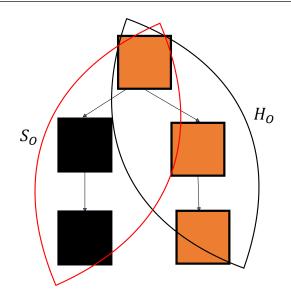


Fig. 1 Scenario of a miner working as an honest miner when mine as a public block in a honest branch and work as a selfish miner, mine the block as a private in a selfish block.

acts as a selfish, the miner can create another private branch, which is the selfish branch, S_O .

We consider the blocks arriving in the system by Poisson distribution.

$$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{1}$$

Where X is a random variable following a Poisson distribution. k is the number of times an event occurs (here, the block is generated), P(X = k) is the probability that an event will occur k times. ϵ is Euler's constant, which is approximately 2.718. λ is the average number of times an event occurs in a particular time interval. ! is the factorial function.

For the sake of simplicity, we make the following assumptions. The blockchain mining environment has the following conditions.

- The total hash power of the blockchain system is normalized as a unit. Then, the hash power of a miner is represented as a fraction of the total.
- The honest miner who finds a valid block will release it immediately. One miner may release its blocks strategically by forcing another miner into wasting his computation. When two miners, for example, A and B are selfish miners, the interaction between two private chains becomes more complicated because none of them knows the other's state. So, a dynamic learning algorithm will be helpful in this perspective.
- $-\alpha_1, \alpha_2, \alpha_3$ and α_4 are the hash powers of agents A, B, C and D respectively. It can be denoted as $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$.

3.1 Algorithm for Selfish Miner Detection

We propose a method based on Deep Q-Learning for learning the best incentive/reward for the agents in the system. We consider the consensus mechanism for detecting the selfish miners in the system and also to neutralize the impact of selfish miners. Agents will cooperate by sharing the Q-values to achieve a consensus. We prefer Deep Q Learning over classical Q learning and State-Action-Reward-State-Action (SARSA) learning since they cannot deal with the contiguous states. On the other hand, we do not choose a complex Actor-Critic Learning Algorithm for the sake of simplicity.

Reinforcement Learning (RL) consists of a set of states, actions, a reward function, and policies based on performed action. In RL, an agent performs an action a at each time step and shifts from one state s to another s'. For shifting from one state to another, the agents receive a reward for performing the action at that particular time step. In this way, agents build a policy π set, which tries to maximize the revenue over time steps [5].

3.1.1 States

We consider the states by a vector with the following variables:

- $-P_0$, the number of blocks on the public branch.
- $-H_0$, the number of blocks on an honest branch.
- $-S_0$, the number of blocks on a selfish branch.
- $-S_r$, the rewards gained by selfish miners.

3.1.2 Actions

Add to the honest branch or add to the selfish branch.

3.1.3 Reward

The reward for the agent per time step is as follows:

$$r = \frac{\alpha_r b_r N_b c_r}{T_h} \tag{2}$$

Here, α_r is the hash rate of the particular agent. b_r is the block reward based on the action. c_r is the coin price. N_b is the number of new blocks. T_h is the total hash rate of the agents.

Now, the reward function for the selfish miners can be calculated as follows:

$$S_r = \frac{s\alpha_r sb_r sN_b sc_r}{T_h} \tag{3}$$

Here, $s\alpha_r$ is the hash rate of the selfish agent. sb_r is the block reward for the selfish miners after performing the action to add a block in the selfish block.

French Regional Conference on Complex Systems - Le Havre, May 31 June 2, 2023

Algorithm 1 Proposed Method based on Deep Q Learning

Parameters and Initialization: ϕ - Parameters of the policy network Q_{ϕ} ϕ_{targ} - Parameters of the target network $Q_{\phi_{targ}}$ σ - Learning rate B - Batch size T_{u} - Period of target updates Initialization: $\phi_{targ} = \phi$ for t = 1 to T do Choose the action based on $Q_{\phi}(s_t, a_t)$ Execute the action and save transition (s_t, a_t, r_t, s'_t) in the buffer Calculate the Q-value **Consensus Step:** Send the Q-value to the neighbors Update the Q-value of agent m as follows considering the neighbor's impact: $Q_m(t+1) = Q_m(t) + \varsigma \sum_{j \in G_m} (Q_j(t) - Q_m(t))$ End of Consensus Step Calculate target Q-values for each sample in the batch: $y_m = r_m + \gamma_m \cdot a' \max Q_{\phi_{targ}}(s', a')$ Calculate the loss: $L(\phi) = \frac{1}{\mathcal{M}} \cdot \sum_{m \in \mathcal{M}} (y_m - Q_\phi(s_m, a_m))^2$ Update the network's parameters: $\phi = \phi - \sigma \cdot \gamma_m \cdot L(\phi)$ if $t \mod T_u = 0$ then Update the target network: $\phi_{targ} \leftarrow \phi$ end if end for

 sN_b is the number of new blocks for the selfish block. sc_r is the price for the coins. T_h is the total hash rate of the agents.

Algorithm 1 shows our proposed method based on Deep Q-Learning. The algorithm aims to build the action policy π by performing actions over time to maximize the cumulative reward over a particular time interval T. So, a particular agent/miner's reward function can be as follows:

$$R = \sum_{t=0}^{T} \sum_{m \in \mathcal{M}} \gamma_m \cdot r_m^t \tag{4}$$

The reward for the selfish miner S_r can be defined as the same above, where they will try to maximize their utility.

This policy is defined if we know a function that estimates the expected return based on the current state and next performed action considering the actions being taken by following the policy:

$$Q^{\pi}(s,a) = E_{s,a}[R] \tag{5}$$

The policy can be defined as follows to maximize the return:

$$\pi(s) = \arg\max_{a} Q(s, a) \tag{6}$$

We can combine the above expressions based on the Bellman equation as follows [6]:

$$Q^{\pi}(s,a) = r + \gamma \max_{a'} Q(s',a') \tag{7}$$

where s' and a' are the next state and the action taken in that state, respectively. If we estimate the Q-function using an approximator, then the quality of the approximation can be measured using the difference as follows:

$$L(\phi) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} (y_m - Q_\phi(s_m, a_m))^2$$
(8)

This value is called the temporal difference error, which is the loss function.

Algorithm 1 shows the procedure step by step about how the system works and training has been done.

3.1.4 Consensus Procedure

An agent/miner m communicates the Q-value to all its neighbors $j \in G_m$. G_m denotes the set of neighbors of miner m. All agents update their Q-values through a linear combination of their own values and the information of neighbors received in the previous time step. The procedure can be written as:

$$Q_m(t+1) = Q_m(t) + \varsigma \sum_{j \in G_m} (Q_j(t) - Q_m(t))$$
(9)

So, in other words, each miner update its state by using the disagreement of states with all its neighbors, scaled by a factor of ς . Thus the convergence rate of this algorithm depends on the scaling factor used. Convergence is guaranteed as long as the following constraint is met [6]:

$$0 < \varsigma < \frac{1}{d_{max}} \tag{10}$$

where d_{max} denotes the maximum degree among all nodes in the network. The constraint can be fulfilled by an upper bound on the maximum possible neighbors for any node.

4 Results and Discussions

We consider four miners connected in a mesh formation. We consider two of the agents to be selfish and add more blocks to the selfish branch. Blocks are generated using Poisson distribution. Our learning algorithm helps to learn the selfish miners, and the consensus part helps to detect and neutralize the effects of adversaries. We perform the simulation using Python. Table 1 shows

Table 1	Simulation	parameters
---------	------------	------------

Parameters	Value
Number of miners, \mathcal{M}	4
Learning rate, σ	0.5
Number of Episodes, T	5000
Batch size, B	512
Period of target updates, T_u	20
Discount factor, γ_m	0.5
Degree of a node, d_{max}	3
Neighboring factor, ς	0.3
Hash generation rate, α_1 , α_2 , α_3 , α_4	0.25
Block reward, b_r	5
Selfish block reward, sb_r	-5
Coin price, c_r , sc_r	1000 - 2000 Unit

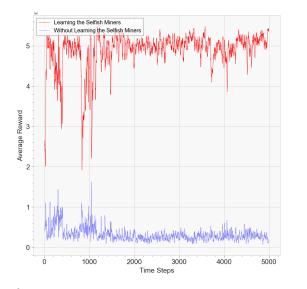
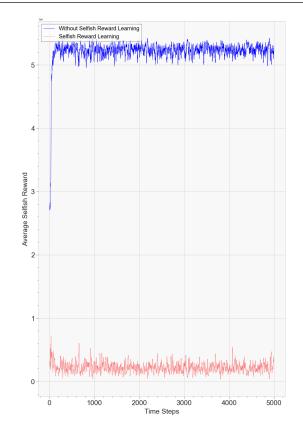


Fig. 2 Average reward over time steps.

the parameter values used for the simulation. The values are set empirically and also based on the existing literature in this study.

As we apply a method based on Deep Q learning, we consider a total of 150 hidden nodes with three layers (50 nodes per layer). We exploit the Adaptive Moment Estimation (Adam) algorithm to optimize the Neural Network (NN) weights. Rectified Linear Unit (ReLU) acts as an activation function to activate a particular input.

Figure 2 shows the graph for average reward over time steps. We compare the average reward with the learning mechanism and without the learning mechanism. Here, without learning means without the consensus part of the algorithm. Here, we can see that with our proposed algorithm, the average reward initially has a higher increment and decrement. We can observe a considerable downfall close to 1000 time steps. Then we can observe the saturation and convergence at 4050 time steps. The reason for this rapid increment and decrement is for the exploration/exploitation and also the training phases.



French Regional Conference on Complex Systems - Le Havre, May 31 June 2, 2023

Fig. 3 Average selfish reward over time steps.

On the other hand, without learning about the selfish miners, we can observe that there are some increments initially. Then for the selfish miners, there are decrements and finally remains so low for the selfish miners. Since there is no consensus in the learning algorithm, the average reward becomes too low compared with our proposed one.

Figure 3 compares selfish reward learning with our learning algorithm and without learning. We can observe that by applying our method, the selfish reward is too low, but it is too high without learning. By applying our proposed method, we can observe that the selfish reward is too low. Initially, there is an increment, and later, it is converged. On the other hand, without learning/without consensus part, the selfish reward is too high.

5 Conclusion

We propose a deep reinforcement learning-based method for detecting selfish miners in a blockchain-based application, which will be beneficial, especially for making a system secure. Our target is to propose a method to identify the selfish miners which will create a secure system and can be useful for adaptive resource allocation in any applications.

Acknowledgment

This work was funded by Norwegian Research Council through the 5G-MODaNeI project (no. 308909).

- M. N. M. Bhutta, A. A. Khwaja, A. Nadeem, H. F. Ahmad, M. K. Khan, M. A. Hanif, H. Song, M. Alshamari, and Y. Cao, "A survey on blockchain technology: evolution, architecture and security," IEEE Access, vol. 9, pp. 61 048–61 073, 2021
- 2. S. Lee and S. Kim, "Rethinking selfish mining under pooled mining," ICT Express, 2022.
- 3. I. Eyal and E. G. Sirer, "Majority is not enough: Bitcoin mining is vulnerable," Communications of the ACM, vol. 61, no. 7, pp. 95–102, 2018.
- S. G. Motlagh, J. Mi^{*}si c, and V. B. Mi^{*}si c, "Analysis of selfish miner behavior in the bitcoin network," in ICC 2021-IEEE International Conference on Communications. IEEE, 2021, pp. 1–6.
- 5. R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.
- M. Siami, S. Bolouki, B. Bamieh, and N. Motee, "Centrality measures in linear consensus networks with structured network uncertainties," IEEE Transactions on Control of Network Systems, vol. 5, no. 3, pp. 924–934, 2017.



Blockchain for the maritime: A modular proposition

Rim Abdallah · Cyrille Bertelle · Jérôme Besancenot · Claude Duvallet · Frédéric Gilletta

Keywords Blockchain \cdot Maritime sector \cdot Modular blockchain

1 Introduction

Modular blockchain deployment in the maritime sector refers to the implementation of blockchain technology in a flexible, adaptable, and scalable manner, with the ability to integrate different modules or components as needed. A modular approach can be more adept at addressing the complexity of the maritime ecosystem, a globally spread environment with multitudinous entities. Traceability and transparency are often deemed challenging within the maritime ecosystem having a heterogeneous infrastructure and superstructure. This approach allows for the segregation of different blockchain components, which enables stakeholders to address specific challenges without disrupting the entire supply chain. For instance, nodes can be segregated based on the type of data they hold, while consensus mechanisms can be segregated based on the level of security required for different transactions. This segregation helps maintain business process integrity and governance, as well as promote interoperability and collaboration among stakeholders. By breaking down the global supply chain into smaller, more manageable components, this research aims to develop an information system designed to handle the complexities of port passage and the multiple logistics actors that are involved to optimize and enhance transparency and improve traceability. In the context of the maritime industry, the use of modular blockchain deployment enables stakeholders to customize the blockchain solution according to their specific needs and requirements. This means that they can choose the blockchain modules or components that are most relevant to their operations and integrate them into their existing systems and processes. This modular approach promotes

R. Abdallah

E-mail: rim.abdallah@haropoaport.com

the adoption of the new technology on a per-component basis, allowing stakeholders to address specific challenges without disrupting the entire supply chain. The use of modular blockchain deployment also allows for greater interoperability and collaboration among stakeholders in the maritime sector. Instead of attempting to deploy a full-fledged blockchain solution in one go, the approach involves sectoring the process into easily assimilable blockchain modules. This enables stakeholders to address specific challenges and benefit from blockchain's advantageous use, allowing for greater interoperability and collaboration among stakeholders in the maritime sector. Blockchain technology can offer several advantages, such as increased transparency, traceability, and security, which are essential in the current global supply chain landscape. Moreover, the implementation of blockchain technology can help reduce the risk of fraud, enhance regulatory compliance, and improve the efficiency of the supply chain.

This research is a collaboration between the LITIS lab, the University of Le Havre, and HAROPA PORT. Its objective is to introduce blockchain technology to the maritime sector. The research not only focuses on the theoretical benefits of the technology but also includes the conceptualization and implementation of a modular approach. The goal is to make maritime traffic more secure and efficient, specifically at a major entry point such as HAROPA PORT integrating the port of le Havre, Rouen and Paris. The proposed blockchain solutions are aimed at covering a wide range of interactions between these actors, which form a complex network. The objective is to ensure that all the logistics actors continue to interact seamlessly with each other while maintaining the highest levels of security and efficiency. This approach can offer several benefits, including greater flexibility, improved efficiency, and enhanced transparency and traceability across the supply chain. As the maritime ecosystem continues to evolve and become more complex, a modular approach can help stakeholders to stay competitive and meet the demands of the global economy.

2 How to implement a modular blockchain solution

2.1 Challenges assessment

The deployment of blockchain technology can provide significant added value to various industries and sectors, including finance, healthcare, and notably supply chain management and the maritime sector. By leveraging the features of decentralization, immutability, and transparency, blockchain technology can enable secure and efficient transactions and information sharing among multiple parties. Additionally, blockchain technology can help reduce transaction costs, improve data quality, and enhance trust and accountability. However, while blockchain deployment offers many benefits, it is also crucial to address the challenges associated with the technology. Addressing these challenges will be critical to realizing the full potential of blockchain technology and ensuring its widespread adoption across different industries and sectors. There are several challenges that need to be overcome when implementing a blockchain solution in the shipping industry. These challenges include:

Challenge	targeted solution	reflection tracks
Integration with existing infrastructure	Smaller pilot projects	
	Demonstrate blockchain benefits	Identify key factors
	Gradually expand use	
Data standardization	Standardization where smart contracts	Cross system standardization
	can be used to automate	Impacts on data quality and integrity
	and notarize data exchanges	Sustainable and viable standards
Data privacy and security	Strong encryption and authentication mechanisms	Limit access to sensitive data
	Robust access control policies	Manage the ecosystem's competitiveness
Regulatory compliance	Respect and meet all requirements	Comply with regulations
		Data can be audited and verified to ensure compliance
Collaboration and governance	Establish a governance framework	Identify key stakeholders and roles
	Defines roles, responsibilities, and decision-making processes	Facilitate open and secure communication
	Provides a mechanism for dispute resolution and coordination	Identify and overcome different governance models

 ${\bf Table \ 1} \ {\rm Challenges \ assessment \ and \ reflection \ tracks}$

2.2 Technical considerations

Implementing a modular blockchain deployment in the maritime sector requires certain hardware and software requirements to be met. The following are some of the essential hardware and software requirements that need to be considered.

2.2.1 Hardware requirements

Servers To participate in a blockchain network, companies need to have reliable and powerful servers that can help ensure the network's integrity and handle high volumes of transactions. Research has found that mining can be profitable in the long run [4].

Storage Adequate storage capacity is needed to ensure smooth operations for a blockchain-based network.

Networking infrastructure A robust networking infrastructure is necessary to support the communication between the nodes in the blockchain network which might be very challenging in the maritime complex ecosystem.

Security Despite blockchain's secure characteristics such as hashing, time stamping and encryption [3], open data exchange is generally frowned upon in the shipping industry.

2.2.2 Software requirements

Blockchain platform The choice of a suitable blockchain that accomodates key features such as scalability, interoperability, security, privacy, and smart contract support.

Node software The node software provides necessary blockchain functionalities. [5].

Wallet software Wallet software is used to manage digital assets, such as cryptocurrencies or other tokens, in the blockchain network. It allows users to send and receive digital assets and manage their account balances.

Smart contract development tools To develop smart contracts, companies need to use suitable development tools and programming languages [6]. Platforms such as Solidity and Vyper are commonly used for smart contract development.

Security and compliance tools To ensure security and compliance, companies can use specialized tools such as encryption software, access control tools, and compliance monitoring solutions.

3 Conclusion and future work

Blockchain deployment in the maritime sector offers several benefits, including increased efficiency, cost savings, and improved transparency and traceability[1]. A full-fledged blockchain solution propagating over the global supply chain seems difficult to achieve due to the complex nature of both the technology itself and the ecosystem[2].

A modular approach can be a lot more feasible. The modular approach suggests creating a blockchain environment with multiple blockchain solutions co-existing and integrating pre-existing technological systems. This environment can serve as an infrastructure for notarizing relevant data for stakeholders in the maritime industry. Future work includes the following: the choice of a blockchain type adequate for maritime widespread use, the conceptualization of a viable solution for the sector that successfully integrates existing systems, and the successful implementation of the solution and its testing.

- 1. Abdallah, R., Bertelle, C., Duvallet, C., Besancenot, J., Gilletta, F.: Blockchain potentials in the maritime sector: A survey. In: Proceedings of the ICR'22 International Conference on Innovations in Computing Research. pp. 293–309. Springer (2022)
- Abdallah, R., Besancenot, J., Bertelle, C., Duvallet, C., Gilletta, F.: Assessing blockchain challenges in the maritime sector. In: Blockchain and Applications, 4th International Congress. pp. 13–22. Springer (2023)
- Dib, O., Brousmiche, K.L., Durand, A., Thea, E., Hamida, E.B.: Consortium blockchains: Overview, applications and challenges. Int. J. Adv. Telecommun 11(1), 51–64 (2018)
- Islam, N., Marinakis, Y., Olson, S., White, R., Walsh, S.: Is blockchain mining profitable in the long run? IEEE Transactions on Engineering Management 70(2), 386–399 (2023). https://doi.org/10.1109/TEM.2020.3045774
- 5. Knirsch, F., Unterweger, A., Engel, D.: Implementing a blockchain from scratch: why, how, and what we learned. EURASIP Journal on Information Security **2019**, 1–14 (2019)
- Mohanta, B.K., Panda, S.S., Jena, D.: An overview of smart contract and use cases in blockchain technology. In: 2018 9th international conference on computing, communication and networking technologies (ICCCNT). pp. 1–4. IEEE (2018)



Automation and fluidity of logistics transactions through blockchain technologies

Maxence Lambard · Cyrille Bertelle · Claude Duvallet

1 Introduction

The fast execution of logistics transactions is an important factor in logistics management. This speed is linked to the fluidity of transactions. Thus, an important aspect of logistics management is the development of digital solutions to automate and certify the transactions that need to be performed between a large number of actors.

Traditional contracts that use documentary credit as a means of payment are known to suffer from moral hazard and payment timing problems. In this article, we are interested in guaranteeing the proper execution of the contract between the partners using SLCs. In this article, we present a model for a new generation of logistics transaction system that manages blockchain-based escrow and validation mechanisms to ensure the proper execution of the contract.

2 Blockchain Technology

Blockchain technology is a system for storing and transmitting digital information that operates in a decentralized, transparent and secure manner. It was first introduced in 2008 by Satoshi Nakamoto with the famous cryptocurrency Bitcoin [8]. Its operation is based on a decentralized network of computers

Cyrille Bertelle

Claude Duvallet

Maxence Lambard

LITIS, Université Le Havre Normandie, LITIS (UR 4108), Le Havre, 76600 E-mail: maxence.lambard@univ-lehavre.fr

LITIS, Université Le Havre Normandie, LITIS (UR 4108), Le Havre, France, 76600 E-mail: cyrille.bertelle@univ-lehavre.fr

Université Le Havre Normandie, LITIS (UR 4108), Le Havre, France, 76600 E-mail: claude.duvallet@univ-lehavre.fr

that record and verify all transactions made on the blockchain. Each block of transactions is encrypted and linked to the previous one, forming a blockchain that cannot be changed without detection.

Other blockchains have appeared, notably Ethereum [3] which proposes to certify all types of transactions and thus introduce the notion of SC. A SC is a computer program that runs once all the conditions written in it are met. In order for a SC to use data external to the blockchain, it is possible to use an oracle [2]. This is a device or program that provides data external to the blockchain from various sources so that SCs automate processes that depend on this data.

One of the limitations of the SC is that it has no legal validity [1]. However, there is a solution to this problem, the SLCs. It is a SC with all the legal requirements of an agreement.

3 Contribution

In this section, we present the beginning of a solution based on a generic SLC developed on the Ethereum blockchain in the context of smart logistics. Our contribution consists in showing how the use of a SLC via the blockchain can resolve conflicts between the actors of a contract.

The blockchain creates an audit trail that documents the provenance of an asset at every stage of its journey. This audit trail acts as evidence and can also expose vulnerabilities in any supply chain. Furthermore, traditional paper-based processes are time-consuming, prone to human error and often require third-party mediation. By leveraging blockchain to streamline these processes, validation and transaction processing can be much faster.

As a result, this SLC will enable automatic, fast and low-cost execution of logistics transactions.

3.1 The model

As you can see in Fig. 1, our model is partitioned in 2 parts. The first one describes the events happening on the blockchain. This includes the storage of contractual information, the history of escrows or the rights of use for the different actors of the contract.

The second one represents the events happening outside the blockchain and allows to transmit information available outside the blockchain. In order to record the events emitted by the SLC but also to interact with it through an oracle to validate the conformity of the goods for example.

We consider that the SLC is accessible by different actors such as buyers, sellers, oracles and lawyers. They are represented by Ethereum addresses in the SLC itself.

French Regional Conference on Complex Systems - Le Havre, May 31 June 2, 2023

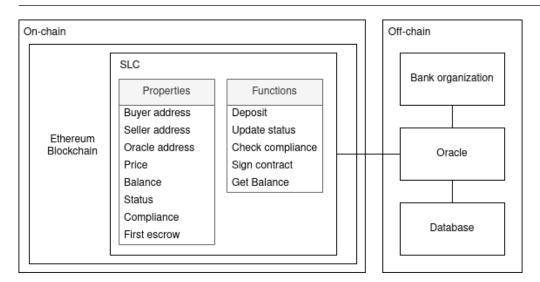


Fig. 1 A representation of the model architecture with a basic SLC

3.2 Contract registration

The seller wishing to sell his merchandise enters the actual contract from an agreement with the buyer(s). The seller must provide the following information: the address of the buyer(s), the minimum escrow for goods, the desired price, the desired currency for payment (ether, euro, USD), the minimum compliance for validation by the oracle... Once these data are entered and correct, a transaction is performed on the blockchain to record the contract data in an immutable and unforgeable way.

3.3 Sequestration

In collaboration with several economist and legal colleagues, we were able to develop a documentary credit [9] system with a progressive sequestration. This has the effect of not freezing all of the buyer's assets.

There are several ways to make an escrow in our SLC. Indeed, if the selected currency is ether then the buyer will have to pay the goods in ether. On the other hand, if the currency is a physical currency such as euro, then he will have to go through a traditional banking institution.

In order for the goods to be shipped by the seller, the buyer must deposit the minimum escrow required. The buyer also has the possibility to deposit assets as he wishes. Some steps can be blocking if the requested escrow is not sufficient, especially during the validation of the goods.

3.4 Delivery and validation of the goods

When the goods are ready to be shipped, the oracle will come and change its status on the SLC in order to track the goods. At each delivery stage, the oracle updates the status of the contract in the blockchain according to data that can come from sensors, handlers...

Once the goods have arrived at the dock, a conformity check can be performed by qualified agents to determine the conformity rate of the goods. The result of this control is sent to the oracle which will provide this information to the SLC in order to determine if the contract can be validated so that the final payment of the buyer is possible.

4 Conclusion and future works

The use of the blockchain is intended to provide confidence in the automation and fluidity of transactions of financial flows in a logistical operation. In case of conflict, a legal authority will then be able to reliably consult the entries in this non-falsifiable digital register if they are compatible with contractual rules, as proposed by the concept of SLC.

It would be interesting to port this model to a hybrid blockchain, i.e. a public blockchain to store the hashes of the blocks and a private blockchain to perform the transactions. This would reduce the cost of transactions performed on a public blockchain and increase the speed of transactions.

We will consider the payment timing problem in the SLC model following Gans solution [5]. Moreover, we will try to solve the problems of non-execution of contract. Finally, a web application allowing to configure and interact with this SLC is under development.

Acknowledgements We would like to thank the Normandy region and the Fonds National d'Aménagement et de Développement du Territoire (FNADT) for supporting this work in the framework of the Catalyse project.

- Levi, B. Stuart, D. (2018). An Introduction to Smart В. 1. Alex. Con-Potential Their and Inherent Limitations. Accessed tracts and on 28/02/2023]. Available : https://corpgov.law.harvard.edu/2018/05/26/ on an-introduction-to-smart-contracts-and-their-potential-and-inherent-limitations/
- 2. Beniiche, A. (2020). A Study of Blockchain Oracles. pp. 1-3.
- 3. Buterin, V. (2015). Ethereum White Paper : A next generation smart contract & decentralized application platform. pp. 13-19.
- 4. Ethereum Foundation. (2022). What is a smart contract ?. [Accessed on 15/02/2022]. Available on : https://ethereum.org/fr/smart-contracts/
- 5. Gans, J. (2019). The Fine Print in Smart Contracts.
- 6. IBM. (2023). Benefits of Blockchain. [Accessed on 07/11/2022]. Available on : https://www.ibm.com/fr-fr/topics/benefits-of-blockchain
- 7. Law Commission. (2021). Summary on the subject of Smart Legal Contracts. [Accessed on 13/12/2022]. Available on : https://s3-eu-west-2.amazonaws.com/ lawcom-prod-storage-11jsxou24uy7q/uploads/2021/11/6.7776_LC_Smart_Legal_ Contracts_2021_Final.pdf
- 8. Nakamoto, S. (2008). Bitcoin : Peer-to-Peer Electronic Cash System. pp. 1-4.
- 9. Nordea. (2009). Documentary credits in practice (2nd Edition). pp. 17-21.



Privileging permissioned blockchain deployment for the maritime sector

Rim Abdallah · Cyrille Bertelle · Jérôme Besancenot · Claude Duvallet · Frédéric Gilletta

Keywords Blockchain \cdot Maritime sector \cdot Permissioned blockchain

1 Introduction

The maritime sector is a vital component of global trade, responsible for transporting goods and commodities across oceans and waterways. The massification of global trade has put significant pressure on the maritime industry to modernize and optimize its operations. This, coupled with the increasing complexity of the supply chain, has led to the transformation of how business is conducted for improved efficiency. It required systems that were faster, more efficient, and less prone to errors, which traditional paper-based systems were unable to fully satisfy. Hence the introduction of digitization, a game changer for the sector. One of the key areas of digitization in the maritime sector is the implementation of Port Community Systems (PCS) and Cargo Community Systems (CCS). These systems are designed to streamline communication and data exchange between the numerous parties involved in the port and cargo operations. And, As ports and logistics hubs became busier, PCS and CCS have become increasingly popular worldwide, helping ports and logistics companies to increase efficiency, reduce costs, streamline processes, improve the overall customer experience, stay competitive and meet the growing demands of the global economy [5].

While Port Community Systems (PCS) and Cargo Community Systems (CCS) have been successful in improving communication and information sharing between stakeholders in the maritime industry. The maritime sector is a complex ecosystem with numerous stakeholders, including shipping lines, freight forwarders, port authorities, customs, and others. And, this complexity

R. Abdallah

E-mail: rim.abdallah@haropoaport.com

has made it challenging to achieve traceability and transparency in the maritime and port field. A taxonomy of CS identifies improvement possibilities [6] that can be provided by blockchain technology[7].

In collaboration with the LITIS lab, the University of Le Havre, and HAROPA PORT, this research justifies privileging permissioned blockchain in the maritime sector. We highlight the complexity of the maritime ecosystem, and its technological resilience and propose blockchain as the next technological innovation that contributes to more fluidified and secure interactions between its various actors. We anchor the characteristics of a permissioned blockchain and argue the added benefits of blockchain for the industry. This research serves as a preliminary study for the application of a permissioned blockchain framework that integrates existing solutions at HAROPA PORT. It handles the complexities of port passage and the diverse logistics actors involved in this framework while maintaining optimal levels of security and efficiency.

2 Blockchain for the maritime

Blockchain is a decentralized digital ledger that enables secure, transparent, and immutable record-keeping of transactions. In the maritime industry, blockchain technology could be used to provide end-to-end visibility and traceability of cargo movements, as well as to reduce fraud, errors, and delays in the supply chain. One of the key advantages of blockchain technology is that it provides a single source of truth that can be accessed by all authorized parties, enabling more efficient and secure information sharing. This could reduce the need for intermediaries, such as customs brokers, and streamline the clearance process. Additionally, blockchain technology could enable the automation of certain processes, such as customs clearance and payments, further reducing costs and improving efficiency. Blockchain-based smart contracts could also facilitate greater collaboration and trust between stakeholders, as they would provide a transparent and enforceable framework for agreements. While PCS and CCS have been successful in improving the efficiency of the maritime industry, blockchain technology has the potential to take this a step further by providing a more secure, transparent, and efficient platform for managing cargo movements and supply chain operations[1].

2.1 What is a permissioned blockchain?

A permissioned blockchain network restricts access and participation to a defined group of participants, unlike a public blockchain which is open to anyone. In a permissioned blockchain, participants must be authenticated and authorized to access the network, and transactions are validated by a select group of nodes. This makes it suitable for use cases where confidentiality, privacy, and control over data access are important[4]. 2.2 Unlocking potentials in the Maritime Sector through Permissioned Blockchain Technology

Permissioned blockchain technology has gained significant attention in recent years as a potential solution to the challenges faced by the maritime sector. In this section, we will discuss the arguments for the use of permissioned blockchain in the maritime sector based on literature analysis and case studies. The use of permissioned blockchain networks in the maritime sector offers several advantages over public ones, particularly in contrast to the financial sector.

Confidentiality The maritime industry deals with sensitive data, such as cargo and vessel information, which requires confidentiality and security. The use of permissioned blockchain networks, which limit access to authenticated and authorized participants, reduces the risk of unauthorized access and data breaches. Public blockchain networks, which are open to anyone, are less suitable for industries with sensitive data requirements[3].

Control Permissioned blockchain networks offer enhanced control over the network, resulting in more efficient and secure operations. The maritime industry involves a multitude of stakeholders, and a permissioned blockchain network provides a secure way for these stakeholders to share data and coordinate activities while maintaining control over who can access and validate the data. Public blockchain networks, on the other hand, operate in a decentralized and uncontrolled environment, which can lead to inefficiencies and security risks [3] [9].

Efficiency Permissioned blockchain networks can operate more efficiently and with lower energy costs compared to public blockchain networks. This is because the number of nodes validating transactions is limited, leading to faster processing times and reduced energy consumption. In the maritime sector, where large volumes of data need to be processed quickly, a permissioned blockchain network can provide significant efficiency gains[8].

Despite the renowned use of public blockchain particularly its application in the financial sector: Bitcoin, Ethereum, and others. We present a contrast between the two sectors. Where the financial sector often requires the transparency and accessibility provided by public blockchain networks. Public blockchain networks can provide the necessary transparency and accessibility for these requirements. However, in the maritime sector, confidentiality and control over data access are more important, making permissioned blockchain networks a more suitable option.

3 Conclusion

In conclusion, the use of permissioned blockchain networks offers significant advantages for the maritime industry, given its sensitive and confidential data

requirements. The ability to restrict access to a defined group of authenticated and authorized participants provides greater control over the network, enabling more efficient and secure operations. The maritime industry's complex network of stakeholders can benefit from the secure sharing of data and coordination of activities, facilitated by permissioned blockchain technology. Future work is an architectural proposal for a permissioned blockchain deployment in the maritime sector including its implementation, benchmarking, and testing to assess its suitability and viability as a solution. Technological limitations have been carefully assessed in [2] and possible deployment solutions have been identified. The implementation stage involves deploying the proposed architecture and integrating it with existing systems to ensure compatibility and interoperability. Once implemented, benchmarking can be performed to measure its performance against established metrics and standards, such as transaction throughput, latency, and fault tolerance. Testing can also be conducted to validate the system's security and confidentiality features, ensuring that sensitive information remains protected from unauthorized access and breaches. By thoroughly evaluating the proposed permissioned blockchain solution, the maritime sector can confidently adopt this technology as a viable and secure solution for its data management and coordination needs.

- 1. Abdallah, R., Bertelle, C., Duvallet, C., Besancenot, J., Gilletta, F.: Blockchain potentials in the maritime sector: A survey. In: Proceedings of the ICR'22 International Conference on Innovations in Computing Research. pp. 293–309. Springer (2022)
- Abdallah, R., Besancenot, J., Bertelle, C., Duvallet, C., Gilletta, F.: Assessing blockchain challenges in the maritime sector. In: Blockchain and Applications, 4th International Congress. pp. 13–22. Springer (2023)
- Ashraf, I., Park, Y., Hur, S., Kim, S.W., Alroobaea, R., Zikria, Y.B., Nosheen, S.: A survey on cyber security threats in iot-enabled maritime industry. IEEE Transactions on Intelligent Transportation Systems (2022)
- Dib, O., Brousmiche, K.L., Durand, A., Thea, E., Hamida, E.B.: Consortium blockchains: Overview, applications and challenges. Int. J. Adv. Telecommun 11(1), 51–64 (2018)
- Posti, A., Häkkinen, J., Tapaninen, U.: Promoting information exchange with a port community system-case finland. Int Supply Chain Manag Collab Pract 4, 455–473 (2011)
- Wallbach, S., Coleman, K., Elbert, R., Benlian, A.: Multi-sided platform diffusion in competitive b2b networks: inhibiting factors and their impact on network effects. Electronic Markets 29, 693–710 (2019)
- Xinyi, Y., Yi, Z., He, Y.: Technical characteristics and model of blockchain. In: 2018 10th International Conference on Communication Software and Networks (ICCSN). pp. 562–566 (2018). https://doi.org/10.1109/ICCSN.2018.8488289
- 8. Yang, C.S.: Maritime shipping digitalization: Blockchain-based technology applications, future improvements, and intention to use. Transportation Research Part E: Logistics and Transportation Review **131**, 108–117 (2019)
- Zhang, P., Wang, Y., Aujla, G.S., Jindal, A., Al-Otaibi, Y.D.: A blockchain-based authentication scheme and secure architecture for iot-enabled maritime transportation systems. IEEE Transactions on Intelligent Transportation Systems (2022)



Secure access control to data in off-chain storage on blockchain-based consent systems by cryptography

Mongetro Goint · Cyrille Bertelle · Claude Duvallet

Abstract _

Keywords blockchain \cdot data access \cdot data encryption \cdot distributed ledger

1 Introduction

Data sharing is considered one of the most beneficial elements of cloud computing. This can promote business collaboration, improve quality of scientific research, improve services offered to users of a system, etc. Data must however be managed effectively. Aspects such as privacy, security and transparency in data management are of increasing concern to data owners. Consent for data access is moreover a legal requirement¹.

Blockchain technology is widely used to build consents management systems for data access today [6–9,11]. However, a blockchain is not an ideal ledger for storing large data. Because, the scalability is often relatively limited and the transaction costs in a blockchain are relatively high compared to the volume of data to be stored. For this, the blockchain is often coupled with off-chain storage systems, to record large data. Therefore, it is necessary to apply strict security mechanisms for data protection in the off-chain system.

Cyrille Bertelle Université Le Havre Normandie, LITIS, UR 4108, Le Havre, F-76600 E-mail: cyrille.bertelle@univ-lehavre.fr

Claude Duvallet

Mongetro Goint

Université Le Havre Normandie, LITIS, UR 4108, Le Havre, F-76600 E-mail: mongetro.goint@univ-lehavre.fr

Université Le Havre Normandie, LITIS, UR 4108, Le Havre, F-76600 E-mail: claude.duvallet@univ-lehavre.fr

 $^{^{1}\} https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679$

This article proposes a securing mechanism, based on data encryption, to secure user data stored in off-chain systems. The protocol uses a symmetric key encryption mechanism [12], to encrypt data stored in off-chain storage. While the symmetric key is anchored securely in the blockchain, it is used, after verification of consent established, to decrypt data in off-chain storage and make them readable by data accessor.

The rest of this article is organized as follows: section 2 presents an overview of blockchain and smart contracts. Section 3 presents how blockchain participates in consents management systems. Section 4 shows the need for data security in off-chain storage. Section 5 presents the mechanism for securing data access in off-chain systems. Finally, section 6 concludes the article.

2 Blockchain overview

Blockchain was first proposed in 2008 by Satoshi Nakamoto [1]. A blockchain is a distributed and secure ledger to which all participants have access. The public register shared between its participants contains all the transactions that have been carried out there since its creation, while participants keep a copy of the ledger. Secure, verified by the entire network using cryptographic protocols and consensus algorithms, transactions in a blockchain are put into blocks and then arranged chronologically to form a chain.

Blockchain technology is going well beyond its original use, cryptocurrencies, allowing data sharing in a decentralized and secure way. This approach was initiated in Ethereum Blockchain [2], thanks to smart contract². A smart contract refers to a program that runs on blockchain when the clauses that have been defined there are met. Ethereum promotes development of decentralized applications (DApps). Blockchain has been used for years in many fields such as Finance, Insurance, Agriculture, Health, Internet of Things, etc.

3 Consents management with blockchain

Consent is a clear positive act by which the data subject manifests in a free, specific, informed and unequivocal way his agreement to the processing of personal data concerning him. It is fundamental in data management.

Several works have shown the usefulness of blockchain for consent management, for data access in different domains [6-10].

4 Access control to data in off-chain storage with encryption

To solve certain problems such as scalability and transaction costs, blockchain is often coupled with off-chain storage systems to build consent management systems. A blockchain, as a consent certification notary, does not necessarily

² https://ethereum.org/en/developers/docs/smart-contracts

guarantee the security of data stored outside the blockchain. To secure offchain data in a consent management systems, we propose a protocol based on cryptography for data encryption.

Cryptography is a process of converting data from readable to unreadable form in order to meet security requirements [12]. There are several cryptographic mechanisms to encrypt data: Asymmetric encryption that uses a set of two keys (public key for encryption and private key for decryption of data); Symmetric encryption (the one used in our protocol) that uses the same unique key to encrypt and decrypt data.

Fig. 1 shows a general overview of our architecture, divided into three parts: the off-chain system, the blockchain and the users.

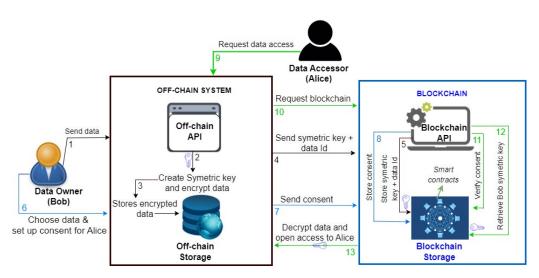


Fig. 1 Generic architecture of the data control system based on data encryption.

In our protocol, large data will be encrypted and stored in the off-chain system; the blockchain ledger will be used to securely record everything needed for data access (access key and consent). Moreover, as data is supposed to be shared with several actors, we propose a model based on symmetric keys encryption that can be used in consents systems, with several actors.

• How it works ?

After the user registers, when storing data, a symmetric encryption key is generated and encrypt the data, before it is sent to the off-chain storage (cf. Fig 1). The symmetric key along with the data ID is then sent to the blockchain, via a smart contract, while the encrypted data is stored in offchain storage. So, data owner decide to grant consent to another user. Then, he establishes a consent with elements like the data owner's ID, the data accessor's ID, the data ID and also the data access conditions. The consent is recorded in the blockchain, and used to access the user's data by the accessor.

Fig. 1 shows how Alice accesses Bob's data after receiving his consent: Alice requests access to Bod's data; the off-chain API ask the blockchain system to

verify the existence of consent; the blockchain retrieves Bob symetric key via a smart contract; finally, the symmetric encryption is used to decrypt Bob's data, which will be available to Alice.

Using data encryption process constitutes a second layer of security. If despite everything a hacker manages to use tricks to access the data stored in the off-chain storage, he will not be able to read them.

5 Conclusions and future work

Secure data access control remains a critical factor in data management.

This article presented a protocol based on symmetric key systems, coupled with consent management mechanisms, for data security. This approach provides multi-level security for data in off-chain systems linked to a blockchain system. This ensures privacy, security and data integrity.

One perspective would be the one proposed in [14]. This is to allow accessors to perform calculations on the data in the off-chain system and get the final results, instead of letting an actor observe data.

- 1. Nakamoto, S. Bitcoin: A peer-to-peer electronic cash system. Bitcoin.org, 1-4 (2008).
- 2. Buterin, V. Ethereum white paper: A next generation smart contract and decentralized application platform, 1-4 (2013).
- 3. Tenopir, C., Allard, S., Douglass, K. et al. Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, 6, (2011).
- 4. GDPR. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natu- ral persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec. Official Journal of the European Union, 1–88 (2016).
- Biswas, K. and Muthukkumarasamy, V. Securing smart cities using blockchain technology. HPCC/SmartCity/DSS, 1392–1393 (2016).
- Mamo, N., Martin, G., Desira, M. et al. Dwarna: a blockchain solution for dynamic consent in biobanking. *European Journal of Human Genetics*, 28, 1–18 (2019).
- Agarwal, R. R., Kumar, D., Golab, L. et al. Consentio : Managing consent to data access using permissioned blockchains. In 2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC), 1–9 (2020).
- 8. Rantos, K., Drosatos, G., Demertzis, K., et al. Advocate: A consent management platform for personal data processing in the iot using blockchain technology. in: *SecITC*, 1–16 (2018).
- Jaiman, V. and Urovi, V. A consent model for blockchain-based health data sharing platforms. *IEEE Access*, 8, 143734–143745 (2020).
- 10. Michelin, R. A., Dorri, A., Lunardi, R. C. et al. Speedychain : A framework for decoupling data from blockchain for smart cities. *CoRR*, 145–154 (2018).
- Makhdoom, I., Zhou, I., Abolhasan, M. et al. Privysharing : A blockchain-based framework for privacy-preserving and secure data sharing in smart cities. *Computers & Security*, 88, 101653 (2019).
- 12. Saranya, K., Mohanapriya, R. and Udhayan, J. A Review on Symmetric Key Encryption Techniques in Cryptography. Int. J. Sci. Eng. Technol, vol. 3, no. 3, 539–544 (2014).
- 13. Aldred, N., Baal, L., Broda, G. et al. Design and implementation of a blockchain-based consent management system. (2019).
- 14. Zyskind, G., Nathan, O., et al. Decentralizing privacy : Using blockchain to protect personal data. In 2015 IEEE Security and Privacy Workshops, 180–184 (2015).

Learning on Graphs



Are Networks Really Useful? — Interplay Between Structures and Vector Representations <i>Tsuyoshi Murata</i>
 Quantile regression: an approach based on GEV distribution and machine learning Lucien M. Vidagbandji, Laurent Amanton, Alexan- dre Berred and Cyrille Bertelle
DyHANE: Dynamic Heterogeneous Attributed Network Embed- ding Liliana Martirano, Roberto Interdonato, Dino Ienco and Andrea Tagarelli



Are Networks Really Useful? — Interplay Between Structures and Vector Representations

Tsuyoshi Murata

Abstract Because of the boom of deep learning, several attempts have been made for network embedding (or representation learning), which transforms nodes in a network into vectors (or latent representation) so that proximity of the nodes in the network will be preserved. On the other hand, graph structure learning is an opposite direction. Networks generated from real-world data often contain noise which degrade the performance of machine learning tasks. Graph structure learning is to transform vectors into structures for alleviating such noise, and it is also an emerging research topic of machine learning. In this abstract, we discuss the interplay between structures and vector representations of networks. Structures and vector representations are both important information obtained from real-world structured data. As the first step, we show the approaches for transforming structures to vectors, and vice versa.

Keywords network embedding \cdot representation learning \cdot graph structure learning

1 Network Embedding

It is often said that networks (or graphs) is quite important for representing relations among entities in the real world. When we analyze networks, it is often the case that the networks are given in advance. However, it is not obvious how to create networks from given non-structured data. In addition, it is still controversial whether graph structure is really useful for analyzing data.

Tsuyoshi Murata

Department of Computer Science, School of Computing

Tokyo Institute of Technology

W
8-59 2-12-1 Ookayama, Meguro, Tokyo, 152-8552 Japan

Tel./Fax.: +81-3-5734-2684

E-mail: murata@c.titech.ac.jp

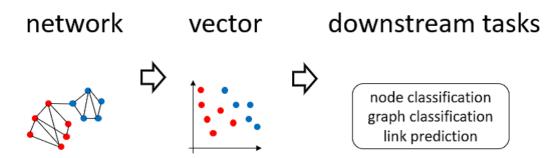


Fig. 1 Network Embedding

Because of the boom of deep learning, several attempts have been made for network embedding (or representation learning), which transforms nodes in a network into vectors (or latent representation) so that proximity of the nodes in the network will be preserved (Fig. 1).

Performing network embedding is expected to alleviate some of the problems of traditional network representation. Vectors are suitable as the input to machine learning methods (especially for deep learning). In addition to that, two or three dimensional network embedding can be regarded as embedding in 2D plane or 3D space, so we can use it for network visualization.

There are many methods for network embedding even for simple homogeneous networks with no node/edge types, such as matrix factorization (SVD, non-negative matrix factorization), random walk-based methods (DeepWalk, node2vec), neural network-based methods (SDNE, SDAE), and so on. Regarding network embedding, survey paper [3][11], tutorial [5] and online resourses [1][9] are available.

However, there are still many challenges for network embedding. Zhang et al. pointed out some of the potential research directions such as task-dependent embedding, dynamics of networks and robustness [11].

2 Graph Structure Learning

Learning structures from data has a long history. For example, Bayesian networks are one of the graphical models for representing causal relations based on probabilities. Topological data analysis (TDA) is an approach to the analysis of datasets using techniques from topology. This is because networks are versatile and they can be used for representing several types of relations among entities.

Recently, graph structure learning (GSL) (Fig. 2) becomes popular in the field of machine learning. Graph neural networks (GNNs) are for performing deep learning on graphs. Because of the excellent expressive power, GNNs can be used for several machine learning tasks such as node classification, graph classification and link prediction. GNNs are also used for graph structure learning [2], which aims to jointly learn an optimized graph structures and corresponding graph embeddings. It often happens that given networks

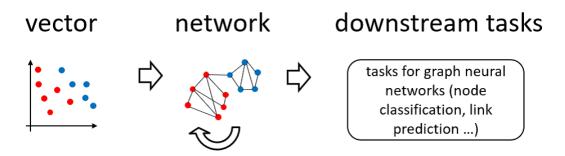


Fig. 2 Graph Structure Learning

are incomplete and noisy, which poses a great challenge for applying GNNs to real-world problems. Because of the mechanism of GNNs (computing node embeddings by recursively aggregating information from neighboring nodes), deliberate perturbation (or adversarial attacks) in graph structures can easily result in wrong predictions for most GNNs. Therefore, high-quality graph structure is often required for GNNs to perform network embedding.

The purpose of graph structure learning is to refine given networks so that they will be good enough for applying GNNs. The inputs of graph structure learning are a network (adjacency matrix) with node attributes (node feature matrix), and the outputs are modified adjacency matrix and corresponding network embedding that are optimized for certain downstream tasks (such as node classification and graph classification).

The primary goal of graph structure learning is the performance improvement of GNN-based machine learning. At the same time, GSL can be regarded as a variation of link prediction. It will enhance the interpretability of networks through visualization when the size of the networks are rather small. Link prediction is one of the popular research topics in network science. Although there are already many attempts for link prediction [7], the goal of GSL is to refine networks for performance improvement of machine learning.

According to [12], applications of graph structure learning are natural language processing, computer vision and medical imaging, and scientific discovery. There are some online resources about graph structure learning [4][10].

3 Interplay between Structures and Vector Representations

It seems a bit strange that these two opposite research directions are popular recently. One is to crash the structure into vectors, while the other is to create structure from given vectors. This is partly because the enthusiastic boom of deep learning.

As mentioned in section 1, network embedding (or representation learning) is studied by many researchers. The main purpose of network embeddding is to transform networks to vectors so that ready-made machine learning tools (including deep learning) can be utilized for structured data. It also eases the combination of network information and features attached to nodes and edges.

Graph structure learning is just the opposite direction of network embedding. It creates structures from non-structured data. The main purpose of graph structure learning is to improve the peformance of machine learning tasks, such as node classification, graph classification and link prediction. Another purpose is to enhance interpretability through network visualization.

There are two types of representations: networks and vectors. We cannot say that one is always better than the other. These are complementary. We need to think when one is better than the other, and how one can be transformed to the other.

4 Concluding Remarks

According to Holme [6], the requirements for general criteria for network science are as follows: (1) nodes and edges should have concrete interpretations, (2) there should be self-evident mechanisms for how nodes can indirectly influence one another, and (3) the number of edges should not be almost zero or almost maximal. He also pointed out that network-based approaches are more versatile and weather patterns can be predicted by networks more accurately than with spatial methods [8]. What network is really useful for? Its answer is left for our future work.

- 1. Chih-Ming Chen, awesome-network-embedding, https://github.com/chihming/awesome -network-embedding (2020)
- Yu Chen, Lingfei Wu, "Graph Neural Networks: Graph Structure Learning", https: //graph-neural-networks.github.io/gnnbook_Chapter14.html, Chapter 14, pp. 297-321, Graph Neural Networks, Springer (2022)
- 3. Peng Cui, Xiao Wang, Jian Pei, Wenwu Zhu, "A Survey on Network Embedding", IEEE Transactions on Knowledge and Data Engineering, Vol. 31, No. 5, pp. 833-852 (2019)
- 4. graph-structure-learning, https://github.com/topics/graph-structure-learning (2022)
- William L. Hamilton, Rex Ying, Jure Leskovec, Rok Sosic, "Representation Learning on Networks", WWW-18 Tutorial, http://snap.stanford.edu/proj/embeddings-www/ (2018)
- Petter Holme, "The Network Scientist's Survival Kit", https://petterhol.me/2021/11 /25/survival-kit/ (2021)
- Linyuan Lu, Tao Zhou, "Link Prediction in Complex Networks: A Survey", Physica A, Vol. 390, Issue 6, pp. 1150-1170 (2011)
- Josef Ludescher, Maria Martin, Niklas Boers, Armin Bunde, Catrin Ciemer, Jingfang Fan, Shlomo Havlin, Marlene Kretschmer, Jürgen Kurths, Jakob Runge, Veronika Stolbova, Elena Surovyatkina, Hans Joachim Schellnhuber, "Network-based forecasting of climate phenomena", PNAS Vol. 118, No. 47, e1922872118, pp.1-10 (2021)
- Alan Tang, Awesome-Network-Embedding, https://github.com/tangzhenyu/Awesome-Network-Embedding (2019)
- Vectory, "Awesome Graph Structure Learning", https://github.com/blacker521/awe some-graph-structure-learning (2022)
- 11. Daokun Zhang, Jie Yin, Xingquan Zhu, Chengqi Zhang, "Network Representation Learning: A Survey", IEEE Transactions on Big Data, Vol.6, Issue 1, pp.3-28 (2020)
- 12. Yanqiao Zhu, Weizhi Xu, Jinghao Zhang, Yuanqi Du, Jieyu Zhang, Qiang Liu, Carl Yang, Shu Wu, "A Survey on Graph Structure Learning: Progress and Opportunities", https://arxiv.org/abs/2103.03036 (2021)



Quantile regression: an approach based on GEV distribution and machine learning

Lucien M. Vidagbandji · Laurent Amanton · Alexandre Berred · Cyrille Bertelle

1 Introduction

Many complex natural or artificial systems are characterised by critical evolutions that are difficult to predict and control. It is therefore relevant to look at the risk assessment of extreme events, which requires an accurate estimation of high quantiles that can sometimes exceed the observation range. Thanks to the asymptotic results of the extreme value theory, extrapolation beyond the data range is possible. When the extreme event depends on some characteristic variables, quantile regression is a statistical tool allowing to have these extreme quantiles conditional to the dependent variables. The classical methods available for modelling conditional quantile regression fail mainly when the structure between the variable representing the extreme event and the characteristic variables is complex or when the size of the characteristics is large. Recent literature has seen the development of machine learning approaches for the analysis of extreme events. These algorithms are used to produce fast predictions in the context of extremes for high dimensional data, moreover they are able to capture much more complex structures in the data. In this work, statistical learning would be used in this context.

Two approaches are mainly used in extreme value theory: the Peaks-Over-Threshold (POT) approach and the block maxima (BM) approach. Existing methods combining extreme value theory and machine learning for modeling extreme quantile regression have used the POT approach. In contrast, the BM method is more suitable for modeling extremes in several domains, for example in hydrology and meteorology (Coles

L. M. Vidagbandji · A. Berred

Le Havre Normandy University, LMAH, Le Havre, France,

Tel.: +33773446863

E-mail: mahutin-lucien.vidagbandji@univ-lehavre.fr alexandre.berred@univ-lehavre.fr

C.Bertelle \cdot L. Amanton

Le Havre Normandy University, LITIS, Le Havre, France,

E-mail: laurent.amanton@univ-lehavre.fr cyrille.bertelle@univ-lehavre.fr

et al. (2001)[2]) . Dombry (2013)[3] justified the use of the maximum likelihood method for the BM approach, and comparative studies of the two approaches are performed by Ana Ferreira and Dombry (2019)[4] and by Bücher *et al.* (2018)[1]. In this work we will propose a conditional extreme quantile regression model by combining the BM approach of extreme value theory and machine learning methods. According to the block maxima approach, we approximate the conditional extreme distribution by the generalized extreme value distribution (GEV) whose parameters depend on the features of the model. These parameters will be estimated using statistical learning methods.

2 Quantile regression

In classical regression, for example linear regression, we are looking for a relationship of the type $Y = f(X) + \varepsilon$ between a variable Y called response variable depending on certain characteristics $X \in \mathbb{R}^p$ and the goal of being able to estimate the value of y for a given x again. Classical regression predicting the conditional mean E[Y|X = x]while the mean summarizes the behavior of Y|X = x in the center of its distribution. Quantile regression provides information on this and allows a richer description since it is interested in all the conditional distributions of the variable of interest and not only in its average. It is shown that this model allows better estimation even in the presence of extreme values.

The conditional quantile function of order τ is defined by:

$$\mathscr{Q}_{\tau}(y|X) = \inf\{y : F_{Y|X}(y) \ge \tau\}$$

And the quantile regression model (Koenker (1978)[6]) is given by: $Y = \mathcal{Q}_{\tau}(y|X) + \varepsilon$ with:

$$\mathcal{Q}_{\tau}(y|X=x) = \arg\min_{q\in\mathbb{R}} E(\rho_{\tau}(Y-q)|X=x)$$

and $\rho_{\tau}(c) = c(\tau - \mathbb{I}_{c < 0})$

The quantile regression is based on the estimation of:

$$\mathscr{Q}_{\tau}(y|X=x) = F_{Y|X=x}^{-1}(\tau) \tag{1}$$

for all $x \in \mathbb{R}^p$.

3 Quantile regression using the GEV approach

Several parametric, non parametric and machine learning based models are proposed for estimating this quantity, but these methods encounter difficulty mainly when the structure between Y and X is complex or the dimension of the dependent variable X is large. To circumvent these problems and perform modern applications with complex data, machine learning methods are useful because of their modeling flexibility and robustness in higher dimensions. Several methods have been proposed in the last few years using the combination of machine learning methods and extreme value theory following the POT approach, we can cite the works of: Velthoen *et al.* (2021)[8] using the gradian boosting method, Gnecco *et al.* (2022) [5] using the random forest and Pasche *et al.* (2022)[7] using neural networks.

Since we are interested in extreme quantiles, we will use the block maxima method by approximating the conditional distribution Y|X = x of the equation 1 by a generalized extreme value distribution (GEV). We will thus have a GEV distribution whose parameters depend on the characteristics *x*.

Setting $\Theta(x) = (\varepsilon(x), \sigma(x), \mu(x))$, this distribution is given by:

$$G(x; \boldsymbol{\Theta}(x)) = \exp\left(-\left(1 + \varepsilon(x)\frac{x - \mu(x)}{\sigma(x)}\right)_{+}^{-\frac{1}{\varepsilon(x)}}\right) \text{ if } \varepsilon(x) \neq 0$$

The conditional quantile of GEV is given for τ close to 1 by:

$$\mathscr{Q}_{\tau}(y|X=x) = \mu(x) + \frac{\sigma(x)}{\varepsilon(x)} \left(\left(\ln(\frac{1}{\tau}) \right)^{-\varepsilon(x)} - 1 \right) \text{ if } \varepsilon(x) \neq 0$$
 (2)

and

$$\mathscr{Q}_{\tau}(y|X=x) = \mu(x) + \sigma(x)\ln\left(-\ln\left(\tau\right)\right) \text{ if } \varepsilon(x) = 0 \tag{3}$$

The estimation of the conditional quantile, therefore, amounts to an estimation of $\Theta(x)$. Referring to the works of Gnecco *et al.* (2022)[5], Pasche *et al.* (2022)[7] and Velthoen *et al.* (2021) [8], the order from which the quantile would be extreme would be considered as $\tau_0 = 0.8$ and so we consider $\tau \in [\tau_0, 1[$ in the expressions (2) and (3).

We will estimate the GEV parameter vector by $\hat{\Theta}(x) = (\hat{\varepsilon}(x), \hat{\sigma}(x), \hat{\mu}(x))$ using the maximum likelihood method. This choice in the BM approach is based on the work of Dombry (2013)[3], Ana Ferreira and Dombry (2019)[4] and Bücher *et al.* (2018)[1]. We have:

$$\hat{\Theta}(x) = \arg\max_{\varepsilon,\sigma,\gamma} L_n(x;\varepsilon,\sigma,\gamma) \tag{4}$$

where $L_n(x; \varepsilon, \sigma, \gamma)$ is the likelihood associated to the sample of maximums per block. The structure between Y and X being complex and the size of the features is large in our context, so the resolution of (4) will be performed using statistical learning algorithms and then the conditional quantile estimate is given by:

$$\hat{\mathscr{Q}}_{\tau}(y|X=x) = \hat{\mu}(x) + \frac{\hat{\sigma}(x)}{\hat{\varepsilon}(x)} \left(\left(\ln(\frac{1}{\tau}) \right)^{-\hat{\varepsilon}(x)} - 1 \right) \text{ if } \hat{\varepsilon}(x) \neq 0$$
(5)

and

$$\hat{\mathscr{Q}}_{\tau}(y|X=x) = \hat{\mu}(x) + \hat{\sigma}(x)\ln\left(-\ln\left(\tau\right)\right) \text{ if } \hat{\varepsilon}(x) = 0 \tag{6}$$

This contribution questions the way to approach conditional quantile estimates characterising extreme events. Our work is part of a recent approach that links extreme value theory and machine learning. We do not yet address any application to real data sets. We are exploring an application concerning the impact of climate change on port infrastructures. This topic requires careful handling of the data available to date, which is a work in progress.

- 1. Axel Bücher et Chen Zhou, A Horse Race between the Block Maxima Method and the Peak–over–Threshold Approach, Statistical Science, 36(3) (2021)
- 2. Stuart Coles, An Introduction to Statistical Modeling of Extreme Values. Springer, London (2001)
- 3. Clément Dombry, Maximum likelihood estimators for the extreme value index based on the block maxima method, arXiv preprint arXiv:1301.5611, (2013)
- 4. Clément Dombry et Ana Ferreira, Maximum likelihood estimators based on the block maxima method, The annals of Statistics (2019).
- 5. Nicola Gnecco, Edossa Merga Terefe et Sebastian Engelke, Extremal Random Forests, arXiv:2201.12865 [stat], (2022)
- 6. Roger Koenker et Gilbert Bassett, Regression Quantiles, Econometrica, 46(1):33 (1978)
- 7. Olivier C. Pasche et Sebastian Engelke, Neural Networks for Extreme Quantile Regression with an Application to Forecasting of Flood Risk, arXiv preprint arXiv:2208.07590, (2022).
- 8. Jasper Velthoen, Clément Dombry, Juan-Juan Cai et Sebastian Engelke, Gradient boosting for extreme quantile regression, arXiv preprint arXiv:2103.00808, (2021)



DyHANE: Dynamic Heterogeneous Attributed Network Embedding

Liliana Martirano · Roberto Interdonato · Dino Ienco · Andrea Tagarelli

1 Introduction

As real world scenarios are inherently dynamic, graph continual learning is a machine learning paradigm gaining increasing popularity in recent years. Furthermore, most applications involve a multiplicity of entities and relationships with associated attributes, which should be captured and exploited effectively. This work considers the open problem of learning representations of changed nodes, i.e., new or updated nodes at current timestamp, without re-training the model from scratch or negatively affecting the learned representations of existing nodes at the previous timestamps.

As noted in [1], methods based on Graph Neural Networks (GNNs) are able to provide more refined graph representations, higher flexibility in leveraging attributes and generalization to unseen nodes, although they might suffer from more restrictive computational requirements with consequent impact on scalability w.r.t. shallow approaches. In dynamic scenarios, assuming a complete graph representation before the training process begins, is not applicable; pretrained models exploiting the inherent induction capability of GNNs fail in integrating new knowledge; also, retraining the model only on the changed nodes, although using the parameters learned at the previous timestamp to initialize, does not guarantee the preservation of previous knowledge. Several GNN-based works store all the past history of nodes and apply an additional recurrent architecture or an attention mechanism to update the node repre-

Università della Calabria, Rende, Italy E-mail: liliana.martirano@dimes.unical.it

Roberto Interdonato

Dino Ienco INRAE, UMR TETIS, Montpellier, France E-mail: dino.ienco@inrae.fr

Andrea Tagarelli Università della Calabria, Rende, Italy E-mail: andrea.tagarelli@unical.it

Liliana Martirano

CIRAD, UMR TETIS, Montpellier, France E-mail: roberto.interdonato@cirad.fr

sentations [2–5]. Literature lacks continual learning approaches for feature-rich heterogeneous networks capable of harnessing the expressive power of GNNs.

To address the above problem, we propose a novel framework, namely Dy-HANE (Dynamic Heterogeneous Attributed Network Embedding) framework, whose key idea is to learn at each new timestamp t an up-to-date node representation by incrementally updating model parameters, based on nodes of both the current and the previous timestamps, avoiding the need to store the history of nodes. To this aim, we propose a novel strategy to detect and integrate new knowledge, i.e., the changed nodes, and combine rehearsal and regularization approaches by employing experience nodes replay, i.e., nodes of previous timestamps stored in memory as experience and replayed at current timestamp, and model regularization for existing knowledge consolidation. DyHANE is designed for networks that may be dynamic, heterogeneous and attributed at the same time. We take into account node/edge addition and removal on feature-rich heterogeneous networks, i.e., networks showing multiple types of nodes and/or edges, and having external content associated to nodes.

2 Proposed framework

We define a dynamic heterogeneous attributed graph at a generic timestamp t as $G^t = \langle \mathcal{V}^t, \mathcal{E}^t, A, R, \phi, \varphi, \mathcal{X}^t \rangle$, where \mathcal{V}^t and \mathcal{E}^t are the sets of nodes and edges at timestamp t, A and R are the (fixed) sets of node and relation types, with |A| + |R| > 2, $\phi : \mathcal{V}^t \to A$ and $\varphi : \mathcal{E}^t \to R$ are the node- and edge-type mapping functions, and \mathcal{X}^t is the matrix storing node attributes at time t. We define the network evolution over time as a set of events at each timestamp t, corresponding to the set of changed edges $\mathcal{E}_c^{-t} = \bigcup_{i,j} e_{ij}$, with $e_{ij} = (i, x_i, j, x_j, r, s)$. In our formulation, s = 1 (resp. s = 0) denotes the addition (resp. removal) of edge of type $\varphi(e) = r$ between v_i and v_j ; x_i and x_j denote the attribute vectors of v_i and v_j . New nodes are initialized with their corresponding attribute vector, while nodes already existing in the network can be updated. The attribute vector is null if there are no changes. To handle attribute changes of isolated nodes, i.e. nodes not involved in any addition or removal, we define the corresponding event as a particular self loop.

Assuming a discrete set of timestamps $\{t_1, t_2, ..., t_T\}$, each of which corresponding to a set of events and consequent changes in graph data, our goal is to incrementally learn $\{\theta^{t_1}, \theta^{t_2}, ..., \theta^{t_T}\}$ where θ^t are the GNN parameters at timestamp t able to generate good representations $z_i^t \forall v_i \in G^t$.

The core of DyHANE consists of three main steps reiterated at each new timestamp t (Algorithm 1): (i) identification of the minimum set of nodes affected by changes, hereinafter referred to as influenced nodes \mathcal{I}^t , (ii) update of the GNN parameters θ^t using both the influenced nodes \mathcal{I}^t and the replay nodes stored in the memory buffer, hereinafter referred to as memory buffer or experience buffer \mathcal{B}^{t-1} , (iii) update of the memory buffer \mathcal{B}^t from the influenced node set \mathcal{I}^t . We point out that the proposed algorithm is relatively flexible, as it is not constrained to a single GNN architecture or to a unique im-

Algorithm 1 DyHANE - updating

- **Require:** Set of events, i.e., changed edges \mathcal{E}_c^t at current timestamp t, experience node buffer \mathcal{B}^{t-1} stored at previous timestamp t-1, GNN parametrized by θ^{t-1} learned at previous timestamp t-1, number of epochs num_epochs.
- **Ensure:** GNN parametrized by θ^t learned at current timestamp t, updated experience buffer \mathcal{B}^t .
- 1: Obtain influenced node set \mathcal{I}^t from $\mathcal{E}_c{}^t$
- 2: Load the experience buffer \mathcal{B}^{t-1}
- 3: Compute parameter importance by estimating F Fisher Information matrix
- 4: i = 0
- 5: while i $< num_epochs$ do

6: Calculate loss function $\mathcal{L}_{tot} = \gamma \mathcal{L}_{new} + (1 - \gamma) \mathcal{L}_{ex}$

- 7: Update parameters using SGD
- 8: i = i + 1
- $9:~\mathbf{end}~\mathbf{while}$
- 10: Update the experience buffer \mathcal{B}^t using the influenced node set \mathcal{I}^t
- 11: return θ^t

plementation of individual steps. While proposing tailored strategies, we also suggest possible alternatives to be investigated according to specific needs. In the following, we will describe in detail the steps of the proposed approach.

Identification of influenced nodes. To identify the minimum set of influenced nodes \mathcal{I}^t , we classify the new events in the graph (i.e., added or removed edges) into strong and weak events, according to their impact on the network topology. Inspired by a recent work on graph representation learning for dynamic homogeneous networks [6] and motivated by the well-established superiority of using meta-path based models to capture heterogeneous information even in large networks [7,8], we mark an event as strong or weak according to the impact of the corresponding edge on the meta-path based adjacency matrix of target type, i.e., the node type targeted for a task at hand. A meta-path is a path of the form $a_1 \xrightarrow{r_1} a_2 \xrightarrow{r_2} \dots \xrightarrow{r_x} a_{x+1}$ describing a composite relation between two nodes of the same type, i.e., $\phi(a_1) = \phi(a_{x+1})$. We thus define for the selected target node type a unique (homogeneous) meta-path based graph obtained as union of all meta-path instances with terminal nodes of that type, by removing the intermediate nodes and establishing a link between the terminal nodes weighted on the number of meta-path instances connecting them. The new event is said *strong* if it adds a new entry to the corresponding metapath based adjacency matrix, and the set of influenced nodes will include, in addition to the terminal nodes of the edge, also their one-hop neighborhood and meta-path based neighborhood; otherwise, it is said *weak* event. Notice that other strategies relying on the graph structure, such as triadic closure or open processes proposed by [9] for homogeneous networks and extended by [10] to multiple node/edge types, can be employed for differentiating between strong and weak events.

Update of the memory buffer. To update the memory buffer \mathcal{B}^t , we borrow from [11] the idea of selecting the most representative nodes for each class among the influenced node set, which has been proven to improve the effective-

ness and stability of data replaying, i.e., the experience of previous timestamps used at the current one. Since the influenced node set is expected to be unbalanced, the idea is to store in the buffer the same amount of information for each node class, ensuring stable distribution of memory categories. For nodes of target type, both strategies of leveraging node attributes (mean of features and coverage maximization) and of estimating node influence (influence maximization) proved to be effective in the homogeneous case [11]. Dealing with heterogeneous networks, we further propose an extension to handle nodes of types different from the target one. Specifically, we select for each class the nodes that appear most frequently as intermediate nodes in meta-paths instances with terminal nodes of that class. We also foresee a further balancing to ensure the presence of all types of nodes in memory, giving more importance to the nodes of target type, and representing the other types proportionally to their presence in the influenced node set. We plan to test also the opposite strategy of selecting nodes with attributes different from their neighborhood and storing in memory the nodes located at the class boundary, assuming they contribute more to the gradient [12]. We also do not exclude the investigation of further strategies independent of the final classification task.

Update of GNN parameters. Regardless of the specific strategies employed, we incrementally train the new model parameters on both nodes updated at the current timestamp and nodes updated at previous timestamp, i.e., the training set at timestamp t is $\mathcal{I}^t \bigcup \mathcal{B}^{t-1}$. We want to weigh the two contributions similarly to [11] and extend the weight factor to the heterogeneous case taking into account all the node types $a \in A$: $\gamma^a = \frac{|\mathcal{B}^a|}{|\mathcal{I}^a| + |\mathcal{B}^a|}$. Note that the number of nodes in \mathcal{I} , varying at each timestamp, is usually significantly larger than the size of the experience buffer, which is instead of fixed size. To solve the overfitting problem caused by the small number of replayed nodes, we introduce a weighted regularization method for model parameters based on Fisher Information matrix, that can be approximated based only on nodes stored in memory [12]. We aim to guarantee that the distance between the current and the historical model parameters will not deviate further, giving different importance to different parameters to keep small the changes of GNN parameters that are important to the past network while the others can be updated more drastically. For this purpose, the Elastic Weight Consolidation (EWC) regularization-based method [13] has shown to be effective.

Summary We have proposed DyHANE, a continual learning framework able to generate up-to-date node representations for dynamic, heterogeneous and attributed network. The proposed framework incrementally updates GNN parameters integrating new knowledge while consolidating the existing one.

We selected some real-world scenarios for the experimental evaluation of the proposed framework, including citation networks, and we are currently developing DyHANE by instantiating the various strategies specifically for a node classification task on a closed-world setting, i.e., with fixed and known number of node classes, such as authors' research fields.

- 1. S. Khoshraftar, A. An, CoRR **abs/2204.01855** (2022). DOI 10.48550/arXiv.2204.01855. URL https://doi.org/10.48550/arXiv.2204.01855
- A. Sankar, Y. Wu, L. Gou, W. Zhang, H. Yang, in <u>WSDM</u> '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February <u>3-7, 2020</u>, ed. by J. Caverlee, X.B. Hu, M. Lalmas, W. Wang (ACM, 2020), pp. 519–527. DOI 10.1145/3336191.3371845. URL https://doi.org/10.1145/3336191.3371845
- 3. Y. Ma, Z. Guo, Z. Ren, J. Tang, D. Yin, in Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, ed. by J.X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, Y. Liu (ACM, 2020), pp. 719–728. DOI 10.1145/3397271.3401092. URL https://doi.org/10.1145/3397271.3401092
- 4. L. Yang, Z. Xiao, W. Jiang, Y. Wei, Y. Hu, H. Wang, in <u>Advances in Information</u> <u>Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal,</u> <u>April 14-17, 2020, Proceedings, Part II, Lecture Notes in Computer Science, vol. 12036,</u> ed. by J.M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M.J. Silva, F. Martins (Springer, 2020), <u>Lecture Notes in Computer Science, vol. 12036, pp. 425–432.</u> DOI 10.1007/978-3-030-45442-5_53. URL https://doi.org/10.1007/978-3-030-45442-5_53
- H. Xue, L. Yang, W. Jiang, Y. Wei, Y. Hu, Y. Lin, CoRR abs/2004.01024 (2020). URL https://arxiv.org/abs/2004.01024
- 6. R. Trivedi, M. Farajtabar, P. Biswal, H. Zha, in <u>7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019</u> (Open-Review.net, 2019). URL https://openreview.net/forum?id=HyePrhR5KX
- 7. Y. Dong, N.V. Chawla, A. Swami, in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017 (ACM, 2017), pp. 135–144. DOI 10.1145/3097983.3098036. URL https://doi.org/10.1145/3097983.3098036
- J. Shang, M. Qu, J. Liu, L.M. Kaplan, J. Han, J. Peng, CoRR abs/1610.09769 (2016). URL http://arxiv.org/abs/1610.09769
- 9. L. Zhou, Y. Yang, X. Ren, F. Wu, Y. Zhuang, in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February <u>2-7, 2018</u>, ed. by S.A. McIlraith, K.Q. Weinberger (AAAI Press, 2018), pp. 571–578. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16572
- R. Bian, Y.S. Koh, G. Dobbie, A. Divoli, in <u>Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, ed. by B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, F. Scholer (ACM, 2019), pp. 861–864. DOI 10.1145/3331184.3331273. URL https://doi.org/10.1145/3331184.3331273
 </u>
- 11. F. Zhou, C. Cao, in <u>Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021</u>, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, <u>EAAI 2021</u>, Virtual Event, February 2-9, 2021 (AAAI Press, 2021), pp. 4714–4722. URL https://ojs.aaai.org/index.php/AAAI/article/view/16602
- 12. J. Wang, G. Song, Y. Wu, L. Wang, CoRR **abs/2009.10951** (2020). URL https://arxiv.org/abs/2009.10951
- J. Kirkpatrick, R. Pascanu, N.C. Rabinowitz, J. Veness, G. Desjardins, A.A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell, CoRR abs/1612.00796 (2016). URL http://arxiv.org/abs/1612.00796

Index by Author

Abbas Haidar, 46 Abdelkader Sbihi, 364 Adnan Yassine, 364 Adriana Iamnitchi, 17 Akrati Saxena, 409 Alexander Belyi, 298, 309 Alexandre Berred, 459 Alexandre Nicolas, 197 Ali Yassin, 32, 46 Amina Azaiez, 64 Ana Maria Jaramillo, 306 Andrea Russo, 219 Andrea Tagarelli, 463 Anna Traveset, 289 Annick Vignes, 369 Anthony Perez, 215 Antoine Houssard, 60, 357 Antoine Huchet, 326 Antonio Picone, 219 Aurélie Charles, 28 Aymeric Henard, 119 B. Ambrosio, 131 Barbara Polo, 244 Benjamin Renoust, 42 Bruno Marnot, 244 Bruno Pinaud, 51, 223 Catherine Annen, 109 Cesar A. Hidalgo, 360 Cesar Ducruet, 244 Charlie Joyez, 373 Cherifi Hocine, 146 Christian Wolff, 302 Christophe Cruz, 206, 227, 249, 302Claude Duvallet, 438, 442, 446, 450 Clémence Magnien, 293 Cruz Christophe, 146 Cyrille Bertelle, 438, 442, 446, 450, 459Cyrine Chenaoui, 151

Céline Rozenblat, 20 Dafnis Batalle, 140 Damien Calais, 403 Damien Olivier, 266 Davide Coppes, 23 Devashish Khulbe, 309 Dino Ienco, 463 Emanuele Brugnoli, 144 Enrico Gerding, 285 Eric Fotsing, 412 Eric Sanlaville, 36 Erkinai Derkenbaeva, 322 Etienne Peillard, 119 Eveline van Leeuwen, 322 Fabrice Lecuyer, 334 Fatiha Najm, 136 Federico Pilati, 60 Flann Chambers, 249 Floriana Gargiulo, 354, 357 Frederic Gilletta, 438, 446 Frédéric Guinand, 240, 266, 274 Gaëtan Laziou, 256 Georgios Panaviotou, 223 Gert Jan Hofstede, 322 Giacomo Kahn, 28 Gilles Coppin, 119 Ginestra Bianconi, 15 Giovanna Di Marzo Serugendo, 249 Giulio Prevedello, 144 Guillaume Bouleux, 28 Guillaume Deffuant, 56, 89 Guillaume Hacques, 36 Guillaume Moinard, 318 Guy Melancon, 51 Hamida Seba, 32, 46 Hocine Cherifi, 32, 42, 46, 302, 338, 342Hussam Ghanem, 227

Isabel Donoso, 289 Jean-Franccois Moyen, 109 Jean-Loup Guillaume, 326 Jeonghwa Kang, 179 Jerome Besancenot, 438 Juan Luis Jiménez Laredo, 266 Juste Raimbault, 179 Jérémy Rivière, 119 Jérôme Besancenot, 446 Katherine Birch, 140 Kittichai Lavangnananda, 240 Laurent Amanton, 459 Liliana Martirano, 463 Loïs Naudin, 112 Luca Maria Aiello, 14 Lucas Lacasa, 289 Lucien M. Vidagbandji, 459 Ludovic Seifert, 36 M. Maama, 131 M.A. Aziz Alaoui, 136 M.A. Aziz-Alaoui, 131 Majda Lafhel, 42 Mar Cuevas-Blanco, 289 Marcellin Julius Antonio Nkenlifack, 412 Marco Tolotti, 347 Maria Tartari, 60 Mariana Macedo, 306, 360, 409 Marion Le Texier, 256 Markus Brede, 201, 285 Markus Schaffert, 302 Martina Galletti, 144 Martina Reif, 51 Marwa Samrout, 364 Massinissa Atmani, 227 Matteo Magnani, 223 Matthieu Latapy, 293, 318 Maxence Lambard, 442 Md Muhidul Islam Khan, 429 Mehdi Naima, 293 Melanie Oyarzun, 360 Michel Dubois, 354 Mohammed El Hassouni, 42, 342 Mongetro Goint, 450

Muhammad Arslan, 206 Nathalie Verdière, 112 Nicolas Dugué, 215 Nicolas Marilleau, 151 Olivier Bonin, 260 Olivier Togni, 32, 46 Ondrej Mikes, 298, 309 Paola Tubaro, 313, 354 Paolo Cermelli, 23 Paulin Héleine, 266 Perrette Benjamin, 146 Pierluigi Sacco, 60 Pietro Gravino, 144 Quentin Bourgeais, 36 Quentin Rossy, 51 Radouane Yafia, 136 Rebecca Hoyle, 201 Remy Cazabet, 109 Riccardo Gallotti, 60 Rim Abdallah, 438, 446 Roberto Interdonato, 463 Roberto Weinberg, 109 Robin Salot, 64 Rodolphe Charrier, 36 Roman Bauer, 116, 140 Ronaldo Menezes, 306 Rosario N. Mantegna, 18 Rémi Lemoy, 256 S.M. Mintchev, 131 Sanaa Hmaida, 342 Sandra Hervías-Parejo, 289 Sarah Alahmadi, 201 Sarah J Berkemer, 313 Sasha Piccione, 347 Severin Vianey Kakeu Tuekam, 412 Simon Guillot, 215 Simon Mendez, 197 Slimane Ben Miled, 151 Solmaria Halleck Vega, 322 Stanislav Sobolevsky, 298, 309, 330 Stephany Rajeh, 338 Supharoek Chattanachott, 240 Sylvain Fontaine, 354

Sylvain Mignot, 369 Sébastien Kubicki, 119 Sébastien Orange, 112

Thibault Prouteau, 215 Thomas Souvignet, 51 Tommaso Venturini, 357 Tsuyoshi Murata, 455

Umar Abubacar, 116

Victor Martínez Eguiluz, 289

Vincent Bridonneau, 274 Vincenzo Miracula, 219 Vittorio Loreto, 144 Víctor M. Eguíluz, 16

Yacine Ghamri-Doudane, 326 Yerali Gandica, 56 Yoann Pigné, 274 Yuri Bogomolov, 298

Zhongqi Cai, 285

