COMPLEX NETWORKS 2020

THE 9TH INTERNATIONAL CONFERENCE ON COMPLEX NETWORKS AND THEIR APPLICATIONS

December 1 - 3, 2020

Online

BOOK OF ABSTRACTS

COMPLEX NETWORKS 2020
The 9th International Conference on Complex Networks & Their Applications
December 1 - 3, 2020 Madrid, Spain - Online
Published by the International Conference on Complex Networks & Their Applications

Editors

Rosa María Benito
Universidad Politécnica de Madrid, Spain

Hocine Cherifi
University of Burgundy, France

Chantal Cherifi
University of Lyon, France

Esteban Moro
Universidad Carlos III, Spain

Luis Mateus Rocha
Indiana University, USA

Marta Sales-Pardo
Universitat Rovira i Virgili, Spain

COMPLEX NETWORKS 2020
e-mail: hocine.cherifi@u-bourgogne.fr

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use. While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained her.

ISBN: 978-2-9557050-4-9

Preface

This 2020 edition of the International Conference on Complex Networks & their Applications is the ninth of a series that began in 2011. Over the years, this adventure has made the conference one of the major international events in network science.

Network science continues to trigger a tremendous interest among the scientific community of various fields such as Finance and Economy, Medicine and Neuroscience, Biology and Earth Sciences, Sociology and Politics, Computer Science and Physics. The variety of scientific topics ranges from Network Theory, Network Models, Network Geometry, Community Structure, Network Analysis and Measure, Link Analysis and Ranking, Resilience and Control, Machine Learning and Networks, Dynamics on/of Networks, Diffusion and Epidemics, Visualization. It is also worth mentioning some recent applications with high added value for current trend social concerns such as Social and Urban Networks, Human Behavior, Urban Systems - Mobility, or Quantifying Success. The conference brings together researchers that study the world through the lens of networks. Catalyzing the efforts of this scientific community, it drives network science to generate cross-fertilization between fundamental issues and innovative applications, review the current state of the field, and promote future research directions.

Every year, researchers from all over the world gather in our host venue. This year's edition was initially to be hosted in Spain by Universidad Politécnica de Madrid. Unfortunately, the COVID-19 global health crisis forced us to organize the conference as a fully online event.

Undoubtedly, the success of this edition relied on the authors who have produced high quality papers, as well as the impressive list of keynote speakers who delivered fascinating plenary lectures:

– Leman Akoglu (Carnegie Mellon University, USA): "Graph-Based Anomaly Detection: Problems, Algorithms and Applications"
– Stefano Boccaletti (Florence University, Italy): "Synchronization in Complex Networks, Hypergraphs and Simplicial Complexes"
– Fosca Giannotti (KDD Lab, Pisa, Italy): "Explainable Machine Learning for Trustworthy AI"
– János Kertesz (Central European University, Hungary): "Possibilities and Limitations of using mobile phone data in exploring human behavior"
– Vito Latora (Queen Mary, University of London, UK): "Contagion and synchronization in systems with higher-order interactions"
– Alex 'Sandy' Pentland (MIT Media Lab, USA): "Human and Optimal Networked Decision Making in Long-Tailed and Non-stationary Environments"
– Nataša Pržulj (Barcelona Supercomputing Center, Spain): "Untangling biological complexity: From omics network data to new biomedical knowledge and Data-Integrated Medicine"

The topics addressed in the keynote talks allowed a broad coverage of the issues encountered in complex networks and their applications to complex systems.

For the traditional tutorial sessions prior to the conference, our two invited speakers delivered insightful talks. David Garcia (Complexity Science Hub Vienna, Austria) gave a lecture entitled "Analyzing complex social phenomena through social media data", and Mikko Kivela (Aalto University, Finland) delivered a talk on "Multilayer Networks". Each edition of the conference represents a challenge that cannot be successfully achieved without the deep involvement of many people, institutions and sponsors.

First of all, we sincerely gratify our advisory board members, Jon Crowcroft (University of Cambridge), Raissa D'Souza (University of California, Davis, USA), Eugene Stanley (Boston University, USA) and Ben Y. Zhao (University of Chicago, USA), for inspiring the essence of the conference.

We record our thanks to our fellow members of the Organizing Committee. José Fernando Mendes (University of Aveiro, Portugal), Jesús Gomez Gardeñes (University of Zaragoza, Spain) and Huijuan Wang (TU Delft, Netherlands) chaired the Lightning sessions. Manuel Marques Pita (Universidade Lusófona, Portugal), José Javier Ramasco (IFISC, Spain) and Taha Yasseri (University of Oxford, UK) managed the Poster sessions. Luca Maria Aiello (Nokia-Bell Labs, UK) and Leto Peel (Université Catholique de Louvain, Belgium) were our Tutorial chairs. Finally, Sabrina Gaito (University of Milan, Italy) and Javier Galeano (Universidad Politécnica de Madrid, Spain), were our Satellite chairs.

We extend our thanks to Benjamin Renoust (Osaka University, Japan), Michael Schaub (MIT, USA), Andreia Sofia Teixeira (Indiana University, USA), Xiangjie Kong (Dalian University of Technology, China), the Publicity chairs for advertising the conference in America, Asia and Europa, hence encouraging the participation.

We would like also to acknowledge Regino Criado (Universidad Rey Juan Carlos, Spain) as well as Roberto Interdonato (CIRAD - UMR TETIS, Montpellier, France) our Sponsor chairs.

Our deep thanks go to Matteo Zignani (University of Milan, Italy), Publication chair, for the tremendous work he has done at managing the Submission system and the Proceedings publication process.

Thanks to Stephany Rajeh (University of Burgundy, France), Web chair, in maintaining the Website.

We would also like to record our appreciation for the work of the Local Committee chair, Juan Carlos Losada (Universidad Politécnica de Madrid, Spain), and all the Local Committee members, David Camacho (UPM, Spain), Fabio Revuelta (UPM, Spain), Juan Manuel Pastor (UPM, Spain), Francisco Prieto (UPM, Spain), Leticia Perez Sienes (UPM, Spain), Jacobo Aguirre (CSIC, Spain), Julia Martinez-Atienza (UPM, Spain), for their work in managing online sessions. They greatly participated to the success of this edition.

We are also indebted to our partners, Alessandro Fellegara and Alessandro Egro from Tribe Communication, for their passion and patience in designing the visual identity of the conference.

We would like to express our gratitude to our partner journals involved in the sponsoring of keynote talks: Applied Network Science, EPJ Data Science, Social Network Analysis and Mining, and Entropy.

Generally, we are thankful to all those who have helped us contributing to the success of this meeting. Sincere thanks to the contributors, the success of the technical program would not be possible without their creativity.

Finally, we would like to express our most sincere thanks to the Program Committee members for their huge efforts in producing high-quality reviews in a very limited time.

Rosa M. Benito
Hocine Cherifi
Chantal Cherifi
Esteban Moro
Luis Mateus Rocha
Marta Sales-Pardo

Table of Contents

Tutorials

Invited Speakers

I Biological Networks

II Community Structure

III Diffusion and Epidemics

IV Dynamics on/of Networks

V Ecological Networks and Food Webs

VI Link Analysis and Ranking

VII Machine Learning and Networks

VIII Modeling Human Behavior

IX Network Analysis

X Network Models

XI Networks in Finance and Economics

XII Social Networks

XIII Urban Systems and Networks

Tutorials

Analyzing complex social phenomena through social media data

David Garcia

TU Graz, Austria

The wealth of data generated by our digital society, when combined with computational methods like agent-based modeling and natural language understanding, provides a new window to study human behavior at new scales and resolutions. This enables the analysis of complex social phenomena in which temporal dynamics and network structures require the use of large and detailed data. I will present an overview of complex social phenomena that have been analyzed through social media data, one of the most accessible and powerful data sources in our digital society. First, I will show how social media data can be used to analyze collective emotions and their long-term effects in terms of solidarity. Second, social media data can capture states of multidimensional polarization that can be explained by cognitive science models. Third, social media data can capture gender inequality across countries and illustrates the role of positive externalities, also known as network effects. And fourth, I will show how the presence of intelligent technologies in online platforms generate the phenomenon of complex privacy, by which the individual decision to share data with an online platform is affected by the decisions of others.

David Garcia has been a group leader at the Complexity Science Hub Vienna since September 2017. The aim is to build a research group funded by WWTF (Vienna Research Groups for Young Investigators Call). He holds computer science degrees from Universidad Autonoma de Madrid (Spain) and ETH Zurich (Switzerland). David did a PhD and Postdoc at ETH Zurich, working at the chair of systems design.

David's research focuses on computational social science, designing models and analysing human behaviour through digital traces. His main work revolves around the topics of emotions, cultures, and political polarization, combining statistical analyses of large datasets of online interaction with agent-based modeling of individual behaviour. David's work lies at the intersection of various scientific disciplines, combining methods from network science, computer science, and statistical physics to answer questions from psychology, economics, and political science. His interdisciplinary collaborations span more than 50 co-authors in 12 countries. David has published more than 20 journal articles, 15 conference papers, and five book chapters, and serves as reviewer for prestigious journals and as program committee member of numerous computer science conferences.

Multilayer Networks

Mikko Kivela

Aalto University, Finland

Network science has been very successful in investigations of a wide variety of applications from biology and the social sciences to physics, technology, and more. In many situations, it is already insightful to use a simple (and typically naive) representation as a simple, binary graph in which nodes are entities and unweighted edges encapsulate the interactions between those entities. This allows one to use the powerful methods and concepts for example from graph theory, and numerous advances have been made in this way. However, as network science has matured and (especially) as ever more complicated data has become available, it has become increasingly important to develop tools to analyse more complicated structures. For example, many systems that were typically initially studied as simple graphs are now often represented as time-dependent networks, networks with multiple types of connections, or interdependent networks. This has allowed deeper and more realistic analyses of complex networked systems, but it has simultaneously introduced mathematical constructions, jargon, and methodology that are specific to research in each type of system. The concept of « multilayer networks » was developed in order to unify the aforementioned disparate language (and disparate notation) and to bring together the different generalised network concepts that included layered graphical structures. In this tutorial talk, I will introduce multilayer networks and discuss how to study their structure. Generalisations of the clustering coefficient for multiplex networks and graph isomorphism for general multilayer networks are used as illustrative examples.

I am a network scientist and an assistant professor at the Aalto University, where I also obtained my doctoral degree. Before coming back to Aalto I was a postdoctoral scholar at the Mathematical Institute at the University of Oxford. My research area is the relatively new field of "network science" or "complex networks". This means that I'm interested in complex systems with a large number of elements that are interacting with each other in some non-trivial way and possibly leading to some emergent phenomena. Social systems are a good example: they consist of multiple elements (people) that are interacting with each other (social relationships) and lead to some very complex emergent behaviour (social groups, societies, conflicts, etc.). Other such complex systems include transportation systems, gene-regulatory systems in cells, ecological systems and many more. I see all of these systems as networks that can be studied with the similar sets of tools and theories.

Invited Speakers

Graph-Based Anomaly Detection : Problems, Algorithms and Applications

Leman Akoglu

Carnegie Mellon University, USA

Graphs provide a powerful abstraction for representing non-iid data, capturing immediate as well as long-range dependencies between entities. The study of the structure and dynamics of real-world graphs has been a central theme of research across various communities. Graph-based anomaly detection focuses broadly on identifying those "constructs" that do not "fit" the expected relational patterns.

This talk involves vignettes from my decade-long research on anomaly detection using graph-based techniques. I will introduce various scenarios in which graphs can be used in a natural way — both to formalize concrete anomaly detection problems, and to develop algorithmic anomaly detection methods. These will be motivated by real-world applications of anomaly detection in the wild; including opinion fraud, accounting anomalies, and host-level intrusion.

Leman Akoglu joined the Heinz College faculty as an Assistant Professor in Fall 2016. She also holds a courtesy appointment in the Computer Science Department (CSD) and the Machine Learning Department (MLD) of School of Computer Science (SCS). Akoglu is the Heinz College Dean's Associate Professor of Information Systems. Prior to joining Heinz College, she was an Assistant Professor in the Department of Computer Science at Stony Brook University since receiving her Ph.D. from CSD/SCS of Carnegie Mellon University in 2012.

Dr. Akoglu's research interests span a wide range of data mining and machine learning topics with a focus on algorithmic problems arising in graph mining, pattern discovery, social and information networks, and especially anomaly mining; outlier, fraud, and event detection. At Heinz, Dr. Akoglu directs the Data Analytics Techniques Algorithms (DATA) Lab. Dr. Akoglu's research has won 7 publication awards; Best Research Paper at SIAM SDM 2019, Best Student Machine Learning Paper Runner-up at ECML PKDD 2018, Best Paper Runner-up at SIAM SDM 2016, Best Research Paper at SIAM SDM 2015, Best Paper at ADC 2014, Best Paper at PAKDD 2010, and Best Knowledge Discovery Paper at ECML PKDD 2009. She also holds 3 U.S. patents filed by IBM T. J. Watson Research Labs.

Untangling biological complexity: From omics network data to new biomedical knowledge and Data-Integrated Medicine

Natasa Przulj

Barcelona Supercomputing Center, Spain

We are faced with a flood of molecular and clinical data. We are measuring interactions between various bio-molecules in a cell that form large, complex systems. Patient omics datasets are also increasingly becoming available. These systems-level network data provide heterogeneous, but complementary information about cells, tissues and diseases. The challenge is how to mine them collectively to answer fundamental biological and medical questions. This is non-trivial, because of computational intractability of many underlying problems on networks (also called graphs), necessitating the development of approximate algorithms (heuristic methods) for finding approximate solutions. We develop methods for extracting new biomedical knowledge from the wiring patterns of systems-level, heterogeneous biomedical networks. Our methods uncover the patterns in molecular networks and in the multi-scale network organization indicative of biological function, translating the information hidden in the network topology into domain-specific knowledge. We also introduce a versatile data fusion (integration) framework to address key challenges in precision medicine from biomedical network data: better stratification of patients, prediction of driver genes in cancer, and re-purposing of approved drugs to particular patients and patient groups, including Covid-19 patients. Our new methods stem from novel network science algorithms coupled with graph-regularized non-negative matrix tri-factorization, a machine learning technique for dimensionality reduction and co-clustering of heterogeneous datasets. We utilize our new framework to develop methodologies for performing other related tasks, including disease re-classification from modern, heterogeneous molecular level data, inferring new Gene Ontology relationships, aligning multiple molecular networks, and uncovering new cancer mechanisms.

Prof. Przulj initiated extraction of biomedical knowledge from the wiring patterns (topology, structure) of "Big Data" real-world molecular (omics) and other networks. That is, she views the wiring patterns of large and complex omics networks, disease ontologies, clinical patient data, drug-drug and drug-target interaction networks etc., as a new source of information that complements the genetic sequence data and needs to be mined and meaningfully integrated to gain deeper biomedical understanding. Her recent work includes designing machine learning methods for integration of heterogeneous biomedical and molecular data, applied to advancing biological and medical knowledge. She also applies her methods to economics.

She is a member of the Editorial Boards of Bioinformatics (Oxford Journals), Scientific Reports (Nature Publishing Group) and Frontiers in Genetics (Frontiers), and an Associate Editor of BMC Bioinformatics (BioMed Central). Prof. Przulj a member of the Scientific Advisory Board of the Helmholtz Centre for Infection Research (HZI / Braunschweig, Germany) and GSK. She is a Proceedings / Area Chair of Protein Interactions, Molecular Networks and Network Biology tracks at the ISMB/ECCB 2015, ISMB 2016 and ISMB/ECCB 2017, elected Chair of NetBio COSI (ISCB, ISMB) since 2019.

The keynote is sponsored by Entropy

Explainable Machine Learning for Trustworthy AI

Fosca Giannotti

University of Pisa, Italy

Black box AI systems for automated decision making, often based on machine learning over (big) data, map a user's features into a class or a score without exposing the reasons why. This is problematic not only for the lack of transparency, but also for possible biases inherited by the algorithms from human prejudices and collection artifacts hidden in the training data, which may lead to unfair or wrong decisions. The future of AI lies in enabling people to collaborate with machines to solve complex problems. Like any efficient collaboration, this requires good communication, trust, clarity and understanding. Explainable AI addresses such challenges and for years different AI communities have studied such topic, leading to different definitions, evaluation protocols, motivations, and results. This lecture provides a reasoned introduction to the work of Explainable AI (XAI) to date, and surveys the literature with a focus on machine learning and symbolic AI related approaches. We motivate the needs of XAI in real-world and large-scale application, while presenting state-of-the-art techniques and best practices, as well as discussing the many open challenges.

Fosca Giannotti is a director of research of computer science at the Information Science and Technology Institute "A. Faedo" of the National Research Council, Pisa, Italy. Fosca Giannotti is a pioneering scientist in mobility data mining, social network analysis and privacy-preserving data mining. Fosca leads the Pisa KDD Lab – Knowledge Discovery and Data Mining Laboratory, a joint research initiative of the University of Pisa and ISTI-CNR, founded in 1994 as one of the earliest research lab on data mining. Fosca's research focus is on social mining from big data: smart cities, human dynamics, social and economic networks, ethics and trust, diffusion of innovations.

Fosca has coordinated tens of European projects and industrial collaborations. She is currently the coordinator of SoBigData, the European research infrastructure on Big Data Analytics and Social Mining an ecosystem of ten cutting edge European research centres providing an open platform for interdisciplinary data science and data-driven innovation. Recently she is the PI of ERC Advanced Grant entitled XAI – Science and technology for the explanation of AI decision making. On March 8, 2019 she has been features as one of the 19 Inspiring women in AI, BigData, Data Science, Machine Learning by KDnuggets.com, the leading site on AI, Data Mining and Machine Learning.

This keynote is jointly sponsored by Applied Network Science, EPJ Data Science, and Social Network Analysis and Mining

Possibilities and Limitations of using mobile phone data in exploring human behavior

János Kertesz

Central European University, Hungary

Big Data as provided by modern communication systems provide unprecedented opportunities for research. Mobile phones have become almost like a new organ in additional to our biological ones and we practically never get rid of them, hence the analysis of CDR-s (Call Detail Records) are particularly important in gaining information about the whereabouts, contacts and activity patterns of people. We will review some of the results from such analyses, including large scale structure of the society, mobility patterns, gender and age dependence of interactions, bursty character of the activity. We will show that sometimes extremely precise information can be obtained and applied to support theories of social anthropology e.g., about family relationships. However, CDR data should be used with care, as bias could occur since information from one communication channel is considered only. We analyze this aspect and suggest a general description of such biases.

János Kertész obtained his PhD in Physics 1980 from Eötvös University. He worked at the Research Institute of Technical Physics of the Hungarian Academy of sciences, at the Cologne University and at Technical University Munich. He has been professor since 1992 at the Budapest University of Technology and Economics, and since 2012 at the Department of Network and Data Science of the Central European University. He was visiting scientist in Germany, US, France, Italy and Finland.

János Kertész is interested in statistical physics and its applications, including percolation theory, phase transitions, fractal growth, granular materials and simulation methods. During the last 15 years his research has focused on multidisciplinary topics, mainly on complex networks as well as on financial analysis and modeling. He has published more than 200 scientific papers. He has been on the editorial boards of Journal of Physics A, Physica A, Fluctuation and Noise Letters, Fractals, New Journal of Physics. His work has been awarded by several recognitions, including the Hungarian Academy Award, the Szent-Györgyi Award of the Ministry of Education and Culture, the Széchenyi Prize and the title of Finland Distinguished Professor.

This keynote is jointly sponsored by Applied Network Science, EPJ Data Science, and Social Network Analysis and Mining

Simplicial model of social contagion

Vito Latora

Queen Mary, University of London, UK

Complex networks have been successfully used to describe the spread of diseases in populations of interacting individuals. Conversely, pairwise interactions are often not enough to characterize social contagion processes such as opinion formation or the adoption of novelties, where complex mechanisms of influence and reinforcement are at work. I will first discuss a higher-order model of social contagion in which a social system is represented by a simplicial complex and contagion can occur through interactions in groups of different sizes. Numerical simulations of the model on both empirical and synthetic simplicial complexes highlight the emergence of novel phenomena such as a discontinuous transition induced by higher-order interactions. I will show analytically that the transition is discontinuous and that a bistable region appears where healthy and endemic states co-exist. This result can help explaining why critical masses are required to initiate social changes. I will then show how the presence of higher-order interaction can affect the stability of a synchronised state in a simplicial complex of coupled dynamical systems.

I am Professor of Applied Mathematics, Chair of Complex Systems, and Head of the Complex Systems and Networks Research Group in the School of Mathematical Sciences of QMUL. I am editor of the Journal of Complex Networks, Fellow of the Turing Institute, and External Faculty of the Complexity Hub Vienna. I study the structure and the dynamics of complex systems, using my background as theoretical physicist and some of the methods proper to statistical physics and non-linear dynamics, to look into biological problems, to model social systems, and to find new solutions for the design of man-made networks. I have coauthored more than 150 scientific publications, including papers in PRL, PNAS, Nature Comm, Science and Physics Reports. See the complete list of my publications here or from Google Scholar. My recent grants include: EU LASAGNE (2012-15), EPSRC GALE (2013-16) and EPSRC LoBaNet (2016-2019). I currently hold a Research Fellowships from the Leverhulme Trust to work on the network components of creativity and success.

The keynote is sponsored by Entropy

Human and Optimal Networked Decision Making in Long-Tailed and Non-stationary Environments

Alex 'Sandy' Pentland

MIT Media Lab, USA

Human social networks frequently give rise to long-tailed and non-stationary information spreading, but most methods of analysis and decision making typically assume stationary, concentrated distributions. Similarly, wisdom of the crowd phenomena are usually analyzed as a single trial with a fixed information sharing network whereas dynamic networks and importance of long-term repeated-trial performance are major feature of human societies. I will discuss new theoretical results on optimal tuning of information sharing networks while accounting for long-tailed distributions. Finally, I will show that these new theoretical results provide a good model for how humans tune their social networks for better performance in non-stationary and long-tailed environments.

Professor Alex "Sandy" Pentland directs MIT Connection Science, an MIT-wide initiative, and previously helped create and direct the MIT Media Lab and the Media Lab Asia in India. He is one of the most-cited computational scientists in the world, and Forbes recently declared him one of the "7 most powerful data scientists in the world" along with Google founders and the Chief Technical Officer of the United States. He is on the Board of the UN Foundations' Global Partnership for Sustainable Development Data, co-led the World Economic Forum discussion in Davos that led to the EU privacy regulation GDPR, and was central in forging the transparency and accountability mechanisms in the UN's Sustainable Development Goals. He has received numerous awards and prizes such as the McKinsey Award from Harvard Business Review, the 40th Anniversary of the Internet from DARPA, and the Brandeis Award for work in privacy.

He is a member of advisory boards for the UN Secretary General and the UN Foundation, and the American Bar Association, and previously for Google, AT&T, and Nissan. He is a serial entrepreneur who has co-founded more than a dozen companies. He is a member of the U.S. National Academy of Engineering and leader within the World Economic Forum.

Over the years Sandy has advised more than 70 PhD students. Together Sandy and his students have pioneered computational social science, organizational engineering, wearable computing (Google Glass), image understanding, and modern biometrics. His most recent books are Social Physics, published by Penguin Press, and Honest Signals, published by MIT Press.

Synchronization in Complex Networks, Hypergraphs and Simplicial Complexes

Stefano Boccaletti

Carnegie Mellon University, USA

All interesting and fascinating collective properties of a complex system arise from the intricate way in which its components interact. Various systems in physics, biology, social sciences and engineering have been successfully modelled as networks of coupled dynamical systems, where the graph links stand for pairwise interactions. This is, however, too strong a limitation, as recent studies have revealed that higher-order many-body interactions are present in social groups, ecosystems and in the human brain, and they actually affect the emergent dynamics of all these systems. I will discuss a general framework that allows to study coupled dynamical systems accounting for the precise microscopic structure of their interactions at any possible order. Namely, I will conider an ensemble of identical dynamical systems, organized on the nodes of a simplicial complex, and interacting through synchronization-non-invasive coupling function. The simplicial complex can be of any dimension, meaning that it can account, at the same time, for pairwise interactions (networks), three-body interactions and so on. In such a broad context, a recent collaboration of mine has shown that complete synchronization, a circumstance where all the dynamical units arrange their evolution in unison, exists as an invariant solution, and has given the necessary condition for it to be observed as a stable state in terms of a Master Stability Function. This generalizes the existing results valid for pairwise interactions (i.e. graphs) to the case of complex systems with the most general possible architecture. Moreover, we show how the approach can be simplified for specific, yet frequently occurring, instances, and we verify all our theoretical predictions in synthetic and real-world systems. Given the completely general character of the method proposed, our results contribute to the theory of dynamical systems with many-body interactions and can find applications in an extremely wide range of practical cases.

Stefano Boccaletti got his PhD in Physics at the University of Florence on 1995. In October 1998 he was awarded the individual EU grant “Marie Curie” n. ERBFMBICT983466. He is Senior Researcher at the CNR-Institute for Complex Systems, and Honorary Professor of the Weizmann Institute of Science, the Tel Aviv University, the University of Bar Ilan, the University of Navarre, and the Technical University of Madrid. In 2015, he was awarded the PhD honoris causa by the University Rey Juan Carlos of Madrid. Currently, he is the Scientific Attache’ at the Italian Embassy in Israel.

Stefano Boccaletti is Author of publications in Physics Journals, which have been cited more than 14,000 times, Editor of 4 books, and Author of other 3, Editor in Chief of the Elsevier Journal Chaos Solitons and Fractals, and member of the Editorial Board of several other International journals of physics and applied mathematics. He has been invited to about 85 International Conferences and Seminars as a plenary lecturer or keynote speaker, and he directly organized 15 Workshops.

Part I

Biological Networks

Multiomics-based inference of cell type-specific regulatory networks in early human embryos

Gregorio Alanis-Lobato[1], Thomas E. Bartlett[2], and Kathy K. Niakan[1]

[1] Human Embryo and Stem Cell Laboratory, The Francis Crick Institute, London, UK
[2] Department of Statistical Science, University College, London, UK
Contact: gregorio.alanis@crick.ac.uk

1 Introduction

Cells respond to environmental changes by activating gene expression programmes that allow them to maintain cellular homeostasis [1, 2]. In multicellular organisms, these programmes differ from cell type to cell type and are thus an important component of a cell's identity [3]. Therefore, mapping the molecular networks responsible for gene expression regulation is key to understand cell type specification, as well as homeostatic maintenance and failure [4].

Although next-generation sequencing has allowed for the development of high-throughput assays to measure genome-wide gene expression, chromatin accessibility and methylation levels [5], the determination of regulatory interactions with methods like ChIP-seq or CUT&RUN [6] is restricted to a few transcription factors (TFs) with good quality antibodies [7]. This has prompted the development of computational methods to infer regulatory relationships between TFs and their target genes based mainly on gene expression data [8].

As part of the fifth edition of the Dialogue on Reverse Engineering Assessment and Methods (DREAM) challenge, a comprehensive comparison of more than 30 network inference methods was performed using *in silico* and real gene expression datasets [8]. This benchmark showed that even the methods with the best overall performance predict a considerable number of false positive interactions. Most of these false predictions originate from indirect associations: a path $a \rightarrow b \rightarrow c$ can result in the prediction of $a \rightarrow c$ even if there is no direct link between those nodes.

Here, we assessed whether the integration of other types of omics datasets with transcriptomic-based predictions could help with the removal of indirect TF-gene relationships and produce more reliable regulatory networks. In particular, we used chromatin accessibility data to ensure that there were regions of open chromatin in the vicinity of target genes and that those regions were enriched for motifs that the potential regulators can bind. We did this using data from early human embryos at the blastocyst stage (Fig. 1a). This has the advantage of being a biological context with only three well-defined cell types [9], good quality omics data [9–12], sufficient domain knowledge to evaluate the plausibility of our predictions [9], as well as many open questions that can be addressed with insights from regulatory network models.

2 Methods

We integrated single-cell RNA sequencing (scRNA-seq) data from three different studies [9–11] focusing on the three cell types present at the late blastocyst stage (Fig. 1b). Alignment to the reference genome (GRCh38) and calculation of gene counts were performed on each dataset separately with nf-core/rnaseq v1.4.2 [13]. The resulting gene expression matrices were integrated and normalised using Bioconductor tools [14]. Chromatin accessibility profiles from the blastocyst stage were obtained with the LiCAT-seq technique [12]. Alignment to GRCh38, peak calling and annotation were performed with nf-core/atacseq v1.1.0 [13]. Then, we carried out TF motif enrichment analysis in the regions of open chromatin using rgt-hint v0.13.0 [15]. Finally, we associated the TFs that exhibited over-represented motifs in these regions with the closest promoters (Fig. 1c).

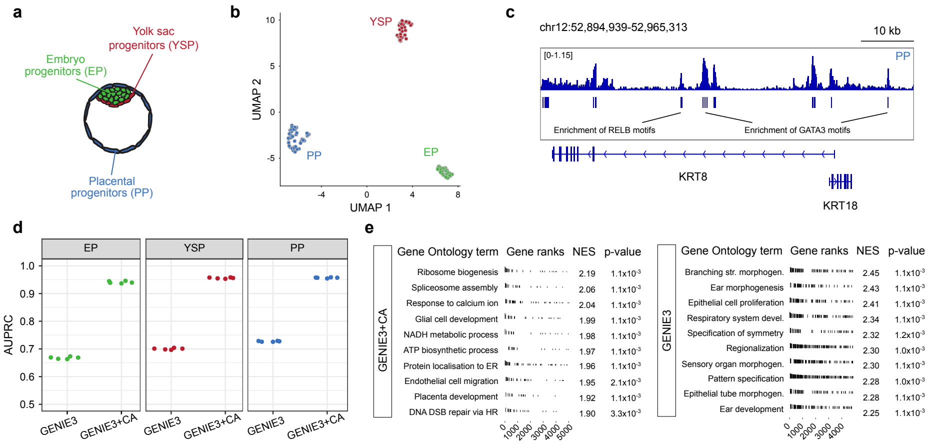
For network inference, we employed the best performing strategy in the DREAM5 challenge, GENIE3 [16], which is based on random forests. The putative regulatory links predicted by this method using the scRNA-seq data were taken as is (GENIE3) or were subjected to a filtering process in which only TF-gene associations supported by the chromatin accessibility data were considered in the final network (GENIE3+CA). We evaluated the performance of each approach using 5-fold cross-validation on each cell type to quantify the extent to which interactions inferred from a training set coincide with interactions inferred from a reference test set, using the area under the precision-recall curve (AUPRC) as our metric. Finally, we applied gene set enrichment analyses (GSEA) to the target genes from each network using fgsea v1.14.0 [17]. The rationale was that regulated genes should be significantly associated with the biological processes known to be active in each blastocyst cell type.

3 Results

Fig. 1d shows that GENIE3+CA outperforms GENIE3 in our *in-silico* benchmarks. This indicates that GENIE3 becomes a more robust predictor with the chromatin accessibility refinement, as it consistently prioritises regulatory links that were highly ranked in the reference test set by using information from the training set only. Moreover, when we assessed the biological relevance of the regulated genes from each network via GSEA, we found a better agreement between the GENIE3+CA predictions and their corresponding cell type. For example, the target genes in the network inferred by GENIE3+CA for the placental progenitor cells are enriched in biological processes associated with placenta development, whereas the target genes in the GENIE3 network are enriched in terms that are more indirectly associated with this tissue (Fig. 1e).

4 Conclusion

Our comparative analysis of regulatory network inferences highlights the value of refining expression-based predictions with complementary context-specific omics datasets. Our goals now are to evaluate the performance of other prediction strategies (e.g. regression or mutual information), study the structure and properties of the inferred regulatory



Gene Ontology term	Gene ranks	NES	p-value
Ribosome biogenesis		2.19	$1.1x10^{-3}$
Spliceosome assembly		2.06	$1.1x10^{-3}$
Response to calcium ion		2.04	$1.1x10^{-3}$
Glial cell development		1.99	$1.1x10^{-3}$
NADH metabolic process		1.98	$1.1x10^{-3}$
ATP biosynthetic process		1.97	$1.1x10^{-3}$
Protein localisation to ER		1.96	$1.1x10^{-3}$
Endothelial cell migration		1.95	$2.1x10^{-3}$
Placenta development		1.92	$1.1x10^{-3}$
DNA DSB repair via HR		1.90	$3.3x10^{-3}$

Gene Ontology term	Gene ranks	NES	p-value
Branching str. morphogen.		2.45	$1.1x10^{-3}$
Ear morphogenesis		2.43	$1.1x10^{-3}$
Epithelial cell proliferation		2.41	$1.1x10^{-3}$
Respiratory system devel.		2.34	$1.1x10^{-3}$
Specification of symmetry		2.32	$1.2x10^{-3}$
Regionalization		2.30	$1.0x10^{-3}$
Sensory organ morphogen.		2.30	$1.1x10^{-3}$
Pattern specification		2.28	$1.0x10^{-3}$
Epithelial tube morphogen.		2.28	$1.1x10^{-3}$
Ear development		2.25	$1.1x10^{-3}$

Fig. 1. **(a)** Human embryo at the blastocyst stage. **(b)** Low-dimensional representation of the scRNA-seq data. Samples cluster by cell type based on expressed genes. **(c)** Regions of open chromatin in the placental progenitor cells at the KRT8/KRT18 locus. Potential regulators of these placental-associated keratins are highlighted. **(d)** Performance evaluation of the GENIE3+CA and GENIE3 predictions with AUPRC. Each point is a fold from a 5-fold cross-validation benchmark. **(e)** Gene set enrichment analysis of target genes in the top-10,000 TF-gene interactions by inferred GENIE3+CA and GENIE3 for the placental progenitor cells. Normalised enrichment scores and p-values are indicated.

networks to identify the interactions with the most prominent roles in each cell type and experimentally validate our results.

References

1. Blais, A., Dynlacht, B.D.: Constructing transcriptional regulatory networks. Genes & Development. 19, 1499–1511 (2005).
2. Ernst, J., Vainas, O., Harbison, C.T., Simon, I., Bar-Joseph, Z.: Reconstructing dynamic regulatory maps. Molecular Systems Biology. 3, (2007).
3. The FANTOM Consortium and the RIKEN PMI and CLST: A promoter-level mammalian expression atlas. Nature. 507, 462–470 (2014).
4. Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., Bergmann, S.: Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nature Methods. 13, 366–370 (2016).
5. Hasin, Y., Seldin, M., Lusis, A.: Multi-omics approaches to disease. Genome Biology. 18, (2017).
6. Klein, D.C., Hainer, S.J.: Genomic methods in profiling DNA accessibility and factor localization. Chromosome Research. 28, 69–85, (2019).
7. Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., Collins, J.J.: CellNet: Network biology applied to stem cell engineering. Cell. 158, 903–915 (2014).
8. Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., The DREAM5 Consortium, Kellis, M., Collins, J.J., Stolovitzky, G.: Wisdom of crowds for robust gene network inference. Nature Methods. 9, 796–804 (2012).
9. Blakeley, P., Fogarty, N.M.E., del Valle, I., Wamaitha, S.E., Hu, T.X., Elder, K., Snell, P., Christie, L., Robson, P., Niakan, K.K.: Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. Development. 142, 3151–3164 (2015).
10. Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., Huang, J., Li, M., Wu, X., Wen, L., Lao, K., Li, R., Qiao, J., Tang, F.: Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. Nature Structural & Molecular Biology. 20, 1131–1139 (2013).
11. Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., Lanner, F.: Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. Cell. 165, 1012–1026 (2016).
12. Liu, L., Leng, L., Liu, C., Lu, C., Yuan, Y., Wu, L., Gong, F., Zhang, S., Wei, X., Wang, M., Zhao, L., Hu, L., Wang, J., Yang, H., Zhu, S., Chen, F., Lu, G., Shang, Z., Lin, G.: An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. Nature Communications. 10, 364 (2019).
13. Ewels, P.A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M.U., Di Tommaso, P., Nahnsen, S.: The nf-core framework for community-curated bioinformatics pipelines. Nature Biotechnology. 38, 276–278 (2020).
14. Amezquita, R.A., Lun, A.T.L., Becht, E., Carey, V.J., Carpp, L.N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., Waldron, L., Pagès, H., Smith, M.L., Huber, W., Morgan, M., Gottardo, R., Hicks, S.C.: Orchestrating single-cell analysis with Bioconductor. Nature Methods. 17, 137–145 (2019).
15. Li, Z., Schulz, M.H., Look, T., Begemann, M., Zenke, M., Costa, I.G.: Identification of transcription factor binding sites using ATAC-seq. Genome Biology. 20, (2019).
16. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., Geurts, P.: Inferring regulatory networks from expression data using tree-based methods. PLoS ONE. 5, e12776 (2010).
17. Korotkevich, G., Sukhov, V., Sergushichev, A.: Fast gene set enrichment analysis. bioRxiv. 060012 (2019).

Adaptive rewiring evolves brain-like structure in directed networks

Ilias Rentzeperis[1], Steeve Laquitaine[1], and Cees van Leeuwen[1,2]

[1] KU Leuven, Belgium,
ilias.rentzeperis@gmail.com,
[2] University of Technology Kaiserslautern, Germany

1 Introduction

The wiring diagram of the brain changes continually through development, learning, and recovery from injury. Changes following statistical dependencies in activity between neural components can be modeled as adaptive rewiring. Adaptive rewiring gradually transforms randomly connected networks into structured ones with small-worldness, modular connectivity, and rich-club organization, akin to brain anatomy networks. Adaptive rewiring studies so far have been focused on undirected networks. While mathematically convenient, such networks are physiologically unrealistic. Directionality of connections is important for establishing a processing hierarchy at the level of motifs, circuits, layers, clustered axonal branches, and differentiating functional regions.

We previously presented an adaptive rewiring model, in which network activity propagation was represented by heat diffusion [1] [2] . During each rewiring iteration, a connection with low diffusion is cut and used for an unconnected pair of nodes with high diffusion. Here we use two variants of heat diffusion to model adaptive rewiring in directed networks: advection and consensus. With advection, networks evolve hub nodes that receive information (convergence), while for consensus the evolved networks contain hubs that propagate information (divergence). Including a proportion of random rewiring or combining advection and consensus decreases path length between nodes and increases the number of connected nodes while maintaining a hub structure. These are characteristics that are pervasive in brain networks.

2 Methods

$$Advection: \alpha(t) = e^{L_{out}t} \tag{1}$$

$$Consensus: c(t) = e^{L_{in}t} \tag{2}$$

where $L_{out} = D_{out} - A$ and $L_{in} = D_{in} - A$. D_{out} is a diagonal matrix with the out-degrees of the nodes in its diagonal entries. Similarly, D_{in} has the in-degrees in its diagonal entries. A is the adjacency matrix representing a binary digraph, i.e. A_{ij}=1 indicates that *(j,i)∈E (edge)*, while A_{ij} =0 that (j,i)∉E (Fig 1A). Before the onset of the adaptive rewiring algorithm, the initial network is $A = A_{random}$, a network with N = n nodes and a predetermined number of nonzero connections randomly assigned to pairs of nodes, with the only exception that the node cannot point to itself. Adaptive rewiring proceeds as follows for in-degree connections (and analogously for out-degrees):

Step 1. Select with uniform probability a node k from the nodes with nonzero, but not n-1, in- and out-degrees.

Step 2. Delete edge (i_2, k) and add edge (i_1, k) with the same weight as the previously connected edge (i_2, k). With probability p_{random} select i_1 and i_2 based on step 2.1 (random rewiring) otherwise select them based on step 2.2 (instructed rewiring).

Step 2.1. i_1 is selected randomly from the set (i,k)∉E, i.e. nodes that are not in the in-degree neighborhood of k. i_2 is selected randomly from the set (i,k)∈E, i.e. nodes that are in the in-degree neighborhood of k.

Step 2.2. Calculate the kernel of the algorithm used, f(t). From the set of nodes (i,k))∉E, i_1 is the one with the highest concentration transfer with k. From the set (i,k)∈E, i_2 is the one with the lowest concentration transfer with k. Mathematically, this is expressed as follows (f(τ) function represents either $\alpha(\tau)$ or c(τ)):

$$i_1 = argmax_{(i,k)\notin E, i\neq k} f_{ik}(\tau) \tag{3}$$

$$i_2 = argmin_{(i,k)\in E, i\neq k} f_{ik}(\tau) \tag{4}$$

Step 3. Go back to step 1 until r edge rewirings have been reached.

We show results for τ = 1.

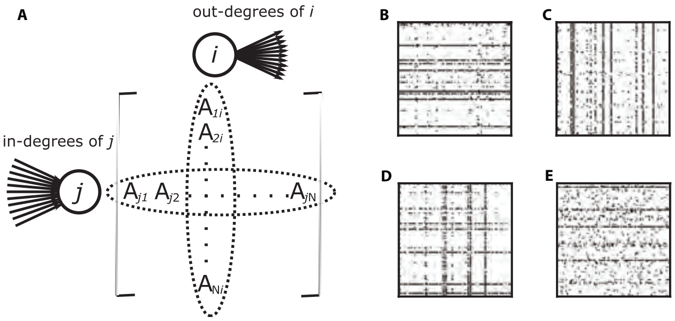


Fig. 1. (A) Schematic representation of the adjacency matrix and how it relates to network structure. The i-th column indicate the out-degrees of the i-th node, and the j-th row the in-degrees of the j-th node. (B) Evolution of a network using advection on the out-degree neighborhood of candidate nodes (C) Evolution of a network using consensus on the in-degree neighborhood (D) Evolution of a network using alternatively consensus and advection (E) Same as (B) but with probability 0.6 we rewire randomly

3 Results

In the context of neuronal dynamics, both consensus and advection (and diffusion in the undirected case) act as homeostatic factors aimed at normalizing the differences in activity between the neurons. Advection, rewiring the out-degree, results in topological patterns as in Fig 1B: a small subset of nodes, acting as hubs, receive connections from all other nodes. Consensus, rewiring the in-degree, results in hubs sending out connections to all other nodes (Fig 1C). Combining advection and consensus in a rewiring scheme results in a pattern of both receiving and broadcasting hubs (Fig 1D). High proportions of random rewiring maintain to an extent the overall hub structure (Fig 1E)

We show results using consensus; we get identical results from advection. Introducing to the in-degree an increasing proportion of random rewiring leads to a decrease in the average path length (Fig 2A) and an increase in the connectivity of the network (Fig 2B), as with random shortcut connections in structured undirected networks [3]. To quantify this effect, we measured the number of hubs in the rewired network; i.e. nodes with an above-threshold number of out-degrees. With increasing probability of random rewiring, the number of hubs reaches a maximum before it declines (Fig 5C).

Ideally, we want networks with small path length that also exhibit the structural properties effected by consensus and advection. However, as we increase random rewiring and reduce the path length, we also reduce the number of hubs. To quantify, the effects of those two opposing forces we devised a structure efficiency metric that takes into account the path length. It is defined as the ratio of the number of hubs over the path length. We find that random rewiring increases structure efficiency up to a point and then it is detrimental (Fig 2D).

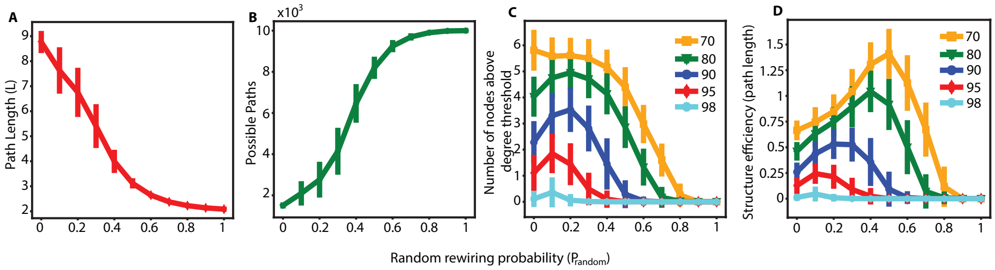


Fig. 2. Random rewiring decreases the average path length, increases the number of connected components, and diminishes the structure obtained by consensus (and advection)(A) Average path length (L) as a function of random rewiring (B) Number of connected nodes as a function of random rewiring (C) Number of hubs as a function of random rewiring. Hubs are defined as nodes with an above-threshold number of out-degrees (from 70 to 98). (D) Structure efficiency metric that uses path length as a function of random rewiring

References

1. Jarman, Nicholas, Erik Steur, Chris Trengove, Ivan Y. Tyukin, and Cees van Leeuwen. ”Self-organisation of small-world networks by adaptive rewiring in response to graph diffusion.” Scientific Reports 7, no. 1 (2017): 1-9.
2. Rentzeperis, Ilias, and Cees van Leeuwen. ”Adaptive rewiring evolves brain-like structure in weighted networks.” Scientific reports 10, no. 1 (2020): 1-11.
3. Watts, Duncan J., and Steven H. Strogatz. ”Collective dynamics of ‘small-world’networks.” nature 393, no. 6684 (1998): 440-442.

The effective graph: a weighted graph that captures nonlinear logical redundancy in biochemical systems

Alexander J. Gates[1], Rion Brattig Correia[2,3], Xuan Wang[4], and Luis M. Rocha[2,4]

[1] Network Science Institute and Department of Physics, Northeastern University, Boston, Massachusetts 02115, USA,
[2] Instituto Gulbenkian de Ciência, Oeiras, Portugal
[3] CAPES Foundation, Ministry of Education of Brazil, Brasília DF, Brazil
[4] School of Informatics, Computing & Engineering, Indiana University, Bloomington IN, USA
a.gates@northeastern.edu, rionbr@gmail.com, rocha@indiana.edu

The ability to map causal interactions underlying genetic control and cellular signalling has led to increasingly accurate models of the complex biochemical networks that regulate cellular function [10, 9, 1]. However, the traditional representation of biochemical networks as static and binary interaction graphs fails to accurately represent an important dynamical feature of these multivariate systems: some pathways propagate control signals much more effectively than others [5] (see Fig. 1A & B). Such heterogeneity of dynamical interactions reflects *canalization*, as the system is robust to interventions in redundant pathways, but responsive to interventions in effective pathways. The simplest way to model such causal, interdependent nonlinear dynamics is with multivariate, discrete dynamical systems; for instance, Boolean Networks (BN) are canonical models of complex systems which exhibit a wide range of dynamical behaviors [2]. BN provide a convenient modelling framework to explore general properties of complex systems, such as self-organization, criticality, causality, canalization, robustness and evolvability [10, 8, 6, 11].

To capture the nonlinear logical redundancy present in biochemical network regulation, signalling, and control, we present the *effective graph*. The effective graph is a weighted, directed graph that statistically integrates all dynamical redundancy present in the BN dynamics, thus revealing the most important interactions in determining state-transitions, as well as very redundant pathways. In this talk we present a summary of key results derived from more than 40 systems biology models analyzed, including that: i) redundant pathways are prevalent in biological models of biochemical regulation (see Fig. 1D & E); ii) the effective graph provides a statistical but precise characterization of multivariate dynamics in a causal graph form (see Fig. 1B & C); and iii) the effective graph provides an accurate explanation of how perturbation and control signals propagate in biochemical regulation, such as those induced by drug therapies on Cancer. See Fig. 1C, and note how cancer drugs (purple nodes) lose their pathway to *Apoptosis* (cell death; green nodes), a desired control outcome in this *ER+* breast cancer model. Overall, our results indicate that the effective graph provides an enriched description of the structure and dynamics of networked multivariate causal interactions. We demonstrate that it improves explainability, prediction, and control of complex dynamical systems in general, and biochemical regulation in particular.

All simulations and code to support the findings are freely available in the CANA python package [4].

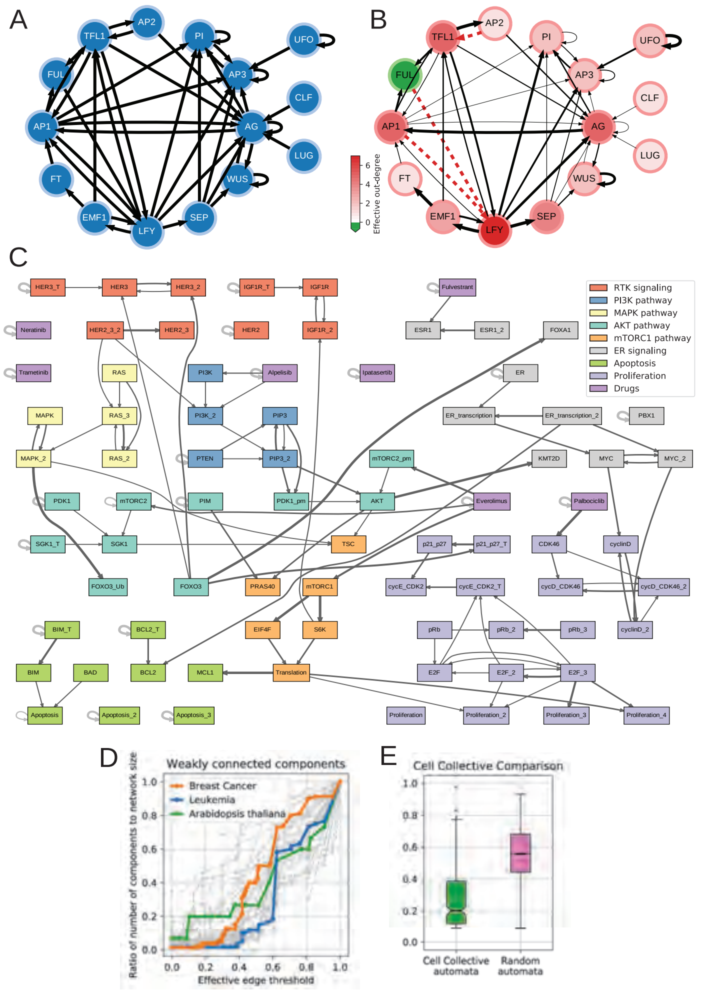


Fig. 1. A. The interaction graph for the *Arabidopsis Thaliana* BN [3]. **B**. The effective graph for the *Arabidopsis Thaliana* BN, in which edge thickness denotes effectiveness, with fully canalized edges shown in dashed red. Node color intensity denotes the node effective out-degree; green nodes denote cases of null effective out-degree. **C**. The effective graph for the BN model of ER+ breast cancer [12], in which edge thickness denotes its effectiveness, thresholded to show only effectiveness edges $e_{ij} > 0.4$ for $e_{ij} \in [0,1]$. **D**. Ratio of the number of weakly connected components to network size in relation to the effective edge threshold for a variety of biochemical BN. The *ER*+ breast cancer (orange), leukemia (blue), and *Arabidopsis thaliana* (blue) networks shown highlighted. **E**. Edge effectiveness of the 240 incoming edges (interactions) to 40 automata with degree $k = 6$ in Cell Collective [7] models (green) compared to a bias-matched sample of random Boolean automata (pink).

References

1. Albert, R., Thakar, J.: Boolean modeling: a logic-based dynamic approach for understanding signaling and regulatory networks and for making useful predictions. Wiley Interdisciplinary Reviews: Systems Biology and Medicine 6(5), 353–369 (2014)
2. Bornholdt, S.: Boolean network models of cellular regulation: prospects and limitations. Journal of Royal Society Interface 5, S85–S94 (2008)
3. Chaos, Á., Aldana, M., Espinosa-Soto, C., León, B.G.P., Arroyo, A.G., Alvarez-Buylla, E.R.: From Genes to Flower Patterns and Evolution: Dynamic Models of Gene Regulatory Networks. Journal of Plant Growth Regulation 25(4), 278–289 (2006)
4. Correia, R.B., Gates, A.J., Wang, X., Rocha, L.M.: Cana: A python package for quantifying control and canalization in boolean networks. Frontiers in Physiology 9 (2018)
5. Davidson, E., Levin, M.: Gene regulatory networks. Proceedings of the National Academy of Sciences 102(14), 4935–4935 (2005)
6. Gershenson, C.: Guiding the self-organization of random boolean networks. Theory in Biosciences 131(3), 181–191 (2012)
7. Helikar, T., Kowal, B., McClenathan, S., Bruckner, M., Rowley, T., Madrahimov, A., Wicks, B., Shrestha, M., Limbu, K., Rogers, J.A.: The cell collective: Toward an open and collaborative approach to systems biology. BMC Systems Biology 6, 96 (Aug 2012)
8. Kauffman, S.A.: The origins of order: Self-organization and selection in evolution. OUP USA (1993)
9. Li, F., Long, T., Lu, Y., Ouyang, Q., Tang, C.: The yeast cell-cycle network is robustly designed. PNAS 101, 4781–4786 (2004)
10. Marques-Pita, M., Rocha, L.M.: Canalization and control in automata networks: body segmentation in Drosophila melanogaster. PloS ONE 8(3), e55946 (2013)
11. Reichhardt, C.J.O., Bassler, K.E.: Canalization and symmetry in boolean models for genetic regulatory networks. Journal of Physics A: Mathematical and Theoretical 40(16), 4339 (2007)
12. Zañudo, J., Scaltriti, M., Albert, R.: A network modeling approach to elucidate drug resistance mechanisms and predict combinatorial drug treatments in breast cancer. Cancer Convergence 1(1) (2017)

Extraction of overlapping modules in networks via spectral methods and information theory

Rion Brattig Correia[1,2,*], Paulo Navarro Costa[1,3], and Luis M. Rocha[1,4,*]

[1] Instituto Gulbenkian de Ciência, Oeiras, Portugal
[2] CAPES Foundation, Ministry of Education of Brazil, Brasília DF, Brazil
[3] ISAMB - Instituto de Saúde Ambiental, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal
[4] Luddy School of Informatics, Computing & Engineering, Indiana University, Bloomington IN, USA
rionbr@gmail.com, rocha@indiana.edu

Networks are a common method to model multivariate interactions in a variety of complex systems found in nature and society. The interactions captured by networks can include a multitude of complex phenomena occurring at various levels of observation and intensity, which are difficult to disentangle automatically. For instance, functionally-relevant gene modules in a network of gene interactions, or disease-related terminology in networks derived from drug and symptom mentions in social media [2] typically have overlapping clusters of widely varying size and strength. To address these and other similar questions, a variety of community structure algorithms have have been proposed in the literature [1, 7, 3]. However, most modularity algorithms seek an optimal partition of the network where each node must belong to a module. This hard-boundary assumption does not match the fluid phenomena often found in biomedical complexity. Indeed, many genes are involved in different biochemical pathways depending on their expression levels and which other biochemical species are present. Hence, the same gene can participate in distinct modules. In these biomedical problems it is more reasonable to assume that network variables can map to multiple, partially overlapping functional communities. Thus, for biomedical applications of network science there has been much recent interest in spectral, overlapping clustering methods [9].

Here we propose a new spectral method to automatically extract overlapping clusters from networks. It is based on two steps: 1) the *Singular Value Decomposition* (SVD) of weighted graph adjacency matrices, in a process akin to *Principal Component Analysis* (PCA) of gene expression data [10], and 2) automatic extraction of overlapping modules using information theory and a polar coordinate projection of data onto singular vector (or component) subspaces. In the first step, when the original network is a bipartite graph relating two distinct sets of variables (e.g. genes vs assays in time, or disease codes vs social media user timelines), we compute the SVD of the bipartite adjacency matrix. If the network is a weighted graph of a single set of variables (e.g. genes), we perform the PCA of the (covariance-normalized) adjacency matrix (see [10] for the difference between SVD and PCA). Fig. 1A depicts the eigenvector variance spectrum of a *Drosophila* gene interaction network obtained via PCA, where the first eigenvector (or component) explains 20% of the variance in the gene co-expression data, and is associated with a large module involving most genes and their regular expression patterns (e.g. cell division, housekeeping and cell cycle).

In functional analysis we are most interested in biochemical processes involving smaller modules which have specific regulatory functions beyond the regular cell operations captured by the first component [8]. Thus, in the second step of the method, we target subsets of lower components. Fig. 1B depicts a biplot of all genes in the network projected as points onto components 2 and 3 of the spectrum. The majority of points is (randomly) projected at the origin of the biplot, showing that they are not correlated with the phenomena captured by either component. Therefore, we want to identify those points (genes) that most protrude and cluster away from the origin, as those are most correlated with the target components. To do so, we transform the Cartesian coordinates of every point to polar coordinates (see Fig. 1D), apply a moving window over the range of radiuses, and compute the Shannon entropy of the distribution of points over angle bins in each radius window. We use overlapping bins (or fuzzy intervals [5]) for both radiuses and angles, meaning that a node at a particular polar coordinate can contribute to more than one angle and radius bin simultaneously (see Fig. 1C & E, respectively).

Computing the Shannon entropy (see red line in Fig. 1D) allows us to track when the distribution of points in polar angle bins transitions from a random to a more structured arrangement. Because points near the origin (radius close to zero) are uncorrelated with the components of interest, the distribution of polar angles tends to be uniformly random, as seen in Fig. 1B,D. As the radius increases, points tend to cluster near specific angles, leading to lower Shannon entropy of the angle distribution. Thus, the goal of the second step of the algorithm is to identify the radius where important transitions in Shannon entropy occur, especially where the distribution of polar angles moves away from a uniform distribution (see blue lines in Fig. 1B,D). Naturally, several entropy transitions may occur, as some clusters are more correlated with components of interest than others—and thus have a higher radius. In other words, identifying the best clusters becomes a multi-objective optimization problem. Several measures can be used to optimize, but we exemplify the method with the rank-sum of radius and entropy to identify the radiuses that maximize the number of points selected while simultaneously minimizing the entropy value. Once a radius is selected, we retrieve only the points that lay beyond the circle it defines. The distinct clusters are then formed by the circle segments that contain similar polar angles; see red polygon in Fig. 1B with radius ≥ 4 selected by rank-sum. In our example, the red module corresponds to genes involved in protein regulation via the proteasome complex, as characterized via gene ontology enrichment analysis (GOEA) [6]. Finally, it important to stress that the clusters thus identified, contain genes that may overlap with clusters found in other component subspaces. In other words, the same network nodes can contribute to overlapping modules associated with distinct phenomena.

In the talk, we will discuss variations of the entropy and multi-objective optimization measures, and apply the method to data from four examples: (i) synthetic networks; (ii) a gene interaction network from transcriptomic data (RNAseq) from Drosophila intestinal cells (Fig. 1); (iii) a knowledge network of drug and symptom terms extracted from social media user timelines [2]; and (iv) a workspace social interaction network collected using radio-frequency identification (RFID) [4].

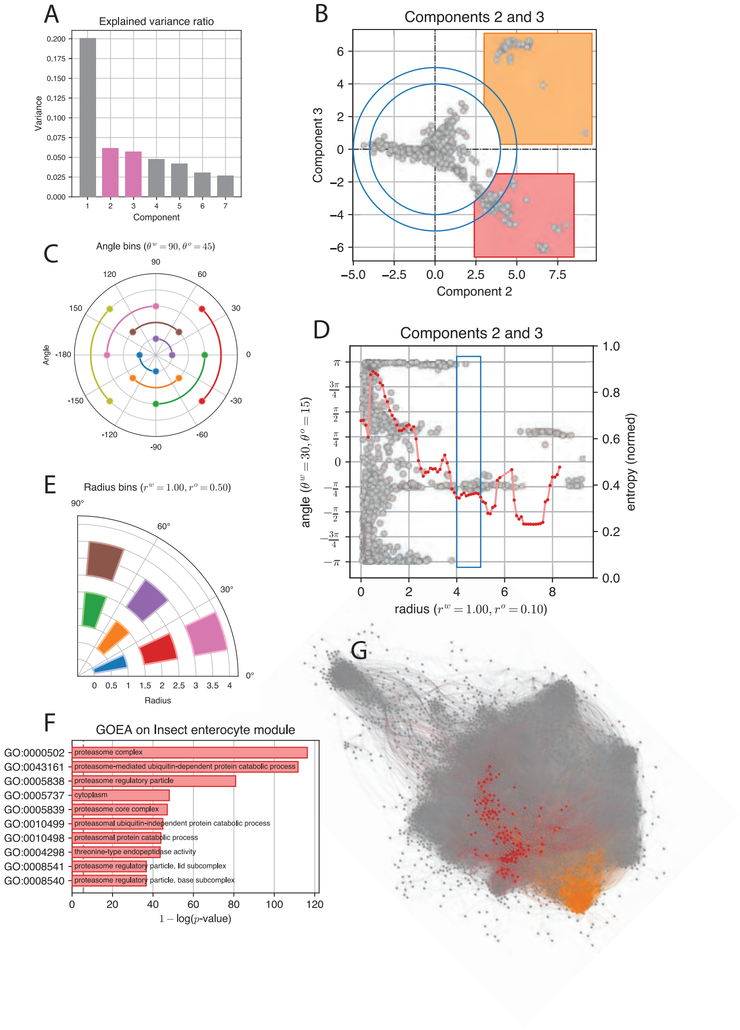


Fig. 1. Gene interaction network of insect (*Drosophila melanogaster*) intestinal cells. A. Spectrum of PCA components of the gene interaction network adjacency matrix, ordered by proportion of explained covariance. **B**. Projection of genes (network nodes) onto biplot of PCA components 2 and 3. Two network modules are highlighted in red and orange. Blue circles shows the minimum entropy window selected (also in D). **C**. Angle bins used in analysis, with width of $r^w = 90$ and overlap of $r^o = 45$. Bins positioned at varying radiuses for easier visualization. **D**. Radius (horizontal) and polar angle (vertical) of same points as in B (subspace of components 2 and 3). Red line and points show the normalized entropy values for each radius window computation ($\theta^w = 30$, $\theta^o = 15$; $r^w = 1.0$, $r^o = 0.1$). Blue rectangle shows the minimum entropy window selected (also in B). **E**. Radius bins used in analysis with width of $\theta^w = 1$ and overlap of $\theta^o = 0.5$. **F**. Gene ontology enrichment analysis (GOEA) of the identified red module (see B). Top 10 significant GO terms shown. **G**. Insect gene interaction network with red and orange modules identified (also in B).

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (10 2008)
2. Correia, R.B., Li, L., Rocha, L.M.: Monitoring potential drug interactions and reactions via network analysis of instagram user timelines. In: Pacific Symposium on Biocomputing, vol. 21, pp. 492–503 (2016)
3. Fortunato, S.: Community detection in graphs. Physics Reports 486(3), 75–174 (2010)
4. Génois, M., Vestergaard, C.L., Fournet, J., Panisson, A., Bonmarin, I., Barrat, A.: Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. Network Science 3(3), 326–347 (9 2015)
5. Klir, G., Yuan, B.: Fuzzy sets and fuzzy logic, vol. 4. Prentice Hall, New Jersey (1995)
6. Klopfenstein, D.V., Zhang, L., Pedersen, B.S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C.J., Yunes, J.M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., Tang, H.: Goatools: A python library for gene ontology analyses. Scientific Reports 8(1), 10872 (2018)
7. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74, 036104 (9 2006)
8. Rechtsteiner, A.: Multivariate analysis of gene expression data and functional information: Automated methods for functional genomics. Ph.D. thesis, Portland State University (2005)
9. Van Lierde, H., Chow, T.W.S., Chen, G.: Scalable spectral clustering for overlapping community detection in large-scale networks. IEEE Transactions on Knowledge and Data Engineering 32(4), 754–767 (2020)
10. Wall, M.E., Rechtsteiner, A., Rocha, L.M.: Singular value decomposition and principal component analysis. In: A practical approach to microarray data analysis, pp. 91–109. Springer (2003)

Viral Infection From a Complex Networks Perspective

Isabel Pérez-Jover[1] and Jacobo Aguirre[2]

[1] Universidad Autónoma de Madrid, Facultad de Ciencias, Ciudad Universitaria de
Cantoblanco, Madrid 28049, Spain,
ipjover13@gmail.com
[2] Centro de Astrobiología CSIC-INTA, Ctra de Ajalvir km 4,
Torrejón de Ardoz, Madrid, Spain

1 Introduction

Complex network theory has been extensively used to analyze a wide variety of real-world systems, including biological organisms. Recently, the study of protein-protein interaction networks (PPINs) has become widespread, since physical interactions between molecules underlie almost every process in a cell. For this reason, understanding protein-protein interactions is essential for describing cellular physiology, and also for developing new drugs to accurately modify selected interaction targets. The rapidly growing knowledge of PPINs for a wide range of organisms developed in the last decade, including viruses, has allowed the study of the viral infection through network-based models. Thus, the study of emerging topological and functional properties of the host-virus PIN leads to the detection of cellular functions that are essential for the viral cycle completion. However, network approaches to the study of host-virus PINs have been mostly restricted to characterize superficially the network topology, and the study of connected nodes and their relative importance in the network have usually been reduced to connectivity analysis [1].

A different widely extended network approach to study the perturbation caused on the cellular function by the virus is the construction of host gene co-expression networks [2], where the dynamic transcriptional response of host genes across the infection is measured by collecting host transcriptomic data at different infection times. To interpret the dynamic host transcriptional response, gene transcription profiles are correlated so that highly correlated genes are connected through edges, and constitute a given subnetwork within the co-expression network. Furthermore, there is evidence that the connectivity obtained through co-expression network analysis is highly related to the biological function [3]. However, this methodology has only been applied to host transcriptional response to infection, and there has not been any previous attempt to construct a host-virus protein co-expression network (PCN).

2 Results

In this work, we describe the infection of a human cell by Epstein-Barr virus, focusing on two different network models: PPINs and PCNs. The host-virus PPIN was reconstructed collecting high-confidence direct interaction data from IntAct database, and the host-virus PCN was built using the temporal proteome of Epstein-Barr virus and its host at different infection times, data previously obtained by Ersing *et al.* [4].

The structural and dynamical features of both networks were compared, and the importance of nodes, i.e. their centrality, was also analyzed through the calculation of node degree (the number of neighbors of the node) and node eigenvector centrality, which is defined as the i-th component of the eigenvector $\vec{u}_1$ associated with the largest eigenvalue λ_1 of the network adjacency matrix $\mathbf{A}$. Our findings confirm that viral proteins are preferentially attached to high-centrality host nodes in host-virus protein-protein interaction networks, while low-centrality host nodes are unexpectedly targeted in host-virus protein co-expression networks (Fig. 1 and Table 1).

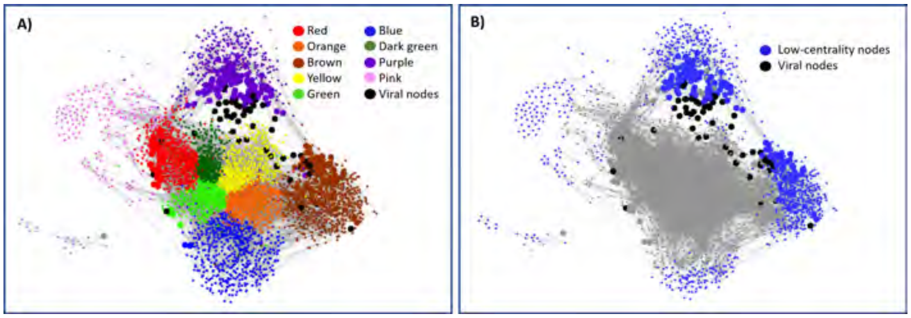


Fig. 1. Viral nodes in human-EBV protein co-expression network (PCN) interact with functionally enriched modules. A) Graphical representation of the host-virus PCN. The main structural modules with size above 10 nodes are named after their representative color (arbitrarily assigned). Viral nodes are highlighted in black, and both viral and human connector nodes display larger node size. B) Low-centrality host nodes, exhibiting an eigenvector centrality below 10^{-10}, are highlighted in blue. Viral nodes (in black) mostly interact with low-centrality host nodes belonging to the purple module represented in A), remarkably less connected to the network core (in gray).

Our results obtained for host-virus protein-protein interaction networks are in agreement with previous work [5, 6], which proved that given a dynamical system of two interconnected networks, the most beneficial connecting strategy for the weakest one in a competition for eigenvector centrality is creating hub-hub interlinks with the stronger one, while connecting through peripheral-peripheral connections becomes especially harmful. In contrast, results showing viral attachment to host peripheral nodes in the reconstructed host-virus protein co-expression network are unexpected. It could be interpreted that viral nodes do not share their expression patterns with host hub proteins; rather, viral expression patterns would be more related to host peripheral nodes located in the most peripheral modules and to the so-called inter-modular nodes.

This potential connecting strategy could be justified from a network science perspective if the viral attachment to the boundaries of two connected host modules obstructed any diffusive process occurring between them, a fact that should be assessed in future research.

Network	Avg. degree	Avg. eig. centrality
HHPIN	6.67	0.0045
HHPIN (connectors)	28.04	0.0244
VVPIN	3.82	0.0897
VVPIN (connectors)	3.40	0.0808
HHCN	95.21	0.0065
HHCN (connectors)	13.31	1.5e-06
VVCN	1.53	0.0489
VVCN (connectors)	1.53	0.0489

Table 1. Comparison of average degree and eigenvector centrality of connector nodes in viral and host protein interaction networks (VVPIN and HHPIN) and viral and host co-expression networks (VVCN and HHCN). Host connector nodes in HHPIN have a higher average degree and eigenvector centrality than those of an average node in total HHPIN. In contrast, host connector nodes in HHCN display the opposite trend having a lower average degree and centrality than an average node. Regarding viral networks, neither VVPIN connector nodes nor VVCN connector nodes have a significantly different average degree or centrality to an average node of the corresponding network.

Summary. The host-virus protein-protein interaction network (PPIN) and the protein co-expression network (PCN) of Epstein-Barr virus and its host cell have been characterized from the perspective of complex network theory. Our findings are pointing towards a different and unexpected connecting strategy of the viral network to the host nodes in the host-virus PCN, opening new questions about the biological meaning of this network model applied to the host-virus system.

References

1. Meyniel-Schicklin, L., de Chassey, B., André, P. and Lotteau, V.: Viruses and interactomes in translation. Molecular & Cellular Proteomics 11(7):M111.014738 (2012).
2. Li, C., Bankhead, A., Eisfeld, A. J., Hatta, Y., Jeng, S. *et al.*: Host regulatory network response to infection with highly pathogenic H5N1 avian influenza virus. Journal of Virology 85(21), 10955–10967 (2011).
3. Zhang, B. and Horvath, S.: A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology 4(1), article 17, doi: 10.2202/1544-6115.1128 (2005).
4. Ersing, I., Nobre, L., Wang, L. W., Soday, L., Ma, Y. *et al.*: A temporal proteomic map of Epstein-Barr virus lytic replication in B cells. Cell Reports 19(7), 1479–1493 (2017).
5. Aguirre, J., Papo, D. and Buldú, J. M.: Successful strategies for competing networks. Nature Physics 9(4), 230–234 (2013).
6. Buldú, J. M., Pablo-Martí, F. and Aguirre, J.: Taming out-of-equilibrium dynamics on interconnected networks. Nature Communications 10(1), 1–9 (2019)

Hierarchical properties and control of ageing-related methylation networks

Gergely Palla[1], Péter Pollner[1], and István Csabai[2]

[1] MTA-ELTE Statistical and Biological Physics Research Group,
H-1117 Budapest, Pázmány P. stny. 1/A, Hungary
[2] Dept. of Physics of Complex Systems, Eötvös University, H-1117 Budapest, Pázmány P. stny. 1/A, Hungary

1 Introduction

An ancient desire of humanity is to understand, slow, or even halt and reverse ageing. In related studies, it was soon realised that certain biomarkers can rather precisely predict the functional capability of tissues, organs and even patients. One of the well-known biomarkers of ageing is provided by DNA methylation, and a prominent example of methylation-based age estimators is the so-called Horvath's clock [1], which is based on 353 CpG dinucleotides, and shows high correlation with chronological age across multiple tissue types. We studied the network between these CpG positions, obtained by applying a regularised regression to methylation data, to locate the most important CpGs (and related genes) that may have a large influence on the rest of the system. According to our analysis based on the Global Reaching Centrality [2], the structure of this network is way more hierarchical compared to what we would expect based on the configuration model. We also studied the control properties of the network using the concept of control centrality [3], and the results show that top nodes according to the hierarchy also seem to have higher control centrality values.

Besides the analysis of the network structure, we also examined how would the perturbation of the methylation levels change the estimated biological age (also called as the DNAm age). When propagation of the change over the network is also taken into account, a unit change in the methylation level of the top CpGs according to the hierarchy tend to have a larger effect on the estimated age compared to the average. By adjusting the methylation of the most influential single CpG site and following the propagation of methylation level changes we can reach up to 5.74 years age reduction, which is significantly larger compared to the results without taking into account the network effects. A flow chart illustration of our study is given in Fig.1.

2 Results

In Fig.2a we show that the GRC value measured in the methylation network is significantly higher compared to what we would expect in random graphs with the same degree distribution. According to that, this network is strongly hierarchical. In Fig.2b we display the 3d scatter plot of the expected change in the estimated age $|\Delta a|$ under a

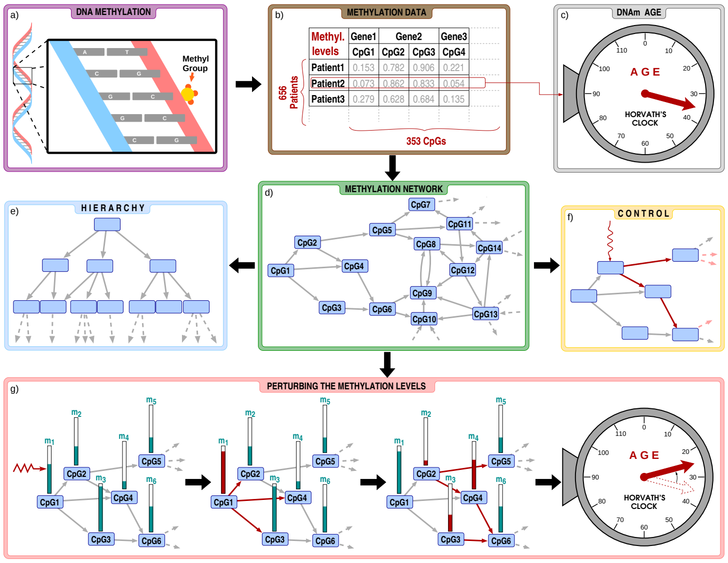

a)
DNA METHYLATION
A T
C G
G C
G C
C G
Methyl Group
b)
METHYLATION DATA
Methyl. levels
Gene1
Gene2
Gene3
CpG1 CpG2 CpG3 CpG4
Patient1 0.153 0.782 0.906 0.221
Patient2 0.073 0.862 0.833 0.054
Patient3 0.279 0.628 0.684 0.135
656 Patients
353 CpGs
c)
DNAm AGE
AGE
HORVATH'S CLOCK
0 10 20 30 40 50 60 70 80 90 100 110
d)
METHYLATION NETWORK
CpG1 CpG2 CpG3 CpG4 CpG5 CpG6 CpG7 CpG8 CpG9 CpG10 CpG11 CpG12 CpG13 CpG14
e)
HIERARCHY
f)
CONTROL
g)
PERTURBING THE METHYLATION LEVELS
m_1 m_2 m_3 m_4 m_5 m_6
CpG1 CpG2 CpG3 CpG4 CpG5 CpG6
AGE
HORVATH'S CLOCK

Fig. 1. Flow chart of our analysis. a) Our study is based on cytosine methylation, a phenomenon where a methyl group is attached to a CpG dinucleotide in the DNA. b) We focus on the methylation level of the 353 CpG positions appearing in Horvath's clock, using the data from Ref.[4], listing altogether 656 patients. c) By plugging in the methylation levels of a given patient into Horvath's clock, we obtain the DNAm age, which is in strong correlation with the chronological age, but is also affected by e.g., the health status. d) Using Lasso-regression, we construct a methylation network between the CpG positions. In order to seek for key influential nodes in the system we analyse the hierarchical (panel e) and control properties (panel f) of the network. g) In addition, we also investigate how would the change of the methylation levels affect the estimated age when the perturbations are transmitted over the methylation network.

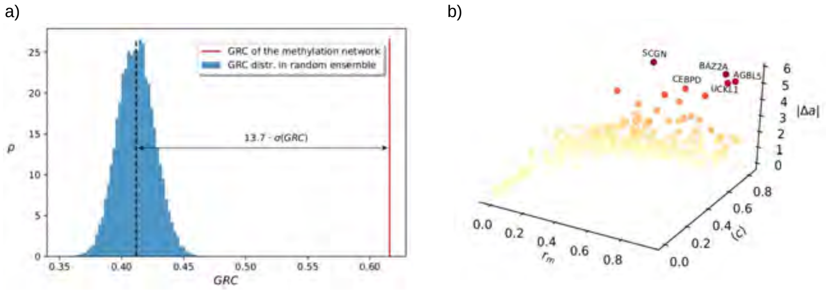


Fig. 2. Hierarchy, control and expected age reducement a) The GRC measured for the methylation network at the optimal link weight threshold (red) together with probability density ρ(GRC) of the corresponding values in a link randomised ensemble. b) Scatter plot of the expected change in the estimated age $|\Delta a|$ as a function of the the node reach r (corresponding to the standing in the hierarchy) and the control centrality c of the nodes averaged for the network realisations obtained at different link weight thresholds.

small, constant perturbation of the methylation level of the individual nodes as a function of the node reach (determining the node position in the hierarchy) and the control centrality. Based on the moderate increasing tendency of the point cloud, the nodes with a larger reach and/or higher control centrality are good candidates for achieving a larger $|\Delta a|$. In Fig.3. we show a hierarchical layout of the network in which the node size and node colour indicates the expected age reduction value.

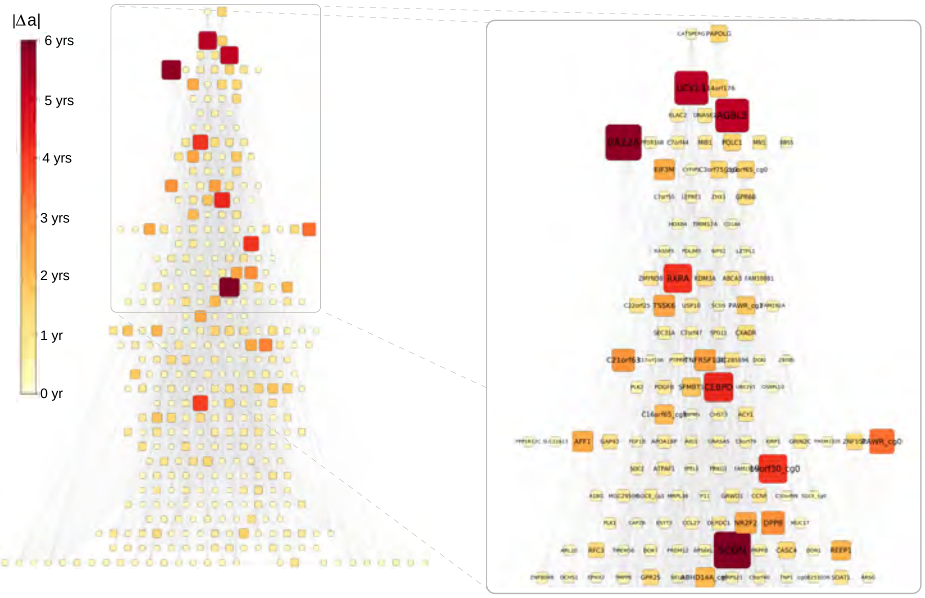


Fig. 3. Top levels of the hierarchy. The shading of the nodes indicates their age reduction value $|\Delta a|$ (with darker shades corresponding to higher values).

References

1. Horvath, S.: DNA methylation age of human tissues and cell types. Genome Biol. 14, R115 (2013)
2. Mones, E., Vicsek, L., Vicsek, T.: Hierarchy Measure for Complex Networks. PLoS ONE 7, e33799 (2012)
3. Liu, YY., Slotine, JJ., Barabási, A.-L.: Controllability of complex networks. Nature 473, 167–173 (2011)
4. Hannum, G., et al.: Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Molecular Cell 49(2), 359 - 367 (2013)

Analysis of psychostimulant-induced group behaviours using network based framework in *Drosophila melanogaster*[*]

Milan Petrović[1], Ana Filošević Vujnović[2], Rozi Andretić Waldowski[2], and Ana Meštrović[1]

[1] University of Rijeka, Department of Informatics, R. Matejčić 2, 51000 Rijeka, Coratia
[2] University of Rijeka, Department of Biotechnology, R. Matejčić 2, 51000 Rijeka, Coratia
{milan.petrovic, ana.filosevic, randretic, amestrovic}@uniri.hr

1 Introduction

Addiction is a complex brain disease reduced to the study of easily measurable addiction-related endophenotypes in the laboratory environment [3]. Large number of mechanistic studies using animal models do not incorporate a critical trait of human addiction, volitional choices between drug use and social interaction. Recent studies on rats suggests that social interaction can change the activity of specific neuronal circuits that control drug craving and relapse [1,2]. To address influence of psychostimulant on social interaction networks before and after the administration of psychostimulants we used *Drosophila melanogaster*, which has been successfully used as a model organism for the study of behaviours related to addiction: alcohol, nicotine and cocaine [5]. There are few studies of the social interaction networks in *D. melanogaster* due to difficulties in objective data collection of social interactions. In one recent study authors introduced a computer vision pipeline for tracking and analysis *D. melanogaster* to compare social interactions of isolated to controlled population of flies [4]. In our research we use similar approach: an open space arena and open source software Flytracker to identify touch events among flies to collect the data. In our study we focused on the analysis of the psychostimulant-induced group behaviours. To analyse group behaviours of adult *D. melanogaster* males, we performed a quantitative analysis of social interaction networks structure on the global and local network level. We constructed two classes of weighted networks: (i) CTR networks based on the social interactions of the group of flies raised in group on the regular fly food and (ii) COC networks based on the social interaction of flies that were raised in group and were orally administrated 0.5 mg/mL of cocaine for 24 hours before tracking. Nodes are related to flies, links refer to the touch interactions of two flies and weights denote the number of touch interactions of each two flies during one experiment.

[*]This research was fully supported by the Croatian Science Foundation through the Project "Influence of redox status on psychostimulant-induced neural plasticity" under Grant IP-2018-01-2794

2 Results

In the Table 1 we report network measures for five CTR and five COC networks (some measures are left out due to the limited space). High modularity score $0,31 \pm 0,05$ in CTRL group is characterised by many links within and few links between groups, while COC treated group had lower modularity $0,20 \pm 0,03$ with few links within and many links between groups. Although this feature is not statistically significant (p= 0.08), finding is supported with total number of interactions, which was higher in COC compare to CTR network (p=0.037). Global network parameters for COC group, characterised by average degree and strength, resulted in an increased overall network interaction density (p=0.002), when compared to CTR flies. Diameter values for COC and CTR group are not significantly different (p=0.07) indicating that the shortest distance between the two most distant flies in the network are similar, but smaller for COC group. Flies exposed to the COC had lower path distance walked compared to CTR (p=0.04), which can be explained by the lower arousal to novelty and anxiety in the COC treated flies. Values of modularity are on average higher for CTR networks, which suggests that CTR populations have higher tendency to form communities than COC populations. Similarly, CTR networks have more disconnected components than COC networks. The results of the local-level analysis are shown as box plots diagrams in 1. Average values for the degree and closeness centrality are almost twice as high for COC networks, while the measures of eigenvector centrality and betweenness across COC and CTR networks are on average equal.

type	CTR	COC	CTR					COC				
measure	average	average	CTR1	CTR2	CTR3	CTR4	CTR5	COC1	COC2	COC3	COC4	COC5
N	30.40	27.40	26	26	32	32	36	29	29	29	25	25
K	43	72.20	40	35	42	36	62	82	88	96	40	55
$< k >$	2.82	5.19	3.08	2.69	2.62	2.25	3.44	5.66	6.07	6.62	3.20	4.40
$< s >$	4.25	7.83	4.85	4.08	4.19	3.12	5.00	9.03	7.79	10.34	5.20	6.80
d	0.10	0.19	0.12	0.11	0.08	0.07	0.10	0.20	0.22	0.24	0.13	0.18
L	2.63	2.09	2.24	2.29	3.28	3.03	2.32	2.09	1.99	2.00	2.25	2.12
D	6.00	4.20	4.00	5.00	8.00	8.00	5.00	4.00	4.00	4.00	5.00	4.00
E_{glob}	0.27	0.47	0.31	0.28	0.22	0.22	0.31	0.46	0.51	0.58	0.37	0.43
c	0.21	0.30	0.21	0.16	0.34	0.16	0.15	0.32	0.31	0.31	0.30	0.25
T	0.26	0.41	0.29	0.24	0.40	0.18	0.18	0.45	0.41	0.37	0.39	0.41
d_{het}	0.85	0.72	0.98	0.74	0.76	0.89	0.87	0.71	0.66	0.60	0.94	0.68
r	-0.12	-0.01	-0.10	-0.16	0.10	-0.34	-0.11	0.11	-0.01	-0.01	-0.25	0.12
N_C	8.40	3.40	7.00	8.00	9.00	10.00	8.00	4.00	3.00	1.00	5.00	4.00
GCC	22.6	25	20	19	23	23	28	26	27	29	21	22
Q	0.31	0.20	0.16	0.26	0.40	0.41	0.30	0.28	0.13	0.23	0.20	0.14

Table 1. Global measures in social interaction networks for COC and CTR populations: number of nodes N, number of links K, avg degree $< k >$, avg. strength $< s >$, density d, avg. path length L, diameter D, global efficiency E_{glob}, clustering coeff. c, transitivity T, heterogeneity d_{het}, asortativity r, number of components N_C, size of giant component GCC and modularity Q.

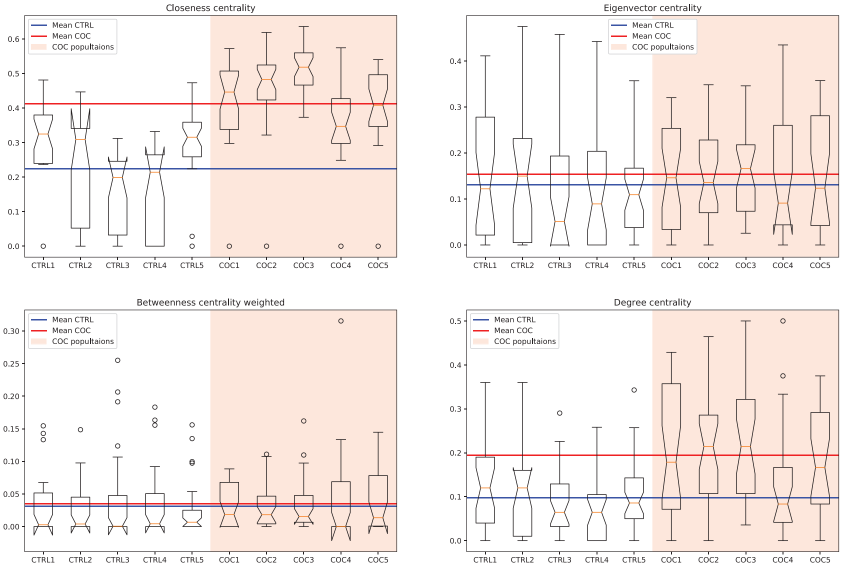


Fig. 1. Closeness, eigenvector, betweenness and degree centrality measures distributions across CTR and COC social interaction networks.

Summary. There are several differences between structures of the COC and CTR networks. COC networks have more links than CTR networks because psychostimulant treatment increased the activities of flies. Consequently, as expected, all global and local network measures related to the number of links differentiate between COC and CTR networks. Interestingly, the values of modularity are higher in the CTR network than in the COC networks. This property may indicate that increased activity in COC networks did not increased interactions within the communities (groups of flies) in population, but on the contrary, the communication is spreading outside the communities.

3 References

[1] Roland Bainton et al. Dopamine modulates acute responses to cocaine, nicotine and ethanol in Drosophila. In: Current biology : CB 10 (Mar. 2000), pp. 187–94.

[2] Jamie L. Catalano et al. Behavioral features of motivated response to alcohol in Drosophila. In: bioRxiv (2020).

[3] K.Kaun, A. Devineni, and U. Heberlein. Drosophila melanogaster as a model to study drug addiction. In: Human genetics 131 (Feb. 2012), pp. 959–75.

[4] Guangda Liu et al. A simple computer vision pipeline reveals the effects of isolationon social interaction dynamics in Drosophila. In: PLOS Computational Biology 14.8 (Aug. 30, 2018). Ed. by Aldo A Faisal.

[5] M. Venniro and Y. Shaham. An operant social self-administration and choice model in rats. In: Nature Protocols 15 (Mar. 2020) default

The Metric Backbone in the Human Connectome and across Lifespan

Andreia Sofia Teixeira[1,2,3], Joshua Faskowitz[4,5], Olaf Sporns[2,4,5], and Luis M. Rocha[1,2,6]

[1] Center for Social and Biomedical Complexity, School of Informatics, Computing, & Engineering, Indiana University, Bloomington IN, USA
[2] Indiana University Network Science Institute, Indiana University, Bloomington IN, USA
[3] INESC-ID, Lisboa, Portugal
[4] Department of Psychological and Brain Sciences, Indiana University, Bloomington IN, USA
[5] Program in Neuroscience, Indiana University, Bloomington IN, USA
[6] Instituto Gulbenkian de Ciência, Oeiras, Portugal
anmont@iu.edu

1 Introduction

Network backbones have been widely used to study the core structure and dynamics of different complex systems, including brain networks [1]. Another key component of network studies is the concept of shortest path, which plays an important role in the optimization of communication. Here we present a preliminary study of the metric backbone – the invariant sub-graph under distance closure that contains all edges that contribute to any shortest path [2, 3] – in human brain structural connectivity of two different cohorts: the Human Connectome Project (HCP) [4] and the Nathan Kline Institute study (NKI) [5]. The HCP is a high-quality dataset of healthy young adults and the NKI provides a community sample with a wide age range, which enables us to track data trends across the lifespan.

Ours is the first study of the metric backbone of human connectome networks. Our preliminary results show that it comprises a surprisingly small subgraph of the original networks, that its size decreases in the earlier decades of human life, and that it accounts for a significantly outsized proportion of brain connection cost.

2 Methods

To estimate structural connectivity for each subject in each dataset, diffusion magnetic resonance images were preprocessed, resulting in maps of white matter tract orientation [6]. Probabilistic tractography [7] was performed on these maps, rendering streamline estimates of white matter anatomical architecture. Structural connectivity was measured by counting the streamlines between brain regions, and normalizing for region volume. Using an atlas with 200 functionally-associated regions [8] resulted in a structural connectivity matrix with 200 nodes.

Following [2, 3], we compute the Metric-Backbone as the invariant subgraph under distance closure, which is sufficient to compute all shortest paths. It contains all metric edges of an original distance (or weighted) graph. An edge is metric if it is the shortest distance between its two nodes, otherwise it is semi-metric because it breaks the triangle inequality – there is a shorter distance between the nodes via an indirect path. A simple s parameter allows us to discriminate these edges and is defined as $s(e_{ij}) = \frac{e_{ij}}{p(i,j)}$, where e_{ij} denotes the (distance) weight of the edge between nodes i and j (the direct distance between nodes), and $p(i, j)$ is the shortest distance path between the same nodes (the smallest sum of distance weights on a path between i and j). To obtain $p(i, j)$ for all edges we compute the all-pairs shortest path (APSP) problem . Edges with $s = 1$ define the metric backbone (they are invariant to distance closure [2]) and are sufficient to compute any shortest path in original graph. Edges with $s > 1$ are semi-metric and do not contribute to any shortest path. Moreover, s can vary widely for edges not on the backbone, and characterizes how much they break the triangle inequality. For example, the pairs of nodes with $s_{i,j} > 2$ means that at least one indirect path between i and j is at least half as short as the direct distance between these nodes.

The brain structural connectivity adjacency matrices entries denote a proximity between nodes. To calculate the shortest paths as above we convert proximity to distance via the nonlinear transformation $\frac{1}{x} - 1$, after normalizing to the interval $[0, 1]$ as suggested in [2].

3 Preliminary Results and Discussion

We study the characteristics of the metric backbone first in the adult HCP cohort, and then compare the same network measures across the lifespan, in the NKI cohort. In this preliminary work we present the results for the fraction of edges in the metric-backbone and the fraction of the connection cost – here defined as the product of the length of the streamline and the streamline count [9] – that it covers. In Figure 1 upper left panel we present the distribution of the fraction of edges in the metric-backbone for the HCP dataset, while in the upper right panel we present the same measurement across the lifespan in the NKI data. The analysis reveals that the metric backbone is comprised of very small subgraph of the original network – around $10 - 13\%$. In other words, there is a lot of redundancy in these connectome networks, whereby only $10 - 13\%$ are needed to compute all shortest paths. Moreover, in the NKI dataset we observe a significant decline in the size of the backbone before the median – 40 years old – (Spearman correlation $\rho = -0.392$ and $p - value = 1.923e^{-10}$). For the remaining lifespan there is no evidence of any strong trend correlated with age ($\rho = -0.064$ and $p - value = 0.321$), that is, the size of the backbone stabilizes after the first decades.

Interestingly, even though the metric-backbone comprises only a small fraction of edges ($\approx 11\%$), on average, it accounts for about half of the total connection cost. This suggests that the metric backbone represents a significant "investment" in connecting streamlines and thus that shortest paths are important for communication in brain networks. Our results are validated by null models constructed from a population of 1000 random connected subgraphs of the same original networks.

References

1. O. Sporns, G. Tononi, and R. Kötter, "The human connectome: a structural description of the human brain," *PLoS Comput Biol*, vol. 1, no. 4, p. e42, 2005.

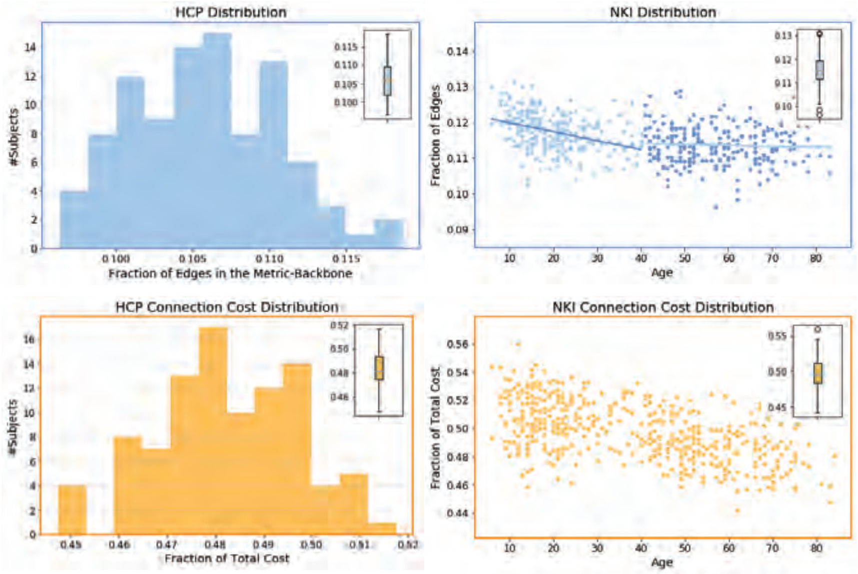


Fig. 1. Characteristics of the metric-backbones regarding the fraction of edges and the fraction of connection cost they support. In left panel is the HCP dataset, and in the right panel the NKI dataset. For the upper right panel, we also present the regression lines for ages between [6,40] ($R^2 = 0.1431$, $slope = -0.0002$) and between [41,84] ($R^2 = 0.0027$, $slope = -2.6789e^{-05}$).

2. T. Simas and L. M. Rocha, “Distance closures on complex networks,” *Network Science*, vol. 3, no. 2, pp. 227–268, 2015.
3. R. Brattig Correia, A. Barrat, and L. M. Rocha, “The metric backbone preserves community structure and is a primary transmission subgraph of contact networks in epidemicspread models,” *Under Review*, 2020.
4. M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, *et al.*, “The minimal preprocessing pipelines for the human connectome project,” *Neuroimage*, vol. 80, pp. 105–124, 2013.
5. K. B. Nooner, S. Colcombe, R. Tobe, M. Mennes, M. Benedict, A. Moreno, L. Panek, S. Brown, S. Zavitz, Q. Li, *et al.*, “The nki-rockland sample: a model for accelerating the pace of discovery science in psychiatry,” *Frontiers in neuroscience*, vol. 6, p. 152, 2012.
6. J.-D. Tournier, C.-H. Yeh, F. Calamante, K.-H. Cho, A. Connelly, and C.-P. Lin, “Resolving crossing fibres using constrained spherical deconvolution: validation using diffusion-weighted imaging phantom data,” *Neuroimage*, vol. 42, no. 2, pp. 617–625, 2008.
7. E. Garyfallidis, M. Brett, B. Amirbekian, A. Rokem, S. Van Der Walt, M. Descoteaux, and I. Nimmo-Smith, “Dipy, a library for the analysis of diffusion mri data,” *Frontiers in neuroinformatics*, vol. 8, p. 8, 2014.
8. A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo, “Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri,” *Cerebral cortex*, vol. 28, no. 9, pp. 3095–3114, 2018.
9. M. P. van den Heuvel, R. S. Kahn, J. Goñi, and O. Sporns, “High-cost, high-capacity backbone for global brain communication,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 28, pp. 11372–11377, 2012.

Unraveling the paradox of weak links in structural brain connectivity

Gorka Zamora-López[1,2] and Matthieu Gilson[1,2]

[1] Center for Brain and Cognition, Pompeu Fabra University, Barcelona, Spain.
e-mail: gorka@zamora-lopez.xyz,
[2] Department of Information and Communication Technologies, Pompeu Fabra University,
Barcelona, Spain.

1 Introduction

Over the last three decades the study of structural brain connectivity has revealed several features which are found consistently across species: structural brain connectomes are modular [1] and small-world [2], with the cross-modular pathways centralised through a set of densily interconnected hubs, which form a rich-club [3][4]. These conclusions have been mainly achieved by considering the unweighted – binarised – connectivity matrices derived from tract-tracing or tractography. However, if the strength of individual interactions were heterogeneous these network properties could be seriously altered. In practice, the connectivity maps derived from both tract-tracing and tractography assign individual weights to the identified connections. These link weights are largely heterogeneous with values spanning over orders of magnitude from the strongest to the weakest; and their weight rapidly decays with the length of the fiber [6].

Taken together, the network properties reported to govern the organization of the anatomical connectivity and the prominent decay of link weights with fiber length result paradoxical. On the one hand graph analysis – of binary matrices – indicates that brain connectivity obeys the small-world property [2]. For that, the long-distance fibers need to act as shortcuts facilitating that information can travel quickly across distant regions of the brain. On the other hand, the rapid decay of link weights with fiber length implies that longer fibers are orders of magnitude weaker than the shorter ones, thus rendering the shortcuts irrelevant at the eyes of any dynamical process propagating along the brain.

2 Results

Here, we aim at resolving this paradox. For that, we study whole-brain effective connectivity derived from brain activity at rest via functional MRI in healthy participants. Effective connectivity has the ability to estimate the strength of interactions between brain regions from the observed activity, encompasing many unknown parameters that would be required for a direct evaluation. We compare the properties of both structural (SC), functional (FC) and effective (EC) connectivity matrices, Fig. 1A. In this case functional connectivity is quantified as the Pearson correlation between the BOLD signals of two regions of interest (ROIs). Effective connectivity is an estimate of the strength and the directionality of the interactions between two brain regions. The estimation is performed fitting empirical data – the BOLD activity of the two ROIs – by mean of a generative model of the signals. In our case, we assume the multivariate Ornstein-Uhlenbeck process as the underlying generative dynamics which is a linear noise propagation model [7].

We find that, in accordance with previous tract-tracing and tractography studies, the weight of structural links sharply decreases with the (euclidean) distance between the brain regions connected by the link, Fig. 1B. However, this behaviour is not corresponded in functional or in effective connectivity where the strength of the links turns to be independent of the distance between the brain regions. For ROIs separated by any distance, both weak and strong functional or effective links can be found. Further, we perform a comparative network analysis of the anatomical and effective connectivities using dynamic communicability [8], a recent method to study complex networks which naturally incorporates the coupling strength of the connections into the analysis. This approach employs the spatio-temporal propagation of perturbations over the system in order to extract topological information of the network. Dynamic communicability $\mathscr{C}_{ij}(t)$ represents the temporal response of region j after a perturbation is applied on region i at time $t = 0$. The total communicability of a network, Fig. 1C, quantifies the global excitability of a network after all nodes are simultaneously perturbed. We see that the connectivity based on SC gets larger response than the same process for EC, however, it reaches its peak later. In this context, the distance d_{ij} between two regions can be defined as the time that region j needs to reach its peak response, given an input is applied to region i. The resulting pair-wise distance matrix, Figs. 1D and E, resembles for weighted networks, the pair-wise pathlength between nodes in a binary graph. We find that the average weighted pathlength derived from EC is shorter than that of SC, evidencing that the loss of shortcuts in SC due to the sharp decay of weights with fiber length makes the structural network inneficient for the propagation of activity on the brain, which contradicts the notion of brain connectivity as small-world networks.

Our results evidence that the interpretation of link weights returned by tractography as the coupling strength of those connections is misleading. In the absence of empirical techniques which could directly and reliably quantify the strength of interactions between ROIs, for now, we advocate to the use of effective connectivity as our source for both network analysis and constraining whole-brain models. Even if effective connectivity does not represent the ultimate solution for this problem, since EC estimation depends on the generative dynamical model of choice, the weights estimated by effective connectivity are closer to the concept of coupling strengths than the weight values returned by tractography for the links.

References

1. C.C. Hilgetag, G.A.P.C. Burns, et al.: Anatomical connectivity defines the organization of clusters of cortical areas in the macaque monkey and the cat. Phil. Trans. R. Soc. Lond. B 355, 91–110 (2000).
2. O. Sporns and J.D. Zwi: The small world of the cerebral cortex. Neuroinformatics 2, 145–162 (2004).

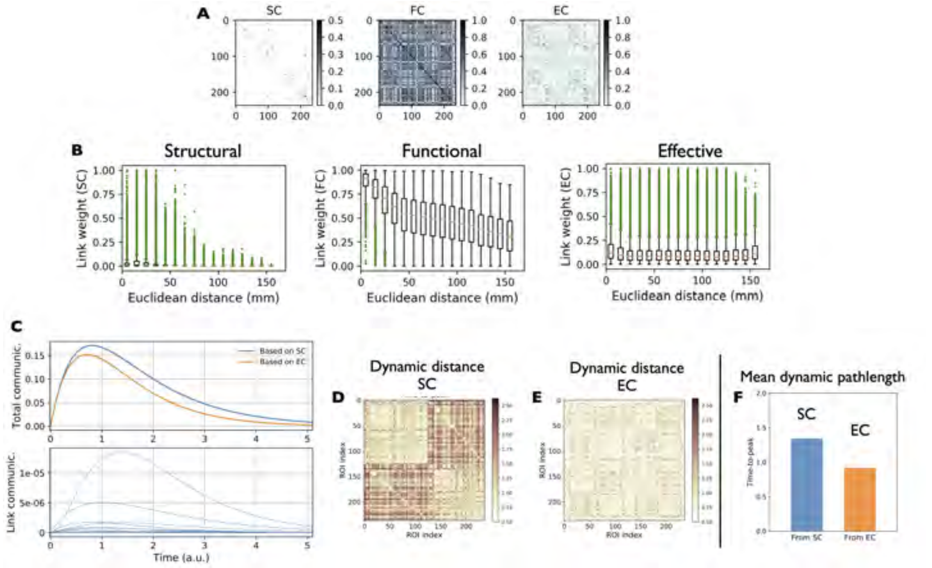


Fig. 1. Comparison of structural, functional and effective link weights and their network properties. A Population averaged structural (SC), functional (FC) and effective (EC) connectivity matrices. **B** Dependence of link weights on the euclidean distance between the regions of interest (ROI) connected. While structural weights derived from probabilistic tractography sharply decay with distance (or fiber length), the strength of functional or effective connections is independent of the distance between ROIs. **C** Total dynamic communicability over time for SC and EC based connectivities. (Top) Curves represent the sum of all values in the $\mathscr{C}(t)$ matrices. (Bottom) Sample responses of various ROIs to a stimulus applied to primary visual cortex at time $t = 0$ shows heterogeneous responses both in amplitude and timing. **D** and **E** Dynamic distance matrices for all ROIs with dynamic communicability is based on structural or effective connectivity respectively. Here, distance d_{ij} is characterised as the time a node j takes to reach the peak response after an initial perturbation is applied on i. **F** Average "pathlength" for SC and EC based dynamic communicability. Results are the mean values for the matrices in D and E.

3. G. Zamora-López, C.S. Zhou and J. Kurths: Cortical hubs form a module for multisensory integration on top of the hierarchy of cortical networks. Front. Neuroinform. 4:1 (2010).
4. M.P. van den Heuvel and O. Sporns: Rich-Club organization of the human connectome. J. Neurosci. 31(44), 15775-15786 (2011).
5. G. Zamora-López, C.S. Zhou and J. Kurths: Graph analysis of cortical networks reveals complex anatomical communication substrate. Chaos 19:015117 (2009).
6. N.T. Markov, A.R. Ribeiro Gomes et al.: The role of long-range connections on the specificity of the macaque interareal cortical network. Proc. Acad. Sci. USA 110(13), 5187-5192 (2013).
7. M. Gilson, R. Moreno-Bote et al.: Estimation of directed effective connectivity from fMRI functional connectivity hints at asymmetries of cortical connectome. PLoS Comput. Biol. 12(3) e1004762 (2016).
8. M. Gilson, N. E. Kouvaris, et al.: Network analysis of whole-brain fMRI dynamics: A new framework based on dynamic communicability. NeuroImage 201, 116007 (2019).

Signed distance correlation to generate weighted and thresholded gene coexpression networks

Javier Pardo-Diaz[1,*], Philip Poole[2], Mariano Beguerisse-Diaz[3], Charlotte M Deane[1], and Gesine Reinert[1]

[1] University of Oxford, Department of Statistics, OX1 3LB, UK,
[2] University of Oxford, Department of Plant Sciences, OX1 3RB, UK,
[3] University of Oxford, Mathematical Institute, OX2 6GG, UK
jdiaz@stats.ox.ac.uk,
https://github.com/javier-pardodiaz

1 Introduction

Genes which are expressed similarly under a variety of experimental conditions are more likely to share biological function than random groups of genes. The corresponding data of gene expression under different conditions are often summarised in a gene coexpression network. Gene coexpression networks are networks in which nodes represent genes and edges correlation between their expression across different samples.

The construction of such gene coexpression networks is far from unique. Correlations can be assessed for example using Pearson's correlation, or using signed distance correlation (Pardo-Diaz et al., 2020). Signed distance correlation is a signed variation of distance correlation. Distance correlations compares two sets vectors by comparing their within-set distance matrices. It is constructed to measure the dependence, both linear and nonlinear, between the two vectors and it is zero if and only if the vectors are statistically independent. Signed distance correlation is already successfully employed to generate unweighted gene coexpression networks from gene expression data in Pardo-Diaz et al, 2020. These networks are more robust and capture more biological information than networks obtained using Pearson correlation but the networks are obtained through thresholding and thus may ignore important information.

Indeed, often correlations are thresholded to obtain a relatively sparse unweighted network which contain edges only between those pairs of genes whose correlation value exceeds a chosen threshold. This network construction method ignores detailed information about the strength of the correlation. Complete weighted networks which have edges connecting all pairs of edges make it difficult to distinguish false positives from true positives. Here, we present a method to construct weighted and thresholded gene coexpression networks for which the sparsity can be controlled by assigning weighted edges only to those pairs of genes with a correlation of their expression higher than a given threshold. The weights of the edges correspond to the correlation values.

We compare weighted and thresholded gene coexpression networks constructed from gene expression data using either Pearson correlation or signed distance correlation. We also compare these networks to the corresponding unweighted networks using the three datasets employed in Pardo-Diaz et al., 2020. For each dataset, we construct a correlation matrix using signed distance correlation and another one using Pearson correlation, as in Pardo-Diaz et al., 2020. We then select the optimal threshold for constructing a weighted and thresholded gene coexpression network from each matrix, choosing the threshold using a variation from COGENT. COGENT (Bozhilova et al., 2020) is an algorithmic method that evaluates the self-consistency of methods to construct gene coexpression networks. COGENT iteratively splits the expression data in possibly overlapping datasets and compares the networks resulting from both sets of data. The more similar the networks are, the more self-consistent is the method and the more robust are the networks. We evaluate the networks by comparing their robustness, assessed with COGENT, and also the amount of biological information they capture, assessed with STRING (Szklarczyk et al., 2019).

2 Results

For each dataset and correlation matrix, we evaluate how the score retrieved using COGENT changes across different thresholds. This score depends on the similarity of the networks constructed at each COGENT iteration. The similarity is adjusted using the overlap expected between each of the networks and random networks with the same edge weights. Fig. 1 illustrates how the signed distance and Pearson scores change for different edge weight values in the *R. leguminosarum* dataset. In all three datasets the curve shows a similar shape and there is an edge weight value for which the scores reach their maxima. The thresholds associated with those edge weights are the optimal thresholds which we choose when constructing the networks. In all cases, the highest scores obtained using signed distance correlation are higher than those for Pearson. Therefore, weighted and thresholded gene coexpression networks based on signed distance correlation are more self-consistent than those based on the Pearson correlation.

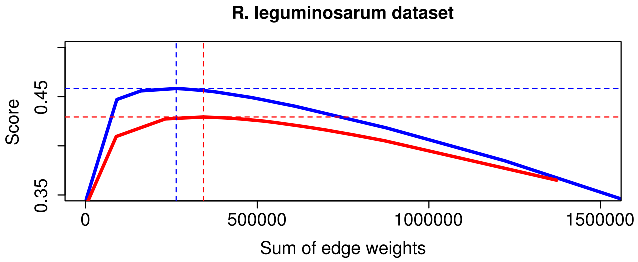


Fig. 1. Self-consistency scores of R. leguminosarum networks for different thresholds. The blue and red lines show the scores of networks obtained using signed distance and Pearson correlations respectively. The dashed vertical lines indicate the optimal sums of edge weights.

We assess the amount of biological information contained in the networks by computing the dot product of the edge weights and the scores provided by STRING to each pair of genes and then dividing by the sum of edge weights to normalise the results. Here we use three different sets of STRING scores: obtained using all the evidence in STRING C, obtained using only coexpression information $C^{\dagger}$, and obtained exluiding coexpression information $C^{\ddagger}$. To compare networks obtained using different correlation measurements, we construct two extra networks with sum of edge weights matching those previously obtained using their optimal threshold. In all the cases and for the three datasets, the networks obtained using signed distance correlation capture more biological information than their competitors. These results suggest that signed distance correlation can be used for generating weighted and thresholded networks from gene expression data, obtaining better results than with Pearson correlation. Table 1 presents the four *R. leguminosarum* studied networks (in italics the optimal network for each correlation), their metrics and their STRING scores.

Table 1. Metrics, COGENT score and STRING scores for the R. leguminosarum optimal networks for each correlation (italics) and networks with matching sum of edge weights.

Correlation	**Score**	**Thr**	**Edges**	**Sum edge weights**	**STRING** *C*	**STRING** *C*†	**STRING** *C*†
Signed distance	*0.458*	*0.6*	*388374*	*263662*	*28.66*	*8.70*	*25.72*
Pearson	*0.429*	*0.55*	*531374*	*342785*	*22.39*	*6.26*	*20.26*
Signed distance	0.456	0.57	523675	342785	25.78	7.33	23.23
Pearson	0.428	0.58	391691	263662	24.43	7.28	22.03

We compare the amount of biological information captured by weighted and thresholded networks, and unweighted networks constructed using signed distance correlation. To this purpose, we use two groups of networks with the same set of edges: optimal set of edges for the weighted and thresholded network and the optimal set for the unweighted network in Pardo-Diaz et al., 2020. The edges in the unweighted networks are assigned the mean weight in the corresponding thresholded and weighted network. The weighted and thresholded networks capture more of the biological information and therefore should be preferred. The results for the *R. leguminosarum* networks are shown in Table 2. The results for the other datasets are similar.

Table 2. STRING scores for the signed distance correlation networks from Table 1 and their paired unweighted thresholded networks. The networks are paired by number of edges.

Correlation	**Network type**	**Edges**	**STRING** *C**	**STRING** *C*†*	**STRING** *C*†*
Signed distance	Weighted	263662	28.66	8.70	25.72
Signed distance	Unweighted	263662	27.31	7.87	24.57
Signed distance	Weighted	313348	30.97	9.88	27.70
Signed distance	Unweighted	313348	29.61	9.02	26.55

References

1. Pardo-Diaz, J. et al. (2020). Robust gene coexpression networks using signed distance correlation. BioRxiv
2. Bozhilova, L. V. et al. (2020). COGENT: evaluating the consistency of gene co-expression networks. Bioinformatics
3. Szklarczyk, D. et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genomewide experimental datasets. Nucleic Acids Research, 47(D1), D607–D613

Inferring eukaryote-prokaryote interactions in microbial communities

Somaye Sheykhali[1], Juan Fernández-Gracia[1] Carlos M. Duarte[2], and Víctor M. Eguíluz[1]

[1] IFISC (CSIC-UIB) Palma de Mallorca - Spain
somaye@ifisc.uib-csic.es,
[2] King Abdullah University of Science and Technology(KAUST), Red Sea Research Center (RSRC), Thuwal, Saudi Arabia

1 Introduction

Microorganisms do not exist in isolation but form complex ecological interaction networks. Such microbial interaction networks derived from a variety of habitats, including soil [1], marine [2], and freshwater communities [3] have been studied for their structural properties. Specially in marine environments the number of different organisms is vast and depends on our sampling and sequencing effort [4, 5], letting place for many interactions to be discovered. Operational taxonomic units (OTUs) can be thought of as groups of microorganisms that are operationally grouped together as a single species. A group of OTUs that tend to co-occur may correspond to taxa that share an ecological niche, or that participate in a symbiotic interaction [6]. Similarly, groups of OTUs that tend to mutually exclude each other may represent competitive interactions within a given niche. In addition, a particular co-occurrence pattern among two OTUs may not necessarily represent a mutualist or competitive interaction, as that pattern can arise in an indirect way, for example resulting from the response of two taxa to an environmental change or to the interaction with a third taxon.

Here we use OTU abundance data to identify statistically significant pairwise similarities which are interpreted as potential ecological interactions. The set of such pairwise interactions among the sampled OTUs can be understood as a microbial ecological network. We infer a co-occurrence multilayer network, which the layers correspond to eukaryote and prokaryote taxa and thus we study both the relations among same type taxa (intra-layer connectivity) and between different types (inter-layer connectivity). With this approach we tackle the question of which marine microorganisms tend to co-occur more than expected by chance and how is this related to which group they belong to (eukaryotes/prokaryotes). Furthermore, we explore two of these networks, one representing the relations among highly prevalent OTUs, present in at least 80% of the samples and one for rare OTUs present only in one sample.

2 Methods

2.1 Data

We analyze the Malaspina Deep-Sea Gene Collection dataset [7, 8], which describes the abundances of 6919 OTUs, including 3017 annotated as eukaryotes (from 18S rDNA sequences) and 3902 as prokaryotes (from 16S rDNA sequences) sampled in deep ocean stations. For each OTU, we consider the D2 taxonomic level as a coarse-graining of different OTUs into similar families.

2.2 Eukaryote-prokaryote co-occurrence multilayer network inference

The microbial networks were obtained from OTU abundance data, where significantly associated objects are connected by edges. The network inference process is divided in preprocessing, initial network computation and assessment of significance. In the first step, a OTU abundance matrix needs to be normalized. In the next step, a similarity measure based on the Jensen-Shannon divergence provides an association strength between every pair of OTUs. In particular, the square root of the JSD converts this into a distance and 1-JSD can be interpreted as similarity measure. The higher is this value, the more similar are the abundance patterns across samples. Only in the case that two species are only present in one sample, and that sample is the same one, the interaction strength will be equal to one. A threshold on the similarity value is selected such that the initial network contains 100 edges. Significance co-occurrence across the samples was assessed using permutations of the original abundance matrix. If any pair of OTUs has a similarity value above the threshold, we call this interaction statistically significant and include it in the interaction network; any correlation below the threshold is discarded. Last, we coarse-grain this network to the D2 taxonomic level by aggregating the weights of edges connecting two OTUs that correspond to different D2 taxonomic classes.

3 Results

We show the eukaryote-prokaryote co-occurrence network for highly prevalent and for rare taxa at the D2 taxonomic level (Fig. 1). The links are weighted by the number of times there is a connection between two OTUs in those D2 classes. The size of a circle segment corresponds to the abundance of a D2 taxonomic class. The most relative abundant OTUs in both networks were generally found to belong to the phyla with the most co-occurred OTUs in both networks (Alveolata from eukaryotes and Gamma/Alphaproteobacteria from prokaryotes). The most connected nodes in the network formed by rare OTUs, imply that some of the low abundant but highly connected OTUs may play a keystone role in network structure. Different OTUs among rare or highly prevalent ones can share the same phylogenetically informative D2 region of the rRNA, so we see some of the same OTUs (e.g. alveolata, rhizata, etc.) in both layers of both networks.

4 Discussion

We are at an incipient stage of applying network analyses to explore the structure of marine eukaryotes and prokaryotes microbial communities. Our results so far have revealed co-occurring and phylogenetic relationships. The resulting co-occurrence multi-

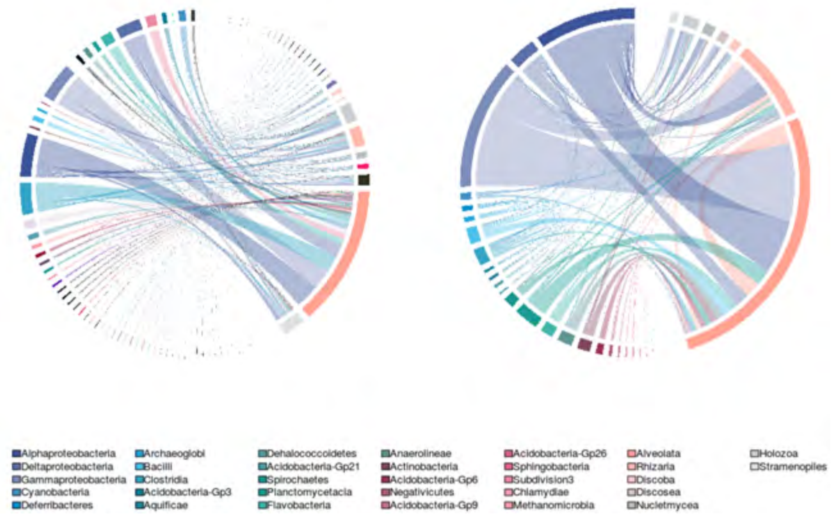


Fig. 1. Circos representations of co-occurrences of the prokaryotes and eukaryotes OTUs. Left: The most prevalent taxa, Right: rare species.

layer network (Fig. 1) permits the visual summary of information about the most significant similarities in abundances at the D2 taxonomic level. In general, dominant phylotypes in eukaryotes (Alveolata and Rhizaria) and Proteobacteria (Gammaproteobacteria and Alphaproteobacteria) are more likely to occur together. Although the analysis is still in progress we believe we can identify keystone taxa, both rare and highly prevalent, as central nodes in the multilayer networks presented here, with possible implications for microbial ecosystems function and robustness.

References

1. Faust, K., & Raes, J.: Microbial interactions: from networks to models. Nat. Rev. Microbiol, 10(8), 538 – 550 (2012)
2. Steele, J. A., et al.: Marine bacterial, archaeal and protistan association networks reveal ecological linkages. ISME J. 5(9), 1414 – 1425 (2011)
3. Kara, Emily L., et al: A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA. ISME J. 7(3), 680 – 684 (2013)
4. Eguíluz, Victor M., et al.: Scaling of species distribution explains the vast potential marine prokaryote diversity. Sci. Rep. 9(1), 1 – 8 (2019)
5. Duarte, Carlos M., et al.: Sequencing Effort Dictates Gene Discovery in Marine Microbial Metagenomes. Environmental Microbiology (2020)
6. Lima-Mendez, Gipsi, et al.: Determinants of community structure in the global plankton interactome. Science 348(6237), (2015)
7. Duarte, C.M.: Seafaring in the 21st century: The Malaspina 2010 Circumnavigation Expedition. Limnol Oceanogr Bull. 24, 11 – 14 (2010)
8. Acinas, Silvia G., et al. : Metabolic Architecture of the Deep Ocean Microbiome. bioRxiv. 635680 (2019)

Part II

Community Structure

The macro-, meso- and micro-structure of individual-based community-wide plant-pollinator networks reflects pollen flow dynamics and plant reproductive success

Alfonso Allen-Perkins[1,3], María Hurtado de Mendoza[2], David García-Callejas[2], Oscar Godoy[2], and Ignasi Bartomeus[1]

[1] Estación Biológica de Doñana (EBD-CSIC), Isla de la Cartuja, 41092, Sevilla, Spain
[2] Universidad de Cádiz, Campus Universitario de Puerto Real, 11510, Cádiz, Spain
[3] alfonso.allen.perkins@gmail.com

1 Introduction

Network ecology has improved the description and interpretation of multitrophic ecological communities, especially for mutualistic bipartite networks which depict two interacting guilds [1]. However, most ecological network research has adopted a species-based approach, despite the fact that ecological processes take place at individual level and depend on individual variation [2]. A common limitation of the few empirical examples of ecological individual-based networks is that only individuals of one species (usually a plant species linked by its pollinators) are considered [3, 4]. Indeed, finding robust methods to tackle the complexity of integrating individual nodes belonging to several species remains an open question.

In this work, we use a node-colored multilayer approach [5] to investigate the patterns of connections between conspecific and heterospecific individuals in 9 well-resolved individual-based plant-pollinator networks. We model the pollen transport among plants mediated by floral visitors as an ensemble of conspecific pollen circuits (or layers) that are coupled through insect species. Specifically, each layer contains the *dependence interactions* among conspecific plant individuals and their floral visitors' species [6], whereas interlayer links account for both, the interspecific movements of insects, and the phenological overlap of coflowering plant species (see Fig. 1). We characterize the overall community structure (i.e. the macroestructure) from a dynamical flow-based perspective [7], and describe node roles (i.e. the microstructure) by using their degree, strength and PageRank metrics [8], as well as their specialization indices [9]. Furthermore, since the mesoscale may be the most relevant level of analysis from a functional point of view [10], we also inspect the composition of triplets, i.e., three node motifs. To take into account the different types of links established with conspecifics and heterospecific plant partners, we introduce here the concept of *homospecific* and *heterospecific motifs*, denoted as HoMs and HeMs, respectively.

Despite its simplicity, this set of metrics allows testing a series of predictions arising from these structures. By relating the topological position of individual plants (i.e., their centrality and motifs metrics) to their fitness (i.e., seed production) with linear

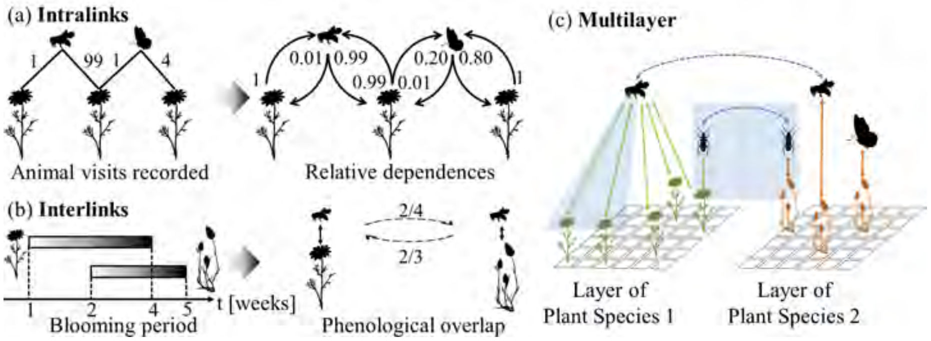


Fig. 1: Example of (a) intralinks and (b) interlinks calculations. (c) Example of multilayer for a plot with 2 plant species and 3 insect species. An example of HoM and HeM motifs are highlighted (left hand side and central shade areas in (c), respectively).

mixed models (LMMs), we analyzed the following hypotheses: (i) being central may entail a decrease in fitness if the plants' main floral visitors act as antagonist (floral robbers) rather than mutualists (pollinators) [4] and viceversa, and (ii) pollen-flows across species (i.e. layers) induced by the movement of insects increases significantly conspecific pollen loss and heterospecific pollen deposition on stigmas.

2 Results

In 2019 we surveyed the plant abundance, floral visits and seed production in 9 plots of 8.5×8.5 m^2 within Doñana NP (Spain), that were separated by an average distance of 150 m. Each of the resulting plot multilayer contains on average 75.89 ± 30.46 nodes, 67.00 ± 26.19 intralinks, 6.56 ± 1.74 interlinks, 3.67 ± 1.22 plant species with floral visitors, 16.11 ± 4.70 species of insect visitors, and 50.89 ± 21.83 focal plants.

Macro-analyses show that there are 13.00 ± 3.77 modules per multilayer on average, and around a third of them spread over various plant species, although a single plant species predominates. At micro-level, the specialization indices confirm that most plants and animal vistors are connectors among modules, whereas peripherals and network hubs are scarce. Remarkably, PageRank and in-strength show that there is more variance within species than among species. Regarding the mesostructure, we see that HoMs are much more abundant than HeMs. The average number of HoMs among plant species is significantly different in all the plots, whereas the number of HeMs is not.

To relate the seed production of visited plant species to the direct and indirect interactions between their individuals, we fit a LMM with plot and plant species as random factors. Our model shows that (i) the more central a focal plant is (i.e., the less conspecific pollen it receives), the smaller its predicted seed set; and (ii) the larger (smaller) the amount of HoMs (HeMs) triplets of a given individual, the larger (smaller) its seed production

Finally, to further explore the influence of the type of floral visitor (i.e., mutualist, antagonist or other) on individuals' fitness, we compare the results for *Leontodon maroccanus* (LEMA) and *Chamaemelum fuscatum* (CHFU), the most abundant plant species. Their main visitors are strikingly different: 46.5% of LEMA's visitors are antagonists, whereas 79.1% of CHFU's insects are non effective pollinators. We fitted one model per plant species and type of visitor, in which HoMs and HeMs were the explanatory variables. Our LMMs show that motifs with non-antagonist visitors increase significantly seed production of both species. In the case of LEMA, such motifs are mostly HeMs, whereas those of CHFU are HoMs. On the other hand, triplets with antagonist visitors reduce LEMA seed production, but do not have a statistically significant influence on CHFU. However, in the case of CHFU, HeMs mediated by non effective pollinators do induce a significant reduction of seed production.

Our results confirm the hypotheses presented in the introduction, and highlight the crucial role that individual interactions and their variability have for the fitness of plant individuals. In particular, both direct and indirect interactions, and the identity of the interacting visitors, are key for understanding plant reproduction success.

Summary. We introduce a framework rooted in multilayer network modeling designed to depict the conspecific and heterospecific pollen flows mediated by floral visitors. By applying this analysis to 9 well-resolved individual plant-pollinator networks, we show that individual-based networks are highly modular, with modules often integrating individuals from different plant species, linked by their animal visitors. Consequently, the individual node position in the network with respect to its conspecifics or to the overall network have contrasting effects on individual plant reproduction. However, considering the mesoscale enhances the description of plant reproductive success, as it integrates all heterospecific and conspecific interactions of a given individual. We provide a simple but robust set of metrics to scale down network ecology from multitrophic communities to the individual level, where most ecological processes take place, hence moving forward the description and interpretation of ecological dynamics across scales.

References

1. Bascompte, J., and Jordano, P.: Mutualistic networks. Princeton University Press (2014).
2. Bolnick, D.I., *et al.*: Why intraspecific trait variation matters in community ecology. Trends in Ecology and Evolution, 26, 183–192 (2011).
3. Gómez, J.M., Perfectti, F., and Jordano, P.: The functional consequences of mutualistic network architecture, PLoS One, 6, e16143 (2011).
4. Gómez, J.M., and Perfectti, F.: Fitness consequences of centrality in mutualistic individual-based networks, Proc. R. Soc. B 279: 1754–1760 (2012).
5. Boccaletti, S. et al. The structure and dynamics of multilayer networks. Phys. Rep. 544: 1–122 (2014).
6. Miele, V., Ramos-Jiliberto, R., and Vázquez, D.P.: Core–periphery dynamics in a plant–pollinator network. Journal of Animal Ecology 89, Issue 7, 1670–1677 (July 2020).
7. Farage, C., *et al.*: A dynamical perspective to community detection in ecological networks. Biorxiv. DOI: 10.1101/2020.04.14.040519 (2020).
8. Gómez, S.: Centrality in Networks: Finding the Most Important Nodes. In: Moscato P., de Vries, N. (eds) Business and Consumer Analytics: New Ideas. Springer, Cham (2019).
9. Olesen, J. M. *et al.*: The modularity of pollination networks. PNAS 11, 2007 104 (50) 19891–19896 (December 2007).
10. Simmons, B.I., et al.: Motifs in bipartite ecological networks: uncovering indirect interactions. Oikos 128: 154–170, (2019).

Sequential and parallel generation of Artificial Benchmark for Community Detection (ABCD) graphs

Bogumił Kamiński[1], Tomasz Olczak[2], and Paweł Prałat[3] François Théberge[4]

[1] SGH Warsaw School of Economics, Poland, bkamins@sgh.waw.pl
[2] SGH Warsaw School of Economics, Poland, tolczak@gmail.com
[3] Ryerson University, Canada, pralat@ryerson.ca
[4] Tutte Institute for Mathematics and Computing, Canada, theberge@ieee.org

1 Introduction

The standard and extensively used method for generating artificial networks is the **LFR** graph generator [6]. This model has some scalability limitations and it is challenging to analyze it theoretically. Moreover, the mixing parameter μ, the main parameter of the model guiding the strength of the communities, has a non-obvious interpretation and so can lead to unnaturally-defined networks (see [4] for a detailed discussion).

We provide an alternative random graph model with community structure and power-law distribution for both degrees and community sizes, the **A**rtificial **B**enchmark for **C**ommunity **D**etection (**ABCD** graph). We show that the new model solves the three issues identified above and more. Indeed, it is fast, simple, and can be easily tuned to allow the user to make a smooth transition between the two extremes: pure (independent) communities and random graph with no community structure.

The research was partially financed with the support of financed by the Polish National Agency for Academic Exchange, grant agreement no. PPI/APM/2018/1/00037.

2 Results

2.1 ABCD Models

We briefly discuss the **ABCD**—full details can be found in [4]. As with **LFR**, for a given number of vertices *n*, we start by generating a power law distribution both for the degrees and community sizes. Those are governed by the power law exponent parameters (γ, β). We also provide extra information to the model, again as with **LFR**, namely, the average and maximum degree, and the range for the community sizes.

For each community, we generate a random *community* subgraph using either the Configuration Model (**CM**, see [2]) which preserves the exact degree distribution, or the Chung-Lu model (**CL**, see [3]) which preserves the expected degree distribution. We also generate a *background* random graph with the same degree distribution. The mixing parameter ξ guides the proportion of edges which are generated via the background graph. In particular, when $\xi = 1$, the graph has no community structure while with $\xi = 0$ we get disjoint communities. In order to generate simple graphs, we may have to do some re-sampling or edge re-wiring, which are described in [4]. This two-step process is similar to the highly scalable **BTER** model [5].

With this process, larger communities will get slightly more internal edges due to the background graph. In order to provide a variant where the expected proportion of internal edges is the same for every community (as with **LFR**), we also provide a "local" version of **ABCD** where the mixing parameter ξ is adjusted for every community.

2.2 Properties

We compare graphs generated with the **LFR** and the **ABCD** benchmarks via some graph statistics: clustering coefficient (the average vertex transitivity), eigenvector centrality, the global transitivity, and the average shortest paths length (approximated via sampling). We generated graphs with 100,000 vertices, average degree 25, maximum degree 500 and power law exponent $\gamma = 2.5$; for the community sizes, we used power law exponent $\beta = 1.5$ with sizes between 50 and 2,000. The mixing parameter for **LFR** is set to $\mu = 0.2$ and, in order to compare similar graphs, for the **ABCD** algorithm we derive the corresponding $\xi = 0.202$.

The experiments with **ABCD** and **LFR** models show high similarity of the generated graphs, in particular, when the configuration model is used. Indeed, some graph parameters that are sensitive with respect to the degree distribution (such as clustering coefficient) are not as well preserved for the Chung-Lu variant of the model, which is natural and should be expected.

2.3 Performance

The computations for **LFR** were performed using the reference implementation[5] and NetworKit implementation (which is faster). For **ABCD**, the Julia 1.5 language implementation was used [1] in order to ensure high performance of graph generation, while keeping the size of the code base small. We tested the **ABCD** model and we see a roughly 100-fold speedup with the **ABCD** models in a sequential processing approach vs. the reference implementation of **LFR** and around 10-fold speedup vs. NetworKit implementation. See [4] for a discussion of theoretical complexity of **ABCD** and **LFR**.

Let us now switch to issues of parallelization of **ABCD**. The model is conveniently designed in such a way that it is relatively straightforward. In order to create **ABCD**, one needs to create community graphs and the background graph. It is important that community graphs can be generated completely independently so their generation can be easily done in a distributed fashion. Generation of the background graph is performed in two stages: a) fully independent phase (as done with community graphs), b) edge conflict resolution phase between background and community graphs. Fortunately, the number of conflicts that need to be resolved in phase b) is very low and does not affect significantly the overall runtime of the algorithm. Finally, before we move to description of the parallelization algorithm let us note that parallel generation of a single **CL** or **CM** graphs is possible following the procedures described, for example, in Section 6.1 and, respectively, Section 6.2 of [7].

The cost of generation of a graph (regardless whether community or background) is generally proportional to the number of edges that are present in this graph. Let us

[5] `github.com/eXascaleInfolab/LFR-Benchmark_UndirWeightOvp`

denote it by e_i, where $i \in [b]$ (there are $b-1$ community graphs and one background graph). Generation of different graphs is perfectly parallel as noted above, however, if we want to generate some single graph on k processors in parallel the expected execution time is reduced by a multiplicative factor $p(k)$. Clearly $p(1)=1$ and this function is typically decreasing, but $k \cdot p(k)$ is typically increasing. E.g., on our test machine we have measured that $p(4) \approx 0.475$ and $4p(4) \approx 1.9$.

In general, we use a greedy knapsack algorithm, in which the number of bags is the count of available processors c, to dynamically allocate processors to computing tasks. This simple algorithm is known to have a relatively good performance in practice. However, a standard greedy knapsack packing assumes that the items (in our case generation of m subgraphs) are indivisible. Therefore we implement a modification to this algorithm that considers splitting one job of size e_i into k jobs of size $e_i \cdot p(k)$ for $k \in [c]$. In order to keep the algorithm fast we also perform this splitting greedily. The procedure maintains a vector k_i which keeps track to how many processors the generation of graph i is split into. Initially, all $k_i = 1$. Then, we sequentially pick the most expensive job $e_i \cdot p(k_i)$ from the jobs for which $k_i < c$ and consider splitting it into $k_i + 1$ jobs. If this split is beneficial, then we perform it and repeat the process. Otherwise, we stop the algorithm. The idea behind such a procedure is that splitting small items would not lead to significant decreases in overall run time anyway. The details of technical implementation of this procedure depend on the target architecture of graph generation that is considered (multithreading on a single CPU, distributed computing, GPU computing). However, our initial tests show that using 4 CPUs on a standard laptop gives us at least 2-fold speedup over a sequential algorithm (with a worst case being a graph consisting only of a background graph).

Summary. We propose a new method for generating graphs with communities that is not only faster but also has a more natural interpretation than the current state of the art.

References

1. Bezanson, J., Edelman, A., Karpinski, S., and Shah, V.: Julia: A fresh approach to numerical computing. SIAM Review, 69, 65–98 (2017)
2. Bollobás, B.: A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. European Journal of Combinatorics, 1, 311–316 (1980)
3. Chung, F. and Lu, L.: Complex Graphs and Networks. American Mathematical Society (2006)
4. Kamiński, B., Prałat, P., and Théberge, F.: Artificial benchmark for community detection (ABCD): Fast random graph model with community structure. pre-print, arXiv:2002.00843 (2020)
5. Kolda, T. G., Pinar, A., Plantenga, T., and Seshadhri, C.: A scalable generative graph model with community structure. SIAM Journal on Scientific Computing, 36(5), C424–C452 (2014)
6. Lancichinetti, A., Fortunato, S., and Radicchi, F.: Benchmark graphs for testing community detection algorithms. Physical Review E, 78 (2008)
7. Penschuck, M., Brandes, U., Hamann, M., Lamm, S., Meyer, U., Safro, I., Sanders, P., and Schulz, C.: Recent advances in scalable network generation, pre-print, 2020.

Unintended communities in hyperbolic networks

Bianka Kovács[1] and Gergely Palla[1,2]

[1] Dept. of Biological Physics, Eötvös Loránd University,
H-1117 Budapest, Pázmány P. stny. 1/A, Hungary
[2] MTA-ELTE Statistical and Biological Physics Research Group,
H-1117 Budapest, Pázmány P. stny. 1/A, Hungary

1 Introduction

A remarkable approach for modelling the statistical properties of real networks is provided by hyperbolic network models, centred around the idea of placing nodes in a low dimensional hyperbolic space and introducing random connections according to probabilities that depend on the hyperbolic distance between the nodes. The random graphs generated by e.g. the popularity-similarity optimisation (PSO) model [1] and the $\mathbb{S}^1/\mathbb{H}^2$ model [2, 3] are known to be small-world, highly clustered and scale-free at the same time, reproducing the most important universal features often seen in real systems. In the present work, we find that the hyperbolic networks obtained from the above models also contain communities for a rather wide range of the model parameters, which comes as a surprise as the appearance of communities was certainly not an intention at the construction of these models. Nevertheless, since communities usually provide very important units at an intermediate level of the structural organisation of real systems as well, these results make the hyperbolic approach even more suitable for modelling real networks than thought before.

We generated hyperbolic random graphs with both the PSO model and the $\mathbb{S}^1/\mathbb{H}^2$ model, working in the native disk representation of the hyperbolic plane. In the PSO model, the N number of nodes are introduced one by one with logarithmically increasing radial coordinates and uniformly random angular coordinates, and a newly appearing node (indexed by i) connects to previous ones (indexed by j) with a probability depending on the hyperbolic distance x_{ij} as $p(x_{ij}) = \left[1 + e^{\frac{\zeta}{2T}(x_{ij}-R_i)}\right]^{-1}$, where ζ is given by the hyperbolic curvature $K = -1$ as $\zeta = \sqrt{-K}$, the temperature T is a model parameter controlling the clustering coefficient, and the cutoff distance R_i is so adjusted that the expected number of connections of the new node i is equal to m, providing a further model parameter (that is equal to the half of the expected average degree $\langle k \rangle$). Roughly speaking, the degree of the nodes is determined by their radial coordinate, and the first appearing nodes close to the centre of the disk (being the most popular) eventually become hubs. However, an important feature of the model is that at the arrival of any node i, the radial coordinate of each previously (at time $j < i$) appeared node $j = 1, 2, ..., i-1$ is increased according to $r_{ji} = \beta r_{jj} + (1-\beta) r_{ii}$, in order to simulate popularity fading. With the help of the parameter β controlling the popularity fading we can set the decay exponent γ of the scale-free degree distribution as $\gamma = 1 + 1/\beta$.

In the $\mathbb{S}^1$ model the N number of nodes are placed on a circle at random angular coordinates and to each node we also associate a hidden variable κ drawn from

$\rho(\kappa) \propto \kappa^{-\gamma}$. Node pairs are connected with a probability depending on both the angular distance $\Delta\theta$ and the hidden variables formulated as $p_{ij} = \left[1 + \left(\frac{N \cdot \Delta\theta_{ij}}{2\pi \cdot \mu \cdot \kappa_i \cdot \kappa_j}\right)^{\alpha}\right]^{-1}$, where the constant μ is given by the model parameters γ, α and $\langle k \rangle$ as $\mu = \frac{\alpha}{2\pi\langle k \rangle} \cdot \sin\left(\frac{\pi}{\alpha}\right)$. In the equivalent $\mathbb{H}^2$ model the κ parameter associated to the nodes is converted into a hyperbolic radial coordinate in the native disk, with which the connection probability becomes similar to the linking probability seen in the PSO model.

The random graphs generated with the PSO model and the $\mathbb{S}^1/\mathbb{H}^2$ model were used as inputs for three very well-grounded community finding methods, namely the asynchronous label propagation [4], the Louvain [5] and the Infomap [6] algorithms. The quality of the communities was measured by the weighted modularity $Q = \frac{1}{2M} \cdot$ $\cdot \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{N} \left[w_{ij} - \frac{s_i s_j}{2M}\right] \delta_{c_i,c_j}$, where the link weights are defined following the practice suggested in Ref. [7] as $w_{ij} = \frac{1}{1+x_{ij}}$, the node strength s_i is simply $s_i = \sum_{\ell=1}^{N} w_{i\ell}$ and $M = \frac{1}{2} \cdot \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{N} w_{ij}$ stands for the total sum of the link weights.

2 Results

We generated random graphs with sizes varying between $N = 100$ and $N = 10,000$ for a wide range of parameter settings and found quite robust community structures with high Q values for the majority of the cases. As an illustration, in Fig. 1. we show an example for the communities in an $\mathbb{S}^1/\mathbb{H}^2$ network. In Fig. 2., we plot Q as a function of T and β in the PSO-model for $N = 10,000$ and $\langle k \rangle = 10$, displaying convincing high values in a relatively large region of the parameter plane. Based on similar plots obtained for the $\mathbb{S}^1/\mathbb{H}^2$ model, we can state that the examined hyperbolic network models yield an inherent community structure for a wide range of their parameters, despite the absence of any intentional community formation mechanisms built into the model construction.

References

1. Papadopoulos, F., Kitsak, M., Serrano, MÁ., Boguñá, M., Krioukov, D.: Popularity versus similarity in growing networks. Nature 489, 537–540 (2012)
2. Serrano, MÁ., Krioukov, D., Boguñá, M.: Self-Similarity of Complex Networks and Hidden Metric Spaces. Phys. Rev. Lett. 100(7), 078701 (2008)
3. García-Pérez, G., Allard, A., Serrano, MÁ., Boguñá, M.: Mercator: uncovering faithful hyperbolic embeddings of complex networks. New J. Phys. 21(12), 123033 (2019)
4. Raghavan, UN., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E 76(3), 036106 (2007)
5. Blondel, VD., Guillaume J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. 2008(10), P10008 (2008)
6. Rosvall, M., Bergstrom, CT.: Multilevel Compression of Random Walks on Networks Reveals Hierarchical Organization in Large Integrated Systems. PLOS ONE 6(4), 1–10 (2011)
7. Muscoloni, A., Thomas, JM., Ciucci, S., Bianconi, G. Cannistraci, CV.: Machine learning meets complex networks via coalescent embedding in the hyperbolic space. Nature Communications 8(1), 1615 (2017)

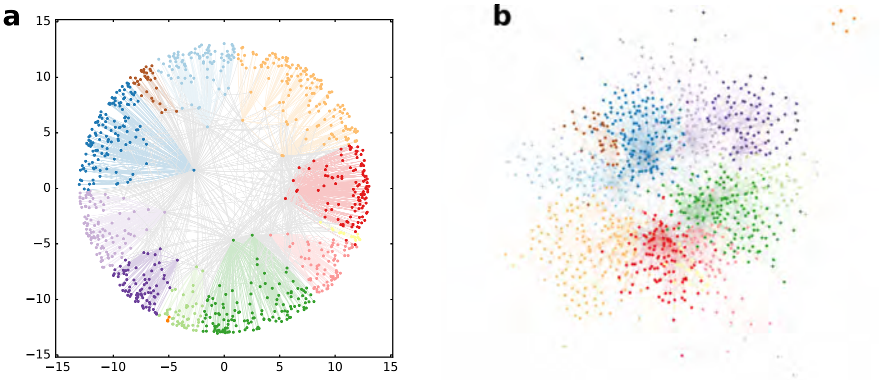


Fig. 1. Example for communities found by the Louvain algorithm in a network generated by the $\mathbb{S}^1/\mathbb{H}^2$ model. The model parameters were $N = 1000$, $\langle k \rangle = 10$, $\gamma = 2.43$ and $\alpha = 5.0$. The modularity of the found community partition reached $Q = 0.738$. a) Layout in the native representation of the hyperbolic plane of curvature $K = -1$, using the hyperbolic coordinates given by the model. b) Force-directed layout in the Euclidean plane.

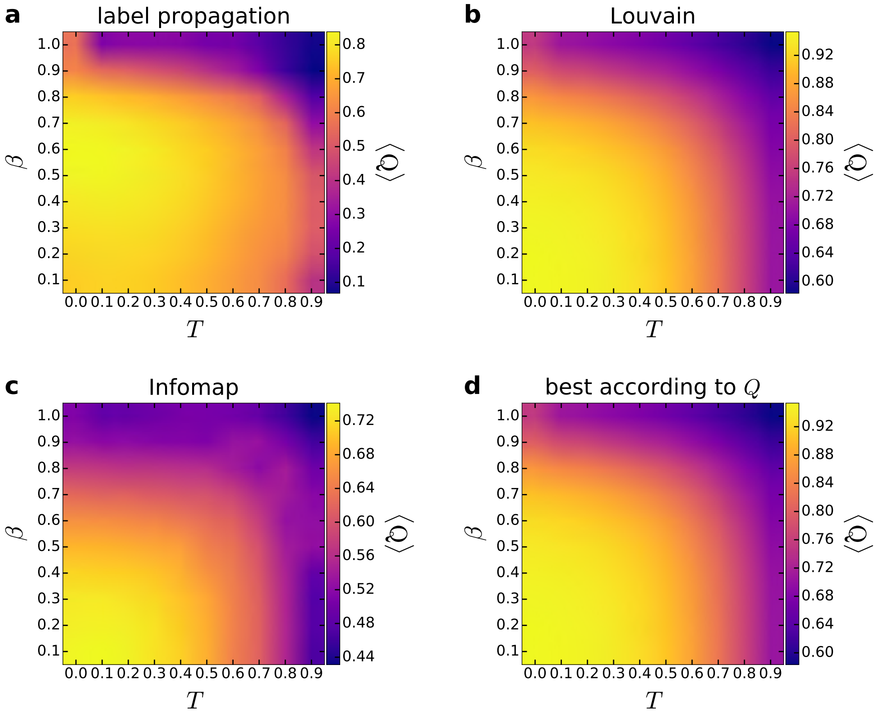


Fig. 2. The modularity as a function of the model parameters in the case of the PSO model. Q was averaged over 100 samples of networks with $N = 10{,}000$ and $\langle k \rangle = 10$ at $\beta = 0.1, 0.2, ..., 0.9, 1.0$ and $T = 0.0, 0.1, ..., 0.8, 0.9$. The panels correspond to the results obtained with asynchronous label propagation (a), Louvain (b), Infomap (c), and when the best community partition is taken from the three methods according to the modularity (d).

A Community-Aware Backbone Extractor for Weighted Networks

Zakariya Ghalmane[1,3], Chantal Cherifi[2], Hocine Cherifi[3], and Mohammed El Hassouni[1]

[1] LRIT, Mohammed V University, Rabat, Morocco,
[2] DISP Lab, University of Lyon 2, Lyon, France
[3] LIB, University of Burgundy, Dijon, France

1 Introduction

Network science provides effective tools to model and analyze complex systems. However, the increasing size of real-world networks becomes a major obstacle in order to understand their structure and topological features. Therefore, reducing its size while preserving its topological features is an important issue. Extracting the so-called backbone of a network is generally solved either by coarse-graining or filter-based methods. Coarse-graining methods reduce the network size by grouping similar nodes, while filter-based methods map the network by discarding nodes or edges based on a statistical property. In this work [1], we propose and investigate a filter-based method exploiting the community structure [2, 3] in weighted networks. In the so-called "overlapping nodes ego backbone", the backbone is formed with the set of nodes shared by communities called the overlapping nodes [4] and their neighbors. It is inspired from a previous study [7] showing that overlapping nodes and hubs are generally neighbors in real-world networks. Once the sub network made of the overlapping nodes and their neighbors is extracted, links with low weights are removed as long as the biggest connected component is preserved.

Experimental evaluation has been performed with real-world weighted networks originating from various domains (social, co-appearance, collaboration, biological, and technological). Comparisons with the popular disparity filter method [5] demonstrate the greater ability of the proposed method to uncover the most relevant parts of the network.

2 Methods

The algorithm used to extract the backbone is detailed below:

Step 1: Form the overlapping nodes ego sub-network by removing all nodes that do not belong to the set of overlapping nodes or the set of their first neighbors.

Step 2: Remove the edges with low weights from the overlapping ego sub-network. To do so, edges are sorted in decreasing order according to their weights. Low weights are removed as long as the sub-network largest component do not split into two components.

Step 3: Tune the size of the overlapping ego backbone with a parameter s. To this end, all the nodes of the overlapping nodes ego backbone are sorted in decreasing order according to the weighted degree centrality [6]. Nodes with low degrees are removed from the network until the prescribed size is reached.

Table 1. N is the network size. A_n represents the estimated values the proportion of common nodes of two backbones. $< \beta >$ is the average node betweenness. $< w >$ is the average link weight. OE stands for the overlapping nodes ego backbone, while DF stands for the disparity filter backbone.

Network	N	$A_n(\%)$	$<\beta>$		$<w>$	
		OE-DF	**OE**	**DF**	**OE**	**DF**
Zachary's karate club	34	68	0.088	0.079	3.31	3.15
Intra-organisational	46	84.61	0.028	0.013	2.31	2.18
Freeman's EIES	48	80	0.014	0.011	2.53	2.14
Train bombing	62	94.73	0.077	0.047	1.38	1.23
Les Miserables	77	85.11	0.057	0.035	4.89	3.74
Game of thrones	107	68.75	0.053	0.034	16.58	14.98
C.elegans Neural	306	64.04	0.012	0.009	5.32	4.91
Facebook-like Forum	899	61.71	0.005	0.004	6.99	5.61
Facebook-like Social	1899	75.04	0.003	0.002	356.19	313.51
US Power Grid	4941	61.03	0.008	0.005	53.59	49.55
Scientific Collaboration	16726	57.83	0.008	0.006	49.97	48.52

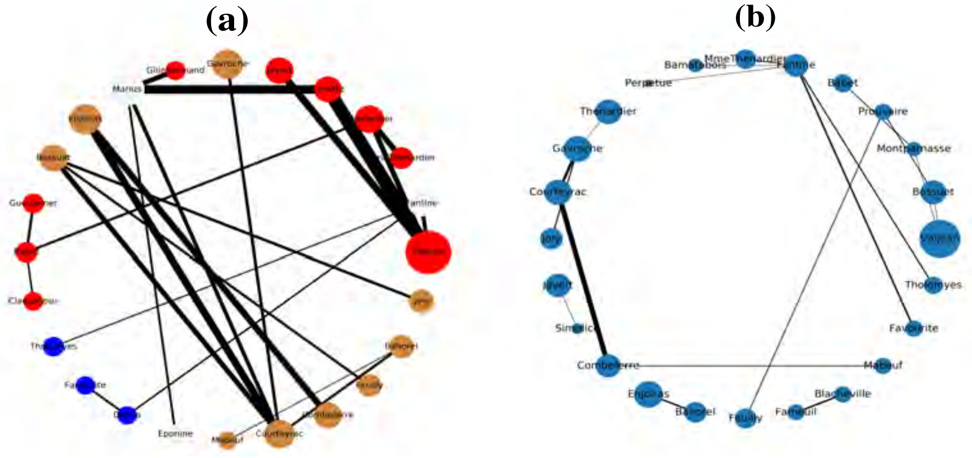


Fig. 1. The backbone extraction of different methods for Les Miserables network. (a) overlapping nodes ego backbone, (b) disparity filter backbone. Nodes with the same color belong to the same community and those in gray are the overlapping nodes. The size of the nodes is proportional to their weighted degree, while the size of links is proportional to their weights.

3 Results

A set of experiments is performed to compare the overlapping nodes ego backbone with the disparity filter which is recognized as one of the most effective alternative method. Parameters values of the backbone extractor are tuned in order to compare backbone of the same size (30% of the size of the original network). The SLPA algorithm is used to uncover the overlapping community structure of the networks. At first, the proportion of common nodes extracted by the two methods is computed. It is defined as the fraction of the size of the intersection between the two sets divided by their size. Table 1 reports the proportion of common nodes computed between the proposed backbone and the disparity filter. Overall, the overlap is more or less pronounced. Les Miserables network reported in Figure 1 is a good illustration that the proposed approach preserves almost all high-connectivity nodes and essential connections. Indeed, one can notice that the disparity filter backbone misses some very important nodes such as "Marius" and "Cosette" that are among the main characters in the Victor Hugo's novel.

Second, the performance is compared by measuring the average betweenness and the average link weight of both backbone extractors. Table 1 reports the results for all networks under test. The average betweenness indicates how much information can pass through the nodes of the backbone. The values of the average betweenness of the proposed backbone are higher than the ones computed in the disparity filter backbone. This implies that the nodes extracted by the overlapping nodes ego backbone act as a better information gateway of the original network as compared to the disparity filter backbone. Experimental results show that the backbones extracted by the proposed method have a higher average link weight as compared to the disparity filter backbones. A higher value of the average link weight demonstrates that the picked links are quite relevant.Thus, some very relevant connections are missed in the disparity filter backbone. All these experiments confirm the ability of the overlapping nodes ego backbone to preserve nodes playing a major role in the network.

References

1. Ghalmane, Z., Cherifi, C., Cherifi, H., & Hassouni, M. E. (2020). Extracting Backbones in Weighted Modular Complex Networks. Scientific Reports,10,1, https://doi.org/10.1038/s41598-020-71876-0
2. Cherifi, H., Palla, G., Szymanski, B. K. & Lu, X., (2019). On community structure in complex networks: challenges and opportunities, Applied Network Science,4,1,SpringerOpen.
3. Gupta, N., Singh, A. & Cherifi, H.(2015). Community-based immunization strategies for epidemic control 7th int. conf. on communication systems and networks, proc. IEEE, 1–6.
4. Jebabli, M., Cherifi, H., Cherifi, C. & Hammouda, A. (2014) Overlapping community structure in co-authorship networks: A case study, 7th Int. Conf. on u-and e-Service, Science and Technology, Proc. IEEE, 26–29.
5. Serrano, M. Á., Boguná, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. Proceedings of the national academy of sciences, 106(16), 6483-6488.
6. Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. Social networks, 32(3), 245-251.
7. Ghalmane, Z., Cherifi, C., Cherifi, H., & El Hassouni, M. (2020). Exploring Hubs and Overlapping Nodes Interactions in Modular Complex Networks. IEEE Access, 8, 79650-79683.

The node2community prediction problem in complex networks

Ilyes Abdelhamid[1,2], Alessandro Muscoloni[1] and Carlo Vittorio Cannistraci[1,3]

[1] Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Cluster of Excellence Physics of Life (PoL), Department of Physics, Technische Universität Dresden, Germany; [2] Lipotype GmbH, Dresden, Germany; [3] Center for Complex Network Intelligence (CCNI) at the Tsinghua Laboratory of Brain and Intelligence (THBI), Department of Bioengineering, Tsinghua University, Beijing, China

1 Introduction

Many real networks have a multifaceted community organization that supports the trade-off between functional segregation and integration [1], which is a crucial feature of complex systems. Connections between a known community and other nodes in the rest of the network are pivotal bridges that might suggest integration with other functional modules. Here, we introduce a novel challenge in network science: the node2community (node2com) prediction. Given only a network topology and the metadata of one or more annotated communities, the problem is to predict whether there are nodes not connected to a certain community that might be candidates for a significant pairing. The pairing prediction can quantify the statistical significance of a potential structural or functional interaction, which does not necessarily imply the prediction of node membership to the community. To address such problem, we propose an algorithm for node-community pairing (NCP) that exploits an ensemble of link predictors and that provides a level of statistical significance for each node-community pair that does not have a direct connection.

2 Results

NCP has several possible variants that have been evaluated on both real and synthetic networks. The results from our analysis confirm that the Cannistraci-Hebb (CH) models [2] are robust predictors across different network topologies, and the adoption of models based on paths of length three (L3) seems more reliable than paths of length two (L2) on most network structural organizations. Finally, as proof of real application in systems biomedicine, we run NCP on the human protein interactome [3] in order to predict unknown associations with the COPD disease module [4], [5], which are then experimentally validated by co-immunoprecipitation and gene silencing.

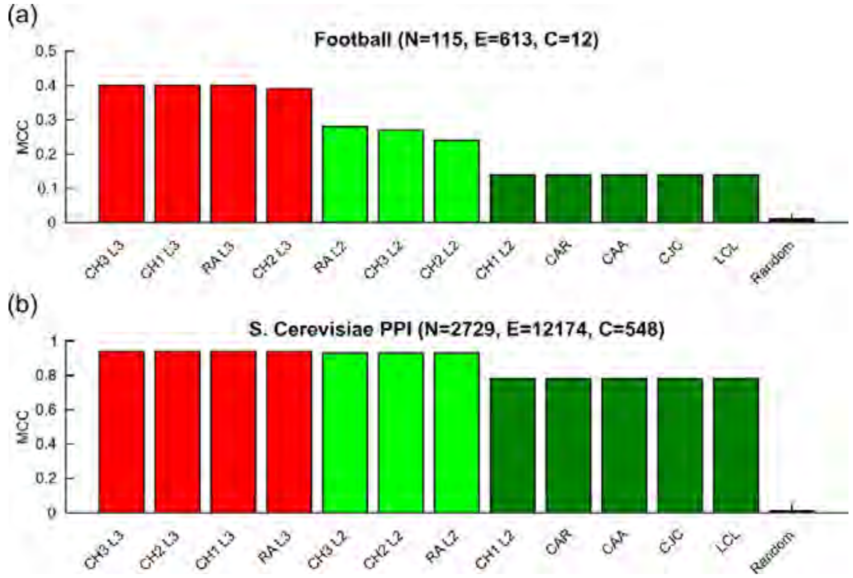


Fig. 1. MCC evaluation on real networks. The figure reports the results on 2 representative real networks out of the 5 that have been evaluated: (a) Football, a social network without overlapping communities and (b) S. Cerevisiae PPI, a biological network with overlapping communities. Number of nodes N, edges E and communities C are reported in brackets. For each community in the network, we define the "removable" nodes, which are the nodes with at least one link internal and one link external to the community members. All the pairs composed of a community and a removable node represent the positive set. For each node-community pair in the positive set, we define a related "negative" pair, which is given by the same node and the community whose members have the highest average shortest path to the node (only considering the communities of which the node is not member and does not have a direct link). All such pairs represent the negative set. For each node-community pair in the positive set, we remove the internal links between the removable node and the community members, and we run the NCP prediction algorithm in order to obtain a p-value for both the communities associated to the node in the positive and negative set pairs. By applying a significance threshold of 0.05 to the p-values, we obtain the positive and negative predictions to the node-community pairs in the positive and negative sets. Then, we evaluate the performance computing the Matthews correlation coefficient (MCC) using the true and predicted labels. The NCP algorithm can be performed adopting several link prediction variants, which are all reported in the barplots, ranked by decreasing MCC. The random predictor is also shown. The results for the other 3 real networks are not reported due to lack of space, however they highlight analogous trends.

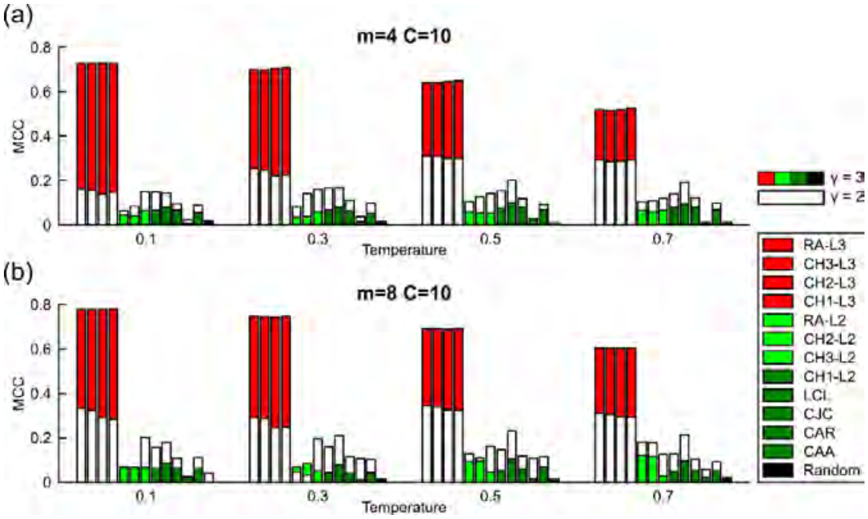


Fig. 2. MCC evaluation on synthetic nPSO networks. We generated synthetic networks using the nPSO model [6] with parameters N = 1000 (nodes), γ = [2, 3] (exponent of the power-law degree distribution), m = [4, 8] (half of the average node degree), T = [0.1, 0.3, 0.5, 0.7] (temperature, inversely related to the clustering) and C = [10, 20] (number of communities). For each network, we have considered all the node-community pairs such that the node is a "candidate" node for the community (it is not a member of the community or a neighbor of the members of the community). For each node-community pair we have assigned a hyperbolic distance equal to the minimum hyperbolic distance between the candidate node and the members of the community. Finally, the positive set is represented by the 5% of the node-community pairs having the lowest hyperbolic distance, whereas the negative set by the 5% of the node-community pairs having the largest hyperbolic distance. For each node-community pair in the positive and negative sets we run the NCP algorithm in order to obtain the associated p-values. By applying a significance threshold of 0.05 to the p-values, we obtain the positive and negative predictions. Finally, we can evaluate the prediction using the MCC. The barplots report the results of the NCP variants over different parameter combinations of the nPSO model. In particular: (a) m = 4 and C = 10, (b) m = 8 and C = 10; the results for different temperature values are shown over the x-axis; the results for different gamma values are reported with white versus colored bars. The results for C = 20 are not reported due to lack of space, however they highlight analogous trends.

References

[1] M. Xu, Q. Pan, A. Muscoloni, H. Xia, and C. V. Cannistraci, "Modular gateway-ness connectivity and structural core organization in maritime network science," *Nat. Commun.*, 2020.

[2] A. Muscoloni, I. Abdelhamid, and C. V. Cannistraci, "Local-community network automata modelling based on length-three-paths for prediction of complex network structures in protein interactomes, food webs and more," *bioRxiv*, 2018.

[3] J. Menche *et al.*, "Uncovering disease-disease relationships through the incomplete interactome," *Science*, 2015.

[4] A. Halu *et al.*, "Exploring the cross-phenotype network region of disease modules reveals concordant and discordant pathways between chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis," *Hum. Mol. Genet.*, vol. 28, no. 14, pp. 2352–2364, 2019.

[5] A. Sharma *et al.*, "Integration of Molecular Interactome and Targeted Interaction Analysis to Identify a COPD Disease Network Module," *Sci. Rep.*, vol. 8, no. 1, 2018.

[6] A. Muscoloni and C. V. Cannistraci, "A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities," *New J. Phys.*, vol. 20, 2018.

Part III

Diffusion and Epidemics

Infectious disease dynamics in homophily-driven dynamic small-world networks: A model study.

Hendrik Nunner[1,2,5], Vincent Buskens[1,2], and Mirjam Kretzschmar[2,3,4]

[1] Department of Sociology/ICS, Utrecht University, Padualaan 14, 3584 CH, Utrecht, The Netherlands
[2] Centre for Complex Systems Studies, Utrecht University, Leuvenlaan 4, 3584 CE Utrecht, The Netherlands
[3] Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands
[4] Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), P.O. Box 1, 3720 BA Bilthoven, The Netherlands
[5] Corresponding author: h.nunner@uu.nl

Introduction & theoretic background

The COVID-19 pandemic has shown how vulnerable our globalized world is to the threat of infectious diseases. To reduce the number of new infections many countries have implemented forms of social distancing. However, not only external interventions alter network structure. Avoidant health behavior, such as avoiding coworkers with symptoms or avoiding sexual contact because a partner has an STD, is known to affect pathways of infection [4]. Further, we know that health behavior is affected by individual risk perceptions making different people behave differently in similar situations [1], while people like to mix with others who behave similarly [5]. Additionally, Coleman describes a dense network as primary source for social capital [3], while Burt describes the benefits of bridging structural holes in the network [2], two mechanisms to explain why our social networks are small-worlds with high clustering and low average path length [7].

These considerations ultimately result in homophilous clusters of heterogeneous health behavior in which agents deliberately distance themselves from a disease. The effect of these social dynamics in small-worlds on the disease dynamics, however, remain unclear. We therefore ask: *How does health behavior homophily shape epidemics in dynamic small-world networks?*

Model & methods

For the current study we build on a previously developed model that integrates theories from sociology (network formation), health psychology (risk perception), and epidemiology (compartmental models) [6]. This ego-centered, utility-based model describes social networking in the context of infectious diseases as a trade-off between the benefits, efforts, and potential health damage a social tie creates. Thus, risk avoiding agents may break ties to infected peers, while risk seeking agents may not. Further, infections can travel between agents along social ties.

To control network clustering, we added benefits for the proportion of closed triads an agent is part of (α). To control homophily, we added a probability that 2 agents similar in risk perception have contact (ω). Consequently, for high α and high ω networks with homophilous interconnected clusters emerge, while low α and low ω result in randomly mixed networks with a single large cluster. Further, the optimum number of ties per agent is constant to reduce the influence of degree on disease dynamics.

An agent-based simulation served to study the theoretic model. A single simulation run consists of three parts: First, agents form ties until a pairwise stable network emerges (no agent benefits from breaking ties; no pair of agents benefits from creating a tie). Second, after a randomly selected agent (index case) is infected, disease dynamics are simulated without network dynamics (static networks). Third, disease states are reset, the index case is reinfected, and disease dynamics are simulated including network dynamics (dynamic networks). That is, agents modify ties depending on benefits, risk perception, disease severity, and transmission probability. To understand the effect of model parameters on the number of infected agents (attack rate), the duration of the epidemic, and the maximum number of simultaneously infected agents (peak size) we analyzed $80,305$ simulation runs with randomized parameter settings.

Results, discussion, & conclusion

The results show that the presence of homophilous clusters (see table rows ”Clustering”, ”Homophily”), reduce attack rate, duration, and peak size of epidemics in dynamic networks. That is because clusters of agents collectively perceiving high risks of infections are hard to be invaded by a disease for two reasons. First, there are only few bridges, thus reducing the probability for infections to enter the cluster. Second, bridging agents in the risk avoiding cluster perceive health risks to be more severe, thus cutting ties quicker than agents in risk seeking clusters. Although clusters of risk seeking agents may become fully infected, the epidemic remains limited to parts of the network.

	Dynamic networks			**Static Networks**		
	Attack rate	**Duration**	**Peak size**	**Attack rate**	**Duration**	**Peak size**
# of network changes	+	+	+++			
Clustering	−−	−−	−−		+	−
Path length		++	−	−	+++	−
Homophily	−−	−	−−−			−
Av. degree	−	−	−	+	−−−	+
Av. risk perception	−	−	−			
Disease severity	−	−	−			
Pr. dis. transm./contact	++++	−	+	+	−−	+
Adjusted R^2	0.88	0.72	0.73	0.76	0.03	0.85
# of observations	80,305	80,305	80,305	80,305	80,305	80,305

Note: table shows summary and comparison of main effects regression models; all variables are standardized; all effects shown are significant at $p < 0.001$; effect sizes (+: 0-20%, …, +++++: 80-100%) are shown in relation to largest effect (+)

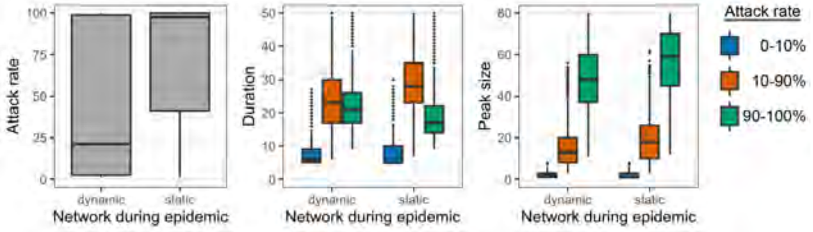


Fig. 1. Epidemics in dynamic and static networks.

Further, we see opposing effects for clustering on duration in dynamic and static networks. In static networks, the disease needs time to travel across bridges to reach other clusters. In contrast, few changes in dynamic networks can isolate infected clusters quickly resulting in lower attack rates and consequently in less time for the disease to disappear from the network.

In line with previous studies, epidemics in dynamic networks have on average lower attack rates, shorter duration, and lower epidemic peaks (Fig. 1). However, when the effect of attack rate is attenuated (attack rates $\geq 90\%$), we see longer duration in dynamic networks. That is, although agents get infected at some point, they delay their infection by distancing themselves from the disease in the early stages of the epidemic.

A finding we cannot confirm, however, is that increases in average path length ought to reduce attack rate [7] (see table row ”Path length”). That is because agents in long path length networks need only a few changes to distance themselves from the disease, making it likely for the disease to have no more pathways to travel along.

In conclusion, we see that homophilous clusters of agents similar in risk perception significantly affect disease dynamics in dynamic small-world networks, affecting attack rate, duration, and peak size of epidemics. Additional findings (opposing or missing effects in dynamic and static networks) further illustrate the importance of considering theoretically sound network dynamics for network models in epidemiology.

References

1. Bish, A., Michie, S.: Demographic and attitudinal determinants of protective behaviours during a pandemic: A review. British Journal of Health Psychology 15(4), 797–824 (nov 2010)
2. Burt, R.S.: Structural holes: The social structure of competition. Harvard university press (2009)
3. Coleman, J.S.: Foundations of social theory. Harvard university press (1994)
4. Funk, S., Gilad, E., Watkins, C., Jansen, V.A.: The spread of awareness and its impact on epidemic outbreaks. Proceedings of the National Academy of Sciences of the United States of America 106(16), 6872–6877 (2009)
5. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology 27(1), 415–444 (aug 2001)
6. Nunner, H., Buskens, V., Kretzschmar, M.E.: A model for the co-evolution of dynamic social networks and infectious disease dynamics, submitted
7. Watts, D.J., Strogatz, S.H.: Collective dynamics of ’small-world’ networks. Nature 393(6684), 440 (1998)

On the Effect of Space on the Spread of Infections

Franz-Benjamin Mocnik

University of Twente, 7500 AE Enschede, the Netherlands
franz-benjamin.mocnik@utwente.nl
https://www.mocnik-science.net

1 Introduction

A good understanding of how pathogens spread is important to mitigate their effect on individuals in terms of diseases and, as a consequence, on society. The recent case of the COVID-19 pandemic illustrates this fact in many ways. The SARS-CoV-2 virus, which causes the COVID-19 disease, can be found in many locations, some of which are unexpected; and mitigation efforts are successful to varying degrees in different places [8, 9]. The explanations given to such observations often refer to the geographical context, which incorporates besides space also cultures, mundane habits, travel behaviour, and many more aspects. While space is of obvious importance for the transmission of a pathogen, it is also one of the most important aspects of the people's travel behaviour, an aspect that is relatively easy to control in case of a pandemic. This short paper raises the question in which ways the structure of space influences the spread of such a pathogen.

Theoretical considerations often focus on the basic reproduction number R_0, i.e., the average number of previously uninfected individuals becoming infected by an infectious person. In case of an exponential growth of infections, this number is constant over time [1]. In practice however, it varies in different contexts and over time, resulting in an effective reproduction number R. When modelling the spread of infections, individuals can be understood as the nodes of a network and their social relations, each of which can lead to the infection of another individual, as edges. In the following, we will examine how the spatial structure of such a network affects the effective reproduction number.

2 Results and Discussion

The effective reproduction number can easily be computed in a spatial network setting, which assumes individuals only to infect others in their spatial vicinity. This property translates to neighbouring nodes in the network to be in the same spatial vicinity, in case of which the network inherits spatial characteristics [2, 7]. Previous epidemiological studies often assume spatial grids [10, 11], in contrast to which this study uses instances of the Mocnik model[1], an example of a non-regular spatial network model [4, 5, 7]. In two-dimensions, each node has $S = \rho^2$ neighbours in average, where ρ is a parameter of the model [4, 5]. This number S will be referred to as the number of social contacts in the following. When choosing S similar to R_0, the spread is prototypically influenced by spatial distance – individuals spread infections only in their direct vicinity.

[1] Other spatial networks yield very similar results. The Mocnik model has been chosen here to blend over between a spatial and a complete network, as is discussed later.

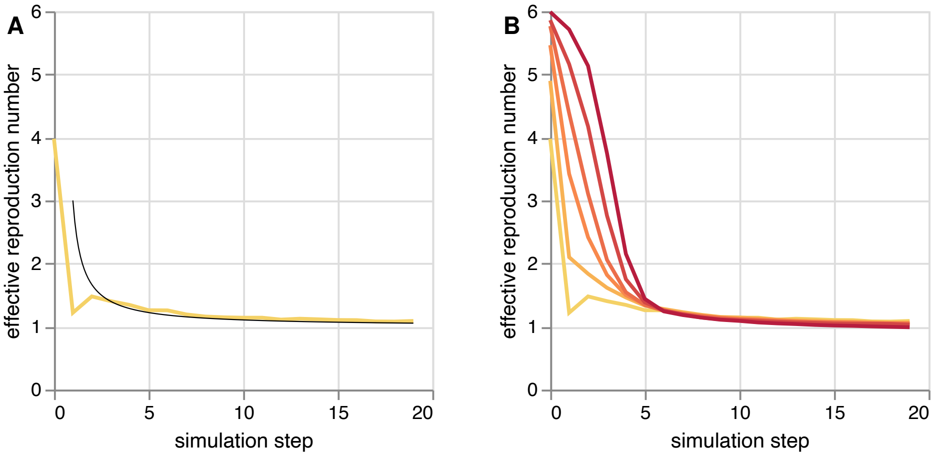


Figure 1: Simulation of infection spread in an instance of the Mocnik model with 100,000 nodes. **A** Spatial case ($R_0 = S = 6$; yellow) compared to the theoretical prediction of Equation 1 (black). **B** Spatial case ($R_0 = S = 6$; yellow) compared with less spatial cases ($R_0 = 6$, $S = 12, 30, 60, 163, 443$; orange to red).

The effective reproduction number R changes during an outbreak in a spatial network. Starting with one node only, more and more nodes get infected over time. As an already infected neighbour of an infected node cannot be infected anew, infections often follow a sub-exponential growth. More specifically, the *Polynomial Volume Law* claims in case of the two-dimensional Mocnik model that $1+n^2$ nodes have statistically been infected until simulation step n if every neighbouring node of an infected one gets infected itself if not yet being infected [4, 5, 7]. Accordingly, $n^2-(n-1)^2$ nodes get infected in step $n-1$. These nodes are infectious in step n and infect $(n+1)^2-n^2$ additional nodes. The statistical average of the effective reproduction number is, in case of this model, thus given by

$$\frac{(n+1)^2-n^2}{n^2-(n-1)^2} = 1+\frac{2}{2n-1}. \tag{1}$$

A comparison to a simulation[2] ($R_0 = S = 6$) confirms this consideration (Figure 1A).

The topologies of real social networks are, despite being spatially influenced, more complex in nature – they borrow characteristics from both spatial and complete networks. As the latter can be considered a special case of a Mocnik model with S approaching the number of nodes, the simulation was run with several values of S for blending over between both types of networks. The simulations show that the effective reproduction factor R increases for larger S but converges to 1 in all cases considered (Figure 1B). Indeed, the circumference of the disk of infected nodes grows only slowly and the ratio of the number of nodes infected in the last step and the ones to infect in the next step converges to 1. At closer inspection, there is little to no influence of S on R after some simulation steps (Figure 2). This suggests that the structure of space has a relevant impact on the spread of infections in a network, also beyond the most prototypical cases.

[2] In the simulation, each node infected in step n is assumed to infect R_0 randomly chosen neighbouring nodes in step $n+1$, or less nodes if the node had less neighbours only.

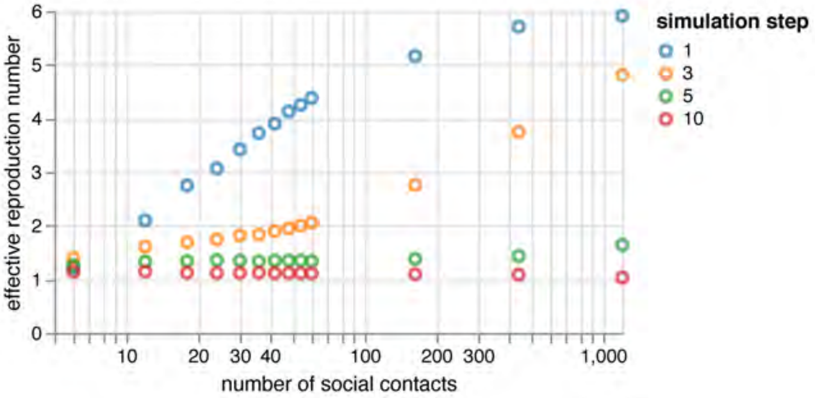


Figure 2: Influence of the number of social contacts on the effective repr. number

3 Conclusion

The study at hand demonstrates that the spread of diseases in a spatial context tends to yield effective reproduction numbers near to 1. The study is, however, limited by the fact that the network topologies of real networks can be more diverse than the one considered here. Hubs [3] and shortcuts in the network have not yet been considered. Both limitations could be addressed by assuming hierarchical Mocnik models [5]. After having addressed these limitations, more realistic scenarios might be considered [6] to understand the (existing or non-existing) impact of closing country borders or motivating people not to travel, which potentially leads to more spatially structured networks. This study is, however, a first indication that the effect of spatial networks on the spread of infections is robust, even in cases where this structure is less dominant.

References

1. Anderson, R.M., May, R.M.: Infectious diseases of humans: dynamics and control. Oxford University Press, Oxford, UK (1991)
2. Barthélemy, M.: Spatial networks. Physics Reports 499(1–3), 1–101 (2011)
3. Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E., Getz, W.M.: Superspreading and the effect of individual variation on disease emergence. Nature 438, 355–359 (2005)
4. Mocnik, F.-B.: A scale-invariant spatial graph model. Ph.D. thesis, Vienna University of Technology (2015)
5. Mocnik, F.-B.: The polynomial volume law of complex networks in the context of local and global optimization. Scientific Reports 8(11274) (2018)
6. Mocnik, F.-B.: An improved algorithm for dynamic nearest-neighbour models. Journal of Spatial Science (2020)
7. Mocnik, F.-B., Frank, A.U.: Modelling spatial structures. Proceedings of the 12th Conference on Spatial Information Theory (COSIT) p. 44–64 (2015)
8. Mocnik, F.-B., Raposo, P., Feringa, W., Kraak, M.-J., Köbben, B.: Epidemics and pandemics in maps – the case of COVID-19. Journal of Maps 16(1), 144–152 (2020)
9. Pearce, N., Lawlor, D.A., Brickley, E.B.: Comparisons between countries are essential for the control of COVID-19. International Journal of Epidemiology (2020)
10. Rüdiger, S., Plietzsch, A., Sagués, F., Sokolov, I.M., Kurths, J.: Epidemics with mutating infectivity on small-world networks. Scientific Reports 10, 5919 (2020)
11. Sun, G.-Q., Jusup, M., Jin, Z., Wang, Y., Wang, Z.: Pattern transitions in spatial epidemics: mechanisms and emergent properties. Physics of Life Reviews 19, 43–73 (2016)

Not all interventions are equal for the height of the second peak

Joost Jorritsma[1], Tim Hulshof[2], and Júlia Komjáthy[1]

[1] Eindhoven University of Technology, Department of Mathematics and Computer Science
P.O. Box 513, 5600 MB Eindhoven, The Netherlands,
[2] Bureau WO, Eindhoven, The Netherlands.
j.jorritsma@tue.nl, timhulshof@bureauwo.nl, j.komjathy@tue.nl.

1 Introduction

When recovering from a disease grants temporary immunity against it, it can happen that an epidemic dies out locally, but survives elsewhere, returning at a later point in time. Immunity being only temporary has been observed for various types of viruses, see [2] and references therein, and the first re-infections of COVID-19 have been determined [7]. As a result we may observe a "second peak", which can also happen when interventions (that may affect traveling behavior) are effectively applied to slow down the spread of a disease locally, but are then lifted. The most popular models for epidemic curve modeling are standard compartmental models, see e.g. [4,8], assuming a perfectly mixed a-geometric and continuum population. Compartmental models can often be described by partial differential equations, making them analytically tractable and suitable for parameter estimation. However, the absence of an underlying geometry makes intuitive modeling of epidemics under interventions in compartmental models harder, if not infeasible. Interventions (e.g. travel restrictions) make the population even less-mixed than under normal circumstances, hence making geometry more important.

We present results from a recent paper [3] where we analyze a spreading model with temporary immunity on geometric inhomogeneous random graphs, a state-of-the-art spatial network model [1]. The spatial embedding of the network allows for intuitive modeling of intervention strategies, such as social distancing, traveling restrictions, and limiting the number of social contacts. We conduct a simulation study to compare the spatio-temporal behavior of epidemics with temporary immunity under these strategies. The simulation results may serve as a *qualitative indication* of possible outcomes of epidemic spread and intervention strategies for infectious diseases, such as the COVID-19 pandemic.

2 Spreading model with temporary immunity on a network

We describe a random spatial network model to model human contact networks. In this network every node u is equipped with a location $x_u \in \mathbb{R}^2$. We think of nodes in the network as individuals and of their location as place of living. The neighbors of a node u in the network G are nodes with a direct connection (also called edge) to u. A connection corresponds to a (recurring) contact event.

Definition 1 (Geometric Inhomogeneous Random Graph (GIRG) [1]). *Fix $N \geq 1$ the number of nodes. Assign to each node $u \in \{1,2,\dots,N\}$ a* fitness *$w_u > 0$, and a uniformly chosen location $x_u \in [0,\sqrt{N}]^2$ independently of the rest. Fix $\alpha > 0$. For any pair of nodes u,v with fixed w_u,w_v,x_u,x_v, connect them by an edge with probability*

$$\mathbf{Prob}(u \textit{ is connected to } v \mid x_u,x_v,w_u,w_v) \asymp \min\left\{\left(\frac{w_u w_v}{\|x_u - x_v\|_2^2}\right)^{\alpha},1\right\}. \tag{1}$$

Examining (1), two nodes that are close to each other are more likely to be connected, leading to strong clustering [1]. Smaller α makes long connections more likely. By prescribing a power-law fitness to every node that describes their *willingness* to make contacts, the degree distribution in the large-network limit decays as a power law with the same exponent τ [1]. Similar statistics have been shown for human contact and activity networks [5, 6]. Contrary to models that assume a lighter tail in the degree distribution or that are non-geometric, the number of nodes $N(u,r)$ within graph distance r of a typical node u grows in GIRGs (depending on (τ,α)) either polynomially, stretched exponentially, or doubly exponentially in r. This ball growth heavily influences the shape of the epidemic curves [3].

Spreading model. We define the simplest spreading process with temporary immunity in discrete time t, a time step corresponding to a day. At any given time a node can be in one of three possible states: *susceptible (S), infected (I) or temporarily immune (T)*. The dynamics of the nodes between the three states are described as follows.
Infecting. Each infected node *infects* each of its neighbors with probability β at every time-step. Infections to its neighbors happen independently.
Healing. When infectious, each node *heals* with probability γ at every time-step, independently of other nodes. Upon healing, the node becomes temporarily immune.
Losing immunity. Each temporarily immune node loses its immunity with probability η at every time-step, independently of other nodes, after which it becomes susceptible again. Hence, the average immune period of a node is $1/\eta$ time steps.

Results. We observe three phases for this S-I-T-S epidemic spread on networks, phase 1 being determined by only β/γ and G. The second or third phases are determined by η, separated by a critical immunity rate $\eta_c = \eta_c(G,\beta,\gamma)$.
1) *Immediate extinction* (β/γ small). The epidemic goes extinct almost immediately.
2) *Extinction after a single peak* (β/γ large, $\eta < \eta_c$). The epidemic has a single peak, after which it immediately goes extinct. Heuristically, after the first epidemic peak, nodes stay immune long-enough to provide barriers in the network that the infection cannot pass. This phase is absent in the analogue continuum compartmental model [3].
3) *Long-time survival of the epidemic* (β/γ large, $\eta > \eta_c$). There is a first major outbreak, followed by smaller peaks that decrease in magnitude, and eventually a *(metastable) stationary proportion* of the population remains infected.

3 Comparison of intervention strategies

We model several intervention strategies by removing edges from the original contact network, and run the spreading model on the remaining network. The parameters for

these strategies are chosen such that the remaining network has comparable average degree, allowing to compare the effectiveness of the strategies. We model
A) *Keeping physical distance*. Randomly remove connections from the network.
B) *Travel restrictions*. Keep an edge (u,v) with a probability that decreases exponentially in its length $\|x_u - x_v\|$ whenever it is at least some $L > 0$.
C) *Limiting the maximal number of contacts per person* up to M. Randomly remove connections for each node u with degree higher than M until M connections remain.

The height of the first peak drops for all interventions. Intervention B is most effective, meaning that the shape of the curve changes from superexponential to linear, elongating the duration and reducing the height of the first peak.

Critical mean immune duration decreases under each intervention for fixed β and γ: even a shorter immune duration leads to extinction after the first peak. In this respect, interventions B and C perform best, intervention A performs poorly.

Higher second peak. For a network with $\tau \in (2,3)$, the epidemic reaches its metastable density of infected nodes immediately after the first peak, which is unaffected under intervention A. However, interventions B and C introduce a second peak and further oscillations. For networks with $\tau > 3$, second and further peaks are present already without interventions, but their height is increased significantly under intervention C.

A possible explanation for the increased height of second peak is coming from the presence of superspreaders (hubs). Once infected, hubs quickly infect a large proportion of their neighbors. Moreover, hubs are much closer to each other (in terms of graph distance) than the average distance in the network, allowing the infection to quickly travel between them. To summarize, hubs synchronize the system. Increasing τ as well as interventions B and C remove hubs from the network and hence oscillations appear/increase.

References

1. Bringmann, K., Keusch, R., Lengler, J.: Geometric inhomogeneous random graphs. Theoretical Computer Science 760, 35–54 (2019)
2. Galanti, M., Shaman, J.: Direct Observation of Repeated Infections With Endemic Coronaviruses. The Journal of Infectious Diseases (07 2020)
3. Jorritsma, J., Hulshof, T., Komjáthy, J.: Not all interventions are equal for the height of the second peak. Chaos, Solitons & Fractals 139, 109965 (2020)
4. Liu, Z., Magal, P., Seydi, O., Webb, G.: A covid-19 epidemic model with latency period. Infectious Disease Modelling 5, 323 – 337 (2020)
5. Muchnik, L., Pei, S., Parra, L.C., Reis, S.D.S., Andrade, J.S.J., Havlin, S., Makse, H.A.: Origins of power-law degree distribution in the heterogeneity of human activity in social networks. Scientific Reports 3(1783) (2013)
6. Newman, M., Barabasi, A.L., Watts, D.J.: The structure and dynamics of networks, vol. 19. Princeton University Press (2011)
7. Parry, J.: Covid-19: Hong Kong scientists report first confirmed case of reinfection. BMJ 370 (2020)
8. Walker, P.G.T., Whittaker, C., Watson, O.J., Baguelin, M., Winskill, P., Hamlet, A., Djafaara, B.A., Cucunubá, Z., Olivera Mesa, D., Green, W., Thompson, H., Nayagam, S., Ainslie, K.E.C., Bhatia, S., Bhatt, S., Boonyasiri, A., Boyd, O., Brazeau, N.F., Cattarino, L., ..., N.M., Ghani, A.C.: The impact of covid-19 and strategies for mitigation and suppression in low- and middle-income countries. Science 369(6502), 413–422 (2020)

Revealing transmission of healthcare-associated infections using network-constrained patient trajectories

Ashleigh C. Myall[1,2], Robert L. Peach[1], Andrea Y. Weiße[2,3], Siddharth Mookerjee[4],
Frances Davies[2,4], Alison Holmes[2,4], and Mauricio Barahona[1]

[1] Department of Mathematics, Imperial College London, London, UK.
[2] Department of Infectious Disease, Imperial College London, London, UK.
[3] Current address: School of Informatics, University of Edinburgh, Scotland, UK.
[4] Imperial College Healthcare NHS Trust, London, UK.

1 Abstract

Contact tracing based on patient overlaps is widely used in hospital epidemiology to analyse outbreaks of healthcare-associated infections. However, missing data and indirect contacts can pose challenges to this methodology resulting in misleading conclusions. We propose a method that mitigates these problems by defining a similarity between network-constrained temporal trajectories of patients, and analysing the ensuing patient trajectory similarity graph. We showcase our methodology through two datasets of trajectories: (i) patients colonised with drug-resistant bacteria, and (ii) patients who acquired covid-19 over their hospital visit. Using graph-based semi-supervised learning, we show that the trajectory similarity graph constructed with our method reveals missing patient interactions that improve the characterisation of disease transmission.

2 Introduction

Healthcare-associated infections (HAIs) acquired during stays in hospitals constitute a tremendous burden to national health systems. To combat HAIs tools such as hygiene measures and, importantly, contact tracing are used. Contact tracing is akin to the analysis of ‘social interaction networks’ constructed from the overlaps of movement of infected individuals. The effectiveness of contact tracing is, however, highly dependent on data quality and suffers from missing data, e.g., due to incomplete screening or false negatives. In addition, the movement of healthcare workers (nurses, doctors, cleaners), as well as contaminated environments where HAIs can persist and later be picked up by patients, can all contribute to misleading conclusions.

To tackle these issues, we present a methodology that takes into account the full patient trajectory as a possible source of contacts between infected patients. Since movement is a key vector for disease spread [1], we hypothesise that even when patient movement histories do not overlap exactly in ward and time, the likelihood of linked cases can be determined by measuring the similarity between time-stamped trajectories across a background network compiling the movement of *all* hospital inpatients across wards. This network-time notion naturally captures the time-varying characteristics of contacts, an important mechanism in temporal interaction data [2].

Below we briefly outline our methodology and present its initial results on two anonymised data sets of patients from a large NHS Trust in London: (i) 110 hospital patients colonised with drug-resistant bacteria (DRB) of which 70 patients had infections sent for whole-genome sequencing enabling semi-supervised learning; (ii) 90 patients who acquired covid-19 during their hospital stay between March and April 2020.

Methods

We consider a set of N patient trajectories $\mathscr{T} = \{T_1, T_2, T_3, ..., T_N\}$. Each trajectory consists of a sequence of ward-time events $T_n = \{l_1, l_2, l_3, ..., l_{k_n}\}$ with $l_i = \{v_i, t_i\}$, where t_i denotes the day on which patient n was at ward v_i. We then introduce a similarity between space-time locations inspired by the network-constrained clustering algorithm ST-TOPOSCAN [3], but generalising it to incorporate both network and temporal distances simultaneously. Specifically, we propose a kernel $\kappa(l_i, l_j)$ that measures the similarity between any two ward-time events by measuring the propagation across a background graph $G = (V, E)$ with hospital wards as vertices $V = \{v_i\}$ and edges $E = w_{ij}$ weighted by the overall average patient flow between wards i and j. The similarity between two trajectories T_n and T_m is then obtained by summing over the pairwise distances between all ward-time events:

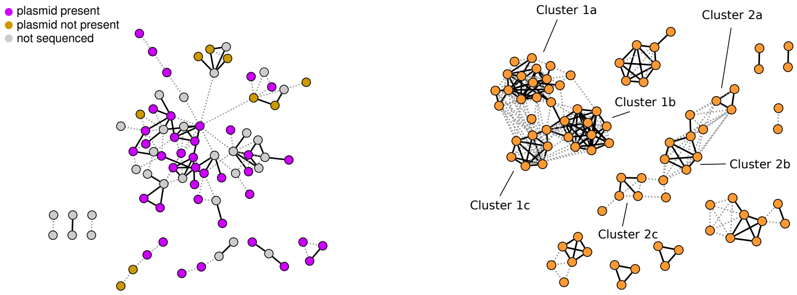
$$\mathscr{S}(T_n, T_m) = \sum_{l_i \in T_n} \sum_{l_j \in T_m} \kappa(l_i, l_j). \tag{1}$$

The trajectory similarities, Eq. (1), are compiled into a patient-to-patient ($N \times N$) matrix $\mathscr{S}$ with elements $\mathscr{S}_{nm} = \mathscr{S}(T_n, T_m)$. This similarity matrix can be thought of as a fully connected weighted graph, which we then sparsify to $\widehat{\mathscr{S}}$ by thresholding edges with weights $\mathscr{S}_{nm} < h$, where h is a parameter. We optimise the parameters in our graph construction by maximising classification accuracy of graph-based semi-supervised learning [4] against ground truth labels of known infection strains collected independently via whole-genome sequencing of DRB plasmids.

Results

Figure 1a shows the sparsified trajectory similarity graph $\widehat{\mathscr{S}}$ for the 110 patients colonised with DRB. The graph links 79 (out of a possible 110) patients and consists of 113 edges, 50 of which correspond to *exact* overlaps, with the remaining 63 edges reflecting indirect contacts mediated through the background patient flows. Without these additional indirect edges, the graph of exact overlaps would consist of only 35 patients, so that 44 patients would have been assigned to separate outbreaks or missed as isolated cases by standard contact tracing methodologies. Our methodology connects cases containing the same plasmid (19/24 connections between nodes with the same label), suggesting that our edges indeed identify missed direct/indirect transmission.

Figure 1b shows the sparsified trajectory similarity graph for the 90 patients who acquired covid-19 during their hospital stay in March-April 2020, with hyperparameters as for the DRB dataset. We found an additional 142 edges (out of a total of 274) that do



(a) DRB

(b) covid-19

Fig. 1: Sparsified patient trajectory similarity graphs for our two data sets. Each node corresponds to a patient who acquired an infection during their hospital stay. Solid edges are *exact* trajectory overlaps, and dotted edges are plausible transmissions uncovered by measuring the network-mediated similarity between trajectories.

not correspond to exact overlaps. If only exact overlaps are used, cluster 1 would appear as 3 mostly disjoint clusters (1a-1b-1c), and cluster 2 appears as three isolated transmission clusters (2a-2b-2c). Such clusters would thus be analysed in isolation curtailing the extent of the outbreak.

Conclusion

We have presented a novel methodology to uncover plausible patient contacts in outbreaks of HAIs by using a similarity measure between network-constrained patient trajectories, where the network captures the overall background flow of patients in the hospital. This approah mitigates problems related to missing or inaccurate data. Our initial analysis of two data sets of patients with bacterial/viral infections suggests that we capture aspects of indirect or missed direct transmission amongst HAIs. Given the differences in epidemiology of bacteria and viruses, we are next exploring validation steps through further sequencing data, as well as applications to additional HAIs.

References

1. R. Pastor-Satorras and A. Vespignani: Epidemic spreading in scale-free networks. Phys Rev Lett Vol. 86, no. 14, p. 3200 (2001)
2. J. Tang, M. Musolesi, C. Mascolo, and V. Latora: Temporal distance metrics for social network analysis. in Proc. 2nd ACM workshop on online social networks. pp. 31–36 (2009)
3. Z. Hong, Y. Chen, and H. S. Mahmassani: Recognizing network trip patterns using a spatio-temporal vehicle trajectory clustering algorithm. IEEE Transactions on Intelligent Transportation Systems. Vol. 19, no. 8, pp. 2548–2557 (2018)
4. R. L. Peach, A. Arnaudon, and M. Barahona: Semi-supervised classification on graphs using explicit diffusion dynamics. Foundations of Data Science. Vol. 2, no. 1, pp. 19–33 (2020)

Optimizing test strategies for epidemic detection in livestock trade network

Sara Ansari[1,2], Jobst Heitzig[1] Laura Brzoska[1], Hartmut H. K. Lentz[3],
Jörg Fritzemeier[4], Mohammad R. Moosavi[2]

Potsdam Institute for Climate Impact Research,
[1] Department of Computer Science and Engineering, School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran
[2] Research Department 4, Complexity Science, Potsdam Institute for Climate Impact Research, Telegraphenberg, 14473 Potsdam, Germany
[3] Institute of Epidemiology, Friedrich-Loeffler-Institut, Südufer 10, 17493 Greifswald-Insel Riems, Germany
[4] Landkreis Osnabrück, Veterinärdienst für Stadt und Landkreis Osnabrück,Am Schölerberg 1, 49082 Osnabrück

Animal trades between farms and other livestock holdings form a complex network which is known as an 'animal trade network'. The movement of animals play a key role in the spread of infectious diseases among farms. The existence of large disease outbreak among animal farms in different countries in the past leads to irreparable impacts on the global economy and public health [1]. Examples of these outbreaks are that of foot and mouth disease in the UK in 2001 [2], the swine fever epidemics in Netherlands and Germany in the 1990s and in 2003 [3] and the recent African swine fever (ASF) outbreak in Eastern Europe [4] and China in 2018 and 2019 [5]. Therefore studies of spreading dynamics are highly necessary for both health and economic reasons.

We propose a temporal network model for the generation of synthetic animal trade data with realistic structure and dynamics which can be used for Monte Carlo simulations in the field of the livestock trade system. Our *random animal trade transmission network model* considers a set N of nodes representing hypothetical farms and traders with different characteristics. It then generates a hypothetical sequence of animal transmissions (t,i,j,x) that may happen between the nodes within some time interval $[0,T]$. Here, x is the number of animals transported from node i to node j at time point t. Figure 1 shows the logic of our model for the application case of the German pig trade network.

Our experimental results show that the model can well reproduce several relevant properties of real trade networks such as the distributions of two novel centrality-like metrics that we specifically designed to indicate nodes that appear promising to test for epidemic outbreaks.

Outlook: Testing strategies for epidemic detection. Authorities need to detect potential disease outbreaks as early as possible in order to implement useful countermeasures [6–8]. For this reason they regularly test a random number of farms for infected animals. Given a fixed budget for tests per year [6], farms and time points for testing should thus be selected to minimize the expected number of animals already infected when an outbreak gets detected. Simulations on synthetic but realistic trade networks can help finding optimal test strategies that make these choices in terms of suitable network-related node properties such as certain centrality-like metrics.

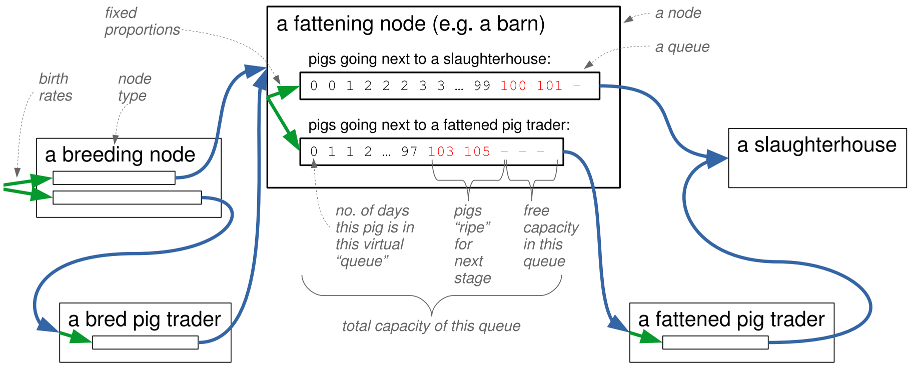

fixed
proportions
birth
rates
node
type
a breeding node
a bred pig trader
a fattening node (e.g. a barn)
pigs going next to a slaughterhouse:
0 0 1 2 2 2 3 3 … 99 100 101 –
pigs going next to a fattened pig trader:
0 1 1 2 … 97 103 105 – – –
no. of days this pig is in this virtual "queue"
pigs "ripe" for next stage
free capacity in this queue
total capacity of this queue
a node
a queue
a slaughterhouse
a fattened pig trader

Fig. 1. Logic of the random animal trade transmission network model, here for the simple model of German pig trade. Each box shows a farm/node which specialize in one of pig growing stage(breeding, fattening or slaughter) and also some nodes act as trader between different farms

References

1. Carpenter, T. E., OBrien, J. M., Hagerman, A. D., McCarl, B. A.: Epidemic and economic impacts of delayed detection of foot-and-mouth disease: a case study of a simulated outbreak in California. Journal of Veterinary Diagnostic Investigation, 23(1), 26–33 (2011)
2. Ferguson, Neil M., Christl A. Donnelly, and Roy M. Anderson: The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions, Science 292.5519, 1155–1160(2001)
3. Fritzemeier, J., Teuffert, J., Greiser-Wilke, I., Staubach, C., Schlüter, H., Moennig, V. : Epidemiology of classical swine fever in Germany in the 1990s. Veterinary microbiology, 77(1-2), 29–41(2000)
4. Aiello, C. M., Nussear, K. E., Walde, A. D., Esque, T. C., Emblidge, P. G., Sah, P., Hudson, P. J. : Disease dynamics during wildlife translocations: disruptions to the host population and potential consequences for transmission in desert tortoise contact networks. Animal Conservation, 17, 27–39(2014)
5. Food and Agriculture Organization of The united Nations,Available from:
www.fao.org/ag/againfo/programmes/en/empres/ASF/situationupdate.html
6. Lison,A. : Evaluating Test Strategies for Early Detection of Diseases in Animal Trade Networks, Bachelor Thesis(2018)
7. Bajardi, P., Barrat, A., Savini, L., Colizza, V. : Optimizing surveillance for livestock disease spreading through animal movements. Journal of the Royal Society Interface, 9(76), 2814–2825(2012)
8. Schirdewahn, F., Colizza, V., Lentz, H. H., Koher, A., Belik, V.,Hövel, P. : Surveillance for outbreak detection in livestock-trade networks. In Temporal Network Epidemiology. Springer, Singapore, 215–240(2017)

Testing framework for proxy-based Influence Maximization algorithms

Balázs R. Sziklai[12] and Balázs Lengyel[12]

[1] Institute of Economics, Centre for Economic and Regional Studies, Budapest, Hungary
sziklai.balazs@krtk.mta.hu,
WWW home page:
https://www.mtakti.hu/en/kutatok/balazs-sziklai/313/
[2] Corvinus University of Budapest
Budapest, Hungary

1 Introduction

Network diffusion models have been in the spotlight of research in the past two decades. Kempe et al. (2003) in their seminal paper formulate the *influence maximization problem* (IMP) as follows: Given a diffusion model D, a network G and a positive integer k, find the k most influential nodes in G. That is, the nodes whose activation would result in the largest spread in the network. The IMP has been studied for various reasons, including the spreading of viruses, innovation and rumours.

Unfortunately, IMP is NP-hard (Kempe et al., 2003). The literature is divided on how to approach this issue. One stream of literature focuses on finding simulation-based heuristics that guarantee good results. However, running simulations on large-scale networks is computationally expensive. Another approach is to use network centralities. The advantage of these so called proxy-based influence maximization algorithms, or proxies, in short, is a lower computational complexity.

The standard way to compare these algorithms is to take a real-life social network, compute the top k choices of the proxies, then check which set fares better in a diffusion simulation. The problem with this approach is that in real life, the highest ranked agents of the network are not always available for various reasons: Some of them are risk-averse unwilling to try the product, some may belong to a group that actively opposes the product for ideological reasons (e.g. anti-vaxxers, religious groups, groups affiliated with a political party), or - and this is the simplest, most common reason - the company does not have access to these agents (they didn't provide contact or give consent to be included).

In reality, the company might have a number of willing individuals at its disposal. The question is which among these should be chosen in a campaign? There is no guarantee that a proxy that is better at predicting the performance of the most popular agents will be equally successful for an arbitrary group of individuals. This calls for a systematic test. Here, we outline one with the help of a novel statistical method, the Sum of Ranking Differences (SRD or Héberger-test). SRD ranks competing solutions based on a reference point (Héberger, 2010; Sziklai and Héberger, 2020). The framework we propose here is somewhat similar to polling. When a survey agency wants to predict the outcome of an election it takes a random, representative sample from the population rather than asking the influencers. Statistical testing and sampling is necessary if we want to uncover the real ranking of influence maximization proxies.

2 Methods and results

The proposed framework takes the following steps:

1. We determine the centralities and IM proxies for each node of the network.
2. We take n samples of size k from the node set.
3. For each set we calculate its average centrality according to each measure.
4. We run a Monte Carlo simulation with diffusion model D for each of the sample sets and observe their performance (*i.e.* their average spread) in the simulation.
5. We rank the sets by centrality for each solution and by avg. spread for the simulation. The latter will serve as a reference ranking.
6. SRD values are obtained by computing the Manhattan distances between the rankings of the solutions and the reference ranking.

SRD scores serve as test statistics in the so-called permutation test which shows whether the rankings are comparable with a ranking taken at random. SRD values follow a discrete distribution that depends on the number of objects (here these are the sample sets). By convention, we accept those solutions that are below 0.05, that is, below the 5% significance threshold. Between 5-95% solutions are not distinguishable from random ranking, while above 95% the solution seems to rank the objects in a reverse order (with 5% significance).

To demonstrate our method we used a sample from a real-life social network, iWiW with 271 913 nodes and 5 425 174 directed edges. We took 21 samples (n) from the node set chosen uniform at random each containing 500 nodes (k). We ran a Monte Carlo simulation with the Linear Threshold model (D) for each of the sample sets and observed their performance. We ran 5000 simulations for the SRD calculation. Some sets contained more potent agents and generated a larger spread on average than others - this induced a reference ranking among the samples. Next we determined the centralities and influence maximization proxies for each node of the network. We included seven measures in our analysis: degree, Harmonic-centrality, PageRank (with damping factor 0.8), LeaderRank (Lü et al., 2011), Generalised Degree Discount (GDD, with spreading factor 0.05) (Chen et al., 2009), k-shell (Kitsak et al., 2010), Linear Threshold centrality (LTC, with parameter 0.7) (Riquelme et al., 2018). We computed the average centralities for each of the sample sets. Again that induced a ranking of the sample sets for each centrality measure.

The results are shown in Table 1. Although all of the methods fall outside the 5% threshold (equivalently nSRD$\leq$ 0.5113), that is, they rank the sample sets more or less correctly, there are notable differences. LeaderRank and Degree centrality tie for the first place. They are closely followed by Linear Threshold centrality and PageRank. Generalized Degree Discount, Harmonic centrality and k-shell perform considerably worse. It shouldn't be surprising that LTC performs well under the Linear Threshold model. Also, we shouldn't hold the relatively high SRD score against the Generalised Degree Discount algorithm, as it was primarily designed for the Independent Cascade model. The take-home message is that despite its simplicity, Degree has still a very good predictive power. On the other hand, k-shell and Harmonic Centrality seem to be the least adequate proxies for this particular environment (running the Linear Threshold model on a social network).

Table 1. Simulation results nSRD stands for normalized SRD score, α denotes the methods parameter.

	PageRank ($\alpha = 0.8$)	k-shell	LeaderRank	Harmonic-centrality	GDD ($\alpha = 0.05$)	LTC ($\alpha = 0.7$)	Degree
nSRD	0.10	0.38	0.05	0.33	0.31	0.09	0.05
Ranking by SRD	4	7	1	6	5	3	1
Ranking by top nodes	5	7	3	6	1	4	2

We carried out an external validation step and compared how the top 500 nodes of each measure perform on a diffusion simulation (last row of Table 1) There is a substantial difference between the ranking induced by the performance of the top choices and the ranking resulting from the SRD analysis. GDD, which was among the poor performers, becomes the very best. On the other hand, LeaderRank which shared the first place with Degree, is now on the 3rd place. The discrepancy between the two rankings suggests that SRD should be a necessary, if not the primary, test for evaluating influence maximisation proxies.

Bibliography

Chen, W., Wang, Y., and Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 199–208, New York, NY, USA. Association for Computing Machinery.

Héberger, K. (2010). Sum of ranking differences compares methods or models fairly. *TrAC Trends in Analytical Chemistry*, 29(1):101 – 109.

Kempe, D., Kleinberg, J., and Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 137–146, New York, NY, USA. Association for Computing Machinery.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., and Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics*, 6:888–893.

Lü, L., Zhang, Y.-C., Yeung, C. H., and Zhou, T. (2011). Leaders in social networks, the delicious case. *PLOS ONE*, 6(6):1–9.

Riquelme, F., Gonzalez-Cantergiani, P., Molinero, X., and Serna, M. (2018). Centrality measure in social networks based on linear threshold model. *Knowledge-Based Systems*, 140:92 – 102.

Sziklai, B. R. and Héberger, K. (2020). Apportionment and districting by sum of ranking differences. *PLOS ONE*, 15(3):1–20.

Nowcasting country-wide headache symptoms from social media traces and air quality

David Martín-Corral[1,2], Nick Obradovich[4] and Esteban Moro[1,3]

[1] Department of Mathematics and GISC, Universidad Carlos III de Madrid, Leganés, Spain.
[2] Zensei Technologies S.L., Madrid, Spain.
[3] Connection Science, Institute for Data Science and Society, MIT, Cambridge, USA.
[4] Center for Humans & Machines, Max Planck Institute for Human Development, Berlin, Germany.

On the World Health Organization's ranking of causes of disability, headache disorders are among the ten most disabling conditions for both genders combined and among the five most disabling for women. Headaches can be a sign of stress or emotional distress [1], or they can result from a medical disorder, such as migraine or high blood pressure [2], anxiety [3], or depression [4–6], or changes in the environment [7, 8]. Previous epidemiological literature has found a link between headaches and environmental factors such as weather and air pollution [9–12]. In the United States annual healthcare costs attributable to migraines have been estimated to approximate $17 billion [9]. In the EU, the total annual cost of headaches was calculated at 173 billion euros [13]. Despite the topic's importance, there are to our knowledge no official data streams to measure headache problems in nearly real time. The best stream of official data is represented only by data on headache-related visits to the emergency room. In this study we leverage the wealth of user-generated data available through social media and air pollution data to estimate the possibility of measuring and tracking headaches in near real time. The motivations of nowcasting headaches from air quality and digital traces are, first, high air pollution episodes have been a major public health problem in many cities around the world, they will come back after the COVID-19 pandemic and further restrictions will be needed in major metropolitan areas worldwide, second, there is only one epidemiological study [12] that makes usage of only air quality (fine particle) to explain headaches, however, chemical compounds, like NO2, need further research, and third, digital traces have been proven to be cost-effective data sources with which to build health proxies, as it has been seen on recent literature on how social media can be used to alert and nowcast public health problems such as epidemics [14] or health emergencies [15].

For this purpose we use the Primary Care Clinical Database from the Spanish health system [16] (BDCAP in Spanish) which contains 110 millions records of codified health problems from 5.515.535 individuals who visited Primary Care (PC) or Emergency Room (ER) centers and who live in 302 basic health units. These data span from January 1st, 2013 to December 31st, 2015. The BDCAP database is updated once a year and is not public: an official research application to the Spanish Ministry of Health is required. In order to validate our hypothesis we propose the following data processing pipeline with the following stages. First, we search for related 'migraines' and 'headaches' keywords in tweets. We classify those tweets as first person headache related mentions using a convolutional neural network sentence classifier which is trained with a sample of 3.704 manually tagged tweets (1.820 negative cases and 1.884 true cases). The classifier achieves a test accuracy of 92%, similar to previous studies [17]. The data points of this stage sum up 3.312.011 first person geolocated headache tweets. Secondly, we extract patients who were diagnosed with a 'headache' or 'migraine' from the BDCAP data, summing up to 206.506 individuals. Third, we build daily and city level time series of first person headache related mentions from Twitter and the BDCAP data. Fifth, we join our headache related time series with air quality data (levels of NO2, NOX, NO, CO, PM10, PM2.5, SO2) from AEMET (Spanish Agency of Meteorology) [18]. Sixth, we generate lags up to 14 days from independent variables. We have two groups of variables, first person headache related mentions from social media and air quality groups. Finally we model daily incidence of headaches at the city level using linear models. To account for potential nation-wide, seasonal or weekly effects we include fixed effects by city, day of the week and week of the year.

The results of our methodology are as follows. We see modest correlations between daily headaches and average NO2 levels, 0.26 (see Figure 1.A.1) and moderately strong correlations, 0.67 (see Figure 1.A.2), between daily headaches captured with our ground-truth measure and first person headache related mentions from social media. Beyond simple correlations, our linear regression models show that daily average NO2 levels and headaches mentions had strong statistical significance to explain and nowcast the headache prevalence in a certain day and in a certain city based on headache related mentions from social media and air quality conditions. As we see in Figure 1.C, we explain 16.44% of headaches incidence variability using only environmental factors without fixed effects as seen in previous studies [9–11], 49.34% using only headache related mentions in social media without fixed effects, 52.71% with environmental factors and social media mentions without fixed effects, and finally, we achieve a R^2_{adj} of 72.25% for headache incidence using air quality, social media and temporal and geographical fixed effects at daily and city level.

Our initial results show that is possible to build operational models to nowcast headaches using social traces that capture individuals' subjective online expressions. This approach can help health public decision makers to measure and understand one of the most common and important health symptoms (headaches), but also the impact of air pollution on headaches in near real time. We have also built the first and biggest epidemiological study of headaches using public health data from the Spanish health system. The results of our study could be of interest to public health decision makers, enabling them to implement a scalable and cost effective epidemiological system to track and monitor in near real time headache incidence. Such a tool may enable policymakers to design behavioural interventions when air pollution levels are high and mobility restrictions are needed in order to reduce pollution levels, alerting of the number of people in a city suffering of headaches due to air pollution. Such data-driven epidemiological systems could also make publicly available an almost real-time data source for developers to build and test new digital health services.

References

1. Paola Perozzo, Lidia Savi, Lorys Castelli, Walter Valfrè, R Lo Giudice, Salvatore Gentile, Innocenzo Rainero, and Lorenzo Pinessi. Anger and emotional distress in patients with

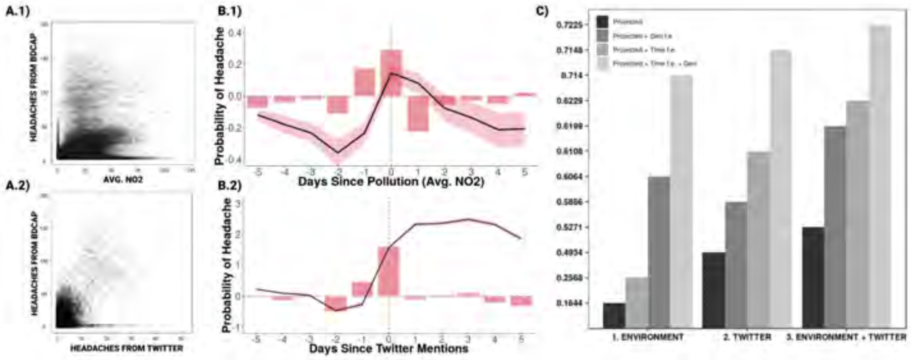


Fig. 1. A.1) Daily headaches against average daily NO2 levels. A.2) Probability of headache against days since high levels of NO2. B) Bars correspond to the probability of change (coefficients) from the model and line to accumulated probability of change. B.1) Probability change of headache against days since pollution episode. B.2) Probability change of headache against days since headache twitter mentions. As we can see, daily twitter mentions are strong predictors of headaches reported in the primary care and emergency room visits. C) R^2_{adj} results for Air pollution, Twitter and both group variables with and without fixed effects.

migraine and tension–type headache. *The journal of headache and pain*, 6(5):392, 2005.

2. WE Waters. Headache and blood pressure in the community. *Br Med J*, 1(5741):142–143, 1971.
3. Justin M Nash, David M Williams, Robert Nicholson, and Peter C Trask. The contribution of pain-related anxiety to disability from headache. *Journal of behavioral medicine*, 29(1):61–67, 2006.
4. Michael J Garvey, Charles B Schaffer, and VB Tuason. Relationship of headaches to depression. *The British Journal of Psychiatry*, 143(6):544–547, 1983.
5. Michael J Garvey, Gary D Tollefson, and Charles B Schaffer. Migraine headaches and depression. *The American journal of psychiatry*, 1984.
6. Daniel Cox and Douglas Thomas. Relationship between headaches and depression. *Headache: The Journal of Head and Face Pain*, 21(6):261–263, 1981.
7. Dan A Svensson. Etiology of primary headaches: the importance of genes and environment. *Expert Review of Neurotherapeutics*, 4(3):415–424, 2004.
8. Deborah I Friedman and Timothy De Ver Dye. Migraine and the environment. *Headache: The Journal of Head and Face Pain*, 49(6):941–952, 2009.
9. Kenneth J Mukamal, Gregory A Wellenius, Helen H Suh, and Murray A Mittleman. Weather and air pollution as triggers of severe headaches. *Neurology*, 72(10):922–927, 2009.
10. Giovanni Nattero and Annalisa Enrico. Outdoor pollution and headache. *Headache: The Journal of Head and Face Pain*, 36(4):243–245, 1996.
11. Mieczyslaw Szyszkowicz. Ambient air pollution and daily emergency department visits for headache in ottawa, canada. *Headache: The Journal of Head and Face Pain*, 48(7):1076–1081, 2008.
12. Chih-Ching Chang, Hui-Fen Chiu, and Chun-Yuh Yang. Fine particulate air pollution and outpatient department visits for headache in taipei, taiwan. *Journal of Toxicology and Environmental Health, Part A*, 78(8):506–515, 2015.

13. Mattias Linde, Anders Gustavsson, Lars Jacob Stovner, Timothy J Steiner, Jessica Barré, Zaza Katsarava, JM Lainez, Christian Lampl, Michel Lantéri-Minet, D Rastenyte, et al. The cost of headache disorders in europe: the eurolight project. *European journal of neurology*, 19(5):703–711, 2012.
14. Michael J Paul, Mark Dredze, and David Broniatowski. Twitter improves influenza forecasting. *PLoS currents*, 6, 2014.
15. Víctor M Prieto, Sergio Matos, Manuel Alvarez, Fidel Cacheda, and José Luís Oliveira. Twitter: a good place to detect health conditions. *PloS one*, 9(1):e86191, 2014.
16. Ministerio de Sanidad de España. Base de datos clínicos de atención primaria.
17. Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. Tweet classification toward twitter-based disease surveillance: new data, methods, and evaluations. *Journal of medical Internet research*, 21(2):e12783, 2019.
18. Alimentación y Medio Ambiente Ministerio de Agricultura y Pesca. Air quality data.

Susceptible-infected-spreading-based network embedding in static and temporal networks

Xiu-Xiu Zhan[1], Ziyu Li[1] Naoki Masuda[2], Petter Holme[3], and Huijuan Wang[1]*

[1] Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands
[2] Department of Mathematics, University at Buffalo, State University of New York, Buffalo, NY 14260-2900, New York, USA
[3] Tokyo Tech World Research Hub Initiative (WRHI), Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8503, Japan

1 Introduction

Classic network embedding algorithms are random-walk-based. They sample trajectory paths via random walks and generate node pairs from the trajectory paths. The node pair set is further used as the input for a Skip-Gram model which embeds nodes into vectors. In this paper, we propose to replace random walk by a spreading process, namely the susceptible-infected (SI) model, to sample paths. Specifically, we propose two SI-spreading-based algorithms, i.e., **S**usceptible-**I**nfected **N**etwork **E**mbedding (*SINE*) on static networks and **T**emporal **S**usceptible-**I**nfected **N**etwork **E**mbedding (*TSINE*) on temporal networks. The performance of our algorithms is evaluated by the missing link prediction task in comparison with state-of-the-art network embedding algorithms. We show that *SINE* and *TSINE* outperform the baselines across all six empirical datasets. We further find that the performance of *SINE* is mostly better than *TSINE*, suggesting that temporal information does not necessarily improve the embedding for missing link prediction. Moreover, we study the effect of the sampling size, quantified as the total length of the trajectory paths, on the performance of the embedding algorithms. The better performance of *SINE* and *TSINE* requires a smaller sampling size in comparison with the baseline algorithms.

2 Method

We name static SI-spreading-based network embedding algorithm that uses Skip-Gram model as *SINE*. We illustrate how to apply SI model on a static network to sample the trajectory paths in the following. *TSINE* can be derived similarly on temporal networks.

The SI spreading process on a static network is defined as follows: each node is in one of the two states at any time step, i.e., susceptible (S) or infected (I); initially, one seed node is infected; an infected node independently infects each of its susceptible neighbors with an infection probability β at each time step; the process stops when no node can be infected further. To derive the node pair set as the input for Skip-Gram, we carry out the following steps: In each iteration or realization of the SI spreading process, a node i is selected uniformly at random as the seed. We perform the SI spreading process from seed node i. The spreading trajectory $\mathscr{T}_i(\beta)$ is the union of all the nodes that finally get infected supplied with all the links that have transmitted infection between node pairs. Hence, edge (i,j) will be included into the spreading trajectories. The spreading trajectories are exactly trees under this assumption. From each of the spreading trajectory $\mathscr{T}_i(\beta)$, we construct m_i trajectory paths, each of which is the path between the root node i and a randomly selected leaf node in $\mathscr{T}_i(\beta)$. The number m_i of trajectory paths to be extracted from $\mathscr{T}_i(\beta)$ is assumed to be given by $m_i = \max\left\{1, \frac{\mathscr{K}(i)}{\sum_{j\in\mathscr{N}}\mathscr{K}(j)} m_{\max}\right\}$, where $m_{\max}$ is a control parameter and $\mathscr{K}(i)$ is the degree of the root node i in the static network.

From each iteration of the SI spreading process that starts from a randomly chosen seed node i, we can generate a spreading trajectory $\mathscr{T}_i(\beta)$ and m_i trajectory paths from $\mathscr{T}_i(\beta)$. We stop such iteration to generate new trajectory paths once the total length of the trajectory paths collected reaches the sampling size $B = NX$, where X is a control parameter, N is the number of nodes in the network. We compare different algorithms using the same B for fair comparison to understand the influence of the sampling size.

3 Results

We summarize the overall performance of the algorithms on missing link prediction in Table 1. For each algorithm, we tune the parameters and show the optimal average AUC. First of all, *tNodeEmbed* with its advanced design like RNN in the learning model performs worse than the algorithms we proposed in 3 out of the 6 networks. This illustrates the importance of exploring the design of the sampling process in embedding algorithms, not only of the learning model. From now on, we will focus on embedding algorithms based on Skip-Gram, to investigate the influence of the sampling process on the performance of the embedding algorithms. Among the static network embedding algorithms, *SINE* significantly outperforms *DeepWalk* and *Node2Vec*. *CTDNE*, *TSINE1* and *TSINE2* are for temporal networks. *TSINE1* and *TSINE2* also show better performance than random-walk-based one (*CTDNE*). Additionally, *TSINE2* is slightly better than *TSINE1* on all data sets. Therefore, we will focus on *TSINE2* in the following analysis. In fact, *SINE* shows better performance than temporal network embedding methods including *TSINE2* on all data sets except for *HT2009*. It has been shown that temporal information is important for learning embeddings [1]. However, up to our numerical efforts, *SINE* outperforms the temporal network algorithms although SINE deliberately neglects temporal information. We study the effect of the sampling size, $B = NX$, on the performance of each algorithm. We evaluate *SINE* and *TSINE2*, and *CTDNE*, because *CTDNE* performs mostly the best among all random-walk-based algorithms. The result is shown in Figure 1. For each X, we tune the other parameters to show the optimal AUC in the figure. Both *SINE* and *TSINE2* perform better than *CTDNE* and are relatively insensitive to the sampling size. This means that they achieve a good performance even when the sampling size is small, even with $X = 1$. This is because the node pair set sampled remains relatively the same when X varies. *CTDNE*, however, requires a relatively large sampling size to achieve a comparable performance with *SINE* and *TSINE2*.

Table 1. AUC scores for link prediction. All the results shown are the average over 50 realizations. Bold indicates the optimal AUC among the embedding algorithms.

Dataset	DeepWalk	Node2Vec	CTDNE	tNodeEmbed	TSINE1	TSINE2	SINE
HT2009	0.5209	0.5572	0.6038	0.5358	0.6740	**0.6819**	0.6726
ME	0.6439	0.6619	0.6575	0.7281	0.7329	0.7462	**0.7744**
Haggle	0.3823	0.7807	0.7796	**0.8702**	0.8051	0.8151	0.8267
Fb-forum	0.5392	0.6882	0.6942	0.6013	0.7104	0.7195	**0.7302**
DNC	0.5822	0.5933	0.7274	**0.9105**	0.7539	0.7529	0.7642
CollegeMsg	0.5356	0.5454	0.7872	**0.8724**	0.8257	0.8321	0.8368

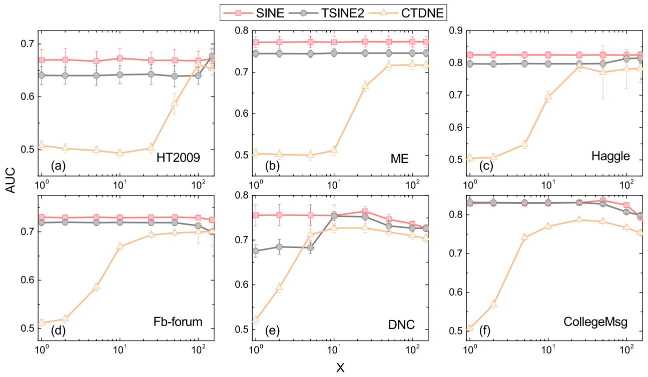


Fig. 1. Influence of the sampling size $B = NX$ on AUC score.

Summary. The key point of an embedding algorithm is how to design a strategy to sample trajectories to obtain embedding vectors for nodes. We used the SI model to this end [2]. The algorithms that we proposed are *SINE* and *TSINE*, which use static and temporal networks, respectively. *SINE* gains much more improvement than state-of-the-art random-walk-based network embedding algorithms across all the data sets. *TSINE1* and *TSINE2* show better performance than the temporal random-walk-based algorithm. For the future study, it is interesting to explore further the performance of the SI-spreading-based algorithms in other tasks such as classification and visualization. Moreover, the SI-spreading-based sampling strategies can also be generalized to other types of networks, e.g., directed networks, signed networks, etc.

References

1. Nguyen, G.H., Lee, J.B., Rossi, R.A., Ahmed, N.K., Koh, E., Kim, S.: Continuous-time dynamic network embeddings. In: Companion of the The Web Conference 2018 on The Web Conference 2018. pp. 969–976. International World Wide Web Conferences Steering Committee (2018)
2. Zhan, X.X., Hanjalic, A., Wang, H.: Information diffusion backbones in temporal networks. Scientific reports 9(1), 6798 (2019)

On an all-Ireland SIRX Network Model for the Spreading of SARS-CoV-2

Philipp Hövel[1], Rory Humphries[1] Mary Spillane[1], Kieran Mulchrone[1], Sebastian Wieczorek[1], and Micheal O'Riordain[1,2]

[1] School of Mathematical Sciences, University College Cork, Western Road, Cork T12 XF64, Ireland,
[2] Department of Surgery, Mercy University Hospital, Grenville Place, Cork, T12 WE28, Ireland

The Republic of Ireland and Northern Ireland have been severely impacted by the recent history of the spreading of the Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2), commonly known as the *coronavirus*, which causes the disease COVID-19. While many efforts in both parts of the island have helped to mitigate the impact of the pandemic, policy, regulations and case numbers on each side of the border have direct implications for the other.

We present results on an all-Ireland network modelling approach to simulate the emergence of COVID-19 across the island of Ireland[1]. Our work contributes to the goal of an island with zero community transmissions and careful monitoring of routes of importation. In the model, nodes correspond to locations or communities that are connected by links indicating travel and commuting between different locations. While this proposed modelling framework can be applied on all levels of spatial granularity and different countries, we consider Ireland as a case study. The network comprises 3440 *electoral divisions* (EDs) of the Republic of Ireland and 890 *superoutput areas* (SOAs) for Northern Ireland, which corresponds to local administrative units below the NUTS 3 regions. The local dynamics within each node follows a phenomenological SIRX compartmental model including classes of Susceptibles, Infected, Recovered and Quarantined (X) inspired from [2]. For better comparison to empirical data, we extended that model by a class of Deaths.

We consider various scenarios including the 5-phase roadmap for Ireland [3], where the parameters are chosen to match the current number of reported deaths. See Fig. 1. After relaxation of the control measures (shaded regions), the number of infected rise again in a second wave (cf. Fig. 1 for a spatial distribution of the infected at the two peaks of Fig. 1), until the pool of susceptibles is sufficiently depleted.

In addition, we investigate the effect of dynamic interventions that aim to keep the number of infected below a given threshold. This is achieved by dynamically adjusting containment measures on a national scale, which could also be implemented at a regional (county) or local (ED/SOA) level. As a simple measure, we define some threshold I_{th} as the maximum number of infected that can be present before we go back into lock-down and go through earlier phases again. See Fig. 3 for $I_{\mathrm{th}} = 10^4$. We find that – in principle – dynamic interventions are capable to limit the impact of future waves of outbreaks, but on the downside, such a strategy can last several years until herd immunity will be reached.

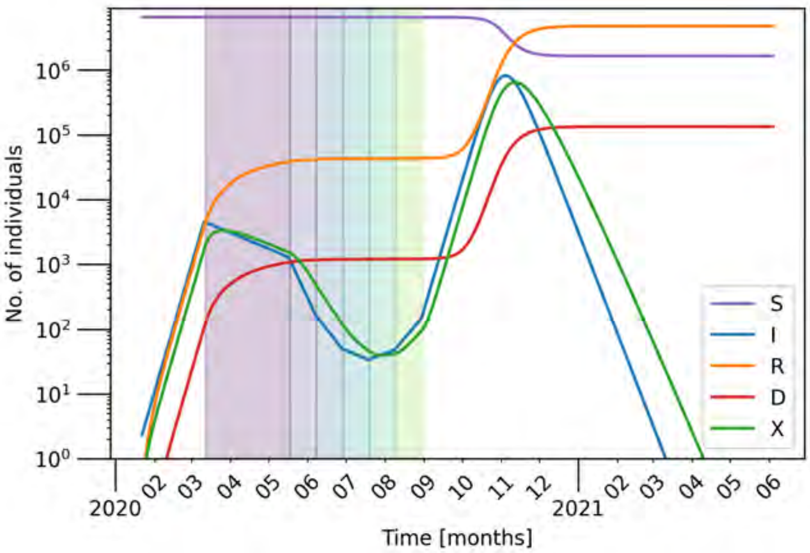


Fig. 1: Number of individuals (aggregated over the entire population/network) belonging to each compartment as stated in the legend over time with a log scale on the y-axis. Each successive lock-down phase is indicated by the differently colored shaded regions on the plot. After the initial lock-down (violet), phases 1 – 5 last 3 weeks each.

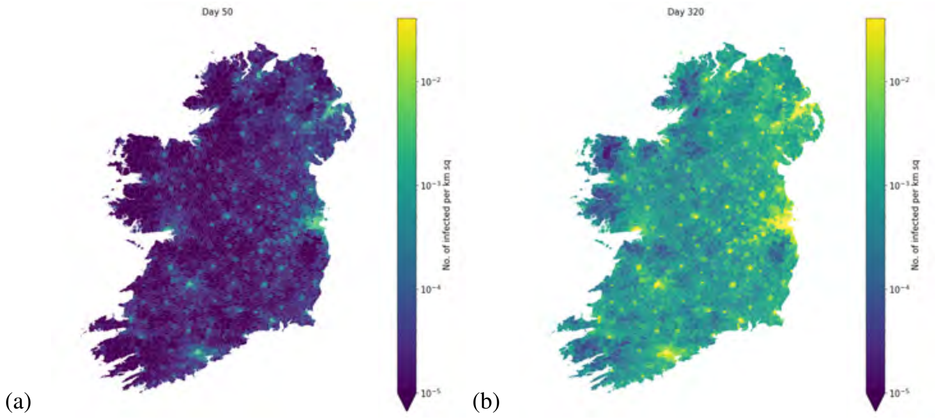


Fig. 2: The spatial distribution of infected individuals per km^2 around the two distinct peaks shown in Fig. 1 at (a) Day 50 (b) Day 320.

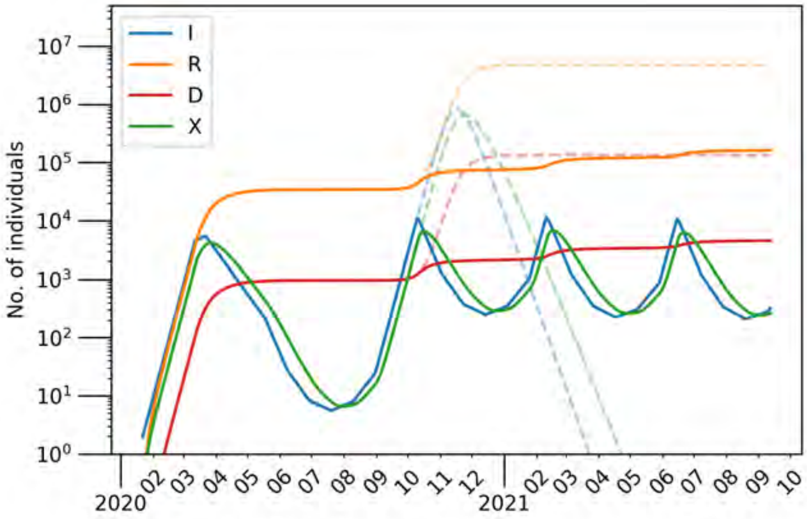


Fig. 3: The number of individuals belonging to each compartment over time under a dynamic lock-down strategy taking $I_{\mathrm{th}} = 10^4$. The dashed lines are the equivalent model without the dynamic lock-down strategy.

References

1. R. Humphries, M. Spillane, K. Mulchrone, S. Wieczorek, M. O'Riordain, and P. Hoevel: *A Metapopulation Network Model for the Spreading of SARS-CoV-2: Case Study for Ireland*, medRxiv:2020.06.26.20140590 (2020).
2. B. F. Maier and D. Brockmann: *Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China*, Science, **368**, 742 (2020).
3. Department of the Taoiseach and Department of Health: Roadmap for reopening society and business, published at: 1 May 2020, last updated: 12 June 2020, https://www.gov.ie/en/news/58bc8b-taoiseach-announces-roadmap-for-reopening-society-and-business-and-u/, https://assets.gov.ie/73722/ffd17d70fbb64b498fd809dde548f411.pdf

Mobility Footprint of Cities Through an Epidemic Diffusion Model

Matteo Zignani[1], Sabrina Gaito[1], and Gian Paolo Rossi[1]

Università degli Studi di Milano
Dipartimento di Informatica
Via Celoria 18, Milan, Italy
matteo.zignani@unimi.it, sabrina.gaito@unimi.it,
gianpaolo.rossi@unimi.it

The on-going SARS-CoV-2 spread is showing how human behaviors can dramatically impact the diffusion of pathogens transmitted by contacts. In particular, the reduction of human mobility[1] due to different severity-level lockdowns has appeared to be the most immediate and practical solution to slow down the diffusion of the epidemic. In this study we start from this strengthened assumption of interplay between human mobility and epidemic processes[2] to evaluate the mobility flow networks of a set of cities all over the world. Specifically, the epidemic process acts as a proxy to evaluate the footprint of each city and to assess the similarity between two cities in terms of similar epidemic curves. We applied the above methodology on urban mobility data gathered from Foursquare and found a pattern in the epidemic curve common to most of cities taken into account. Indeed, we observed a typical period of about 40 days in which the epidemic reaches the peak and similar percentages of infected people during the peak, telling us that the urban mobility networks of these cities are quite similar.

Data Even if in the last decade we have witnessed an increasing interest in the understanding of the human mobility within urban contexts, a certain difficulty in getting highly detailed, comparable and public urban mobility data still persists. In this context, the mobility dataset gathered from Foursquare [4] is quite an exception, since it collects global-scale mobility data from April,2012 to January, 2014. The dataset contains the temporal-annotated sequences of check-ins of more than 100,000 users placed all over the world. Moreover, the GPS position, the venue category[2] and country code are associated to each check-in/venues.

Since our goal focuses on urban mobility network, we aggregated the users' movements into mobility flows among the venues, obtaining a mobility directed network capturing the average people's flow between two venues over 2013 data. Then, to better capture the daily dynamics of the movements [4], we decomposed the daily flows into three periods: morning, afternoon and evening/night. This way, each link of the mobility networks is defined by three values. Finally, we limited our analysis to a subset of venues located in the different cities - London, New York, Tokyo, Los Angeles, Singapore, Paris - that during the first wave of the pandemic experienced different trends in the number of infected.

[1] https://www.google.com/covid19/mobility/

[2] Foursquare venues are organized into a well-defined hierarchy available at https://developer.foursquare.com/docs/build-with-foursquare/categories/

Methods Since urban mobility networks describe the mobility through spatial interacting subpopulations, we adopted a metapopulation model in modeling the diffusion process[3]. Specifically, each patch or subpopulation corresponds to a 250m X 250m square and the connections among them are proportional to the aggregated flows between the venues located in each square. As for the modeling of the epidemic process within the subpopulation, we adopt a SIR model with the infection and recovery rates reported in [1], scaled to take into account the three periods of the day. The stratification of the daily flows impacts also on the definition of the transport operator term in the equation of the variation of the compartments; indeed, in the multinomial process leading the movements among the subpopulations the total number of trials varies according to the period of the day.

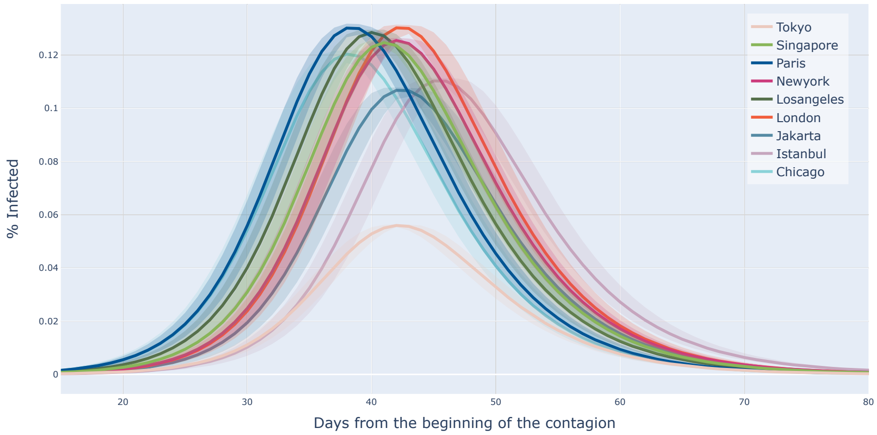


Fig. 1. Epidemic curves of the portion of infected over the population of London, New York, Los Angeles, Singapore, Paris, Tokyo, and their confidence intervals.

Results We compute the epidemic curves by running 50 instances of the aforementioned epidemic model for each city, so to get a confidence interval for the portion of infected. The resulting curves have been reported in Fig. 1, where the lines represent the average portion of infected, and the underlying regions are the 95% confidence intervals. In general, we observe that the ”epidemic footprints” of most of the cities are quite similar in terms of shape and position of the peak of infected. In fact, on average, the epidemic reaches the maximum portion of infected within 40 days. The peak is preceded by a rapid increase in the number of infected and a slightly slower decrease which mostly depends on the recovery rate of the disease. In the common trend we notice subtle differences, indeed in Paris the peak occurs before the other cities (38 days), while London and New York experience the peak a week later. Los Angeles and Singapore have trends close to the average one. In this context, Tokyo epidemic curve shows a different behavior, in fact the portion of infected is considerably lower than in the other cities as well as the increase/decrease of the infected is less pronounced. At first glance, the urban mobility network in Tokyo is different from the other cities. In general, we show how epidemic models based on mobility flows may be a tool to assess similarities of mobility behavior across different cities. As a further work, we would apply this tool to investigate the similarity of the different types of flows since Foursquare data allows us to decompose the flows according to the venue category hierarchy. In fact, the epidemic footprint based on check-in data has also the edge to provide forecasts of the trend of the epidemic in response to targeted containment actions such as containment by neighborhoods, time of day or semantic with respect to functional areas of the city.

References

1. Aleta, A., Martín-Corral, D., y Piontti, A.P., Ajelli, M., Litvinova, M., Chinazzi, M., Dean, N.E., Halloran, M.E., Longini Jr, I.M., Merler, S., et al.: Modelling the impact of testing, contact tracing and household quarantine on second waves of covid-19. Nature Human Behaviour pp. 1–8 (2020)
2. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J.J., Vespignani, A.: Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences 106(51), 21484–21489 (2009)
3. Tizzoni, M., Bajardi, P., Decuyper, A., Kon Kam King, G., Schneider, C.M., Blondel, V., Smoreda, Z., Gonzalez, M.C., Colizza, V.: On the use of human mobility proxies for modeling epidemics. PLOS Computational Biology 10, 1–15 (07 2014)
4. Yang, D., Qu, B., Yang, J., Cudre-Mauroux, P.: Revisiting user mobility and social relationships in lbsns: a hypergraph embedding approach. In: The World Wide Web Conference. pp. 2147–2157 (2019)

A Systematic Framework of Modelling Epidemics on Temporal Networks

Rory Humphries, Kieran Mulchrone and Philipp Hövel

School of Mathematical Sciences, University College Cork, Western Road, Cork T12 XF64, Ireland,
r.humphries@umail.ucc.ie

This work is based on a submitted and accepted paper which is to appear in the Applied Network Science journal. We present a general modelling framework for the spreading of epidemics on temporal networks from which both individual-based [7] and pair-based [5] [1] models can be obtained. The proposed pair-based model that is systematically derived from this framework offers an improvement over existing pair-based models by moving away from edge centric descriptions while keeping the description concise and relatively simple. Our pair-based model continues the work of [1] by extending it to the temporal setting and greatly simplifying the description without sacrificing accuracy by using a particular choice for moment closure. This closure is what is often referred to as the Kirkwood closure and allows us to write higher-order moments in terms of lower orders under an assumption of conditional independence.

For the contagion process, we consider the Susceptible-Infected-Recovered (SIR) model, which is realized on a temporal network with time-varying edges. We show that the shift in perspective from individual-based to pair-based quantities enables exact modelling of Markovian epidemic processes on temporal networks which contain no time respecting non-backtracking cycles. This is equivalent to a tree network when viewed in a static embedding of a supra-adjacency representation. On arbitrary networks, the proposed pair-based model provides a substantial increase in accuracy at a low computational and conceptual cost compared to the individual-based model. We also show under what conditions our pair-based model is equivalent to existing models such as the edge centric contact-based model [4]. In order to investigate the conditions under which the pair based model is justified on arbitrary networks and by how much it fails, we look at how and when echo chambers [6] appear and their effect on the prediction of our pair-based model. This is done by measuring the density of non-backtracing [3] cycles that appear in all time-respecting paths.

From the pair-based model, we analytically find the condition necessary for an epidemic to occur in the SIR process, otherwise known as the epidemic threshold. Because the SIR model has only one stable fixed point, which is the global disease-free state, we identify an epidemic by looking at the initial stability of the model. We derive this stability condition by linearising the pair-based model near the disease-free state and finding a temporal linear operator related to the supra-adjacency matrix which propagates the system forward in time, we then look at the spectral properties of this operator which gives rise to the critical condition for stability of the system.

In order to test our findings, we use several artificial and empirical networks which display varied topological and temporal properties. The first network we look at is a static tree network, this lets us test our finding that the pair-based model is exact on static tree networks (more generally temporal networks with no time respecting non-backtracking cycles). We compare the model to the ground truth which is the average of a large number of Monte-Carlo (MC) realisations of the underlying stochastic process that is being described. As expected we find a perfect agreement between the MC average and pair-based model for such a network.

There are two empirical networks in particular we look at. The first network is a years worth of cattle trade data from 2017 between 111513 herds in the Republic of Ireland with a temporal resolution of one day. The second network [2] is the face-to-face interactions of 405 participants at the SFHH conference held in Nice, France 2009. Each snapshot of the network represents the aggregated contacts in windows of 20 seconds. Both of these networks contain non-backtracking cycles but at different degrees so they are good candidates for testing the pair-based model on arbitrary networks. Again we run many MC realisations on both networks and compare the average to the pair-based model, interestingly, we find excellent agreement on the cattle trade network but a large deviation between the MC average and pair-based model on the face-to-face network. This is explained by the different structure in either network, the face-to-face network is a physical social interaction network where individuals congregate in groups where most or all in the group interact with one another leading to large clusters that give rise to many non-backtracking cycles. However, because the cattle trade network is a production network there exist very few non-backtracking cycles making the network structure highly tree-like in its supra-adjacency embedding, thus the pair-based model is nearly exact when compared to the MC average. This leads us to the conclusion that the pair-based model works extremely well on such production networks where the existence of cycles is inefficient and cost-prohibitive. From this we propose a new measure for the temporal-cyclicity of a network from which we can deduce whether or not the use of a pair-based model on a given network is justified. This measure is based upon the idea of temporal non-backtracking matrices with a memory parameter.

The next result we test is the analytical finding which gives us the theoretical epidemic threshold. This tells us for what combination of parameters an epidemic will occur as local outbreaks no longer die out and propagate through the network. For both of the empirical networks, we run the pair-based model for different combinations of parameters both less than and greater than the critical values required for an epidemic according to the epidemic threshold and record the final outbreak size, which is the total number of recovered once the disease has died out. In both cases, we can see a definite qualitative change in behaviour once the epidemic threshold is surpassed thus showing agreement with the analytical finding, however, the accuracy of the epidemic threshold when compared with MC realisations is dependent on how well the pair-based model performs on the given network.

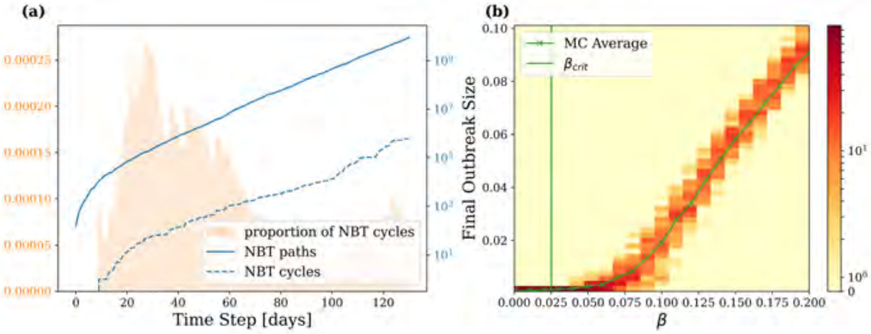


Fig. 1. Panel (a) shows the density of the non-backtracking (NBT) paths in all total paths for the cattle trade data and panel (b) shows the distribution of the final outbreak sizes for a number of simulations on the cattle trade data.

References

1. Frasca, M., Sharkey, K.J.: Discrete-time moment closure models for epidemic spreading in populations of interacting individuals. J. Theor. Biol. 399, 13–21 (Jun 2016), https://linkinghub.elsevier.com/retrieve/pii/S002251931600165X
2. Génois, M., Barrat, A.: Can co-location be used as a proxy for face-to-face contacts? EPJ Data Science 7(1), 11 (Dec 2018), https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0140-1
3. Hashimoto, K.i.: Zeta functions of finite graphs and representations of *p*-adic groups. In: Automorphic Forms and Geometry of Arithmetic Varieties. pp. 211–280. Mathematical Society of Japan, Tokyo, Japan (1989), https://doi.org/10.2969/aspm/01510211
4. Koher, A., Lentz, H.H., Gleeson, J.P., Hövel, P.: Contact-Based Model for Epidemic Spreading on Temporal Networks. Phys. Rev. X 9(3), 031017 (Aug 2019), https://link.aps.org/doi/10.1103/PhysRevX.9.031017
5. Sharkey, K.J., Kiss, I.Z., Wilkinson, R.R., Simon, P.L.: Exact Equations for SIR Epidemics on Tree Graphs. Bulletin of Mathematical Biology 77(4), 614–645 (Apr 2015), http://link.springer.com/10.1007/s11538-013-9923-5
6. Shrestha, M., Scarpino, S.V., Moore, C.: Message-passing approach for recurrent-state epidemic models on networks. Phys. Rev. E 92(2), 022821 (Aug 2015), https://link.aps.org/doi/10.1103/PhysRevE.92.022821
7. Valdano, E., Ferreri, L., Poletto, C., Colizza, V.: Analytical Computation of the Epidemic Threshold on Temporal Networks. Phys. Rev. X 5(2), 021005 (Apr 2015), https://link.aps.org/doi/10.1103/PhysRevX.5.021005

Using pandemic periods to improve now-casting models based on search engine data

Sara Mesquita[1,+], Cláudio Haupt Vieira[2,+], Lília Perfeito[1], and Joana Gonçalves de Sá[1,2,3]

[1] LIP and Physics Department, Instituto Superior Técnico, Lisboa, Portugal
smesquita@lip.pt,
[2] Nova School of Business and Economics, Carcavelos, Portugal
[3] Instituto Gulbenkian de Ciência, Oeiras, Portugal
+ Equal contribution

1 Introduction

Online searches have been used to study different health-related behaviours including identifying disease outbreaks. However, as several reasons can motivate individuals to seek online information, particularly during a pandemic, current models, blind to whether such activity is related to an actual disease or not, are of limited interest. Here we propose a methodology to disentangle search behaviours linked to general information seeking (media driven) reflecting, for example, curiosity or fear, from searchers looking for treatment information (disease driven). The difficulty in making this separation leads to the disregard of pandemic periods for disease surveillance. However, from previous studies, we know that information seeking become less common as the pandemic progresses [1, 2], so we argue that selecting the search terms during the worst possible moment, with highest media hype, can help to understand which are the ones more associated with the disease and the ones that were prompted by media exposure. As a case study, we apply our methodology to two respiratory infectious diseases able to cause a pandemic, 2009-H1N1 and COVID-19.

2 Methods

Online behaviours have proven to be very relevant tools, as health-seeking is a prevalent habit of online users. In fact this methodology has been applied to predict other epidemics, such as Dengue [3], Avian Influenza [4] and for Zika virus surveillance [5]. Online-based surveillance models harness the collective online search activity of flu-infected individuals to provide real-time monitoring. However, 2009 Influenza pandemic has been mainly described as a period when both classical and novel epidemiological tracking methods failed, as it lead to unordinary viral and human activity [6]. Current models cannot distinguish whether online activity is related with flu infection or not, rendering them useless, at least in pandemic settings. We present a methodology that enables this separation and for that we have collected large datasets of flu-related data from both online (such as search trends and social networks) and offline (such as media coverage and actual flu cases) sources, both in pandemic and seasonal settings. We expanded our analysis by applying the same principles to COVID-19 pandemic. Data from United States was considered for flu pandemic and data from Spain was used to study COVID-19 pandemic since is one of the few countries that exhibit a clear second wave of number of cases. Using Google search trends (GT), we selected 49 search terms related with the flu pandemic and 63 with COVID-19 pandemic. To those search terms, we applied hierarchical clustering to understand if different clusters reflected different online behaviours. This step was only applied to search terms extracted during the pandemic period, i.e. from March 2009 to August 2010 for 2009 H1N1 Pandemic and from January 2020 (since the first case in Spain) until August 2020 for COVID-19 pandemic. We obtained, for flu pandemic, 3 clusters: C1 with higher correlation with cases, C2 highly correlated with media and C3, a more noisy one included mainly symptoms. Within each cluster, we selected the top 5 and top 10 search terms more correlated with flu cases during the pandemic period. The chosen search terms were then used to train a model using Random Forest algorithm to predict seasonal influenza from 2005 to 2019 using either all 45 search terms or just these subsets. The model was fed with a data set containing independently extracted search terms volume for each year, having each search term a maximum value of 100 within each year. The model was trained with at least 4 flu seasons and tested with the 5th, in a k-fold process. Regarding COVID-19 pandemic, 3 clusters were also identified with the same online behaviours: one cluster more correlated with cases, a second one strongly correlated with media, and a third one with a less clear pattern. Even though we only had six months of collected data for coronavirus disease, we used the same methodology described above to predict the second wave of cases that started around June in Spain.

3 Results

Our findings indicate that separating online search trends, selected during a pandemic period, that are more sensitive to media activity from search trends related to flu activity increases model performance. It allows us to separate the signal from the noise and aim for more accurate predictions over time. For both flu and COVId-19 pandemics, we obtained better predictions using only search terms included in the cluster that better correlated with cases versus using all data. Our results reveal that our sampling criterion is more often than not better than using all search terms, especially on the long run.

Summary. We were able to show that our methodology can reduce the impact of searches not related with the disease itself, leading to better and more robust predictions over time. Additionally, despite the common intuition that more information is always better, we prove that intentional sampling can help the model performance. This can not only advance disease detection but also pave the way to improve personalized interventions. In practical terms, our system is flexible and general enough to be applied to other diseases, or different phases of the disease like seasonal events, and human activities that spread on networks (real or online). This is particularly timely, helping not only as it brings together new challenges in disease monitoring and a novel way of tackling it.

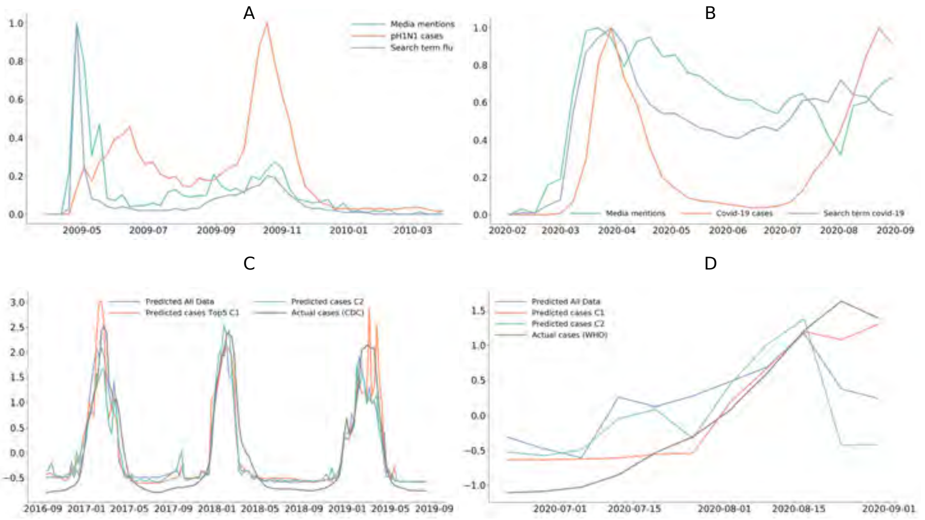


Fig. 1. Media activity versus flu activity (max-scaled) in the United States for flu pandemic 2009 **(A)** and Media activity versus flu activity (max-scaled) in Spain for Covid-19 pandemic **(B)**. This shows a quick increase in media activity in both situations that precedes the number of cases of infection. As we can see in both pandemics, some google searches have a very similar trend to the media activity. **C** shows USA CDC ILI model with a fit of R2=0.82 on average and standard deviation of 0.12 (Cluster 1 Top 5) versus R2=0.71 and standard deviation of 0.17 when using all data (not shown). In **D** we can find the results obtained for Spain with a R2=0.91 using the cluster more correlated with cases (orange) versus a R2=0.43 using all data (not shown).

References

1. Tausczik, Y., Faasse, K., Pennebaker, J. W., & Petrie, K. J. (2012). Public anxiety and information seeking following the H1N1 outbreak: blogs, newspaper articles, and Wikipedia visits. Health communication, 27(2), 179-185.
2. Towers, S., Afzal, S., Bernal, G., Bliss, N., Brown, S., Espinoza, B., & Mamada, R. (2015). Mass media and the contagion of fear: the case of Ebola in America. PloS one, 10(6), e0129179.
3. Chan, E. H., Sahai, V., Conrad, C., & Brownstein, J. S. (2011). Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. PLoS neglected tropical diseases, 5(5), e1206.
4. Lu, Y., Wang, S., Wang, J., Zhou, G., Zhang, Q., Zhou, X.,& Chou, K. C. (2019). An epidemic avian influenza prediction model based on google trends. Letters in Organic Chemistry, 16(4), 303-310.
5. Teng, Y., Bi, D., Xie, G., Jin, Y., Huang, Y., Lin, B., & Tong, Y. (2017). Dynamic forecasting of Zika epidemics using Google trends. PloS one, 12(1), e0165085.
6. Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. Science, 343(6176), 1203-1205.

Mixed Strategies for Improving the Scale and Efficiency of Infection Tests using pool-testing

Yingkai Liu[1] and Gabriel Cwilich[2]

[1] University of Illinois at Urbana-Champaign, Champaign, Illinois, 61801, USA
[2] Yeshiva University, New York, New York, 10033, USA
cwilich@yu.edu

1 Optimal Pool Size of the Tests

It has been pointed out a long time ago [1] that when dealing with a population of N individuals with a prevalence α of the infection (fraction of individuals infected over the whole population) one can diminish the amounts of tests necessary to detect all infected by dividing the population into groups of size n and pooling the samples. For each group, if none of the group members are infected, we need 1 test to make the assertion. If any of the individuals are infected, the pool test will return positive, and we will need instead $n+1$ tests, including the pool test. The optimal number of the pool size to get the maximum reduction in the number of tests has been considered empirically and numerically by many authors [2, 3] and it clearly depends on the prevalence α, and is obtained by minimizing the necessary number of tests $K_1(n,\alpha) = N(1+\frac{1}{n}-(1-\alpha)^n)$ with respect to n, to find an $n_{optimal}$. The result obtained is

$$n_{optimal} = \frac{2}{\ln(1-\alpha)} W_0(-\frac{1}{2}\sqrt{-\ln(1-\alpha)}), \tag{1}$$

where W_o is the Lambert Function and $\alpha < 1-e^{-\frac{1}{e}} \approx 0.30779$ because otherwise $K_1(n_{optimal},\alpha) > N$ and the pooling offers no advantage at higher prevalences.

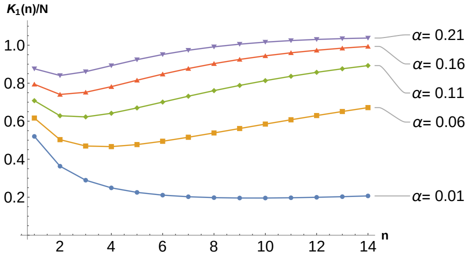


Fig. 1. Number of necessary tests to identify all the infected over the size of the population for different values of the prevalence

Figure1 exhibits the number of necessary tests as a fraction of the size of the population, as a function of the size of the pool, at different levels of prevalence. It clearly

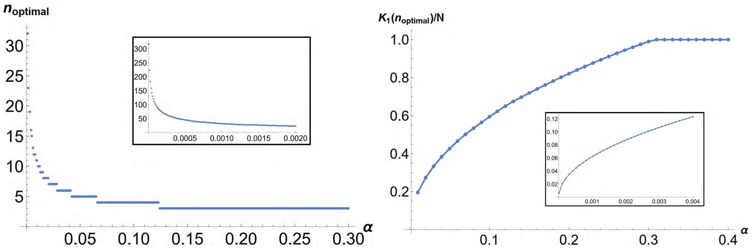


Fig. 2. Left: Optimal size of pool as a function of the prevalence. Right:Number of necessary tests using optimal pooling as a function of the prevalence. The insets in both cases shows the values for very small α

shows that only in the case of small prevalence the advantage of pool testing is considerable. Figure 2a shows the optimal value of the size of the pool which is obtained minimizing the data displayed in Figure 1, while Figure 2b indicates the fraction of necessary tests at the level of optimal size of pools, clearly indicating that for small prevalence the savings achieved by pool testing can be significant, but they diminish drastically with the increase of the prevalence.

One can generalize these calculations in a similar fashion to analyze the effect of implementing more than one level of pooling (which present further savings in the number of tests. (Those results will be presented elsewhere).

2 False results and pooling schemes to reduce them

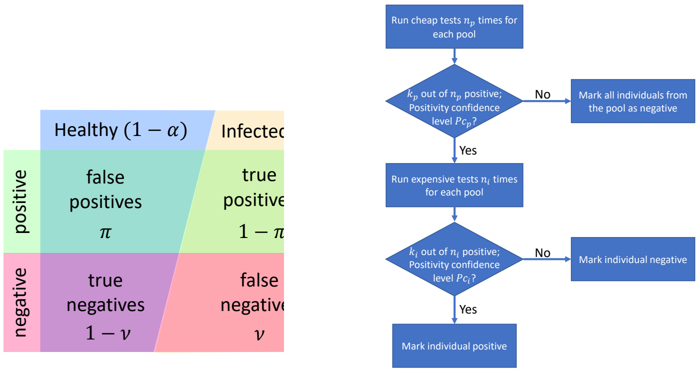

Healthy $(1-\alpha)$
Infected
positive
negative
false positives π
true positive $1-\pi$
true negatives $1-\nu$
false negativ ν
Run cheap tests n_p times for each pool
k_p out of n_p positive; Positivity confidence level Pc_p?
No
Mark all individuals from the pool as negative
Yes
Run expensive tests n_i times for each pool
k_i out of n_i positive; Positivity confidence level Pc_i?
No
Mark individual negative
Yes
Mark individual positive

Fig. 3. Left: The total population of individuals. ν and π are the fraction of false positives and false negatives, which are independent of the prevalence of the disease; Right:Workflow of the scheme. $n_p(n_i)$ are the number of tests we perform on a sample at the pool (individual) level; $k_p(k_i)$ are the number of positive results among them

The problem of false positives or false negatives is one of the great challenges when testing populations. In particular the false negatives, by being misidentified and not being isolated, have an important effect in the propagation of the disease, and need to be minimized. A simple scheme of pool testing might aggravate this problem since when testing a population with $N\alpha$ individuals infected, testing individually all the population will result in $N\alpha\nu$ false negatives as in Figure 3. But it can be shown that one level of test pooling increases this problem, since at the pool level the number of false negatives is $N\alpha\nu$ (the same result that one would obtain if applying the test individually to the whole population) and subsequently during the re-testing of the positive samples $N\alpha\nu(1-\nu)$ more individuals will be mislabeled, almost doubling the total rate.

For that reason pool testing is used when reliable tests with small values of ν are available, which are expensive and slow; on the contrary, we present here a scheme of combining a much more inexpensive and fast, although inaccurate, test for the pooling level, followed by a retest of positive samples with the more reliable test; this scheme can simultaneously reduce the number of tests (and its cost and speed) and yield very high reliability, by diminishing the rate of false negatives, and it is based on using repetitive testing of the fast inexpensive test at the pool level. The scheme is illustrated in Figure 3 (right)

It is a simple probabilistic problem to calculate the probability (for given values of the parameters α, ν and π) that a sample that yields n_k positive results out of n_T attempted tests is infected. That allows us to consider different scenarios in which we change the number of repeated tests at the individual and pool level, and the number of positive results required for a sample to be considered positive, and explore numerically the consequences. We present below a table which illustrates some of the many simulations we performed for samples of 100,000 individuals, for different values of the parameters and the scheme performed. We considered at the pool level a fast and quite unreliable test with high level of false negatives [4] , and a much more precise one for the retests. The parameters of the experiments are presented in Table 1, and the corresponding results in Table 2. Each entry reports the average over 100 independent simulations.

Let us take as an example of the use of the tables by describing experiment number 7: a population of 100,000 individuals out of which 4992 are infected (a prevalence of 5%) ; the pooling size for this prevalence is 10; The pool test has a relatively high level of false negatives (20%). We perform three tests for each pool sample, and we label it as infected if at least one of the tests results positive. Samples that are reported negative the three times are considered uninfected and their individuals marked as healthy. Then we proceed to retest all the individuals in the samples marked as infected with the more expensive and accurate test (which has a false negative rate of just 3%), and we test those individuals only once. This scheme resulted in 90 individual mislabeled as false negatives, which gives an effective false negative rate of 90/4992 = 1.8% (below the 3% rate that would have been expected from the accurate test) The total number of tests performed per capita adding the ones performed at both stages is $0.6+0.29=0.89$. Moreover, only 29% of the slow and expensive tests were performed (as compared to testing the whole populations with that test).

	Population	Pool Size	Infected Ind (Target Prevalence)	Number of Repeted Tests		Minimal k to be marked positive		False Positive Rate (π) (%)		False Negative Rate (ν) (%)	
				Pool	Ind	Pool	Ind	Pool	Ind	Pool	Ind
exp1	100000	10	5003.00 ± 55.6 (5%)	1	1	1	1	3%	3%	30%	3%
exp2	100000	10	4995.48 ± 69.9 (5%)	2	1	1	1	3%	3%	30%	3%
exp3	100000	10	4998.45 ± 64.9 (5%)	3	1	1	1	3%	3%	30%	3%
exp4	100000	10	5001.47 ± 68.8 (5%)	1	1	1	1	3%	3%	20%	3%
exp5	100000	10	5000.80 ± 65.1 (5%)	2	1	1	1	3%	3%	20%	3%
exp6	100000	10	4992.07 ± 62.7 (5%)	2	2	1	1	3%	3%	20%	3%
exp7	100000	10	4992.07 ± 62.7 (5%)	3	1	1	1	3%	3%	20%	3%
exp8	100000	10	4992.07 ± 62.7 (5%)	3	2	1	1	3%	3%	20%	3%
exp9	100000	10	5002.00 ± 72.8 (5%)	3	2	2	1	3%	3%	20%	3%
exp10	99996	6	29996.62 ± 52.2 (3%)	1	1	1	1	0.2%	0.1%	3%	0.2%

Table 1. parameters of the experiments

	Population	Infected Ind (Target Prevalence)	Actual False Negatives	Effective False Negative Rate (%)	Effective False Positive Rate (%)	Required Test Per Capita	
						Pool	Ind
exp1	100000	5003.00 ± 55.6 (5%)	1605.8 ± 40.3	32.10 ± 0.7	0.30 ± 0.07	0.20	0.18 ± 0.002
exp2	100000	4995.48 ± 69.9 (5%)	586.08 ± 24.3	11.73 ± 0.4	0.41 ± 0.09	0.40	0.25 ± 0.003
exp3	100000	4998.45 ± 64.9 (5%)	183.97 ± 14.5	3.68 ± 0.3	0.48 ± 0.08	0.60	0.29 ± 0.003
exp4	100000	4998.45 ± 64.9 (5%)	1038.07 ± 36.7	20.75 ± 0.6	0.32 ± 0.07	0.4	0.26 ± 0.002
exp5	100000	5000.80 ± 65.1 (5%)	248.8 ± 16.3	4.98 ± 0.3	0.44 ± 0.09	0.2	0.22 ± 0.002
exp6	100000	4994.03 ± 66.0 (5%)	201.3 ± 16.6	4.03 ± 0.2	0.85 ± 0.15	0.4	0.53 ± 0.006
exp7	100000	4992.07 ± 62.7 (5%)	90.4 ± 9.3	1.81 ± 0.2	0.50 ± 0.10	0.6	0.29 ± 0.003
exp8	100000	5005.01 ± 67.8 (5%)	40.6 ± 7.3	0.81 ± 0.1	0.97 ± 0.13	0.6	0.58 ± 0.007
exp9	100000	5002.00 ± 72.8 (5%)	519.3 ± 27.8	10.83 ± 0.5	0.64 ± 0.13	0.6	0.41 ± 0.006
exp10	99996	29996.62 ± 52.2 (3%)	118.2 ± 11.4	3.94 ± 0.4	0.45 ± 0.12	0.17	0.16 ± 0.003

Table 2. results of the experiments

The different experiments illustrate different possible schemes, and among them some which, without increasing the overall cost of the testing procedure, can achieve levels of false negativity remarkably smaller than the best test available. All the values of false negativity measured agree quite well with the theoretical predictions of probability theory which for the case of n pool tests and m individual tests gives: rate of false negatives $= \nu_{pool}^{n} + (1 - \nu_{pool}^{n})\nu_{individual}^{m}$

In summary these schemes open to possibility of getting the advantage of testing large populations mostly with inexpensive tests of low reliabilty without increasing prohibitively the rate of false negative results than normally are associated with them.

References

1. Dorfman R. The Detection of Defective Members of Large Populations. Ann Math Statist. 1943;14(4):436-440.
2. Lohse, S.; Pfuhl, T.; Berk-Gttel, B.; Rissland, J.; Geiler, T.; Grtner, B.; Becker, S. L.; Schneitler, S.; Smola, S. Pooling of samples for testing for SARS-CoV-2 in asymp- tomatic people. The Lancet Inf. Dis. (2020); published online April 28, https://doi.org/10.1016/S1473-3099(20)30362-5
3. Hogan, C. A.; Sahoo, M. K.; Pinsky, B. A. Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2. JAMA 2020, 323, 1967-1969.
4. https://www.publichealthontario.ca/-/media/documents/lab/covid-19-lab-testing-faq.pdf?la=en

Part IV

Dynamics on/of Networks

Synchronizability and information transmission in complete dynamical networks of discontinuous maps

J. Leonel Rocha[1] and S. Carvalho[2]

[1] CEAUL. ADM, ISEL-Eng. Superior Institute of Lisbon, IPL
Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa, Portugal
jrocha@adm.isel.pt
[2] CEAFEL. ADM, ISEL-Eng. Superior Institute of Lisbon, IPL
Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa, Portugal
scarvalho@adm.isel.pt

Abstract. This work is dedicated to the study of information measures and synchronization in complete dynamical networks of discontinuous piecewise linear maps with different slopes. It stands out that the networks topologies are characterized by circulant matrices and the conditional Lyapunov exponents are explicitly determined. Some monotony properties related to the amplitude of the network synchronization interval are established, depending on the network order and on the local dynamics. Properties of the mutual information rate and the Kolmogorov-Sinai entropy, depending on the synchronization interval, are discussed. Furthermore, various types of computer simulations show the experimental applications of these results and techniques.

1 Introduction: preliminary notions and local dynamics

The mathematical information theory studies the quantification, storage and communication of information, highlighting the quantities that are known as information measures, their properties and applications. In addition its clear importance in the area of telecommunications, information theory has several applications in other scientific and technological areas such as: biology (computational neuroscience), physics (quantum computing), chemistry (intercellular communication) and mathematics (statistical inference, cryptography, network theory and graph theory). Motivated by the theoretical and practical connection between the information measures and the synchronization phenomenon, our purpose in this work is to analyze the relations between the mutual information rate, the Kolmogorov-Sinai entropy and the synchronization of complete networks of order N, see [1]. The discontinuous local dynamics considered at each node establishes the topological, metrical and chaotic complexity of the network that is being studied. Discontinuous dynamical systems are recurrently found in physical systems and your study in synchronization phenomena has also attracted the attention of several researchers, see [2], [3] and references therein.

In this work we consider complete networks of order $N \in \mathbb{N} \setminus \{1\}$ with size $\frac{N(N-1)}{2}$ and discontinuous local dynamics, which are denoted by K_N. The complete network of N identical chaotic dynamical systems or units is described by a maximal simple unoriented graph G. The dynamics of these N coupled dynamical systems can be written in the discretized form as,

$$x_i(k+1) = f(x_i(k)) + \sigma \sum_{j=1}^{N} l_{ij} f(x_j(k)), \tag{1}$$

where f is a scalar-valued map describing the dynamics of the nodes, $L = [l_{ij}]$ represents the laplacian matrix of the complete graph G, $\sigma > 0$ is the coupling strength or parameter and $i = 1, 2, ..., N$. This equation is also known as a complex dynamical network of maps or network of discrete time systems, see, for example, [4].

The local dynamics in each node is defined by a discontinuous piecewise linear one-dimensional map, $f : I = [b_1, b_2] \subset \mathbb{R} \to I$, with $|I| = 1$ represents the amplitude of the compact interval I, such that there exist points $b_1 = d_1 < d_2 < \ldots < d_n < d_{n+1} = b_2$, where f has slope $|s_i| > 1$ in each subinterval $I_i =]d_i, d_{i+1}[$, with $i = 1, \ldots, n$. Generally, the discontinuous piecewise linear map f is defined by,

$$f(x) = |s_i| x + a_i \text{ (mod 1)}, \ \forall x \in I_i \text{ and } a_i \in \mathbb{R}. \tag{2}$$

So, we consider the following parameters space,

$$\Sigma^{\pm} = \left\{ (N, s_i, \sigma) \in \mathbb{R}^{n+2} : N \in \mathbb{N} \setminus \{1\}, |s_i| > 1, \sigma > 0, \text{ with } i = 1, \ldots, n \right\}. \tag{3}$$

Thus, in general we will carry out our study in the $(K_N, \Sigma^{\pm})$ space, see [2] and [3].

2 Synchronization interval and information measures

Since each complete dynamical network K_N has identical chaotic nodes and the eigenvalues of L are $|\mu_1| = 0$ and $|\mu_2| = |\mu_N| = N$, then the synchronization interval is nonempty, for all $|s_i| > 1$. Let $\bar{p} = (p_1, \ldots, p_m)$ a probability vector, $m \geq n$ be the number of states in the Markov partition and $\bar{\mu}$ be an invariant probability measure, [3].

Corollary 1. *Consider the $(K_N, \Sigma^{\pm})$ space of complete dynamical networks, given by Eq.(1). Let f be the local dynamics, given by Eq.(2). Then, the complete dynamical networks K_N synchronize if the coupling parameter σ verifies,*

$$\sigma_1 = \frac{1 - e^{-\chi_{\bar{\mu}}(f)}}{N} < \sigma < \frac{1 + e^{-\chi_{\bar{\mu}}(f)}}{N} = \sigma_2, \tag{4}$$

where the Lyapunov exponent $\chi_{\bar{\mu}}(f) = \sum_{i=1}^{m} p_i \ln(|s_i|)$.

Notice that, the jacobian matrix J has the eigenvalues $\nu_1 = |s_i|$ and $\nu_2 = |s_i|(1 - N\sigma)$, with multiplicity $N-1$. Thus, the parallel Lyapunov exponent is given by,

$$\lambda_{\|} = \int_I \ln|\nu_1| \, d\bar{\mu} = \sum_{i=1}^{m} p_i \ln(|s_i|), \tag{5}$$

where $|I| = b_2 - b_1 = 1$ and $\bar{p} = (p_1, \ldots, p_m)$ is the probability vector. On the other hand, the transversal Lyapunov exponent is given by,

$$\lambda_{\perp} = \int_I \ln|\nu_2| \, d\bar{\mu} = \sum_{i=1}^{m} p_i \ln(|s_i(1 - N\sigma)|). \tag{6}$$

In the following result we prove that to stabilize the synchronized states of K_N, it suffices to require that the transversal Lyapunov exponent to be negative.

Proposition 1. *Consider the $(K_N, \Sigma^{\pm})$ space of complete dynamical networks, given by Eq.(1). Let f be the local dynamics, given by Eq.(2), I_σ be the synchronization interval, given by Eq.(4), and $I_{\lambda_\perp^-}$ be the interval where $\lambda_\perp < 0$, with $\lambda_\perp$ given by Eq.(6). It is verified that:*

(i) *$\sigma \in I_\sigma$ if and only if $\sigma \in I_{\lambda_\perp^-}$;*

(ii) *there exists $\sigma > 0$ such that the synchronized states of Eq.(1) stabilize exponentially, i.e., $\left|x_i(k) - x_j(k)\right| \to 0$ as $k \to \infty$, for all $i \neq j$ with $1 \leq i, j \leq N$, and $x_i(k) \to e(k)$, where $e(k)$ is a solution of an isolated node (equilibrium point, periodic orbit or chaotic attractor), satisfying $\dot{e}(k) = f(e(k))$.*

Attending to Eqs.(5) and (6), the mutual information rate and the Kolmogorov-Sinai entropy are explicitly written by the following expressions:

$$I_C = \begin{cases} \ln\left(\frac{1}{|1-N\sigma|}\right), & \text{if } \lambda_\perp > 0 \\ \\ \sum_{i=1}^{m} p_i \ln\left(|s_i|\right), & \text{if } \lambda_\perp \leq 0 \end{cases} \tag{7}$$

and

$$H_{KS} = \begin{cases} \sum_{i=1}^{m} p_i \ln\left(s_i^2 |1 - N\sigma|\right), & \text{if } \lambda_\perp > 0 \\ & \qquad\qquad . \\ \sum_{i=1}^{m} p_i \ln\left(|s_i|\right), & \text{if } \lambda_\perp \leq 0 \end{cases} \tag{8}$$

The next proposition establishes some properties of the mutual information rate and the Kolmogorov-Sinai entropy, depending on the synchronization interval I_σ.

Proposition 2. *Consider the $(K_N, \Sigma^{\pm})$ space of complete dynamical networks, given by Eq.(1). Let f be the local dynamics, given by Eq.(2), I_σ be the synchronization interval, given by Eq.(4), and $I_{\lambda_\perp^-}$ be the interval where $\lambda_\perp < 0$, with $\lambda_\perp$ given by Eq.(6). It is verified that:*

(i) *if $\sigma \in I_\sigma$, then $I_C = H_{KS}$;*
(ii) *if $\sigma \notin I_\sigma$ and $\sigma < \sigma_1$, then I_C increases and H_{KS} decreases;*
(iii) *if $\sigma \notin I_\sigma$ and $\sigma > \sigma_2$, then I_C decreases and H_{KS} increases.*

References

1. J. L. Rocha and A. Caneco, Mutual information rate and topological order in networks, Chaotic Modeling and Simulation, Int. J. Nonlinear Sci., **4** (2013), 553–562.
2. J. L. Rocha and S. Carvalho, Information measures and synchronization in regular ring lattices with discontinuous local dynamics, in Proc. of CHAOS 2020, 13th Chaotic Modeling and Simulation International Conference, ISAST, Ed. Christos H. Skiadas, (2020), 1–13.
3. J. L. Rocha and S. Carvalho, Information theory, synchronization and topological order in complete dynamical networks of discontinuous maps, submitted (2020).
4. C. W. Wu, "Synchronization in Complex Networks of Nonlinear Dynamical Systems", World scientific, New Jersey, 2007.

Simplicial Closure in Significant Higher-Order Network among Cooking Ingredients

Masahiro Ikeda[1], Masahito Kumano[1], João Gama[2], and Masahiro Kimura[1]*

[1] Faculty of Advanced Science and Technology, Ryukoku University, Otsu, Japan
kimura@rins.ryukoku.ac.jp
[2] LIAAD, INESC TEC, University of Porto, Porto, Portugal

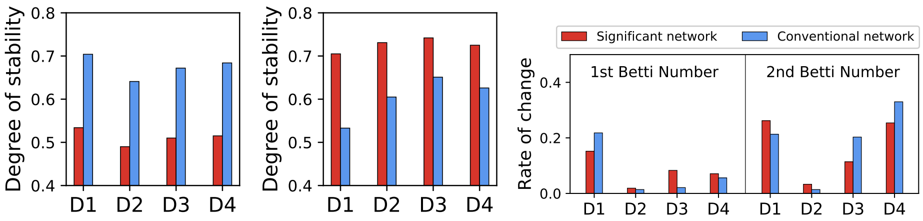
1 Introduction

With the advent of social media for cooking recipes, considerable attention has recently been devoted to food science and computing [4]. From the perspective of complex network science, ingredient pairings in recipes were analyzed [1, 4]. Recently, Benson et al [2] provided a framework for analyzing a dataset recording time-stamped interactions among a set of elements as a temporal higher-order network in terms of *simplicial closure*, which is a distinctive phenomenon of higher-order structure and cannot be captured by traditional network analysis like *triadic closure*. Using real data, they demonstrated that higher-order network evolution is fundamentally different from dyadic network evolution. According to their definition, only one interaction among a set of n nodes causes a (higher-order) n-link. However, it should be unsuitable for a trend analysis of ingredient combinations in recipes on social media since their observation can often contain anomalous noise events. In this paper, aiming to reveal statistically significant properties of their temporal changes, we introduce a novel concept of *significant (higher-order) n-link*. Using recipe datasets from social media, we investigate the characteristics of significant higher-order ingredient networks as compared with conventional ones from the viewpoints of temporal stability and simplicial closure.

Let V be a set of all ingredients to be considered, where each element of V is treated as a node. First, we define *significant* 2*-link*. A subset $\{v_1, v_2\}$ of V is referred to as a *significant* 2*-link* when $p(v_1 \,|\, v_2) > p(v_1)$. Here, for any $v \in V$ and $S \subset V$, $p(v)$ denotes the probability that ingredient v appears in a recipe, and $p(v \,|\, S)$ denotes the conditional probability that v appears in a recipe containing S. Note that $p(v_1 \,|\, v_2) > p(v_1)$ if and only if $p(v_2 \,|\, v_1) > p(v_2)$. The existence of significant 2-link $\{v_1, v_2\}$ means that the cooccurrence of ingredients v_1 and v_2 in a recipe is a significant fact. We straightforwardly extend the concept of significant link to higher-order one as follows: For $\forall n \geq 3$, a subset $\sigma = \{v_1, \ldots, v_n\}$ of V is referred to as a *significant (higher-order) n-link* when its boundary $\sigma \setminus \{v_j\}$ is a significant $(n-1)$-link for $1 \leq \forall j \leq n$ and there exists some $v_i \in \sigma$ such that $p(v_i \,|\, \sigma \setminus \{v_i\}) \; > \; p(v_i)$.

We focus on the simplest case of higher-order interaction (i.e., $n = 3$) according to [2], and split each dataset into training and test sets along time-axis on year granularity

*This work was supported in part by JSPS KAKENHI Grant Number JP17K00433 and Research Support Program of Ryukoku University. The Cookpad dataset we used in this paper was provided by Cookpad Inc. and National Institute of Informatics.



(a) 2-links. (b) 3-links (closed triples). (c) Homology groups.

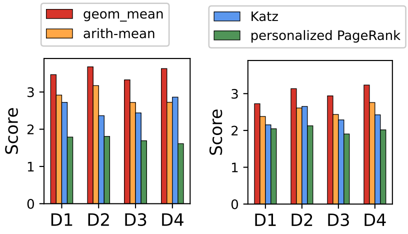
Fig. 1: Results for stability evaluation.

to explore the annual changes of such higher-order links among ingredients. We first provide a comparative analysis of the significant and the conventional networks in terms of stability by employing *computational homology theory* [3] as well. For a triple $\sigma = \{v_1, v_2, v_3\} \subset V$, we say that σ is *open* when its boundary $\sigma \setminus \{v_j\}$ forms a 2-link for any j and σ is not a 3-link. We also say that σ is *closed* when it forms a 3-link. Benson et al [2] defined that a *simplicial closure event* occurs when an open triple in the training set becomes a closed one in the test set. We systematically compare the significant and the conventional networks in terms of simplicial closure events.

2 Results

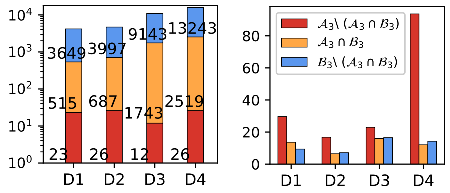
We employed real data from Japanese recipe-sharing site "Cookpad", and explored all recipes posted for its Dessert and Vegetable-Side-Dish categories during 2011-2013. We constructed four datasets D1, D2, D3 and D4 by using two consecutive years for each category, where the former and the latter years correspond to the training and the test sets, respectively. For every dataset, we focused on such ingredients that appeared in five or more recipes within each year. Here, $|V|$ was 178, 205, 255 and 309 for D1, D2, D3 and D4, respectively. We empirically estimated all the probabilities involved.

For the four datasets, we first compared the significant and the conventional higher-order ingredient networks from the perspective of stability (see Fig. 1). Figures 1a and 1b indicate the fractions of such n-links in the training set that continue to be n-links in the test set for $n = 2$ and $n = 3$, respectively. We find that the significant closed triples are more stable than the conventional ones although the reverse is true for the 2-links. This remarkable result suggests that the triples of ingredients forming significant 3-links can be important combinations for recipes. Moreover, to quantify the annual change in the global structure of each 3rd-order ingredient network, we investigated the *homology groups* [3] of the *simplicial complex* derived from all the closed triples for each of the training and the test sets. We observed that the homology groups are *torsion free*, and the 0th *Betti number* is always one, meaning that the polyhedron determined by the simplicial complex is connected. Thus, we examined the rate of change in the 1st and the 2nd Betti numbers (see Fig. 1c). We see that the significant and the conventional 3rd-order networks have a almost similar stability property in terms of this global measure.



(a) Significant. (b) Conventional.

Fig. 2: Prediction performance of simplicial closure events in triples.



(a) Number of cases. (b) Average number of Cooksnaps per triple.

Fig. 3: Analysis results for triples detected by simplicial closure events.

Next, we investigated simplicial closure events in triples for both the significant and the conventional ingredient networks (see Figs. 2 and 3). Figure 2 shows the results for their prediction by basic four models based on *projected graph* (i.e., a pairwise weighted graph derived from co-occurrences of ingredients) in terms of the area under the precision-recall curve (AUC-PR) metric according to [2], where the AUC-PR values relative to the random baseline are displayed. We observe similar results between the significant and the conventional networks, indicating that a suitable use of local features can be more important than the use of global features for these 3-link prediction problems, which is different from traditional 2-link prediction (see [2]). As for Fig. 3, $\mathcal{A}_3$ and $\mathcal{B}_3$ denote the sets of triples detected by going through simplicial closure events for the significant and the conventional networks, respectively. From Fig. 3a, we see that $|\mathcal{A}_3|$ was much smaller than $|\mathcal{B}_3|$, and $\mathcal{A}_3$ was almost covered by $\mathcal{B}_3$, indicating that the simplicial closure events for the significant network were almost explained by those for the conventional network. However, in Cookpad, people can post their "Thank-You" messages, called "Cooksnap", to a recipe if they tried and liked it. Thus, for every triple belonging to $\mathcal{A}_3 \cup \mathcal{B}_3$, we examined the number of Cooksnaps received by recipes including it. From Fig. 3b, we find that triples belonging to $\mathcal{A}_3 \setminus (\mathcal{A}_3 \cap \mathcal{B}_3)$ (i.e., triples being proper to simplicial closure events for a significant network) tend to obtain more Cooksnaps than other triples in $\mathcal{A}_3 \cup \mathcal{B}_3$, suggesting that such triples of ingredients can play an important role on the popularity of a recipe. These results imply the effectiveness of significant higher-order network.

References

1. Ahn, Y.Y., Ahnert, S.E., Bagrow, J.P., Barabási, A.L.: Flavor network and the principles of food pairing. Scientific Reports **1**, 196:1–196:7 (2011)
2. Benson, A.R., Abebe, R., Schaub, M.T., Jadbabaie, A., Kleinberg, J.: Simplicial closure and higher-order link prediction. PNAS **115**(48), E11221–E11230 (2019)
3. Kaczynski, T., Mischaikow, K., Mrozek, M.: Computational Homology. Springer (2004)
4. Min, W., Jiang, S., Liu, L.: A survey on food computing. ACM Computing Surveys **52**(5), 92:1–92:36 (2019)

Modelling spreading process of a wildfire in heterogeneous orography, fuel distribution and environmental conditions – a multi-scale analysis using complex networks

Sara Perestrelo[1], Maria Grácio[1,3] Nuno Ribeiro[1,4], and Luís Lopes[2,3]

[1] Universidade de Évora, Largo dos Colegiais 2, 7000-645 Évora, Portugal,
https://www.uevora.pt/
[2] Instituto Superior de Engenharia de Lisboa, R. Conselheiro Emídio Navarro 1, 1959-007 Lisboa, Portugal,
https://www.isel.pt/
[3] Centro de Investigação em Matemática e Aplicações, Colégio Luís Verney, Rua Romão Ramalho, 59, 7000-671 Évora, Portugal,
http://www.cima.uevora.pt/
[4] Mediterranean Institute for Agriculture, Environment and Development, University of Évora, Mitra Pole, Apartado 94, 7006-554 Évora, Portugal
https://www.med.uevora.pt/

1 Introduction

Forest fires are phenomena that represent a great danger to the population and bring severe environmental consequences. The greatest efforts of direct confrontation to the flames require expensive costs in terms of money and human resources every year. This is partly due to the unpredictability of fire behaviour, whether at the flame development, at a local scale level, whether at its spreading process at a wider scale. This unpredictability is sustained by heterogeneous orography and forest fuel distribution. Factors such as the wind speed and direction, terrain slope, soil humidity and vegetation type are preponderant in the fire development along the forest landscape. To deal more efficiently with this unpredictability, the main goal of our work is to establish an optimal fire break structure in a region that is recurrently affected by the occurrence of ignitions, which develop until wildfires at a rate that is frequently uncontrollable and along a path difficult to monitor. The function of a fire break structure is to block fire propagation through the natural landscape. Selective thinning of vegetation creates strips of bare land that interrupt the density of vegetation typical of that region. Forests, scrubs, but also grass lands are examples of areas susceptible to burn and, because of that, should be divided by such a structure, as a preventative measure to the occurrence of a fire event. Fire breaks must be efficient in the sense that its maintenance cost must allow its permanence and successfulness in stopping propagation of any forest fire. Firebreaks so extensive that their maintenance is too expensive for their continuity easily disappear due to the natural growth of the surrounding vegetation, hence opening a connection between two susceptible forest areas, previously separated by a space devoid of fuel. This preventative approach acts as complement to direct confrontation forces and contributes to reduce material, economic and human losses. For the establishment of an efficient fire break structure, it is necessary to model fire behaviour in a heterogeneous orography and using environmental parameters such as wind direction and speed, fuel type (forests, scrubs, woods, dead fuel matter, etc.), soil humidity, among others. This is a task that requires itself to different acting scales. For the effect, we resort to graph theory and complex networks, in particular, the multilayer network model. Within this model, we aim the construction of a network of networks that allows us to simulate fire spreading at a local scale and at a more global scale, the landscape. In this model, a network is a graph, where each node represents an area susceptible to burn and each edge represents a connection between adjacent areas. Each node is associated to a polygon (every area defined by a closed line – we can see illustrated in fig. 1), which represents different fuel types and, inherently, a different fire propagation velocity – these velocities are tabulated for different fuel models. Parameters such as terrain slope, soil humidity and wind direction and speed contribute to this fire spreading rate. Thus, depending on environmental conditions, orography and fuel distribution, each node is going to burn at different time instants. This time differentiation may allow direct fire combat forces to give priority to certain areas in detriment of others whose risk to population or environment may be not as high.

2 Results

The method in the development of this study consists of performing computational simulations that allow to test fire spreading at the national scale. These fire simulations start with an ignition point and the spreading process develops according to the input parameters. We use several examples of fires occurred in Portugal in previous years, which provide the area and perimeter of the burnt area, to calibrate our simulations. Once accomplished that calibration, we intend to study properties of this network of networks, such as the connectivity and robustness, among others. Finally, the goal, to perform tests to the establishment of the fire-break structure, sequentially eliminating several combinations of edges and evaluating the effect of that elimination in the neutralization of the fire spreading process. This stage of simulations is still in progress and we to obtain results within the present year. Our computational tools are: ArcGIS, a Geographic Information Systems (GIS) software for the construction and visualization of maps with the areas of interest to the study; Python 3, a programming language that is the basis of the ArcGIS software and that serves us for data processing, simulations and graph construction.

This is one of the subprojects within the project CILIFO (Forest Fires Fight and Investigation Center), supported by the international European fund Interreg, which acts in three POCTEP euroregions, Alentejo, Algarve and Andaluzia.

References

1. Grácio, M., Fernandes S., Lopes, L. M.:, Strong generalized synchronization with a particular relationship R between the coupled systems. Journal of Physics: Conf. Series 990 (2018)

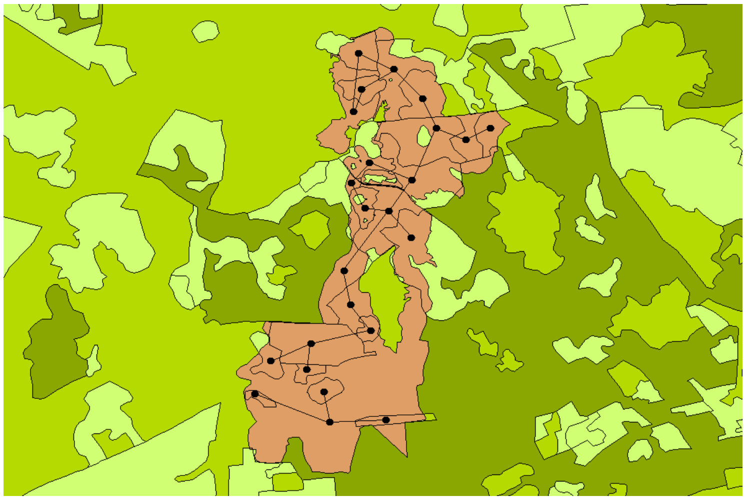


Fig. 1. A map of a small area (approx. 3.5 x 4 km) of Portalegre, in Portugal. Each closed line forms a polygon. The green scale polygons show different types of fuel, whilst the reddish polygons show the burned area (approx. 1 x 3 km) of a fire occurred in 2017. From the burned area we can construct a network, by associating each polygon to a node.

2. Russo, L., Russo, P., Siettos, C.: A Complex Network Theory Approach for the Spatial Distribution of Fire Breaks in Heterogeneous Forest Landscapes for the Control of Wildland Fires. PLoS ONE 11(10): e0163226. doi:10.1371/journal.pone.0163226 (2016)
3. Domenico, M. Granell, C., Porter, M., Arenas, A.: The physics of spreading processes in multilayer networks. Nature Physics vol 12. doi: 10.1038/NPHYS3865 (2016)
4. Boccaletti, S., Bianconi, G., Criado, R., del Genio, C., G´omez-Garde˜nes, J., Romance, M., Sendi˜na-Nadal, I., Wangk, Z., Zanin, M.: The structure and dynamics of multilayer networks. Physics Reports 544 (2014) 1–122 (2014)
5. Kivela, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., Porter, M.: Multilayer Networks, arXiv:1309.7233v4 [physics.soc-ph] (2014)
6. Aleta, A., Moreno, Y.: Multilayer Networks in a Nutshell. Annu. Rev. Condens. Matter Phys., 10:45-62 2019, arXiv:1804.03488 [physics.soc-ph] (2018)
7. Newman, M.: The structure and function of complex networks, arXiv:condmat/ 0303516v1 [cond-mat.stat-mech] (2003)
8. Bianconi, G.: Multilayer Networks – Structure and Function, Oxford University Press, DOI: 10.1093/oso/9780198753919.001.0001 (2018)
9. Urban, D., Keitt, T.: Landscape Connectivity: A Graph-Theoretic Perspective. Ecology, 82(5), pp. 1205–1218 (2001)
10. Barros, A., Pereira, J.: Wildfire Selectivity for Land Cover Type: Does Size Matter?, PLoS ONE 9(1): e84760. doi:10.1371/journal.pone.0084760 (2014)
11. Shekhtman, L., Danziger, M., Vaknin, D., Havlin, S.: Robustness of spatial networks and networks of networks, C. R. Physique 19 233–243, https://doi.org/10.1016/j.crhy.2018.09.005 (2018)
12. Seidl, R., Muller, J., Hothorn, T., Bassler, C., Heurich, M., Kautz, M. : Small beetle, large scale drivers: how regional and landscape factors affect outbreaks of the European spruce bark beetle, Journal of Applied Ecology 2016, 53, 530–540 doi: 10.1111/1365-2664.12540 (2016)

Earthquake complex network analysis for the M_w 8.2 earthquake in Iquique, Chile

Denisse Pastén[1]

Universidad de Chile, Santiago, Chile
denissepasten@uchile.cl

1 Introduction

In recent years, complex networks have had an important improve in their characterization and their applications to real networks [1, 7, 8]. Good examples of this fact are the social networks, internet, and the biological networks [9–11]. Parameters such as the connectivity of a node, k_i, the betweenness centrality or the clustering coefficient are useful tools to find communities and correlations in networks [2, 3, 5–9]. Studies of complex networks using seismic data sets have grown in the last time [2–4, 6, 13], showing that this is a very good track to study seismicity. Time based complex networks have shown to be scale-free and the critical exponent γ for the probability distribution of connectivity has shown a particular behavior in the proximity of a large earthquake, showing a change in the value of this critical exponent before and after a large earthquake [12]. Telesca et al. have found a relationship between the magnitude frequency of earthquakes and the degree of nodes in a Visibility Graph [16]. Following these ideas, we have analyzed the dynamical behavior of γ using a moving window for a time before, during and after the large earthquake M_w8.2 occurred on April 1, 2014, in the city of Iquique, in Chile. We are looking for a relation between the change in the topology of the network reflected in a change in the value of γ and the occurrence of a large earthquake.

2 Data and network

The data set used in this study was measured in the northern Chile. This catalogue was compiled by the Integrated Plate boundary Observatory Chile (IPOC) [14, 15], in the area between 19.0° S and 23.5° S Latitude and between 69.0° W and 71.5° W Longitude, with depth less than 250 km, and from January 01, 2007 until December 31, 2014 and contains the $M_w = 8.2$ earthquake of April 01, 2014. This data set contains 101 602 seismic events.

The seismic data set was measured in the format date, hypocenter, and magnitude. The hypocenter of a seismic event is characterized by (Latitude, Longitude, Depth). To make the complex network analysis, the epicenter (Latitude, Longitude) is converted to kilometers, measured from the positive lower value of Latitude (θ) and Longitude (ϕ). The $i-$th hypocenter of a seismic event is $d_i^{NS} = R(\theta_i - \theta_0)$, $d_i^{EW} = R(\phi_i - \phi_0)\cos(\theta_{\text{av}})$ and $d_i^z = z_i$. Where z_i is the depth in km, θ_{av} is the average of the Latitude, θ_0 is the minimum value of the Latitude, ϕ_0 is the minimum value of the Longitude and R is the radius of the Earth (6370 km [12]).

The space is divided in cells with side size Δ, each cell is a potential node, if a seismic event occurs into a cell, so this cell is called node. The nodes are connected between them temporarily, i.e., if the first seismic event occurs in the node 10 and the second seismic event occurs in the node 450, there is a link from node 10 to node 450. We have used a cubic cell with size 10km×10km×10km based in previous results [2, 3, 12, 13], but the size of the cubic cell needs a further analysis in future works.

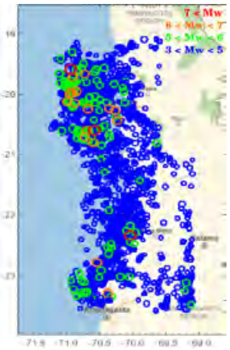


Fig. 1. Map of the zone studied in the northern Chile.

3 Results

The completeness magnitude is M_w 2.1, it means we have 91 548 seismic events that satisfy the Gutenberg-Richter's law.

We develop a dynamical analysis on the probability distribution of connectivity in a directed network. For this purpose, we have used a moving window over the seismic data. Each window contains 10 000 seismic events overlapping 8 000 data. So, we build a directed complex network in each window and we compute the value of the critical exponent γ each time. The result is shown in Fig. 2.

Fig. 2 shows a change along the time in the value of γ. There is a valley between windows 6 and 8, in those windows occurred another earthquake in Tocopilla city with magnitude M_w 7.7. The large earthquake of M_w 8.2 occurred between windows 38 an 42, where we appreciated an increase in the value of γ. The tectonic origin of these two large earthquakes are not the same, while the Tocopilla earthquake was intraplate (inside of the South American plate), the Iquique earthquake was interplate (in the shock of Nazca and South America plates). It seems that the effect of a large earthquake in the complex network is not the same. We are studying how the occurrence of a large earthquake can change the topology of a complex network and 2 suggests a relationship between the occurrence of a large earthquake and a change in the value of the exponent γ.

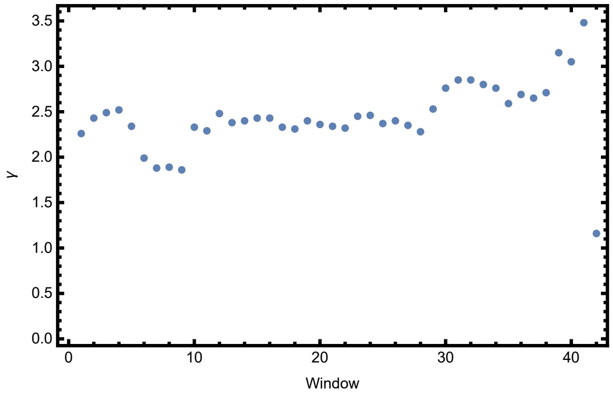


Fig. 2. Evolution of the critical exponent γ along time from year 2007 to 2014 in the northern zone of Chile, in the vicinity of a large earthquake M_w8.2.

Summary. We have performed a dynamical analysis on the behavior of the critical exponent γ for the probability distribution of connectivity in a directed complex network based on seismic data set. The complex network has been built with seismic data events measured in the time/space vicinity of the large earthquake of magnitude M_w 8.2 on April 1, 2014, in the city of Iquique in Chile. We have found a change in the value of the critical exponent γ, which may be due to the accumulation of stress in that region before a main seismic event. Further analysis must be done in the future to verify these results.

References

1. S. Abe, N. Suzuki, Nonlinear Proc. Geophys. **13** (2006) 140-150.
2. S. Abe, N. Suzuki, Europhys. Lett. **65** (2004) 581-586.
3. S. Abe, D. Pastén, N. Suzuki, Physica A **390** (2010) 7.
4. S. Abe, D. Pastén, V. Muñoz, N. Suzuki, Chinese Science Bulletin **56** (2011) 34.
5. B. Aguilar-SanJuan, L. G. Vargas, Eur. Phys. Journal B **86** (2013) 454.
6. M. Baiesi, M. Paczuski, Phys. Rev. E **69** (2004) 066106.
7. A.-L. Barabási, R. Álbert, H. Jeong, Physica A **281** (2000) 69-77.
8. M. Barthelemy, Eur. Phys. Journal B **38** (2004) 163-168.
9. K.-I. Goh, E. Oh, B. Kahng, D. Kim, Phys. Rev. E **67** (2003) 1-4.
10. M. E. J. Newman, Phys. Rev. Lett. **89** (2002) 208701.
11. M. E. J. Newman, M. Girvan, Phys. Rev. E **69** (2004) 1-15.
12. D. Pastén, F. Torres, B. Toledo, V. Muñoz, J. Rogan, J.A. Valdivia, Pure Appl. Geophys. **173**, 2267-2275 (2016). https://doi.org/10.1007/s00024-016-1335-7
13. D. Pastén, Z. Czechowski, B. Toledo, Chaos **28** (2018) 083128.
14. Sippl, C., Schurr, B., Asch, G., Kummerow, J., 2018a). Journal of Geophysical Research Solid Earth Solid Earth **123**, 4063-4087.
15. Sippl, C., Schurr, B., Asch, G., Kummerow, J., 2018b). GFZ Data Services. http://doi.org/10.5880/GFZ.4.1.2018.001
16. Telesca L, Lovallo M, Ramirez-Rojas A, Flores-Marquez L (2014) PLoS ONE **9(8)**: e106233. doi:10.1371/journal.pone.0106233

Convergence towards an Erdős-Rényi graph structure in network contraction processes

Eytan Katzav, Ido Tishby, and Ofer Biham

Racah Institute of Physics, The Hebrew University, Jerusalem 9190401, Israel
eytan.katzav@mail.huji.ac.il

Complex network architectures and dynamical processes taking place on them play a central role in current research. Since the 1960s, mathematical studies of networks were focused on model systems such as the Erdős-Rényi (ER) network [1], which exhibits a Poisson degree distribution of the form $\pi(k|c) = e^{-c}c^k/k!$, where c is the mean degree. In fact, ER networks form a maximum entropy ensemble under the constraint that the mean degree is fixed [2]. In the 1990s, the growing availability of data on large biological, social and technological networks revolutionized the field. Motivated by the observation that the World Wide Web [3] and scientific citation networks exhibit power-law degree distributions, Barabási and Albert (BA) introduced a simple model that captures the essential growth dynamics of such networks [4]. A key feature of the BA model is the preferential attachment mechanism, namely, the tendency of new nodes to attach preferentially to high degree nodes. More specifically, each new node is connected to m existing nodes with a probability that is proportional to the number of links that the existing nodes already have. Using mean-field equations and computer simulations it was shown that the combination of growth and preferential attachment leads to the emergence of scale-free networks with power-law degree distributions [4]. It was subsequently found that a large variety of empirical networks exhibit such scale-free structures, which are radically different from ER networks [5].

In many of these networks the growth phase is not likely to proceed indefinitely. Moreover, networks may be exposed to node deletion processes due to node failures, attacks and epidemics, which may eventually halt the expansion phase and induce the contraction and eventual collapse of the network. Since network growth is a kinetic nonequilibrium processes, it is generically not a reversible process, and indeed the contraction process is not the same as the growth process when played backwards in time. Three generic scenarios of network contraction exist: the scenario of random node deletion that describes the random, inadvertent failure of nodes, the scenario of preferential node deletion that describes intentional attacks that are more likely to focus on highly connected nodes and the scenario of propagating node deletion that describes viral and infectious processes that spread like epidemics (e.g., the closely related random bond percolation process maps into the Susceptible-Infected-Recovered (SIR) dynamics). It was found that scale-free networks are resilient to attacks targeting random nodes, but are vulnerable to attacks that target high degree nodes or hubs. Using the framework of percolation theory, it was shown that when the number of deleted nodes exceeds some threshold, the network breaks down into disconnected components [6]. However, the evolution of the network structure throughout the contraction phase was not addressed.

In this work we analyzed the structural evolution of networks during the contraction process [7]. To this end we derived a master equation for the time dependence of the degree distribution during network contraction via the random deletion, preferential deletion and the propagating deletion scenarios. Using the relative entropy and the degree-degree correlation function we showed that the ER graph structure, which exhibits a Poisson degree distribution and lacking any correlations, is an asymptotic structure for these network collapse scenarios, in analogy to the way in which the scale-free structure is an asymptotic solution for the preferential attachment growth scenario. In a more recent publication [8] we provide a rigorous proof that the ER structure is an attractive solution for the three contraction processes, leading to the conclusion that the ER structure is a universal asymptotic structure for contracting networks.

In Fig. 1 we present the structure of a BA network with $m = 3$ during growth at an intermediate size of $N = 150$ (left) and at the final size of $N = 200$ (middle). At this point the network starts to contract via preferential node deletion. The structure of the network during the contraction process is presented (right), when its size is down to $N = 150$. To emphasize the variation in the degrees of different nodes, each node is represented by a circle whose area is proportional to the degree of the node.

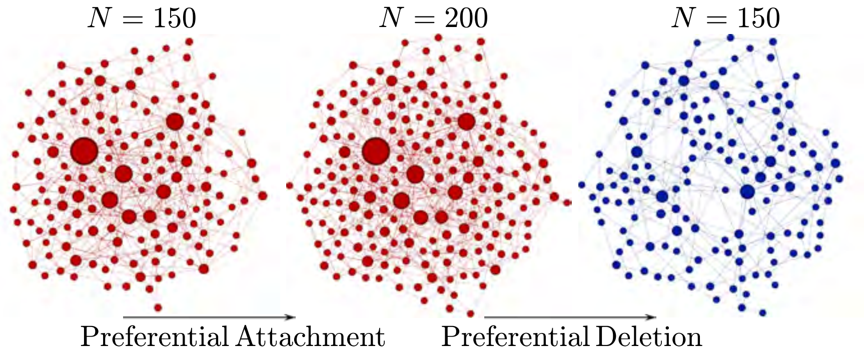


Fig. 1. A graphical demonstration of the network structure, starting with a BA (Barabási-Albert) network (left) that grows via preferential attachment (middle), and then contracts via preferential deletion recovering its original size (right). There is a striking difference between the structures of the growing networks that exhibit large hubs and the contracting network that shows little variation between the degrees of different nodes.

In Fig. 2 we present the degree distributions $P(k)$ (solid lines) of a BA network with $m = 50$, obtained from numerical integration of the master equation that describes the growth process during growth at an intermediate size of $N = 1,300$ (left) and at the final size of $N = 10,000$ (middle). The resulting degree distributions, presented in a log-log scale, follow a straight line that corresponds to $P(k) \sim k^{-\gamma}$, with $\gamma = 3$. They coincide with the degree distributions obtained from computer simulations of the BA growth process (circles). The corresponding Poisson distributions with the same value of the mean degrees, namely $c = \langle K \rangle$, are also shown (dashed lines). They form narrow and nearly symmetric distributions whose peaks are close to c. Clearly, the power-law distribution (solid line) and the Poisson distribution (dashed line) are essentially as different from each other as any two distributions with the same mean degree could be.

Starting from $N = 10,000$ the network contracts via the preferential node deletion scenario. The degree distribution of the contracted network when its size is reduced back to $N = 1,300$ is shown (right). The results obtained from numerical integration of our master equation (solid line) and from simulations (circles) are found to be in excellent agreement with a Poisson distribution with the same mean degree (dashed line).

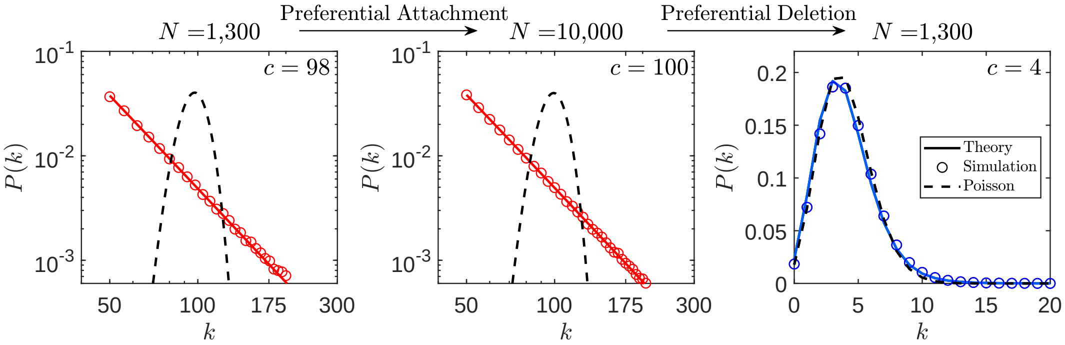


Fig. 2. The evolution of the degree distributions $P(k)$ through a growth phase from $N = 1,300$ (left) to the final size of $N = 10,000$ (middle), and back to size $N = 1,300$ (right) via a contraction phase. It becomes clear that while the growing structure exhibits a scale-free degree distribution, upon contraction it converges to a Poisson distribution.

In summary, complex networks encountered in biology, ecology, sociology and technology often contract due to node failures, infections or attacks. The ultimate failure, taking place when the network fragments into disconnected components was studied extensively using percolation theory. We show that long before reaching fragmentation, contracting networks lose their distinctive features. In particular, we identify that a very large class of network structures, which experience a broad class of node deletion processes, exhibit a stable flow towards a universal fixed point, representing a maximum-entropy ensemble, called the Erdős-Rényi ensemble. This is in sharp contrast to network expansion processes, which lead to diverse families of complex networks, whose structure is highly sensitive to details of the growth mechanism. These results imply that contracting networks in the late stages of node failure cascades, attacks and epidemics reach a common structure, providing a unifying framework for their analysis.

References

1. P. Erdős and A. Rényi, *Publ. Math. Debrecen* **6**, 290 (1959).
2. A. Annibale, A.C.C. Coolen, L.P. Fernandes *et al.*, *J. Phys. A* **42**, 485001 (2009).
3. R. Albert, H. Jeong and A.-L. Barabási, *Nature* **401**, 130 (1999).
4. A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).
5. A.-L. Barabási, *Science* **325**, 412 (2009).
6. R. Cohen, K. Erez, D. ben-Avraham and D. Havlin, *Phys. Rev. Lett.* **85**, 4626 (2000).
7. I. Tishby, O. Biham and E. Katzav, *Phys. Rev. E* **100**, 032314 (2019).
8. I. Tishby, O. Biham and E. Katzav, *Phys. Rev. E* **101**, 062308 (2020).

Spontaneous symmetry breaking of active phase in coevolving nonlinear voter model

Arkadiusz Jędrzejewski[1], Joanna Toruniewska[2], Krzysztof Suchecki[2], Oleg Zaikin[3], and Janusz A. Hołyst[2,3]

[1] Wrocław University of Science and Technology, Wrocław, Poland,
arkadiusz.jedrzejewski@pwr.edu.pl,
[2] Warsaw University of Technology, Warsaw, Poland,
[3] ITMO University, Saint Petersburg, Russia

1 Introduction

A feedback loop between the network topology and dynamical processes that occur between nodes is common in real-world networks [1, 2]. The topology impacts the evolution of node states, which in turn influence the way the structure itself is modified. This feedback is a signature of networks that are called adaptive or coevolutionary [1]. Adaptive networks are especially relevant for social systems, where they can model phenomena such as the emergence of consensus and polarization, opinion formation, group fragmentation, or language diversity [3–5]. These coevolutionary models rely on two basic mechanisms. One accounts for the changes in the node states, whereas the other for the link rewiring. Both of them may be implemented in various ways. The competition between these two mechanisms in adaptive networks leads frequently to a fragmentation transition, where the network splits into smaller components.

The voter model, as a minimalist model of opinion formation process [7, 6], provides the basis for the evolution of state variables in many adaptive networks that represent social interactions. Being analytically tractable, it has played a fundamental role in understanding the process of network fragmentation. This work extends the study in this area by the analysis of one of the nonlinear extensions of the coevolving voter model.

Most of the link rewiring mechanisms in adaptive systems reflect the effect known in sociology as homophily, the tendency of individuals to bond with others who are similar to themselves [3]. Under this paradigm, nodes may remove their links to disagreeing neighbors and form new ones to randomly chosen nodes in the same states. The coevolving nonlinear voter model analyzed in Ref. [8] implements this rewire-to-same mechanism. However, another popular approach to the rewiring dynamics is to form new links to randomly chosen nodes in any state [4]. In this regard, the analysis of the nonlinear version of the coevolving voter model with the rewire-to-random mechanism seems to be interesting not only for comparative but also cognitive reasons since it may potentially reveal some new phenomena related to the network fragmentation. In this work, we carry out such an analysis.

2 Model description

We consider an undirected network of voters. Each node can be in one of two possible states, $j \in \{-1, 1\}$. Let ρ_i denote the concentration of disagreeing neighbors with a randomly selected node labeled i. With probability ρ_i^q, the interactions between the node i and its neighbors cause a change in the system; with complementary probability $1-\rho_i^q$, nothing happens. In case of the change, two events are possible. With probability p, one randomly picked active link (i.e., a link that connects disagreeing voters) of the node i is rewired to another node picked at random from all the nodes in the network. Otherwise, with probability $1-p$, the node i changes its opinion to the opposite. The only difference between this model and the model analyzed in Ref. [8] is that the model from the reference adopts the rewire-to-same mechanism instead of the rewire-to-random mechanism adopted in our study.

3 Results

We analyzed the coevolving nonlinear voter model with the rewire-to-random mechanism by the use of the pair approximation in which the distinction between the average degrees of nodes in different states is made. This approach allowed us to identify two dynamically active phases – the well-known symmetric phase and the asymmetric one, which can arise from spontaneously broken symmetry. The symmetric active phase is characterized by the same numbers of nodes in the opposite states, so none of the states is preferred in the network; see Fig. 1. In the asymmetric active phase, on the other hand, there is a predominance of

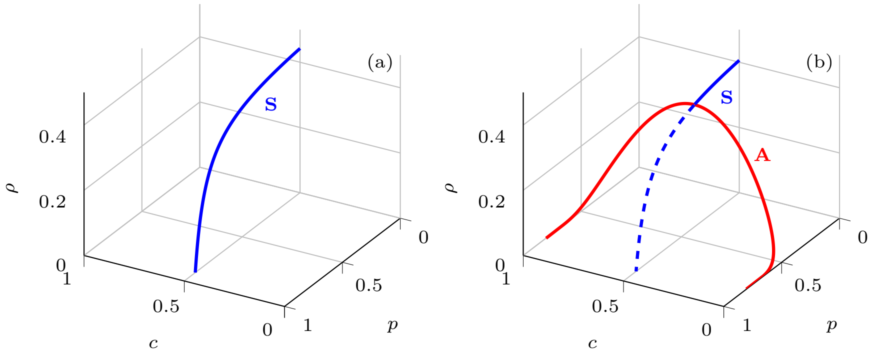


Fig. 1. Stability diagrams for the coevolving nonlinear voter model with (a) rewire-to-same and (b) rewire-to-random mechanism for $q^* \leq q < 1$. In the figures, c is the concentration of voters with $j = 1$, whereas ρ is the concentration of active links. The solid and dashed lines correspond to the stable and unstable states, respectively. The blue line refers to the symmetric active phase (since $c = 0.5$), whereas the red line refers to the asymmetric active phase (since $c \neq 0.5$).

nodes in one state, so the majority opinion can be distinguished; see Fig. 1. Only in the symmetric active phase, the average degrees of nodes in different states are equal to the average node degree of the network.

In the pair approximation, for $0 < q < \bar{q}$, where $\bar{q}$ depends on the average node degree of the network, the coevolving nonlinear voter model with the rewire-to-random mechanism exhibits only continuous phase transitions between the symmetric active and the absorbing phases. Similar behavior is shared by the coevolving nonlinear voter model with the rewire-to-same mechanism for $0 < q \leq 1$ [8] (see Fig. 1). However, for $\bar{q} < q < 1$, the pair approximation predicts much richer phase diagram for the model with the rewire-to-random mechanism than for its rewire-to-same counterpart. In this range of the parameter, the asymmetric active phase emerges. For $\bar{q} < q < q^*$, where $q^* = \frac{1}{6}(\sqrt{13}+1) \approx 0.7676$, discontinuous phase transitions are possible, and a hysteresis loop may be observed as a result of system bistablity. The discontinuous phase transitions may occur between the symmetric active and the absorbing phase or directly between both active phases. On the other hand, for $q^* \leq q < 1$, two continuous phase transitions are predicted. The first transition occurs between the symmetric and the asymmetric active phase. At the transition point to the asymmetric active phase, the symmetry is spontaneously broken, and the majority opinion arises in the network (see Fig. 1). Interestingly, there are different critical exponents on both sides of this transition for $q = q^*$. As p increases further, a continuous phase transition to the absorbing phase takes place. Our Monte Carlo simulations conducted on initially random networks confirm the appearance of the asymmetric active phase in the model with the rewire-to-random mechanism.

References

1. Gross, T., Blasius, B.: Adaptive coevolutionary networks: a review. J. R. Soc. Interface 5(20), 259–271 (2008)
2. Sayama, H., Pestov, I., Schmidt, J., Bush, B.J., Wong, C., Yamanoi, J., Gross, T.: Modeling complex systems with adaptive networks. Comput. Math. Appl. 65(10), 1645 – 1664 (2013)
3. Centola, D., González-Avella, J.C., Eguíluz, V.M., Miguel, M.S.: Homophily, cultural drift, and the co-evolution of cultural groups. J. Confl. Resolut. 51(6), 905–929 (2007)
4. Nardini, C., Kozma, B., Barrat, A.: Who's talking first? consensus or lack thereof in coevolving opinion formation models. Phys. Rev. Lett. 100(15), 158701 (2008)
5. Raducha, T., Gubiec, T.: Predicting language diversity with complex networks. PLOS ONE 13(4), 1–11 (04 2018)
6. Jędrzejewski, A., Sznajd-Weron, K.: Statistical Physics Of Opinion Formation: is it a SPOOF? C. R. Physique 20(4), 244–261 (2019)
7. Redner, S.: Reality-inspired voter models: A mini-review. C. R. Physique 20(4), 275 – 292 (2019)
8. Min, B., San Miguel, M.: Fragmentation transitions in a coevolving nonlinear voter model. Sci. Rep. 7(1), 12864 (2017)

Max-Plus Opinion Dynamics With Temporal Confidence

Daniel Feinstein[1] and Ebrahim Patel[1]

University of Oxford, Andrew Wiles Building, Woodstock Rd, Oxford OX2 6GG
danielfeinstein12@gmail.com

1 Introduction

Often in the setting of human-based interactions, the existence of a temporal hierarchy of information plays an important role in diffusion and opinion dynamics within communities [1]. For example at the individual agent level, more recently acquired information may exert greater influence during decision-making processes [2]. To facilitate further exploration of this effect, we introduce an efficient method for modelling temporally asynchronous opinion updates, where the timing of updates depends on the timing of incoming opinion states received from neighbours. The framework enables the introduction of *information arrival-time lag* by means of *lag-vectors*. These are used to weight the relevance of information received by agents, based on the delay between its receipt and the subsequent opinion update. The temporal dynamics (i.e. the times at which information is transmitted) are governed by an underlying algebraic structure called max-plus algebra ([3], [4]). We investigate the resulting continuous opinion dynamics under the max-plus regime using a modified Hegselmann-Krause model [5], replacing the conventional confidence-interval based on the distance between opinions with one based instead on the recency of information received by agents.

Our model works as follows: at time-step $k = 0$, each agent (represented by a node in a network) is assigned an initial opinion from the interval $[0,1]$ uniformly at random and transmits this value to all neighbours in the network. If an edge from agent i to j exists, the information leaving agent i arrives at j after A_{ji} time units (e.g. minutes), where A is the max-plus adjacency matrix (which is nothing more than the transpose of the conventional adjacency matrix). After sending their current opinion, each agent enters a dormant period where it waits to receive all incoming opinion values. Once received, agents update their opinions before immediately re-sending their new values to all neighbours (possibly at different times), and this process continues for a desired duration.

The event-times (the times at which opinion updates occur) can be conveniently modelled using the max-plus algebra which we denote R_∞ . Let $\vec{x}(k) \in R_\infty^{(n\times 1)}$ denote the vector of the $(k+1)_{st}$ time agents communicate their opinions. Then $\vec{x}_i(k)$ is the time of the $(k+1)_{st}$ transmission of agent i and is defined, in line with the above description, by:

$$\vec{x}_i(k) = \max\{\vec{x}_j(k-1) + A_{ij} : j = 1, \dots n\}, \text{ for all } k \geq 1. \tag{1}$$

In the notation of the max-plus algebra this becomes,

$$\vec{x}(k) = A \otimes \vec{x}(k-1), \text{ for all } k \geq 1. \tag{2}$$

We also conveniently model the number of time units agent i has been sitting on the opinion value received from agent j before its next update:

$$\vec{\xi}(k,i) = \vec{x}_i(k)\vec{I} - \vec{x}(k-1) - A_i^T \tag{3}$$

where $\vec{I}$ is the $(n \times 1)$ unit column vector and A_i^T is the transposed i_{th} row of A. We refer to the vector above as the *lag-vector* for opinions arriving at agent i, having been sent neighbours at time-step $(k-1)$.

To simulate the resulting opinion dynamics, we modify the Hegselmann-Krause model [5] (which from here, we refer to just as the HK model) to incorporate the lag-vectors. At each time-step $(k+1)$, every agent i updates their opinion according the following update-rule:

$$o_i(k+1) = \left|\mathscr{N}\big(i,k\big)\right|^{-1} \sum_{j \in \mathscr{N}\left(i,k\right)} o_j(k), \tag{4}$$

where $\mathscr{N}(i,k) = \{1 \leq j \leq n \big| 0 \leq \vec{\xi}_j(i,k) \leq \varepsilon\}$, i.e. the set of $i's$ neighbours whose opinion values are sat on by i for at most ε time units. Note the standard, unmodified HK model update-rule is given by replacing $\mathscr{N}(i,k)$ with $\mathscr{M}\big(i,k\big) = \{1 \leq j \leq n \big| |o_i(k) - o_j(k)| \leq \varepsilon\}$.

To summarize the entire process for each time-step: each agent waits until it has received all incoming information (modelled by equation 2). On receipt of the final incoming opinion value for the current time-step, agents update their opinion using the modified HK update-rule (equation 4) and send this to all neighbours.

2 Results

We show via extensive computational simulations that the updated HK model (using the temporally bounded confidence-interval in equation 4) supports multi-opinion consensus clusters despite the absence of the conventional confidence-interval based on the distance between neighbouring opinions (Fig 1). This is significant because it demonstrates that opinion fragmentation is possible even with seemingly innocent sorting of content based only on recency considerations. Simulations are carried out on random weighted strongly-connected and directed Barabási–Albert, Erdős–Rényi and Watts-Strogatz graphs consisting in each case of 100 nodes. Furthermore, we examine typical behaviours emerging from varying the threshold beyond which agents fail to take opinions from their neighbours into account, i.e. from varying ε, while keeping all other initial conditions fixed. Two noticeably distinct regimes emerge. The first is a gradual transition from multiple consensus clusters to a single global consensus. As epsilon is increased, the number of consensus clusters grows slowly, until ε becomes large enough for a single global consensus to form. The second regime is of a more abrupt and discrete change. The number of consensus clusters remains fixed before a sudden transition occurs beyond some critical value of $\varepsilon < 1$, where multiple opinion clusters suddenly collapse into global consensus.

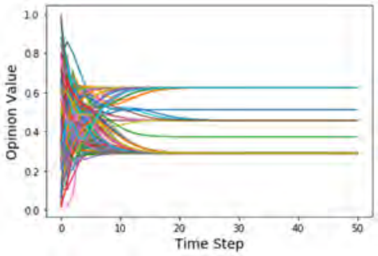


Fig. 1. Multi-opinion consensus clusters emerging using the modified HK update-rule with $\varepsilon = 1$. The simulation was carried on a random weighted strongly-connected directed Barabási–Albert graph with 100 nodes, Barabási–Albert parameter of 2, and edge-weights drawn uniformly at random from $\{1, ..., 20\}$.

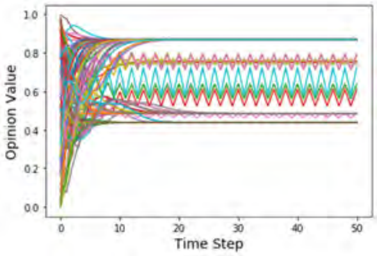


Fig. 2. Oscillatory behaviour of opinions emerging using the modified HK update-rule with $\varepsilon = 1$. The simulation was carried on a random weighted strongly-connected directed Barabási–Albert graph with 100 nodes, Barabási–Albert parameter of 2, and edge-weights drawn uniformly at random from $\{1, ..., 20\}$.

We also uncover a new phenomenon arising from the dynamics using the modified HK update-rule (equation 4), whereby multi-opinion consensus clusters emerge alongside groups of agents exhibiting opinion values which oscillate in time with a regular period (Fig 2). This type of behaviour is not supported by the standard HK model under any circumstance. Using a max-plus periodicity result, we explain this phenomenon analytically by showing that lag-vectors are in fact periodic, with the period being dependent on circuits within the network. Namely, we prove the following: if A is a strongly-connected max-plus adjacency matrix, there exist positive constants k_c and C such that $\vec{\xi}(k+C,i) = \vec{\xi}(k,i)$ for all $k_c \leq k$ and agents i. This provides analytical insight to characterise neighbourhood structures within the network which are susceptible to experiencing periodicity of opinion states.

References

1. Hartmann, Stephan, and Soroush Rafiee Rad. ‘Anchoring in deliberations.’ Erkenntnis (2019): 1-29.
2. Zdep, Stanley, and Warner Wilson. ‘Recency effects in opinion formation.’ Psychological reports 23.1 (1968): 195-200.
3. Hogben, Leslie, ed. Handbook of linear algebra. CRC press, 2006.
4. Heidergott, Bernd, Geert Jan Olsder, and Jacob Van Der Woude. Max Plus at work: modeling and analysis of synchronized systems: a course on Max-Plus algebra and its applications. Vol. 48. Princeton University Press, 2014.
5. Hegselmann, Rainer, and Ulrich Krause. ‘Opinion dynamics and bounded confidence models, analysis, and simulation.’ Journal of artificial societies and social simulation 5.3 (2002).

Dynamic Social Media Network Analysis: an Edge Depreciation Approach

Stefan Katz[1] and Aleksandra Urman[23]

[1] Network Research, polyflow LLC
stefan@polyflow.ch
[2] Institute of Communication and Media Studies, University of Bern, Switzerland
aleksandra.urman@ikmb.unibe.ch
[3] Social Computing Group, Department of Informatics, University of Zurich, Switzerland

1 Introduction

Though there have been several methods proposed for the analysis of temporal networks (see i.e. [2] for an overview), when it comes to social media-based networks (i.e. retweet/follower/friendship networks derived from data from Twitter, Facebook or other platforms) widely used analytic approaches mostly disregard the dynamic nature of such networks. They either discard the temporal aspect altogether collapsing the whole network into a single snapshot [1, 4] or divide the network into several time-based snapshots with fixed start and end dates [5]. While the approach that relies on several temporal snapshots allows authors to trace certain dynamic developments in the network, these developments are only traced within the pre-specified time periods that are chosen somewhat arbitrarily and thus might not cover important developments within the network. In this abstract we propose an approach to the analysis of large dynamic networks that relies on continuous calculation and update of edge strength using what we call edge strength depreciation. We suggest that this approach can be particularly useful for the analysis of large social media-based datasets. To illustrate the potential advantages of the proposed method, we apply it to a social media-based dataset that was analyzed using the snapshot approach[5], and compare the results. The dataset we rely on deals with the data from Telegram. Each Telegram channel and/or group is treated as a node in the network, and each reference (i.e. a mention or a repost) from one channel/group to another is treated as an edge. The details on the data structures and data collection are outlined in the original paper [5].

2 Results

When the analysis is performed on the temporal network snapshots of pre-specified length, the value of any edge in the network is equal within that snapshot. Within a given snapshot, an edge created at the beginning of the time period is treated equally to an edge created at the end of that period. In the case of social media networks such approximation is problematic as it disregards the changing importance of a given message and/or formed or dissolved edge. For instance, in the case of Telegram the importance

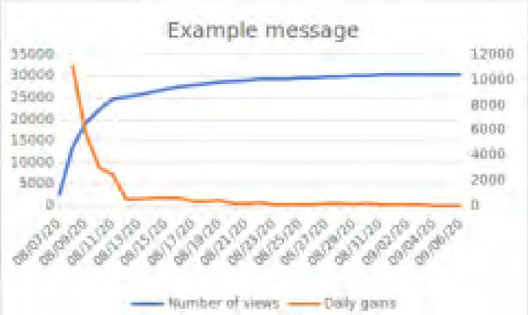


Fig. 1. Change in the number of views overtime for an example message on Telegram.

(by the number of views received) of a given channel mention/repost decreases over time as the attention to this message decreases (see Fig.1).

To account for this decrease in edge value, we suggest to depreciate the value of an edge created on day x by a standard factor. With decreasing values for edges in the past, two nodes only maintain a strong edge by continuous interaction between them. In social media, this means that two channels are in constant interaction and share content, which is what a community should represent. If two channels had very intense interaction a long time ago but then drifted apart, this tie should not be valued as much as more recent ties between two nodes. Self-loops are not taken into account. The edge depreciation is calculated using the following formula:

$$a_t = a_{t-1} * d + e_t$$

Where a is the total edge strength between two nodes at times t and t-1, e is the number of new edges created at time t and d is the depreciation factor (in this case 0.9).

To illustrate the difference in the results obtained on the same network with edge depreciation and without it, we report the authority scores calculated using both approaches for the top channels of the second-largest community in the original network reported in [5] that is largely dominated (in terms of authority scores) by Donald Trump-related channels. We calculated authority scores using both approaches for the top (by static authority) Trump-related channels in this community: realdonaldtrump_bytwitter, realdonaldtrump and trump (see Fig.2). The calculation was performed using HITS algorithm [3]. In Fig.2 the authority scores are shown over the period from February 2019 to February 2020. As described in the original paper [5], from May to July new channels entered the network of the far-right on Telegram. Many of these actors were US-based channels promoting Trump-related content. During these months the dynamic authority (blue line) of these channels was higher than than the static authority (orange line) would indicate. The peaks in dynamic authority show the relative importance of those channels over time. Before the migration, in March, the most used source of these three channels was *realdonaldtrump_bytwitter*. In later stages, this channel's authority decreased and the channel trump became the main source of Trump-related content, with some citing *realdonaldtrump* (name identical to Trump's Twitter handle). These developments that are crucial to the understanding of network evolution are obscured

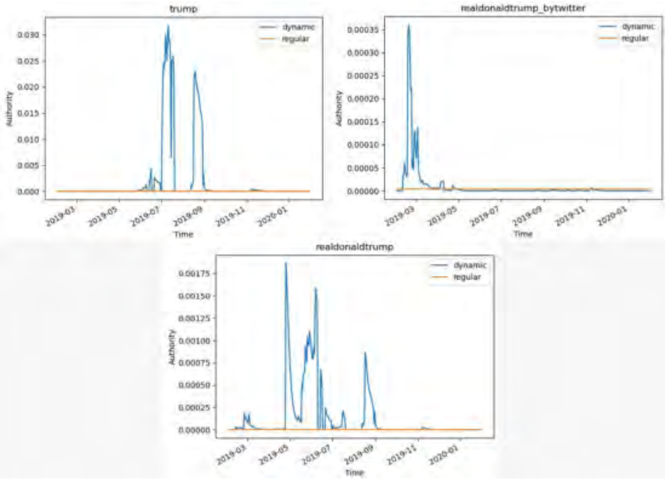


Fig. 2. Channel authority calculated on a network collapsed into a single snapshot vs based on the edge depreciation dynamic calculation

when the snapshot approach is used. Monthly snapshots might combine periods of low authority and periods of high authority that result in averaged authority scores. Using the proposed approach makes evident, which channels have elevated levels of authority in the network at which time without running the risk of not seeing developments due to poorly placed snapshot boundaries. Our approach also shows how different channels interact and replace each other for specific functions in the network. This adds insight to the dynamics of network formation allowing researchers to better understand network evolution while taking into account the changes in the importance of each edge overtime which is particularly important in the context of social media.

References

1. Froio, C., Ganesh, B.: The transnationalisation of far right discourse on Twitter. European Societies 0(0), 1–27 (Jul 2018), https://doi.org/10.1080/14616696.2018.1494295
2. Ghanem, M., Magnien, C., Tarissan, F.: Centrality Metrics in Dynamic Networks: A Comparison Study. IEEE Transactions on Network Science and Engineering 6(4), 940–951 (Oct 2019), conference Name: IEEE Transactions on Network Science and Engineering
3. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. In: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms. pp. 668–677. SODA '98, Society for Industrial and Applied Mathematics, San Francisco, California, USA (Jan 1998)
4. O'Callaghan, D., Greene, D., Conway, M., Carthy, J., Cunningham, P.: An Analysis of Interactions within and between Extreme Right Communities in Social Media. In: Atzmueller, M., Chin, A., Helic, D., Hotho, A. (eds.) Ubiquitous Social Media Analysis. pp. 88–107. Lecture Notes in Computer Science, Springer Berlin Heidelberg (2013)
5. Urman, A., Katz, S.: What they do in the shadows: examining the far-right networks on Telegram. Information, Communication & Society 0(0), 1–20 (Aug 2020), https://doi.org/10.1080/1369118X.2020.1803946, publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2020.1803946

Cluster synchronization on hypergraphs

Anastasiya Salova and Raissa M. D'Souza

University of California, Davis, CA 95616, USA
avsalova@ucdavis.edu

1 Introduction

Cluster synchronization (a type of synchronization where different groups of oscillators in the system follow distinct synchronized trajectories) on networks is a broadly analyzed phenomenon including an important set of behaviors with wide areas of applicability from neuroscience to consensus dynamics [1]. It can be used to understand phenomena such as remote synchronization and chimera states [2]. Ideas from graph and equivariant dynamical systems theory can be applied to deduce admissible patterns of synchronization and simplify their stability analysis. However, higher order interactions may be required to describe many social, biological, and ecological systems [3], making it necessary to go beyond the pairwise interaction analysis to study certain phenomena such as consensus dynamics, epidemic spreading, and metabolic reactions. While complete synchronization and its stability have been analyzed very recently for such systems [4–6], and examples from consensus dynamics, where the system settles on a fixed point, were addressed in Ref. [7], general cluster synchronization has not been considered. To address that, we formulate conditions for cluster synchronization based on the hypergraph structure from equitable partition and symmetry perspective. Then, we show how to reduce the dimensionality of stability calculation based on the hypergraph structure for any specific pattern of cluster synchronization. Our results are an extension of existing cluster synchronization literature to higher order systems and could be of interest for a larger audience.

2 Results

We define a general dynamics of coupled oscillators on a hypergraph using notation similar to Ref. [4]:

$$\dot{x}_i = F_i(x_i) + \sum_{e_j \in \mathscr{E}_i} G_j(x_i, x_{\{e_j\}}). \tag{1}$$

The state of each node of the system is contained in $x_i \in R^n$. The nonlinear function $F_i(x_i)$ describes the evolution of uncoupled nodes. $\mathscr{E}_i$ is the set of edges coming into the node i, and $e_j = j_1, ..., j_m$ is a specific hyperedge of degree m. The function $G_j(x_{\{e_j\}})$ is a coupling function corresponding to the jth hyperedge of the network. While our formalism is applicable to a variety of interaction types (adjacency, non-invasive, directed, etc.), we focus on undirected Laplacian-like coupling $(G_j(x_i, x_{\{e_j\}}) = G(\sum_{k \in e_j}(x_j - x_i)))$ in our example.

We consider homogeneous systems where the node types and coupling types are homogeneous (or consist of several homogeneous blocks). In that case, it is useful to additionally consider the interaction topology, contained in the system's incidence matrix I (or, equivalently, adjacency tensor A) [3].

Cluster synchronization in networked dynamical systems arises from balanced relations on the coupling graph, resulting in equitable and external equitable partitions with partition cells corresponding to fully synchronized nodes. Further analysis of the quotient network can reveal more intricate patterns of synchronization [8]. Equitable partitions divide the network into cells, where the input from each node in C_i to a node in C_j (where $i \neq j$) is uniquely determined by the indexes i and j. Each cell of the network defines a cluster of nodes that could be synchronized.

The same idea is applicable to hypergraphs, where we need to consider input from node i to j from interactions of all orders. To illustrate it, we consider the hypergraph structure of Fig.1(a) (similar to Fig.1 or Ref. [9], added hyperedges) with incidence matrices shown in Fig.1(d). The effective dyadic and triadic interactions that define the structure of cluster synchronization manifold are encoded in quotient (hyper)networks, shown in Fig.1(b-c). We note that symmetries are a subset of balanced equivalence relations, so orbital partitions of nodes with respect to all orders of interaction lead to cluster synchronization patterns in hypergraphs with symmetries. Additionally, we note that for some coupling types, it is helpful to study the projected adjacency matrix [4] to find the admissible cluster synchronization patterns, so methods developed for dyadic interactions (e.g., Ref.[10]) are applicable with a caveat that some patterns obtained on the projected network are not admissible for the original hypergraph. However, Laplacian and noninvasive coupling require explicit considerations of triadic interactions to obtain a full range of admissible cluster synchronization states.

Determining the stability of cluster states is an important step to reveal the conditions under which they can be observed in experiments or natural systems. To simplify the stability calculations for networks with dyadic interactions, one can block diagonalize the system's Jacobian by simultaneously block diagonalizing the set of cluster indicator matrices, together with the Laplacian/adjacency matrix (shown on Fig.1(e)) [1, 9]. We show that a larger set of matrices needs to be block diagonalized to simplify the analysis for higher order systems, since matrices corresponding to a specific cluster synchronization patterns on hyperedges (e.g., *violet-yellow-teal* on Fig.1(a)) have to be additionally considered (Fig.1(f)) to obtain the Jacobian block diagonalization (Fig.1(g)). These additional matrices (effective cluster synchronization Laplacians) can be calculated by using the columns of the incidence matrix corresponding to edges with a specific pattern of synchronization.

Summary. We demonstrate how to deduce admissible synchronization patterns of systems with higher order interactions from the hypergraph structure and the type of interaction dynamics, how this structure manifests itself in stability analysis, and how it can be used to reduce the dimensionality of stability calculations. This extends both the studies of full synchronization on higher order systems and the analysis of cluster synchronization on systems with dyadic interactions. Our next steps will include using this framework to show how higher order interactions stabilize/destabilize cluster states in specific systems with various coupling topologies.

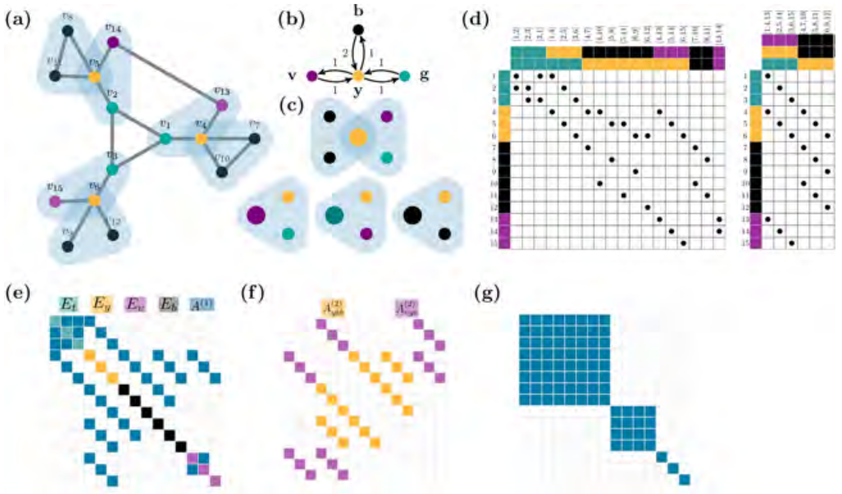


Fig. 1. (a): example cluster synchronization pattern, (b): quotient graph, (c): quotient hypergraph, (d): incidence matrices for dyadic and triadic interactions, (e-f): matrices to simultaneously block diagonalize, (g): block diagonal Jacobian structure.

References

1. Pecora, L.M., Sorrentino, F., Hagerstrom, A.M., Murphy, T.E. and Roy, R., 2014. Cluster synchronization and isolated desynchronization in complex networks with symmetries. Nature communications, 5(1), pp.1-8.
2. Cho, Y.S., Nishikawa, T. and Motter, A.E., 2017. Stable chimeras and independently synchronizable clusters. Physical review letters, 119(8), p.084101.
3. Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.G. and Petri, G., 2020. Networks beyond pairwise interactions: structure and dynamics. Physics Reports.
4. de Arruda, G.F., Tizzani, M. and Moreno, Y., 2020. Phase transitions and stability of dynamical processes on hypergraphs. arXiv preprint arXiv:2005.10891.
5. Mulas, R., Kuehn, C. and Jost, J., 2020. Coupled dynamics on hypergraphs: Master stability of steady states and synchronization. arXiv preprint arXiv:2003.13775.
6. Gambuzza, L.V., Di Patti, F., Gallo, L., Lepri, S., Romance, M., Criado, R., Frasca, M., Latora, V. and Boccaletti, S., 2020. The master stability function for synchronization in simplicial complexes. arXiv preprint arXiv:2004.03913.
7. Sahasrabuddhe, R., Neuhäuser, L. and Lambiotte, R., 2020. Modelling Non-Linear Consensus Dynamics on Hypergraphs. arXiv preprint arXiv:2007.09391.
8. Stewart, I., Golubitsky, M. and Pivato, M., 2003. Symmetry groupoids and patterns of synchrony in coupled cell networks. SIAM Journal on Applied Dynamical Systems, 2(4), pp.609-646.
9. Zhang, Y. and Motter, A.E., 2020. Symmetry-independent stability analysis of synchronization patterns. arXiv preprint arXiv:2003.05461.
10. Kamei, H. and Cock, P.J., 2013. Computation of balanced equivalence relations and their lattice for a coupled cell network. SIAM Journal on Applied Dynamical Systems, 12(1), pp.352-382.

Part V

Ecological Networks and Food Webs

Ecological networks reveal species associations and communities in Urban Forests of South Delhi Ridge, India

Sonali Chauhan[1] , Gitanjali Yadav[2,3][0000-0001-6591-9964] and Suresh Babu[1][0000-0002-3004-3066]

[1] Ambedkar University of Delhi, Delhi 110007, India
[2] National Institute of Plant Genome Research, New Delhi 110067, India
[3] Department of Plant Sciences, University of Cambridge, CB23EA, U.K

1 Introduction

Woodlands and forest remnants in cities consist of unique assemblages of species that range from native species that have survived urbanization, to exotics and ‘escapes’ from cultivation, to invasives that often dominate urban ecosystems. Contemporary literature in ecology even refers to these urban ecosystem as ‘novel ecosystems’ in that they represent unique formations not fully understood by ecosystem ecologists and pose challenges to general principles of community ecology (Hobbs et al., 2009, Aronson et al, 2015). Invasive species are globally considered as major challenge for conservation of biodiversity because they outcompete native species leading to local declines in native species populations. And yet over 60-80% of the urban flora could consist of non-native species. There is increasing evidence that urban nature plays a significant role in modulating the microclimate, and enhancing the quality of life in the cities (Gaston, 2010). These urban woodlands and forests provide significant ecosystem services to urban society and understanding the species composition and unraveling patterns of associations is of considerable importance, for their conservation and management. In this study, we investigate the patterns species associations using a network approach using vegetation data on South Delhi Ridge fragments, with a view to identify communities.

The city of Delhi is constructed over 1100 square kilometers of erstwhile ‘Dry Thorn Scrub Aravali Vegetation’, and agricultural lands and wetlands of the Yamuna river (Champion & Seth 1968; Maheshwari, 1961). Historically, Delhi has been built and rebuilt several times with substantial modification of the landscape during Mughal and British periods. Among these changes were general beautification efforts and creation of public parks and avenues. However, the Delhi Ridge Forest - which is now broadly grouped as the Northern, Central and the South Delhi Ridge - are fragments of the erstwhile Aravali vegetation that survived these transformations. In the last fifty years, the Ridge forest in known to have been overrun by invasive species such as *Prosopis juliflora* and *Lantana camara*, and much of the vegetation now consists of a combination of invasive species, exotics or agricultural escapes, and native Aravali species. The actual relationship between these species however, is less known. Do invasive species decimate native vegetation? If so, whether present forest patches consist of singular stands of invasives? How do native species respond to the propagule pressure of invasive species? In short, what kind of species associations form with a combined impact of urbanization, biological invasions and active use by local communities.

We look at the primary vegetation survey data collected from a standardised ecological survey of six forest fragments in South Delhi region, using the Line Transect Method (Krebs, 1989; Chauhan 2013). In all, the data includes presence/absence and abundance information of 92 species, based on their representation in a total of 73 transects of 200 metres each.

2 Results

2.1 Study Area Networks

We identified over 6000 associations among 92 native, introduced and invasive plant species spread across six study sites in the urban forests of New Delhi, as depicted in the map on Figure 1. Each study site represented a distinct ecological community as established by a rarefaction test, resulting in six individual ecological networks. Species presence-absence matrices were converted to structured information file (SIF) files using our in-house webserver NEXCADE (Yadav & Babu, 2012). The SIF edgelists were exported for visualisation in Cytoscape version 3.6 (Shannon et al., 2003). Any two species found on a given transect were considered 'associated' in the bipartite species association network, to retain both transect and species identity.

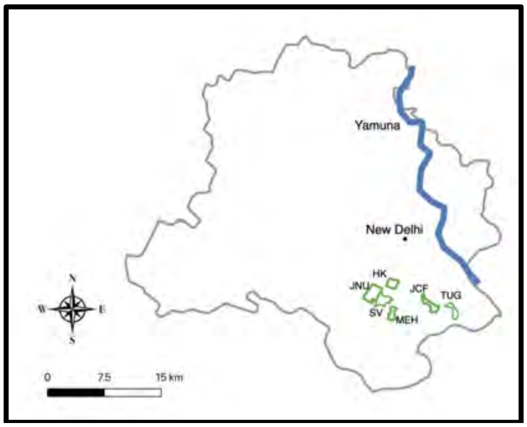
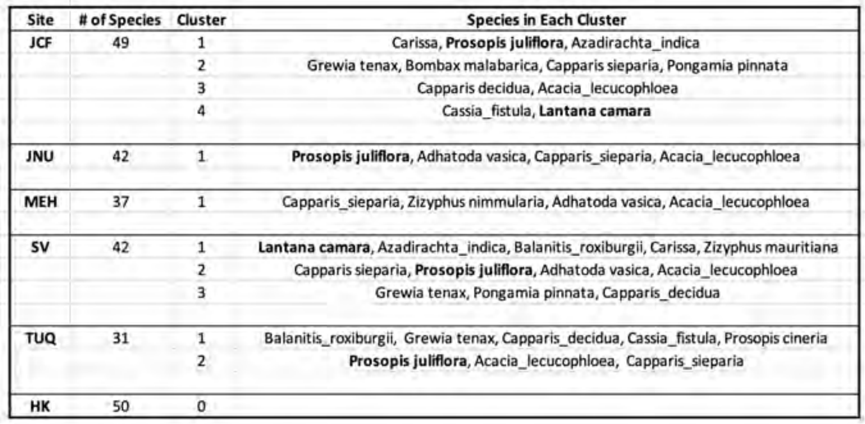


Fig 1: Map of New Delhi indicating the study area represented by the urban forest fragments in South Delhi labelled as: HK - Hauz Khas, JCF - Jahanpanah City Forest, MED - Mehrauli Forest, TUQ - Tughlaquabad Forest, SV- Sanjay Van and JNU - Jawaharlal Nehru University Ridge

2.2 Distinct Native-Invasive Communities

Community detection was performed by parsing each study area network through the MCODE algorithm that identifies densely connected regions using topological parameters (Bader & Hogue 2003). In all, we found 11 species-specific communities across the six urban forest sites. Interestingly, no communities could be detected in the Hauz khas village forest, despite maximum species density, possibly due to the HK network being highly connected. Species membership of clusters is listed in Table 1. The two most highly invasive species *Lantana camara* and *Prosopis juliflora* are members of several communities, but they do not show co-occurrence in the same cliques. Instead, each invasive appears to have evolved a distinct community of native congeners around itself, as evident from species association networks. The largest of these communities were mapped onto the study area networks as depicted in Figure 2.
As can be seen in Figure 2, each study area network has a distinct topology and the largest communities have four to five most connected plants, comprising various

combinations of invasive, native and introduced species. It would be interesting to study the significance of choice of native congeners, for each invasive. This work is currently underway in our group, in terms of species specific networks of invasives in each study area. We are also working towards resolving the differences found here by means of additional layers of information, such as abundance data, and abiotic features of each study site.



Site	# of Species	Cluster	Species in Each Cluster
JCF	49	1	Carissa, **Prosopis juliflora**, Azadirachta_indica
		2	Grewia tenax, Bombax malabarica, Capparis sieparia, Pongamia pinnata
		3	Capparis decidua, Acacia_lecucophloea
		4	Cassia_fistula, **Lantana camara**
JNU	42	1	**Prosopis juliflora**, Adhatoda vasica, Capparis_sieparia, Acacia_lecucophloea
MEH	37	1	Capparis_sieparia, Zizyphus nimmularia, Adhatoda vasica, Acacia_lecucophloea
SV	42	1	**Lantana camara**, Azadirachta_indica, Balanitis_roxiburgii, Carissa, Zizyphus mauritiana
		2	Capparis sieparia, **Prosopis juliflora**, Adhatoda vasica, Acacia_lecucophloea
		3	Grewia tenax, Pongamia pinnata, Capparis_decidua
TUQ	31	1	Balanitis_roxiburgii, Grewia tenax, Capparis_decidua, Cassia_fistula, Prosopis cineria
		2	**Prosopis juliflora**, Acacia_lecucophloea, Capparis_sieparia
HK	50	0	

Table 1: Distinct communities within urban forests show Invasive species (marked in bold) in cliques with sets of distinct native species. Study area codes same as in Fig 1.

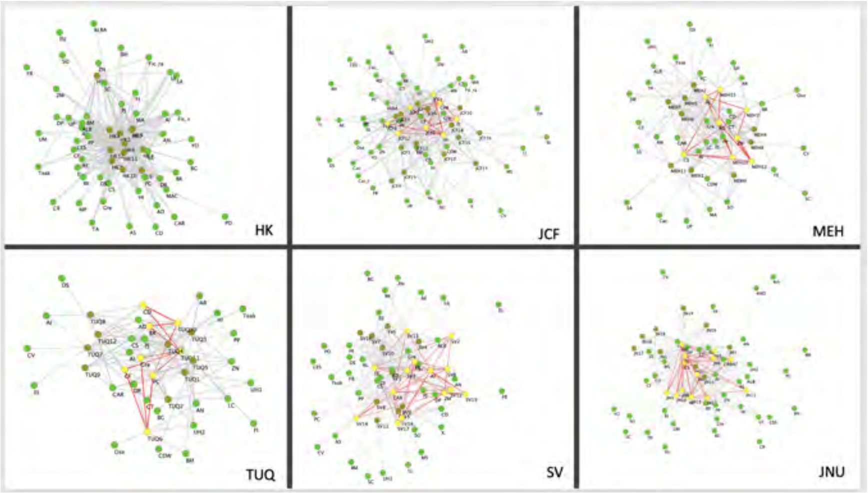


Fig. 2: Study area networks for all six study area sites; Green nodes represent plant species while mint nodes represent location of transects. The largest communities in each study area are highlighted (yellow nodes/ red edges). Study area codes same as in Figure 1.

3 Conclusion

Using a network approach, we identify distinct communities of plants in each of the six urban localities investigated, despite overlaps in overall species composition. There is evidence in the study that indicates formation of native-invasive and native-native associations by major invasive species such as *Lantana camara* and *Prosopis juliflora* in distinct urban forest patches. The associations formed by these invasives are mutually exclusive, in the sense that these two major invaders do not form associations with each other. As such, these invasive-native and invasive-invasive associations seem consistent across different forests, indicating formation of new stable associations in Delhi's woodlands. This is in line with patterns reported urban plant assemblages in other studies, using conventional multivariate vegetation analysis (Aronson et al. 2015, Cilliers and Siebert, 2011). This work indicates that community identification algorithms can find applications in pattern analysis in vegetation ecology and may pave an altogether new way of investigating species associations using networks. Further, this approach may provide an alternate way of visualizing species associations and functional characterization of species based on node attributes such as degree and centrality measures.

References

Aronson M.F.J., Handel S.N., La Puma I.P. and Clemants S.E. (2015) Replacement of native communities with novel plant assemblages dominated by non-native species in the New York metropolitan region. Urban Ecosystems 18: 31–45.

Champion, Harry G.; Seth, S. K. (1968). A revised survey of the forest types of India. New Delhi: Manager of Publications, Government of India.

Chauhan, Sonali (2013). Analysis of woody species in Delhi Ridge: Community Dynamics in Urban Context. MA Thesis *submitted to* School of Human Ecology, Ambedkar University Delhi.

Cilliers, S., Siebert, S. (2011). Urban Flora and Vegetation: Patterns and Processes. *In* Urban Ecology: Patterns, Processes, and Applications. : Oxford University Press.
DOI: 10.1093/acprof:oso/9780199563562.001.0001

Gaston, K., Davies, Z., & Edmondson, J. (2010). Urban environments and ecosystem functions. In K. Gaston (Ed.), *Urban Ecology* (Ecological Reviews, pp. 35-52). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511778483.004

Hobbs, R. J., Higgs, E., & Harris, J. A. (2009). Novel ecosystems: implications for conservation and restoration. *Trends in ecology & evolution*, *24*(11), 599-605.

Krebs, C. J. (1989). Ecological methodology. New York: Harper & Row.

Maheshwari, JK (1961) "The Flora of Delhi," Council of Scientific and Industrial Research, New Delhi

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13:2498–2504.

Yadav G and Babu S (2012) NEXCADE: Perturbation Analysis for Complex Networks. PLoS ONE 7(8): e41827. doi:10.1371/journal.pone.0041827'

Deforestation network fractal analysis in Sumaco biosphere reserve

Andrea Urgilez-Clavijo[1, 2] [0000-0001-5671-6348] and Ana M. Tarquis[1, 3, 4] [0000-0003-2336-5371]

[1] Complex Systems Group (GSC), Universidad Politécnica de Madrid, Madrid, Spain.
[2] IERSE, Universidad del Azuay, Cuenca, Ecuador.
[3] Department of Applied Mathematics, ETSIAAB, Universidad Politécnica de Madrid, Madrid, Spain.
[4] CEIGRAM, ETSIAAB, Universidad Politécnica de Madrid, Madrid, Spain.

Keywords: Connectivity, Local Connected Fractal Dimension, deforestation patterns.

1 Introduction

Landscapes are complex and heterogeneous mosaics constantly transformed by biotic and abiotic factors, disturbances, and human activities such as agriculture, urbanization, and forestry, etc. This transformation is reflected in land use and land cover (LULC) maps at different spatial and temporal scales. These maps are essential inputs to study land use land cover change (LULCC) [1] in pattern identification [2,3], trajectories measures [4], and its relationships [5,6].

In South America, the major LULCC is from primary forest to agricultural land. This forest loss (deforestation) causes landscape fragmentation, isolated patches, habitat and ecosystem services loss, etc. According to the Food and Agriculture Organization of the United Nations [7], Ecuador has maintained the highest deforestation rates of South America with annual rates of 1.5% and 1.8% for 1990 to 2000 and 2000 to 2010 periods, respectively.

Complex networks of dynamical systems have proved their great interest in various studies such as dynamics of forest fragmentation and forest ecosystems [8,9], urban connectivity [10], fish gills analysis [11], etc. Here, we use this complex network framework for studying the connections of cumulative deforested patches identified from LULC maps in *Sumaco* biosphere reserve based on the Local Connected Fractal Dimension (LCFD) frequency distribution. Based on this approach we performed a complementary spatio-temporal characterization of deforestation expansion in Ecuador.

2 Results

Fig. 1 shows LCFD of the cumulative deforestation images for three time-intervals: 1990 to 2000, 1990 to 2008, and 1990 to 2016. The concept of LCFD was used to construct a color-coded dimensional image. Using LCFD values we highlight and discriminate all deforested patches. Green pixels indicate 1,83 ≤ LCFD ≤ 1,88, blue pixels indicate 1,88 < LCFD < 1,93, and red pixels indicate 1,93 ≤ LCFD < 2,00. These LCFD values range across the time intervals suggest an increasing of the complexity in deforested patches and connections between them.

Fig. 2, shows the LCFD distribution of three-time intervals. This confirms an increment of patch network complexity in time. The LCFD of deforested patches distribution increases and exhibit major density in time shifting from 1,83 until 2008 to 1,96 for all period until 2016.

The complexity increasing can be influenced by (1) the growth of pre-existing patches without fusion with the adjacent patches. (i.e., the deforested area grew without increasing the number of patches), (2) the growth and fusion of pre-existing patches (i.e., the pre-existing deforested patches grew until they merged with the adjacent patches, decreasing the number of patches and increasing the area), (3) the appearance of new deforested patches (i.e., due to the fragmentation of pre-existing patches through natural or induced reforestation processes or due to new deforestation). Applying the LCFD provides a spatio-temporal characterization approach to study shape complexity of deforestation process as a complement for landscape ecology metrics.

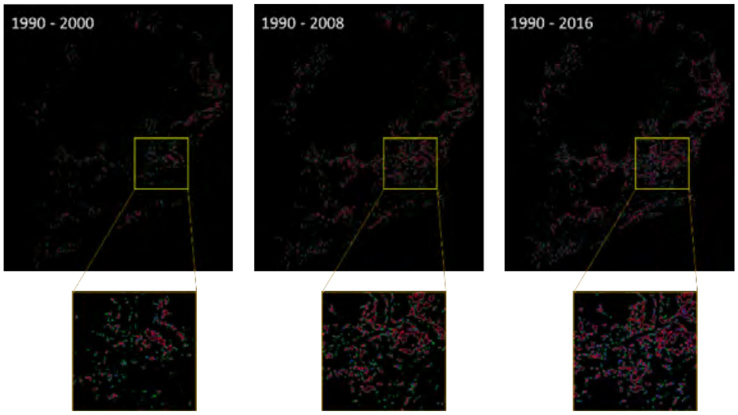


Fig 1. Deforestation network from Local Connection Fractal Distribution (LCFD) in Sumaco biosphere reserve in Ecuador for the three times intervals (1990-2000, 1990-2008 and 1990-2016).

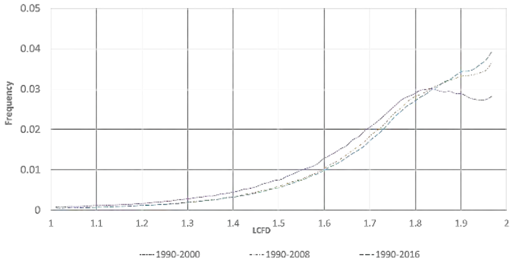


Fig 2. Local connection Fractal Distribution (LCFD) for the three times intervals in Sumaco biosphere reserve in Ecuador.

3 Acknowledgements

The authors acknowledge support from Project No. PGC2018-093854-B-I00 of the Spanish Ministerio de Ciencia Innovación y Universidades of Spain and the financial support from Fundacion Premio Arce project under agreement FPA18PPMAT08.

References

[1] B. Turner, W.B. Meyer, D.L. Skole, others, Global land-use/land-cover change: towards an integrated study, Ambio. Stock. 23 (1994) 91–95.

[2] N. Ramankutty, J.A. Foley, Characterizing patterns of global land use: An analysis of global croplands data, Global Biogeochem. Cycles. 12 (1998) 667–685.

[3] G.F. Curatola Fernández, W.A. Obermeier, A. Gerique, M.F. López Sandoval, L.W. Lehnert, B. Thies, J. Bendix, Land cover change in the Andes of Southern Ecuador—Patterns and drivers, Remote Sens. 7 (2015) 2509–2542.

[4] C.F. Mena, Trajectories of Land-use and Land-cover in the Northern Ecuadorian Amazon, Photogramm. Eng. Remote Sens. 74 (2008) 737–751.

[5] C. Monfreda, N. Ramankutty, T.W. Hertel, others, Global agricultural land use data for climate change analysis, Econ. Anal. L. Use Glob. Clim. Chang. Policy. 14 (2009) 33.

[6] A. Urgilez-Clavijo, J. de la Riva, D. Rivas-Tabares, A.M. Tarquis, Linking deforestation patterns to soil types: A multifractal approach, Eur. J. Soil Sci. (2020). https://doi.org/10.1111/ejss.13032.

[7] FAO, State of the World's Forests, Food and Agriculture Organization of the United Nations, Rome, 2011. https://doi.org/10.1103/PhysRevLett.74.2694.

[8] I. Andronache, R. Fensholt, H. Ahammer, A.M. Ciobotaru, R.D. Pintilii, D. Peptenatu, C.C. Draghici, D.C. Diaconu, M. Radulović, G. Pulighe, A.F. Azihou, M.S. Toyi, B. Sinsin, Assessment of textural differentiations in forest resources

in Romania using fractal analysis, Forests. 8 (2017). https://doi.org/10.3390/f8030054.

[9] G. Cantin, N. Verdière, Networks of forest ecosystems: Mathematical modeling of their biotic pump mechanism and resilience to certain patch deforestation, Ecol. Complex. 43 (2020) 100850. https://doi.org/10.1016/j.ecocom.2020.100850.

[10] B. Swaid, E. Bilotta, P. Pantano, R. Lucente, Thresholding urban connectivity by local connected fractal dimensions and lacunarity analyses, Late Break. (2015) 15.

[11] M. Manera, L. Giari, J.A. De Pasquale, B.S. Dezfuli, Local connected fractal dimension anaysis in gill of fish experimentally exposed to toxicants, Aquat. Toxicol. 175 (2016) 12–19.

Assembling Mutualistic Networks from Adaptive Niche Interactions

Weiran Cai[1], Jordan Snyder[12], Alan Hastings[3,4], Raissa M. D'Souza[14]

[1] Department of Computer Science, UC Davis, Davis, CA 95616
[2] Department of Mathematics, UC Davis
[3] Department of Environmental Science and Policy, UC Davis
[4] Sante Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501

1 Introduction

Mutualism is a vital element that comprises natural and social systems. As evidenced in plant-pollinator relations and designer-contractor partnerships, the actors typically exhibit an ordered pattern of collective interactions. The interspecific relation can be represented by a bipartite network of multiple species (plant and animal guilds for example). Such networks consistently express highly modular and nested structures, compared with their randomized counterparts [1, 2]. Whether and how the two structural properties of mutualistic networks can emerge out of a unified mechanism however remains unclear [3, 4]. Here, we elucidate a unified principle that explains how high-level mutualistic network patterns can emerge from interactions that adapt based on the niches of individual actors. Key dynamical properties are revealed at different time scales, ranging from network stability to environmental impacts [6, 5].We further demonstrate that such adaptiveness is crucial for preserving the network structure under invasions.

2 Dynamic Niche Model

We consider the co-adaptation of species niche relations and the demographic distribution of population under a single incentive of maximizing individual fitnesses. We extend Hutchinson's conception of niche adaptation to a network of cooperative species [7]. Multiple species in two distinct guilds are involved simultaneously in mutualistic interactions with selected partner species in the opposite guild and subject to competition with all rival species within their own guild (Fig. 1a). Each species possesses two fundamental characteristics: its niche and abundance. The niche profile is formulated by a Gaussian function $H_i(s)$, representing the niche distribution on a niche axis. If two species interact, their coupling strength is proportional to their niche overlap H_{ij} [7].

$$\text{mutualistic:} \quad \gamma_{ik} = \Omega_m \cdot \theta_{ik} \cdot H_{ik}; \qquad \text{competitive:} \quad \beta_{ij} = \begin{cases} 1, & i = j \\ \Omega_c \cdot H_{ij}, & i \neq j \end{cases}$$

where the niche overlap is the joint probability of occupying the same position of two species on the niche axis $H_{ij} = \int H_i(s) H_j(s) ds$. The species abundances follow the generalized Lotka-Volterra dynamics with mutualistic functional response. At fixed time intervals, a randomly chosen species attempts to rewire to a different mutualistic partner in the opposite guild $\gamma_{ik} \longrightarrow \gamma_{ik'}$ to maximize its own abundance (Fig. 1b) [4].

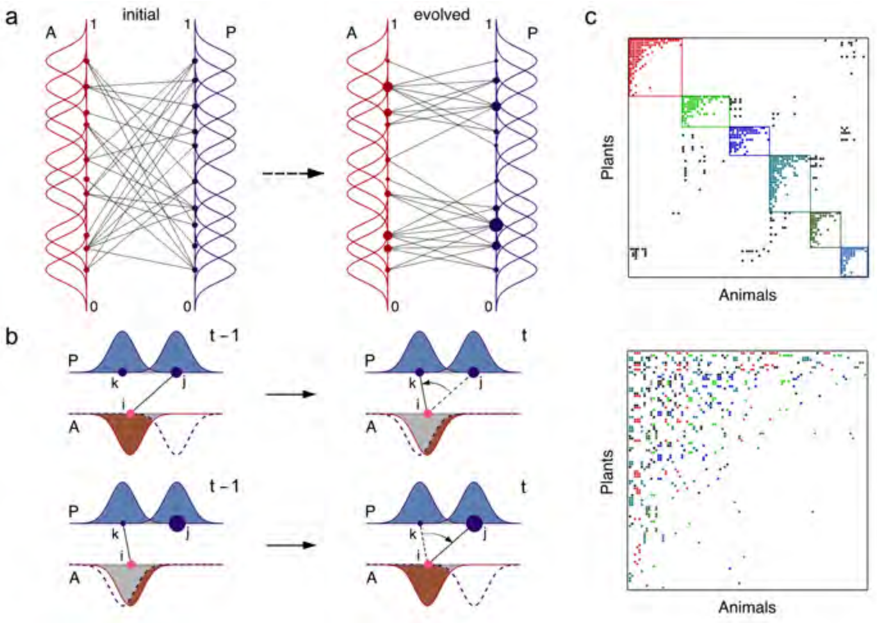


Fig. 1. Dynamic niche model. **a** Adaptation of niche relation. **b** Rewiring in one time interval. **c** Modular and nested interaction pattern in evolved network shown in adjacency matrices.

3 Results

With these premises, we demonstrate that nestedness NODF and modularity Q can emerge concurrently from an initial random structure (exemplified in Fig. 1c). As shown in Fig. 2a, the dyadic measures (NODF, Q) of an ensemble of 300 generated networks (red) show a significant overlap with those of the 144 empirical networks (blue) from the Web of Life Dataset and exhibit a similar negative correlation (dashed lines). The degree distribution evolves into a typical truncated power law when the overall interaction intensity are high (Fig. 2b). Heuristically, the modular and nested structure is formed through a positive feedback of local advantages in the structural and demographic distributions. This process contrasts with the existing models that handle them with separate mechanisms. This type of dynamics belongs to a broad class of localized preferential attachment processes, whereby 'the rich get richer' under the constraints on the potential linkage.

We then delve into the dynamical properties of the network at different time scales. At the ecological scale, we show the decisive role of within-guild competition intensity. An evolved mutualistic network is more or less stable than the randomized counterpart when the overall competition intensity is higher or lower than a transition point (Fig. 2c and 2d), despite the intensity of mutualistic interactions. We further illustrate that the interspecific linking pattern may exhibit a strong history-dependency in response to environmental changes, as shown by the hysteretical trajectory of the structural measures (NODF and Q) with the changing mutualistic interaction intensity (Fig. 2e and 2f). The

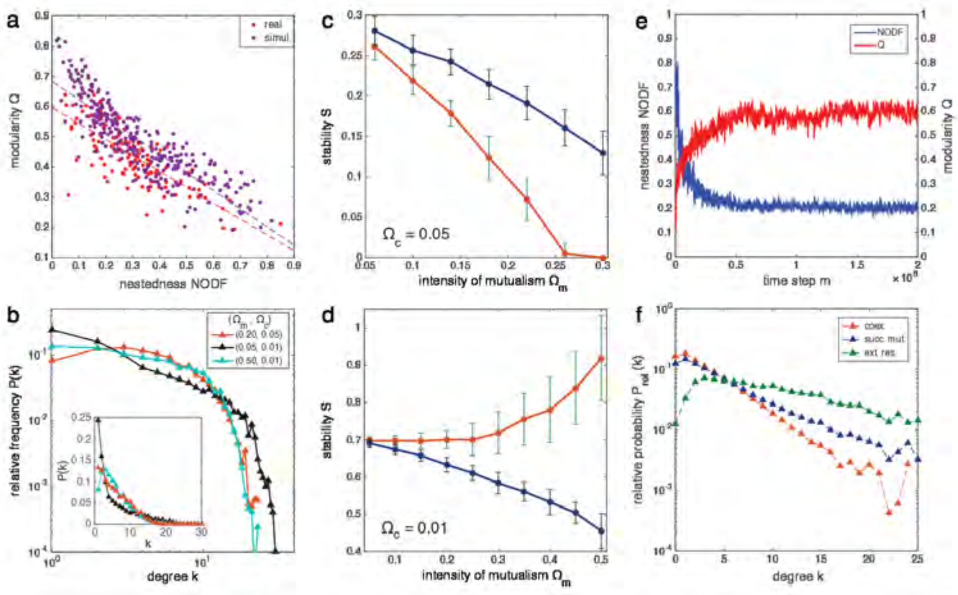


Fig. 2. Structural and dynamical properties. **a** Comparison of dyadic measures $(NODF, Q)$ of generated and empirical networks. **b** Power law degree distribution. **c, d** Role of competition intensity on stability. **e** Structural measures under invasions. **f** Relative probabilities of survival, extinction and coextinction.

structural alteration may be irreversible even if the original environmental condition is recovered. At the evolutionary timescale, we show that a nontrivial nested and modular architecture persists in the presence of repeated invasions and extinctions, by playing out whether the mutant survives, goes extinct, replaces or coexists with the resident.

References

1. Bascompte, J. and Jordano, P. and Melian, C. J. and Olesen, J. M.: The nested assembly of plant-animal mutualistic networks. Proc. Natl Acad. Sci. USA, 100, 9383–9387 (2003)
2. Olesen, J. M. and Bascompte, J. and Dupont, Y. L. and Jordano, P: The modularity of pollination networks. Proc. Natl Acad. Sci. USA 104, 19891–19896 (2007)
3. Saavedra, S. and Reed-Tsochas, F. and Uzzi, B.: A simple model of bipartite cooperation for ecological and organizational networks. Nature. 457, 463–466 (2009)
4. Suweis, S. and Simini, F. and Banavar, J. R. and Maritan, A.: Emergence of structural and dynamical properties of ecological mutualistic networks. Nature 500, 449–452 (2013)
5. Thébault, E. and Fontaine, C.: Stability of ecological communities and the architecture of mutualistic and trophic networks. Science. 329, 853–856 (2010)
6. Okuyama, T. and Holland, J. N.: Network structural properties mediate the stability of mutualistic communities. Ecology Letters 11, 208–216 (2008) (2008)
7. Hutchinson, G. E.: Concluding remarks. Cold Spring Harbor Symp. Quant. Biol. 22, 415–427 (1957); MacArthur, R. and Levin, R.: The limiting similarity, convergence, and divergence of coexisting species. The American Naturalist 101, 377–385 (1967)

Part VI

Link Analysis and Ranking

Random Walk Decay Centrality

Tomasz Wąs[1], Talal Rahwan[2], and Oskar Skibski[1]

[1] Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland
{t.was,o.skibski}@mimuw.edu.pl
[2] New York University Abu Dhabi,
PO Box 129188, Saadiyat Island, Abu Dhabi, United Arab Emirates

1 Introduction

Centrality measures are among the most important tools of the network analysis. Many of them, e.g., *Closeness centrality*, *Betweenness centrality* [2], or *Decay centrality* [3], are based on distances and shortest paths in the graph. It stems from a simplifying assumption that the information in the network always travels through the fastest routes. However, multiple real-life situations clearly violate such assumption. In fact, whether one is modeling the spread of gossip through a social group, the propagation of viruses through a computer network, or the way users surf the Internet, such processes tend to spread more chaotically in practice, e.g., through somewhat random paths as opposed to shortest paths [1].

To better model such scenarios, a number of centrality measures based on the random walks have been proposed in the literature. Arguably, the most famous such a measure is *PageRank* [4]. To see how the PageRank works, consider the following model, which is a modification of the standard Random Surfer model [4]. Imagine a surfer that starts traversing the network in a random node, $w(0)$. Then, in each step, with probability a, she randomly chooses one of the outgoing edges of the node she currently occupies and follows it to the next node. At the same time, with probability $1-a$, the surfer stops the walk all together. The PageRank of a node is the expected number of times the surfer visits this node in her walk, i.e., $PR_v(G) = \sum_{t\geq 0} \mathbb{P}_G(w(t) = v)$ where $w(t)$ is the node visited by the surfer at moment t.

In this paper, we propose a novel centrality measure called *Random Walk Decay centrality* (RWD), which is defined not as the expected number of visits, but as the probability that the node is visited at all, i.e., $RWD_v(G) = \mathbb{P}_G(\{s : w(s) = v\} \neq \emptyset)$. The advantage of such a definition is that the centrality of a node does not depends on its outgoing edges—a property that can be interpreted as a resistance to manipulation. In order to further understand PageRank and RWD, we create an axiomatic characterization of both measures. Specifically, we show that PageRank is the unique centrality measure satisfying six simple properties (axioms): *Locality*, *Sink Merging*, *Directed Leaf Proportionality*, *One-Node Graph*, *Random Walk Property* and *Edge Swap*. In turn, the first five axioms and *Lack of Self-Impact* instead of Edge Swap uniquely characterize RWD. Our analysis highlights the similarities between both measures and their key differences in an easy to grasp and comprehensive way.

2 Results

Let us begin by introducing the axioms that are satisfied by both PageRank and RWD:

> ***Locality (LOC):*** *For two disjoint graphs $G = (V,E)$ and $G' = (V',E')$, the centrality of every $v \in V$ in the union of G and G' is the same as in G.*
>
> ***Sink Merging (SM):*** *For every graph $G = (V,E)$ merging two sinks $u,v \in V$ without common predecessors does not affect the centralities of the remaining nodes in the graph; moreover the centrality of the merged node is the sum of the centralities of nodes u and v in graph G.*
>
> ***Directed Leaf Proportionality (DLP):*** *There exists $a \in (0,1)$ such that for every $G = (V,E)$, sink u and isolated node v, the centrality of v in G with edge (u,v) added is equal to the centrality of v in G plus a times the centrality of u in G.*
>
> ***One-Node Graph (1NG):*** *In a graph that consists of one node and no edges, the centrality of the node is equal to its node weight.*
>
> ***Random Walk Property (RWP):*** *For two graphs $G = (V,E)$ and $G' = (V',E')$ and node $v \in V \cap V'$ if for every $t \geq 0$ and $k \geq 1$ the probability that the random walk visits v in moment t for the k-th time is equal in G and G', then the centrality of v in G and G' is also equal.*

To uniquely characterize RWD we need an additional axiom, *Lack of Self-Impact*, stating that the outgoing edges of a node does not have an impact on its centrality.

> ***Lack of Self-Impact (LSI):*** *For every graph $G = (V,E)$ and edge $(u,v) \in E$ removing edge (u,v) does not affect the centrality of u.*

Theorem 1. *Random Walk Decay centrality is a unique centrality measure that satisfies LOC, SM, DLP, 1NG, RWP, and LSI.*

Now, if we replace Lack of Self-Impact with *Edge Swap* stating that edges coming from nodes with the same centrality and number of outgoing edges are interchangeable, we obtain the unique characterization of PageRank.

> ***Edge Swap (ES):*** *For every graph $G = (V,E)$ and edges $(u,u'),(v,v') \in E$ if nodes u,v have equal centralities and equal number of outgoing edges, then replacing edges (u,u') and (v,v') with (u,v') and (v,u') does not affect the centrality of any node in the graph.*

Theorem 2. *PageRank is a unique centrality measure that satisfies LOC, SM, DLP, 1NG, RWP, and ES.*

In the remainder, we focus on the two axioms that distinguish RWD and PageRank.

Consider Lack of Self-Impact that is satisfied by RWD, but not by PageRank. If we assume that nodes decide upon their outgoing edges (e.g., Twitter users decide who to follow), then the axiom states that the centrality is not affected by the choices of the node. In other words, it assures strategyproofness of a centrality measure. As an

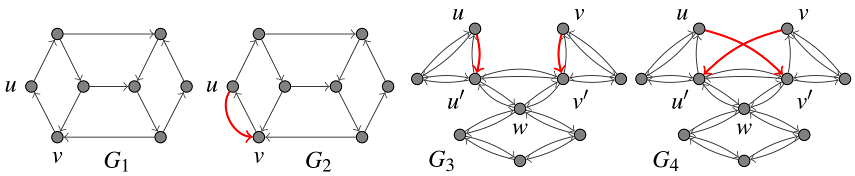


Fig. 1. Four graphs highlighting the differences between PageRank and RWD.

example, consider graph G_1 in Fig. 1. There, for $a = 0.8$ node u is ranked as the fifth both by PageRank and RWD. Now, in G_2, an artificial edge from u to v is added. RWD of u is the same in both graphs and in fact node u is still ranked as fifth. On the contrary, PageRank yields much greater centrality of u in G_2 and ranks it now as the second.

Consider Edge Swap that is satisfied by PageRank, but not by RWD. It states that if two nodes, u, v, have the same centrality and out-degree, then for the centrality of another node it does not matter if it receives an edge from u or v. Hence, other aspects such as graph structure or diversity of the predecessors do not play a role. Consider graph G_3 from Fig. 1. There are three communities connected by three nodes: u', v', and w. According to both PageRank and RWD, w is the most important node and u' and v' are ranked as the second. Moreover, both measures give the same centrality to u and v, which also have the same number of outgoing edges. Hence, based on Edge Swap we know that in graph G_4 PageRank of every node is the same as in G_3. Node w is still the most important. However, RWD does not satisfy Edge Swap and the centrality of u' and v' is higher in G_4 than in G_3, because their sets of predecessors are more diverse. As a result, according to RWD, u' and v' are more central than w in G_4.

The comparison of Random Walk Decay centrality to standard Decay centrality [3] and more details can be found in the full version of the paper [5].

References

1. Borgatti, S.P.: Centrality and network flow. Social Networks 27(1), 55–71 (2005)
2. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry pp. 35–41 (1977)
3. Newman, M.E.J.: A measure of betweenness centrality based on random walks. Social Networks 27(1), 39–54 (2005)
4. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Stanford InfoLab (1999)
5. Wąs, T., Rahwan, T., Skibski, O.: Random walk decay centrality. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). pp. 2197–2204 (2019)

Dependency Networks and Systemic Risk in Open Source Software Ecosystems

William Schueller[1] and Johannes Wachs[1,2]

[1] Complexity Science Hub Vienna
[2] Institute for Information Business, Vienna University of Economics and Business
johanneswachs@wu.ac.at

Trends in the creation and use of software are increasing the macroscopic complexity of open source software ecosystems. Software is increasingly modular: developers write smaller libraries that carry out specialized functions. These compact libraries interact and depend on one another. The resulting ecosystem of libraries is efficient: a developer with an new idea for a new car doesn't have to reinvent the wheel. It also offers an explosion of possibilities for the recombination of code [1].

Like in many networked systems such as the financial system [2] or supply chains [3], failure of a single element in OSS ecosystems can lead to a cascade of failures within the network - presenting a systemic risk. Recent events such as the leftpad incident, the Heartbleed vulnerability in OpenSSL, and the EventStream hack show how issues in single libraries can propagate through much of the network [1]. Failures also frequently occur when library owners stop maintaining their code - a concept known as Lehman's Law in computer science [4].

In this work we examine the network of dependencies of libraries in the Rust programming language ecosystem. We simulate the spread of failures in this system to quantify systemic risk. Using data from Cargo (crates.io), the Rust package manager, we create the dependency networks between Rust libraries at different points in time. In these networks, a library is connected by a directed edge to the packages it depends on (qualified as *upstream*). Specifically, we observe the system on June 30th annually from 2015 to 2020. Rust is a particularly well-suited for our study. As a quite young language (created 2010), the history of almost all of its libraries is available.

Previous work in software engineering has noted that the number of direct dependencies (a library's in-degree in the network) and transitive dependencies (two-step in-degree) are growing over time in several major ecosystems [5,6]. These measures are good first order approximations of the relative importance of individual libraries and can capture broad changes in system connectivity over time. But as we know from studies of interbank networks, supply chains, and powergrids - the complexity of these networks can often obscure their most vulnerable points [7].

To better measure the contribution of individual libraries to systemic risk, we adopt a simulation approach. For each library in a snapshot we simulate its failure 100 times. Failures spread from the library to each direct downstream dependency with fixed probability p, which we set to .1 (our results are robust to a variety of probabilities). Each failure of a library induces the same check at each downstream library. Repeating this simulation for each node in the system, we obtain library-level distributions for failure cascade lengths. We can aggregate these distributions to the system level and observe changes in the global distribution over time.

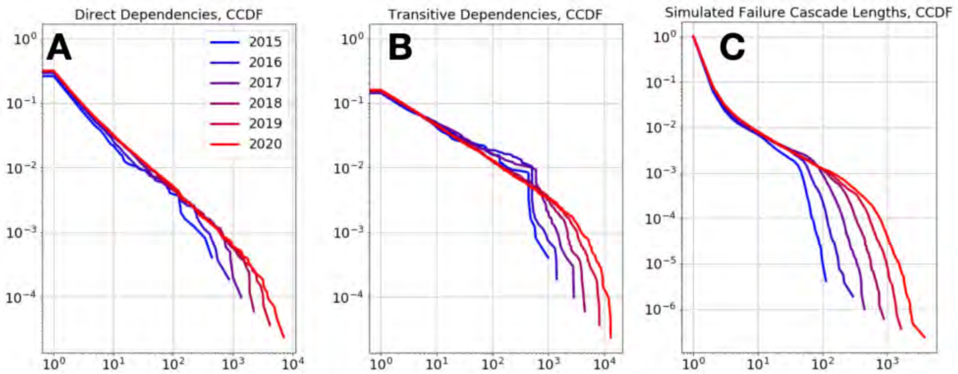


Fig. 1. Complementary cumulative distribution functions of various connectivity statistics of libraries in the Rust ecosystem. A) the direct dependencies (in-degrees) of libraries. B) the transitive (two-step in-degrees) of libraries. C) the lengths of simulated failure cascades starting from each library. Complex connectivity and systemic risk is growing over time.

Library	Direct Deps.	Trans. Deps.	Cascade Length
serde	7146	12486	1737
serde_derive	3732	12189	1358
rand	3881	13036	1036
lazy_static	3301	13147	962
serde_json	5412	10903	962
rand_core	144	4310	895
log	4919	9446	871
cfg-if	377	9350	820
quote	1817	8744	810
rustversion	55	9456	810
trybuild	149	9594	799
serde_test	117	6997	763
syn	1799	7376	726
serde_stacker	4	5418	688
regex	2388	9146	644
sval	3	4919	642
libc	3160	8074	637
proc-macro2	1375	8956	616
serde_bytes	118	6070	607
tokio	1995	5990	601

Table 1: Top 20 packages in the Rust ecosystem, June 2020, by characteristic cascade length.

In Figure 1 we plot the CCDFs of various connectivity measures for the Rust ecosystem observed annually. The system is growing and the number of direct dependencies seems to scale with the size of the system. The number of transitive dependencies is also growing, but the distribution is getting more skewed. Finally, the distribution of simulated failure cascades suggests that systemic risk is increasing over time.

As expected, the number of direct dependencies, transitive dependencies, and the characteristic lengths of cascades[3] originating from specific libraries are significantly correlated (Spearman rank correlations between .5 and .9).

In Table 1 we list the top 20 libraries by characteristic cascade length for the most recent snapshot of the Rust ecosystem. We also report the number of direct and transitive dependencies, noting that there are systemically important libraries with few direct dependencies. This suggests that there are some less visible libraries of systemic importance in the ecosystem.

In future work we propose to widen our analysis by considering more ecosystems, and to go into greater depth by studying social aspects of ecosystem stability. In particular, future work should consider the population of library maintainers, the use of the libraries (measured by downloads), and their visibility (measured by watches or stars on GitHub, volume of Google search, mentions on Stack Overflow).

References

1. Eghbal, N.: Roads and bridges: The unseen labor behind our digital infrastructure. Ford Foundation (2016)
2. Battiston, S., Gatti, D.D., Gallegati, M., Greenwald, B., Stiglitz, J.E.: Liaisons dangereuses: Increasing connectivity, risk sharing, and systemic risk. Journal of economic dynamics and control **36**(8) (2012) 1121–1141
3. Shukla, A., Lalit, V.A., Venkatasubramanian, V.: Optimizing efficiency-robustness trade-offs in supply chain design under uncertainty due to disruptions. International Journal of Physical Distribution & Logistics Management (2011)
4. Lehman, M.M.: On understanding laws, evolution, and conservation in the large-program life cycle. Journal of Systems and Software **1** (1979) 213–221
5. Kikas, R., Gousios, G., Dumas, M., Pfahl, D.: Structure and evolution of package dependency networks. In: 2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR), IEEE (2017) 102–112
6. Decan, A., Mens, T., Grosjean, P.: An empirical comparison of dependency network evolution in seven software packaging ecosystems. Empirical Software Engineering **24**(1) (2019) 381–416
7. Buldyrev, S.V., Parshani, R., Paul, G., Stanley, H.E., Havlin, S.: Catastrophic cascade of failures in interdependent networks. Nature **464**(7291) (2010) 1025–1028

[3] Here we present results using the 90th percentile failure cascade length of each library.

Assessing the Relationship Between Centrality and Hierarchy Measures in Complex Networks

Stephany Rajeh[1], Marinette Savonnet[1], Eric Leclercq[1], and Hocine Cherifi[1]

Laboratoire d'Informatique de Bourgogne - University of Burgundy, 21000 Dijon, France
stephany.rajeh@u-bourgogne.fr

1 Introduction

Identifying influential nodes in complex networks allows to uncover key spreaders for marketing campaigns, finding essential proteins, detecting financial risks, inhibiting diseases, and many more. Centrality measures are one of the main ways of assessing a node's importance. These measures are mainly based on the connections of the nodes and the dynamics of the network [1]. Another way of assessing a node's importance is through hierarchy. Hierarchical structure is pervasive and natural among real-world networks [3]. The hierarchical decomposition of networks using hierarchy measures results in nodes that exist at the core of the network. Even though several studies have investigated the relationship between centrality measures [4] [5], to our knowledge, the interplay between hierarchy and centrality is unexplored. In order to fill this gap, an extensive investigation has been conducted in order to gain a better understanding about the relationships between hierarchy and centrality measures together with basic topological properties of networks [6]. Centrality measures under test incorporate information from the neighborhood (Degree and Local), from the flow of resources (Betweenness and Current-flow Closeness), and from an iterative refinement process of the network structure (Katz and PageRank). Note that Local centrality is Degree centrality extended to the second order neighborhood of the direct nodes. Hierarchy measures are based on nestedness of the network (*k*-core and *k*-truss), flow of resources (LRC [3]), and a mix between nestedness and flow (Triangle participation ratio). Experiments are performed on 28 real-world networks spanning a broad spectrum of real-world applications. Three main questions are investigated:
1) Do hierarchy and centrality measures convey the same information?
2) What is the influence of the macroscopic topological properties on their relationship?
3) Which are the most orthogonal centrality and hierarchy measures?

2 Results

To answer the first question, correlation measures (Pearson, Spearman, and Kendall-Tau) are calculated among the 4 hierarchy α_i and 6 centrality β_j measures for each network. Similarity measures are also considered (Jaccard and Rank-biased Overlap). Results show that there is a high variability of both correlation and similarity measures between hierarchy and centrality for the various real-world networks under test. Nevertheless, analyzing the results of the 28 networks under test 6 main categories emerge.

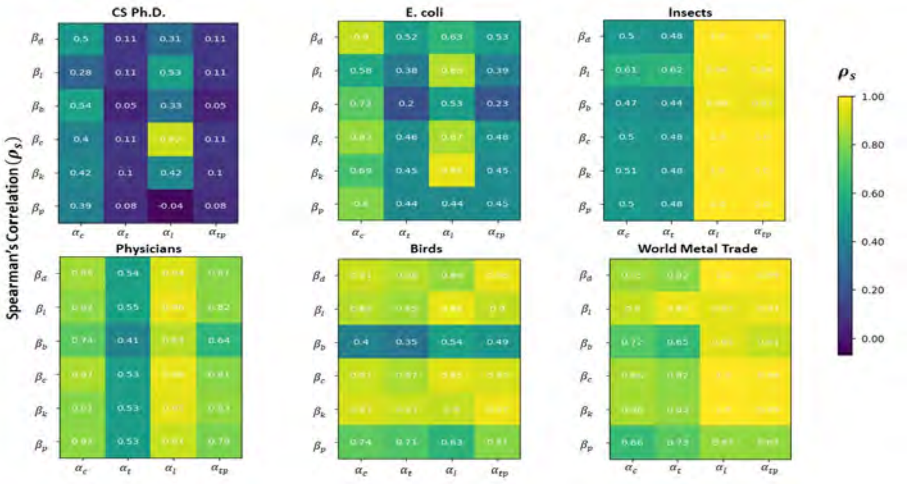


Fig. 1. Heatmaps of the Spearman's correlation for the various combinations of hierarchy α_i and centrality β_j measures of 6 real-world networks. The hierarchy measures are α_c = k-core, α_t = k-truss, α_l = LRC, and α_{tp} = triangle participation. The centrality measures are β_d = Degree, β_l = Local, β_b = Betweenness, β_c = Current-flow Closeness, β_k = Katz, and β_p = PageRank.

They range from low correlation to high correlation between hierarchy and centrality measures. Additionally, k-truss and to a lesser extent k-core appear to be less correlated with centrality measures as compared to other hierarchy measures. Figure 1 reports the Spearman's correlation results of 6 real-world networks exhibiting these behaviors.

To answer the second question two experiments are performed. First of all the correlation and similarity measures are binarized using thresholding (μ >0.7) to distinguish between meaningful and non-meaningful values. A value of μ >0.7 is chosen because it is considered as a high value for meaningful similarity and correlation. Networks are then ranked based on the ratio of meaningful values. Table 1 shows three categories of networks ranked according to their meaningful proportion of correlation. Their main topological characteristics (density, transitivity, and assortativity) are also reported. A clear association between hierarchy, centrality, and network topology can be seen after the reduction of the 6 categories from the first experiment into 3 categories. The first category (high fraction of meaningful correlation) is characterized by high density and transitivity, alongside negative assortativity. The second category (medium fraction of meaningful correlation) exhibit low density, high transitivity, and positive assortativity. The third category (low fraction of meaningful correlation) is made of networks with low density and transitivity with negative assortativity. The second experiment allows to check the consistency of the results. Networks are clustered based on their correlation/similarity feature sets using the k-means algorithm with three clusters. Results show similar clusters of the networks as compared to the network categories obtained after thresholding. Hence, density and transitivity, and to a lesser extent assortativity, play a major role in the redundancy of information among hierarchy and centrality

Table 1. Real-world networks grouped according to their meaningful correlation proportion. The basic topological characteristics are: ν is the density, ζ is the transitivity, and $k_{nn}(k)$ is the assortativity. Two states can be given to the density and transitivity, either high denoted as H or low denoted as L. Two states can be given for assortativity, either positive denoted as P or negative denoted as N.

Network Categories	ν	ζ	$k_{nn}(k)$
Category 1: Adjective Noun, Zachary Karate, Les Misérables, World Metal Trade, U.S. Airports, Madrid Train Bombings, Birds, and Mammals	H	H	N
Category 2: Physicians, Facebook Politician Pages, Facebook Ego, Insects, U.S. States, AstroPh, GrQc, Adolescent Health, Reptiles, and PGP	L	H	P
Category 3: Retweets Copenhagen, Internet A. Systems, NetSci, Human Protein, E. coli Transcription, Mouse Visual Cortex, Yeast Protein, U.S. Power Grids, EuroRoads, and CS Ph.D.	L	L	N

measures. High density and high transitivity induce high correlation among hierarchy and similarity. On the other hand, the information extracted by hierarchy and centrality is less redundant in low density and/or low transitivity networks.

Finally, to answer the third question, the Schulze voting method is used to rank the couples of hierarchy and centrality measures from most orthogonal to least orthogonal. The networks are the voters and the various (hierarchy, centrality) couples are the candidates. The winner is the couple (*k*-core, betweenness) followed by (*k*-truss, betweenness). These results suggest that one can take advantage of the complementary information carried by nested hierarchy measures and betweenness centrality to design effective measures of influence. Future work will investigate the relations between classical and community-aware centrality measures [7–9].

References

1. Lü, L., Chen, D., Ren, X. L., Zhang, Q. M., Zhang, Y. C., & Zhou, T. (2016). Vital nodes identification in complex networks. Physics Reports, 650, 1-63.
2. Borgatti, S. P. (2005). Centrality and network flow. Social networks, 27(1), 55-71.
3. Mones E, Vicsek L, Vicsek T (2012) Hierarchy Measure for Complex Networks. PLoS ONE 7(3): e33799. https://doi.org/10.1371/journal.pone.0033799
4. Li, C., Li, Q., Van Mieghem, P., Stanley, H. E., & Wang, H. (2015). Correlation between centrality metrics and their application to the opinion model. The European Physical Journal B, 88(3), 1-13.
5. Oldham, S., Fulcher, B., Parkes, L., Arnatkevičiūtė, A., Suo, C.& Fornito, A. (2019). Consistency and differences between centrality measures across distinct classes of networks. PloS one, 14(7), e0220061.
6. Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2020). Interplay Between Hierarchy and Centrality in Complex Networks. IEEE Access, 8, 129717-129742.
7. Cherifi, H., Palla, G., Szymanski, B. K & Lu, X. (2019). On community structure in complex networks: challenges and opportunities. Applied Network Science,4,1,1-35.
8. Gupta, N.,Singh, A., & Cherifi, H. (2015). Community-based immunization strategies for epidemic control. 7th int. conf. on communication systems and networks,Proc. IEEE, 1-6.
9. Ghalmane, Z., Cherifi, C., Cherifi, H. & El Hassouni, M. (2019). Centrality in complex networks with overlapping community structure. Scientific reports,9,1,1-29.

Predicting Primary-Specialty Referrals in Health Services using Graph Embeddings

Regina Duarte[1], Qiwei Han[2], and Claudia Soares[1]

[1] Instituto Superior Técnico, University of Lisbon, Portugal
reginaduarte@tecnico.ulisboa.pt, csoares@isr.tecnico.ulisboa.pt,
[2] Nova School of Business and Economics, Universidade NOVA de Lisboa, Campus de Carcavelos, 2775-405 Carcavelos, Portugal
qiwei.han@novasbe.pt

1 Introduction

The medical referral between primary care physicians (PCP) and specialists (SP) represents the formal mechanism in the health system to address the need of patients for specialty care. It may affect many aspects of patient care, such as quality of care, patient satisfaction, health care costs, etc. [2]. The existing literature typically leveraged the patient consultation history extracted from insurance claims data to construct the patient sharing network between physicians based on the shared patients [2]. The patient sharing network essentially operationalizes an informal information-sharing network in which physicians provide care to shared patients. However, this network does not necessarily conform to the formal organizational structure that physicians are affiliated with, and thus may provide valuable insights in explaining the referral mechanism.

In this paper, we hypothesize that medical doctors' informal social networks can influence the referral process. In other words, doctors' referral decisions may be limited to their social contacts and do not always benefit patient-centered care. This implies that network structure metrics derived from the doctors' social network can serve as informative features to boost the predictive performance of a model for referral recommendations [1]. As such, we create two networks: 1) the referral network connecting PCP to SP if a patient consults a PCP and then an SP within a month, and 2) the social network of all doctors according to their similar profiles. Then, we learn the representation of the referral network using a Graph Neural Network (GNN) [4]. The main objective of the study is to uncover hidden mechanisms in the primary-specialty referrals using features extracted from informal social network of doctors, which may help health organizations to improve the referral process through recommendations [5].

2 Data and Network Constructions

We analyze a large-scale patient consultation data from a European private healthcare provider with over 12 million consultations between 1.4 million patients and 3,632 physicians (389 PCP and 1,313 SP and 1,390 with unknown specialty) in 7 hospitals from 2012 to 2017, to understand the primary-specialty referral mechanism. Besides, we obtain additional physician registration data from the Human Resources department of the provider, including demographics, such as gender, age, education, and professional information, such as internship institution, specialty, working hospital. As such, we develop a weighted bipartite network where 294 PCP are connected with 839 SP through 34,249 edges. The edge weight on the referral network represents the number of patients that PCP refers to the SP. Importantly, many physicians do not link to the referral network, which raises potential inefficiency concerns for their lack of involvement in the referral process. We also create a complementary social network of the doctors (for both PCP and SP), where an edge connects two doctors if they shared similar profiles such as they receive the medical degree from the same institution, they perform the residency internship in the same hospital.

We calculated summary network metrics for the resulting referral and social networks, respectively. For example, the degree of nodes in the referral network follows the power-law distribution. This means that: (1) there are few physicians that receive a huge amount of referrals (if the node is an SP) or that refer to a lot of specialty doctors (if the node is a PCP) (2) the vast majority of the SP only receive referrals from few PCP or PCP only refer to a few SP. Such observations are subject to several possible explanations. The SP with high in-degree can be those with high popularity, while PCP will low out-degree may have relatively limited social contacts. Meanwhile, we obtained the average clustering coefficient for the referral network (0.149) to measure the fraction of the number of observed squares to the total number of possible squares in the network. This represents an essential precondition for the referral network to exhibit small-world structure and suggests that physicians in the referral network have a higher tendency to cluster together.

3 Referral Prediction using Graph Embeddings

Node embeddings learned from graph-structured data provide low-dimensional vector representations for each node using its graph neighborhood [3]. It has showed to be very useful for numerous machine learning applications, such as node classification, clustering, and link prediction. We adopt the GraphSAGE model to generate graph embeddings for the referral network, because the model can leverage node attributes to jointly learn the structure of each node's neighborhood together with the distribution of node features in the neighborhood [4].

From the social network of doctors, we computed three centrality measures for each node: betweenness, node, and degree centrality, and added them as features to GraphSAGE to accomplish the link prediction task, based on the referral network. Firstly, we learned an unsupervised graph representation of the referral network. The GNN was trained with 1,134 nodes and 27,743 edges, and tested with the same number of nodes and 30,825 edges. Features used were physician gender and age, plus the centrality features for the social networkaware experiment. We trained for 20 epochs, layer sizes of 20 by 20 and a dropout rate of 0.3. A neural network model was trained for link prediction optimizing with Adam (learning rate of 1e-3) over binary cross-entropy loss. In comparison, we also evaluate the link prediction task on the referral network in the *absence* of social network information.

Table 1: Link prediction task metrics, when social network features are added to referral information.

Social Network	**Test loss**	**Test accuracy**
Without information	1.0629	0.5232
With information	0.6939	0.7072

The results, presented in Table 1, show that the information on social relationships of doctors does improve the predictive power of the model. Accuracy increases about 0.18 with the added information about the social network of doctors. As expected, the loss suffered decreases as well.

4 Future work

As future work and to test the consistency of these results, a few possibilities line up: 1) train a link prediction model with more features that characterize the doctors from both their demographic attributes and social network measures, 2) compare a set of the state-of-the-art GNN models and verify whether the our findings still hold and 3) introduce the explanability model to identify the factors that affect referral process.

Acknowledgements. This work was funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2019, UIDB/00124/2020 and Social Sciences DataLab, PINFRA/22209/2016), POR Lisboa and POR Norte (Social Sciences DataLab, PINFRA/22209/2016).

References

1. An, C., O'Malley, A., Rockmore, D.: Referral paths in the U.S. physician network. Applied Network Science 3(1) (2018)
2. Barnett, M.L., Christakis, N.A., O'Malley, A., Onnela, J.P., Keating, N.L., Landon, B.E.: Physician Patient-Sharing Networks and the Cost and Intensity of Care in US Hospitals. Medical Care 50(2), 152–160 (2012)
3. Grover, A., Leskovec, J.: Node2vec: Scalable Feature Learning for Networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 855–864. KDD '16 (2016)
4. Hamilton, W.L., Ying, R., Leskovec, J.: Inductive representation learning on large graphs. In: NIPS (2017)
5. Han, Q., Ji, M., Martinez de Rituerto de Troya, I., Gaur, M., Zejnilovic, L.: A hybrid recommender system for patient-doctor matchmaking in primary care. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). pp. 481–490 (2018)

Part VII

Machine Learning and Networks

Are there Racial Disparities in Fatal Police Shootings? Exploration with Uniform Manifold Approximation and Projection

Philip D. Waggoner[1]

University of Chicago, Chicago, IL 60637, USA
pdwaggoner@uchicago.edu,
https://pdwaggoner.github.io/

1 Introduction

For many years, though more so in recent months, America and the world have grappled with the phenomenon of and fallout from police shootings. The controversy tends to be strongest in contexts where the suspect is non-white, usually African American, and is also often unarmed.

This phenomenon has captured the attention of many scholarly communities as well, who are interested in understanding and diagnosing this issue from a variety of angles including politically [1], criminologically [2], behaviorally [3], and others [4].

Building these and other recent studies on race and police shootings, I explore this issue from a different angle. I develop a nonlinear, exploratory framework to explore race and police shootings to allow the data to speak as freely as possible. To do so, I use two distinct unsupervised dimension reduction approaches, manifold learning and neural network-based projection, to explore whether non-random, latent patterns emerge along a racial dimension in the context of police shootings in America. The main method in this analysis is uniform manifold approximation and projection (UMAP), with self-organizing maps (SOM) used as validation, given that SOM have been around for a longer period of time. Using recent data on fatal police shootings the United States from 2015 to 2020 [5], I find that there does indeed appear to be separation in the projection space of fatal police shootings among suspects of different races.

The broader take away is that recent advances in complex, nonlinear embedding and projection allow for the picking up of non-random, latent structure in highly consequential sociological data. The result is a deeper understanding of this and related phenomena via largely atheoretic means. These and related methods should encourage researchers who are interested in uncovering patterns and reducing complexity, over seeking confirmation for existing biases on both (and many) sides of such consequential, high-stakes issues like police shootings.

2 Results

In this section, I briefly present my main results from the UMAP procedure. As UMAP is an unsupervised technique making external validation elusive, I validate these patterns using self-organizing maps. Patterns across both UMAP and SOM are stable, suggesting racial patterns in fatal police shootings are likely more signal than noise.

UMAP is a recent nonlinear, unsupervised approach to approximating a high dimensional space and embedding it on a less complex, lower dimensional subspace [6]. UMAP is a computationally efficient algorithm that is able to capture global structure and project it locally. The algorithm assumes some small neighborhood of observations existing along a manifold can be more simply, yet still consistently characterized in a lower dimensional setting, allowing for greater interpretability from complex, high dimensional data spaces. Building on the concept of local neighborhoods reflecting structural similarity at different positions along the manifold, the algorithm searches for an optimal manifold to characterize the raw inputs. The result is a nonlinear, low dimensional manifold that represents true distances between observations, which can also aid in reconstruction of the original space, regardless of size and complexity.

Before jumping into the results, there were several missing observations in these data across several features including, e.g., gender, age, race, whether/how the suspect attempted to flee, and so on. See the patterns of missingness in Figure 1.

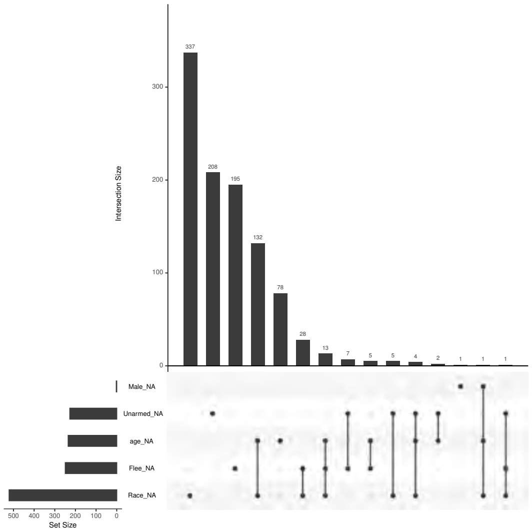


Fig. 1. Pattern of missingness across the full data space. In addition to raw counts of missing cases by feature, the figure also displays intersections of missing cases across all features.

As such, multiple imputation was conducted for all missing values, where the mode was used for categorical features and k-Nearest Neighbors (kNN)-based imputation on all features was used for continuous features. To ensure findings were not due to any choice made during imputation, results from two fits of UMAP on both the imputed data and then the data with missing values simply dropped, are presented side-by-side in Figure 2. These models were tuned with neighborhoods of size 15, minimum lower dimensional distance of 0.1, and 200 epochs each using Manhattan distance. Consistency in patterns in both figures suggest imputation did not negatively impact results.

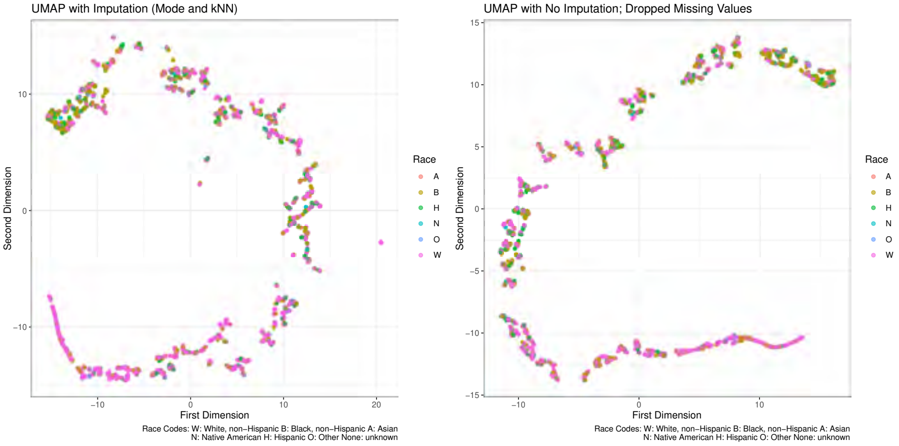


Fig. 2. UMAP results on imputed (left) and non-imputed (right) data.

The results in Figure 2 demonstrate that there is a clear difference in patterns of fatal police shootings by race. This is seen in the pink points (representing white suspects) being mostly grouped together in the projection space on the second dimension, compared to all other races being largely grouped together in a different location in the same space along the same dimension. These patterns suggest that police respond with deadly force differently to different races, with white suspects being treated similarly to each other but notably distinct from their non-white counterparts with wider variance.

Comparison to Self-Organizing Maps. Using the neural network-based technique of self-organizing maps, I validate UMAP results and find similar patterns of fatal police shooting varying by race using the same shootings data. Saying nothing of normative police behavior or motivations underlying these trends, the results across both stages suggest police tend to use deadly force *differently* based on suspects' race.

References

1. McGregor A. Politics, police accountability, and public health: civilian review in Newark, New Jersey. Journal of Urban Health. 2016 Apr 1;93(1):141-53.
2. Ridgeway G. Officer risk factors associated with police shootings: a matched case–control study. Statistics and Public Policy. 2016 Jan 1;3(1):1-6.
3. Patterson GT, Swan PG. Police shootings of unarmed African American males: A systematic review. Journal of human behavior in the social environment. 2016 May 18;26(3-4):267-78.
4. Fryer Jr RG. An empirical analysis of racial differences in police use of force. National Bureau of Economic Research; 2016 Jul 7.
5. Samoshin A. Kaggle Data Police shootings: Database of every fatal shooting in the United States by a police officer. `https://www.kaggle.com/mrmorj/data-police-shootings`. 2020.
6. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. 2018 Feb 9.

Learning on graphs with diffusion

Alexis Arnaudon[1,2], Robert L. Peach[1], and Mauricio Barahona[1]

[1] Department of Mathematics, Imperial College London, London, UK,
[2] Blue Brain Project, EPFL, Geneva, Switzerland,

Diffusion or random-walks are one of the most fundamental dynamical processes on graphs. They have been heavily exploited in many contexts and for many applications, ranging from centrality measures to graph clustering. The success of diffusion-based analyses on networks may be in part due to the fact that graphs do not have simple and mathematically well-defined topologies such as, for example, Riemannian manifolds. Diffusion dynamics thus become a simple and powerful tool to extract relevant pieces of information from graphs structures and relate them to more classical, or natural notions from geometry or topology.

In this work, we extend this diffusion-based approach for graph analysis by constructing measures which exactly correspond to their Riemannian, or more simply, Euclidean counterparts. Our approach is based on the simple observation that diffusion dynamics may create 'overshooting' events. To understand what constitutes an overshooting event, let us take the simplest case, whereby we have a delta initial condition on the real infinite line. Taking another position on the line, different to that of the initial condition, the solution to the heat equation at this new point will first increase in time, to eventually relax to zero as time approaches infinity, the stationary state.

Moving away from the infinite line to a compact space, such as an interval or graph, only a fraction of the positions will see an overshooting peak due to the presence of reflective boundaries (a finite size effect). In Figure 1(a) and (d) we plot the evaluation of the heat equation as a function of time for the interval [0, 1] and the Karate club respectively. The subplots display the evaluation of the heat equation as a function of time for a given position and reveals 'peaks' (green plots), which we refer to as an overshooting event.

The overshooting peaks are sensitive to all topological scales of a graph, and thus analysing their presence and properties, such as position, amplitude and time, can be used to construct different graph measures or algorithms. Herein, we describe two recently published graph theoretic methodologies that have leveraged the presence of overshooting events: (1) Multiscale centrality [1][3] and (2) graph semi-supervised classification [2] [4]. For both applications we have provided an open-source Python package and welcome external contributions.

Multiscale Centrality. The presence or absence of peaks can be used as a proxy for the location of the boundaries, and therefore the center of the space. Because finite graphs are compact spaces, but their centers and thus boundaries are not as well defined as for topological spaces, it is natural to use overshooting events to define the center of the

[3] https://github.com/barahona-research-group/MultiscaleCentrality

[4] https://github.com/barahona-research-group/GDR

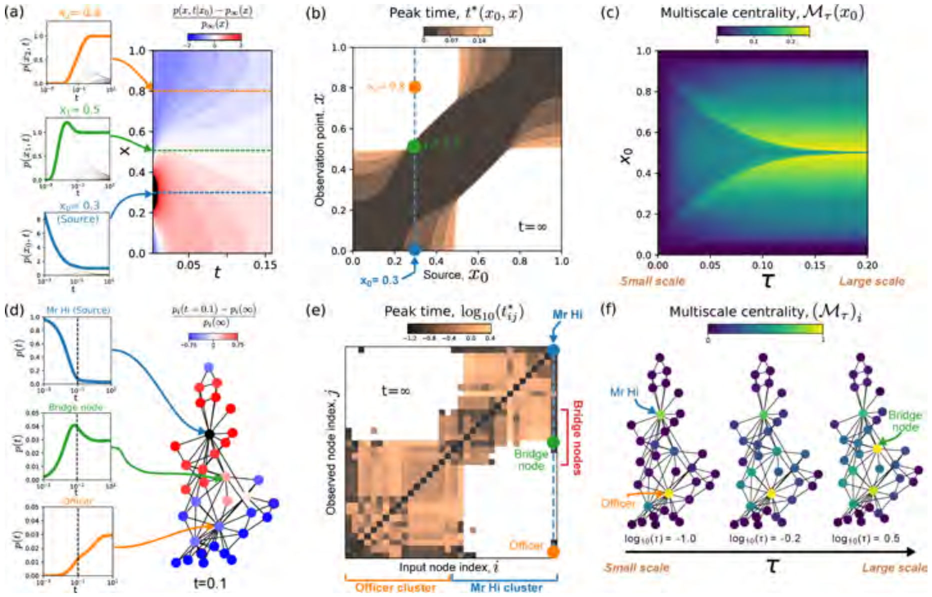


Fig. 1. Figure taken from [1], illustrating the concept of multiscale centrality. (a-c) Diffusion on a segment $[0, 1]$, with overshooting events in (a), times of these events in (b) with respect to the location of the source of diffusion and (c) corresponding multiscale centrality on the segment. (d-f), same set of panels, but on the Karate club graph, with graph diffusion instead of Euclidean heat equation.

boundaries of a graph. We call this measure *Multiscale centrality*, as it a graph centrality measure with an additional scale parameter τ, related to the maximum diffusion time used for the detection of overshooting events [1]. We show that this measure recovers the center of an interval, or a square, for infinite time horizon, and provide relevant information on the possible inhomogeneities in the space (discretisation for example) at small scales (Figure 1). We apply this centrality measure to several classic graphs, such as Karate club (Figure 1f), Dolphin social network, European power grid or directed graphs such as C. Elegans neural network to illustrate its use in various contexts. See Figure 2 for an example on the European Powergrid and the road network of Manhattan. We also show that it strongly correlates with local (degree centrality) and global (closeness centrality) at different scales.

Graph semi-supervised learning. Graph semi-supervised learning algorithms take advantage of an associated graph to improve supervised learning of node classes. Commonly, neighbouring nodes are more likely to be in the same class. Herein, we leverage the amplitude of the ‘overshooting peak’ to gain a probabilistic interpretation of node classes. We developed a methodology called Graph Diffusion Reclassification (GDR), which reclassifies nodes according to the maximum amplitude overshooting peak for any class [2]. From any prior node classifications (given by a standard classifier such as a random forest, for example), GDR uses the class assignment probabilities as initial

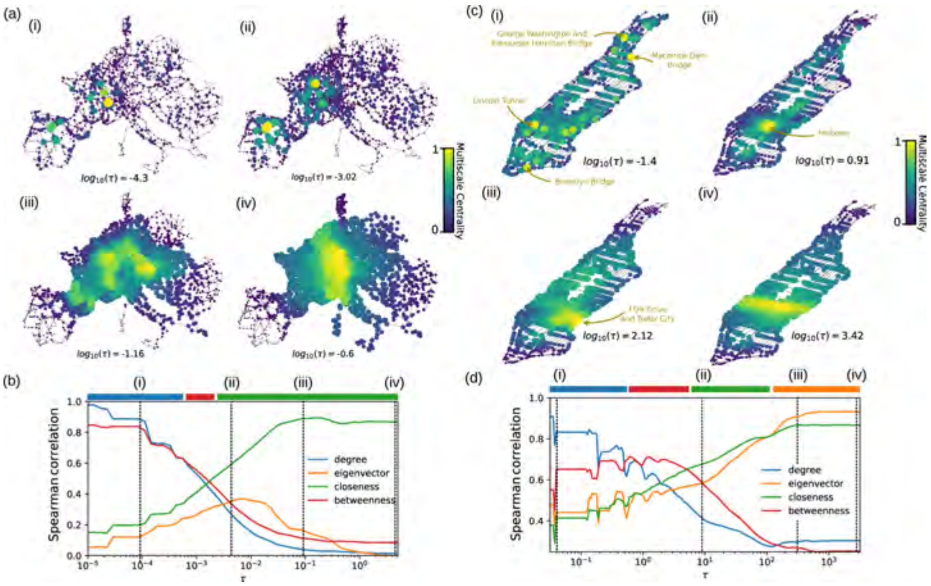


Fig. 2. Figure taken from [1]. Two examples of applications of multiscale centrality on real networks: in (a) the Europeen powergrid network and in (c) the road network of Manhatten. In (b) and (d) we show the Pearson correlation with the four most standard centrality measures, to place multiscale centrality in context: correlation with degree at small scale and correlations with closeness (or other more global centrality measures) at large scales.

conditions for a diffusion process and reclassify nodes according to the largest 'overshooting peak' between the classes. We show that this method almost always improves the classification accuracy of the original classifier on standard datasets such as citations networks or Wikipedia webpages. We have shown in two published examples how 'overshooting' events from diffusion dynamics can be leveraged to reveal interesting topological properties of networks. However, this line of research has provided further insights and methodological development that are not described in detail here. For example, we have used the overshooting events to define a measure of influence within a non-linear dynamical system, such as identifying the most infectious nodes in the SIR epidemic model. As another example, using *both* the time and the amplitude of the overshooting peaks we are able to define a notion of relative and local dimensionality on graphs. These exciting applications make this line of research both fruitful and of interest to the wider community.

References

1. A. Arnaudon, R. L. Peach, and M. Barahona, "Scale-dependent measure of network centrality from diffusion dynamics," *Physical Review Research*, vol. 2, no. 3, p. 033104, 2020.
2. R. L. Peach, A. Arnaudon, and M. Barahona, "Semi-supervised classification on graphs using explicit diffusion dynamics," *Foundations of Data Science*, vol. 2, no. 1, p. 19, 2020.

Comparing the efficacy of embeddings in Hyperbolic and Euclidean geometries with respect to the task of community detection.

Jade Chattergoon*, Aneeqah Hosein*, Daniel Pino*, and Inzamam Rahaman

jade.chattergoon@my.uwi.edu, aneeqah.hosein@my.uwi.edu,
daniel.pino@my.uwi.edu, inzamam@lab.tt
University of the West Indies, STA, TT

1 Abstract

Network representation learning has emerged as a efficacious tool in graph mining. Recent work has suggested that representations embedded in hyperbolic geometry often surpass representations learned in Euclidean geometry on downstream tasks by notable margins. Hitherto, most work has examined tasks such as link prediction, network alignment, and node classification. While hyperbolic embedding for community detection has been investigated to an extent in [5], its performance has yet to be compared to that of its Euclidean counterpart. In this extended abstract, we seek to address this gap.

2 Introduction

Community detection plays a major role in contemporary graph mining. Community detection seeks to detect groups of nodes that are strongly connected among themselves but sparsely connected to nodes in other groups. Through community detection, we can draw interesting and useful conclusions about networks and their underlying dynamics that can allow us to build better experiences for users.

Community detection is often performed by an algorithm that exploits graph traversals or some matrix decomposition. However, representation learning can also be used. In representation learning, a neural network is used to learn mappings from nodes in a graph to some lower dimensional space that is amenable to more conventional machine learning and data mining algorithms. Learnt representations can be fed into conventional clustering algorithms to coax out communities from our graphical data.

Most network representation techniques project onto Euclidean space. However, recent work has suggested that Hyperbolic geometries, being continuous analogues for trees, might be more appropriate embedding targets for graphs. Hyperbolic geometries are non-Euclidean geometries that possess negative curvature and do not respect Euclid's parallel postulate. Work such as Gunel et al. [2] suggests that these properties allow hyperbolic geometries to geometrically encode the latent hierarchies present in graphs. When tasked with reconstructing the original graph data and predicting links

[*] equal contribution

between nodes, hyperbolic embeddings have been shown to outperform Euclidean embeddings. This is also exemplified in terms of overfitting and complexity issues when using Euclidean embeddings with data which exhibit an intrinsic hierarchical structure [6].

This extended abstract seeks to report findings on the suitability of Hyperbolic embeddings vis-a-vis Euclidean embeddings for the task of community detection.

3 Methodology

For our experiments, we used several datasets from the SNAP repository [1]. We considered the Facebook dataset, comprising 4,385 nodes and 37,304 edges; the Twitch dataset comprising 4,039 nodes and 88,234 edges; and the TV-Shows dataset comprising 3,892 nodes and 17,262 edges. Although some of these datasets are weighted, their weights were ignored since current Hyperbolic embedding methods do not consider weights.

We considered three methods: Grover et al.'s [1] Euclidean Node2Vec, Nickel and Kiela's [7] Poincaré ball hyperbolic model embedding method, and Nickel and Kiela's [7] Lorentz embedding model.

Using the code provided by all the authors of the aforementioned three methods, we generated embeddings of various dimensions. We then leveraged **sklearn**'s implementations of agglomerative clustering function (using complete linkage) and DBSCAN on the trained embeddings.

Finally, we used Python's **community** library to calculate the modularity of the clusters produced by the clustering methods using Euclidean and both hyperbolic neural network embeddings for comparison.

4 Experimental Results

Table 1. Modularities obtained testing on the facebook dataset.

	Poincaré			**Lorentz**			**Euclidean**		
Dimensionality	5	32	64	5	32	64	5	32	64
Agglomerative	0.1820	0.1947	0.1742	0.1900	0.1608	0.1659	0.8004	0.8143	0.8048
DBSCAN	0.0371	0.0406	0.0388	0.0374	0.0401	0.0393	0.6504	0.6298	0.5778

[1] `http://snap.stanford.edu/data/index.html`

Table 2. Modularities obtained testing on the twitch dataset.

	Poincaré			**Lorentz**			**Euclidean**		
Dimensionality	5	32	64	5	32	64	5	32	64
Agglomerative	0.0008	0.0036	0.0007	0.0003	0.0004	0.0004	0.1971	0.2329	0.2059
DBSCAN	-0.0018	-0.0016	-0.0012	-0.0014	-0.0016	-0.0017	0.0158	0.0479	0.0411

Table 3. Modularities obtained testing on the TV shows dataset.

	Poincaré			**Lorentz**			**Euclidean**		
Dimensionality	5	32	64	5	32	64	5	32	64
Agglomerative	0.0029	0.0009	0.0023	0.0058	0.0010	0.0020	0.7622	0.8466	0.8340
DBSCAN	0.0004	0.0005	0.0007	0.0003	0.0005	0.0004	0.5422	0.6274	0.6122

5 Conclusion and Discussion

As seen in our experimental results, the Euclidean embeddings outperformed the Hyperbolic embeddings on community detection regardless of clustering method, embedding method, or embedding dimension. Moreover, the difference in modularity scores are rather marked. We conjecture that this might indicate that if Hyperbolic embeddings are to be useful for community detection, at the very least, they require different clustering techniques than conventional ones. We plan to investigate this in future work.

It should also be noted that some datasets were more amenable to the Hyperbolic methods over others. This might be the result of some property of these datasets. In future work, we can utilise the nPSO model [4] and LFR benchmark [3] to generate networks with controlled communities as opposed to real datasets. Complementing this, Normalised Mutual Information may be a more suitable performance metric.

References

1. Grover, A. and Leskovec, J. node2vec: Scalable feature learning for net-works. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016).
2. Gunel, B., Sala, F., Gu, A. and Ré, C. HyperE: Hyperbolic Embeddings for Entities (2018).
3. Lancichinetti, L., Fortunato, S. and Radicchi, F. Benchmark graphs for testing community detection algorithms. Phys. Rev. E 78, 046110 (2008).
4. Muscoloni, A. and Cannistraci C.V. A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities. IOP Publishing Ltd on behalf of Deutsche Physikalische Gesellschaft (2018).
5. Muscoloni, A., Thomas, J.M., Ciucci, S. et al. Machine learning meets complex networks via coalescent embedding in the hyperbolic space. Nat Commun 8, 1615 (2017).
6. Nickel, M. and Kiela, D. Poincaré Embeddings for Learning Hierarchical Representations. 6338–6347 (2017).
7. Nickel, M. and Kiela, D. Poincaré embeddings for learning hier-archical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Infor-mation Processing Systems 30, pages 6338–6347. Curran Associates, Inc. (2017).

Data-Driven Analysis of Complex Networks and Their Model-Generated Counterparts

Marcell Nagy[1] and Roland Molontay[12]

[1] Dept. of Stochastics, Budapest University of Technology and Economics, Hungary
[2] MTA-BME Stochastics Research Group
marcessz@math.bme.hu, molontay@math.bme.hu

1 Introduction

Data-driven analysis of complex networks has been in the focus of research for decades [1, 2]. An important research question is to study how well real networks can be described with a small selection of metrics, furthermore how well network models can capture the relations of graph metrics observed in real networks. In this paper, we apply machine learning techniques to investigate the aforementioned problems [3]. We study 482 real-world networks from different domains such as brain, food, social, protein interaction, web and infrastructural networks, along with 6×482 synthetic networks generated by six frequently used network models with previously calibrated parameters to make the generated graphs as similar to the real networks as possible.

The six models that we consider are: the clustering version of the Barabási–Albert model [4], the duplication-divergence model [5], the 2K-Simple model [6], which captures the joint-degree distribution of a graph, the forest-fire model [7], Kronecker graph model [8], and the (degree-corrected microcanonical) stochastic block model [9] that creates graphs with community structure.

Our approach unifies several branches of data-driven complex network analysis, such as the study of graph metrics and their pair-wise relationships, network similarity estimation, model calibration, and graph classification. We find that the correlation profiles of the structural measures significantly differ across network domains and the domain can be efficiently determined using a small selection of graph metrics. The goodness-of-fit of the network models and the best performing models themselves highly depend on the domains. By solving classification problems, we find that the models lack the capability of generating graphs with high clustering coefficient and large diameter simultaneously.

2 Metric selection

First, we calculate the following rich set of metrics for all the networks: assortativity, average clustering coefficient, average degree, average path length, density, global clustering coefficient, four interval degree probabilities [10], largest eigenvector centrality, maximum degree, maximum edge and vertex betweenness centralities, number of edges and nodes and pseudo diameter. We select a subset of the metrics which still well describes the networks, but has smaller redundancy. Based on the average absolute correlation network (Fig. 1), we select metrics from each connected component of the correlation graph. We note that the metrics of the component containing the size of the network are excluded from the analysis since the high correlation with the size implies significant trivial predictive power with respect to network domains due to the size difference of typical networks from different domains. Moreover, we aim to find distinctive size-independent topological properties. The selected attributes are listed in Table 1.

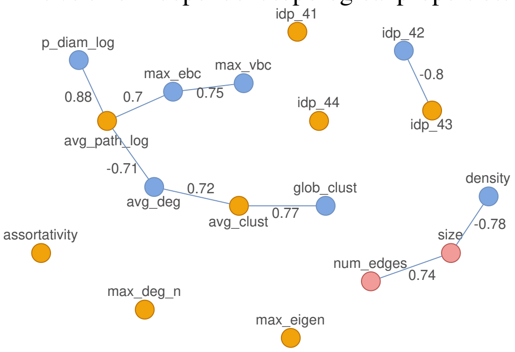


Fig. 1: The correlation network of structural metrics. Two nodes are connected if the domain-averaged absolute Spearman's rank correlation of the corresponding metrics is above 0.65. The actual values of the correlations are written on the edges.

Name	Description
assortativity	The assortativity coefficient
avg_clust	The average local clustering coefficient
avg_path_log	The average path length divided by the logarithm of the size.
idp's (1,3,4)	The interval degree probabilities
max_deg_n	The scaled maximum degree
max_eigen	Maximum eigenvector centrality

Table 1: The selected structural metrics and the nominal variables regarding the graphs' origin.

3 Domain classification

In order to evaluate our metric selection and to gain a better understanding of the different network domains, we predict the domains of the networks, first using all of the metrics, then using only the selected attributes. Our selection is considered ”optimal” if the domains are well separated by these metrics. Fig. 2 suggests that the domains are properly separated, which is also confirmed by the results of the machine learning algorithms (see Table 2).

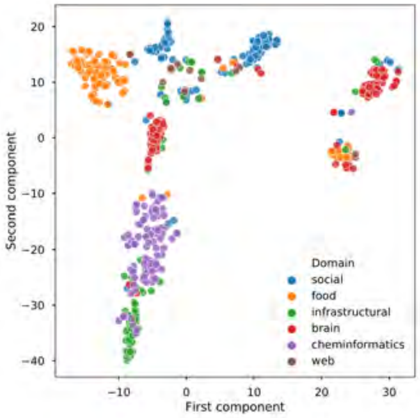


Fig. 2: The t-distributed stochastic neighbor embedding of the 8 selected graph metrics (detailed in Table 1) of the real networks.

Classifier	All attributes	Selected attributes
kNN	78.5%	87.0%
Decision Tree	85.2%	85.5%
Random Forest	89.3%	91.2%

Table 2: The average accurcacy of the models with a 5-fold cross validation.

4 Model calibration and classification

With a given target network G_T (real network), a network model $M(\theta)$ with parameter vector θ and a similarity s (or distance d) function defined on graphs pairs, the goal of model calibration is to find a parameter vector θ^* of the model, such that the generated graphs $G_{M(\theta)}$ (realizations of $M(\theta)$) are as similar (or as close) to the target network as possible. Formally it can be expressed as:

$$\theta^* = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} s\left(G_{M(\theta)}^{(i)}, G_T\right) = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} d\left(G_{M(\theta)}^{(i)}, G_T\right), \qquad (1)$$

where n is the number of the independent identically generated graphs $G_{M(\theta)}^{(1)}, \dots, G_{M(\theta)}^{(n)}$. In this work, the d distance function is defined as the Canberra distance of the previously selected graph metrics. The distance minimization here is carried out with grid search.

After calibrating the network models, we trained machine learning algorithms to distinguish between real graphs and their model generated counterparts. The results suggest that the food webs are easy to model since the classifiers were not able to distinguish the modeled food networks from the original ones. The most distinguishing attribute, i.e. which significantly differs in real and model generated networks is the average path length, followed by assortativity and the average clustering coefficient. Therefore the applied models are unable to capture the exact diameter and clustering of real networks. For example, protein interaction networks are clustered and have relatively large average distances, none of the models were able to capture these properties at the same time [3].

References

1. V. Filkov, Z. M. Saul, S. Roy, R. M. D'Souza, and P. T. Devanbu, "Modeling and verifying a broad array of network properties," *EPL*, vol. 86, no. 2, p. 28003, 2009.
2. N. Attar and S. Aliakbary, "Classification of complex networks based on similarity of topological network features," *Chaos*, vol. 27, no. 9, p. 091102, 2017.
3. M. Nagy and R. Molontay, "Data-driven analysis of complex networks and their model-generated counterparts," *arXiv preprint arXiv:1810.08498*, 2018.
4. P. Holme and B. J. Kim, "Growing scale-free networks with tunable clustering," *Phys. Rev. E*, vol. 65, no. 2, p. 026107, 2002.
5. I. Ispolatov, P. L. Krapivsky, and A. Yuryev, "Duplication-divergence model of protein interaction network," *Phys. Rev. E*, vol. 71, no. 6, p. 061911, 2005.
6. M. Gjoka, B. Tillman, and A. Markopoulou, "Construction of simple graphs with a target joint degree matrix and beyond," in *IEEE INFOCOM*, pp. 1553–1561, Citeseer, 2015.
7. J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *11th SIGKDD Conf.*, pp. 177–187, ACM, 2005.
8. J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker graphs: An approach to modeling networks," *J. Mach. Learn. Res.*, vol. 11, no. Feb, pp. 985–1042, 2010.
9. T. P. Peixoto, "Nonparametric Bayesian inference of the microcanonical stochastic block model," *Phys. Rev. E*, vol. 95, no. 1, p. 012317, 2017.
10. S. Aliakbary, J. Habibi, and A. Movaghar, "Quantification and comparison of degree distributions in complex networks," in *7th Int. Symp. on Telecomm.*, pp. 464–469, IEEE, 2014.

Graph Neural Network Models for Node Classification in Multilayer Networks

Lorenzo Zangari[1], Roberto Interdonato[2], and Andrea Tagarelli[1]

[1] DIMES, University of Calabria, Italy, andrea.tagarelli@unical.it
[2] Cirad, UMR TETIS, Montpellier, France, roberto.interdonato@cirad.fr

1 Introduction

Graph Neural Networks (GNNs) are powerful tools that are nowadays reaching state of the art performances in a plethora of different tasks such as node classification, link prediction and graph classification [9]. One of the main challenges addressed by these methods is to redefine basic deep learning operations, such as convolution, on structures like graph networks, where nodes may have unordered neighborhoods of varying size. A basic solution is proposed in Graph Convolutional Networks (GCNs) [3], that perform convolution on graphs by aggregating the values of each node's features along with its neighbors' features. Graph Auto-Encoders (GAE) [2] perform a node embedding step by exploiting multiple GCN layers as encoders, while the aim of the decoding step is to obtain an adjacency matrix as similar as possible to the original one. Graph Attention Networks (GATs) [7] use a masked self-attention mechanism in order to learn weights between each couple of connected nodes. Self-attention can be seen as a mechanism able to discover the most important/representative parts of the input, and it is a de-facto standard in the context of sequence-to-sequence problems like machine translation and machine reading. It should be noted that the above methods work on simple networks only, i.e., networks modeling a single type of relation among an homogeneous set of nodes. The aim of this work is to extend some of these approaches to the Multilayer network model [4], in order to enable the analysis of networks including an arbitrary number of layers and the existence of intralayer and interlayer relations among the nodes. More specifically, we here propose formulations that extend the GCN and the GAT approaches to the multilayer case. Both approaches are tested on four real world multilayer networks, in the context of a node classification task.

2 Extending Graph Neural Networks to the Multilayer case

Given a set $\mathscr{V}$ of N *entities* (e.g., users) and a set $\mathscr{L} = \{L_1, \cdots, L_\ell\}$ of *layers* (e.g., user relational contexts), with $\ell \geq 2$, we denote a multilayer network with $G_{\mathscr{L}} = (V_{\mathscr{L}}, E_{\mathscr{L}}, \mathscr{V}, \mathscr{L})$, where $V_{\mathscr{L}} \subseteq \mathscr{V} \times \mathscr{L}$ is the set of entity-layer pairings or *nodes* (to denote, e.g., each user is present in which layers), and $E_{\mathscr{L}} \subseteq V_{\mathscr{L}} \times V_{\mathscr{L}}$ is the set of undirected edges between nodes within and across layers.

We represent a multilayer network by a set of adjacency matrices $\mathscr{A} = \{\mathbf{A}_1, \cdots, \mathbf{A}_\ell\}$, with $\mathbf{A}_l \in \mathbb{R}^{n_l \times n_l}$ ($l = 1..\ell$), where $n_l = |V_l|$. Entities may be associated with *features* stored in layer-specific matrices $\mathscr{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_\ell\}$, with $\mathbf{X}_l \in \mathbb{R}^{n_l \times f_l}$ and f_l the number of node features in the l-th layer. In case no side-information is available for $G_{\mathscr{L}}$, each layer-specific feature matrix is assumed to be the identity matrix $\mathbf{I}_l \in \mathbb{R}^{n_l \times n_l}$.

Note that, since we need to account also for inter-layer edges, while performing convolution operations the aggregation over node features should be computed not only with the ones of each node's neighbors, but also with the features of the different nodes coupled to the same entity over the layers of the multilayer network. In this regard, to ease the notation in subsequent equations, we also define the sets $\mathit{\Gamma}$ and $\mathit{\Psi}$ that correspond, respectively, to the intralayer and interlayer neighborhoods of each node. Given a node v in a layer l, we define the set $\mathit{\Gamma}$ of neighbors of v in the same layer l as:

$$\mathit{\Gamma}(v,l) = \{(u,m) \in V_{\mathscr{L}} | ((u,m),(v,l)) \in E_{\mathscr{L}}, m = l\} \tag{1}$$

Similarly, we define the set $\mathit{\Psi}$ of neighbors of v in layers different than l as:

$$\mathit{\Psi}(v,l) = \{(u,m) \in V_{\mathscr{L}} | ((u,m),(v,l)) \in E_{\mathscr{L}}, m \neq l\} \tag{2}$$

At this point, we can exploit this model to extend the propagation rules for the original GCN and GAT frameworks, in order to make them suitable for the multilayer case. The original propagation rule for GCN can be expressed with the following equation:

$$z_i = \sigma\left(\sum_{j \in \mathit{\Gamma}(i) \cup \{i\}} \frac{1}{\sqrt{\tilde{D_{ii}} \tilde{D_{jj}}}} h_j W^T\right) \tag{3}$$

where W is the parameters matrix, D is the degree matrix, $\mathit{\Gamma}$ is the neighborhood of node i, $\sigma(\cdot)$ is the activation function (e.g., ReLU) and h_j is the feature vector for node j. According to the previously introduced network model, Eq. 3 can be easily extended in order to enable the analysis of Multilayer Networks. For a node i in a layer m, this results in the following rule:

$$z_{(i,m)} = \sigma\left(\sum_{(j,l) \in (\mathit{\Gamma}(i,m) \cup \mathit{\Psi}(i,m))} \frac{1}{\sqrt{\tilde{D_{ii}} \tilde{D_{jj}}}} h_{(j,l)} W^T\right) \tag{4}$$

where the aggregation is performed over the complete set of intralayer and interlayer neighbors of i, i.e., the union of sets $\mathit{\Gamma}$ and $\mathit{\Psi}$. Note also that the feature vectors are layer specific (i.e., $h(j,l)$), and may have different size in different layers. Concerning GAT, the original propagation rule is defined as:

$$z_i = \Big\|_{k=1 \dots K} \sigma\left(\sum_{j \in \mathit{\Gamma}(i)} \alpha_{i,j}^k W^k h_j\right) \tag{5}$$

where $\|$ is the concatenation operator, α^k are the attention coefficient for the $k-th$ attention mechanism and W^k is the corresponding linear transformation. Similarly to the solution proposed for GCN, we can extend Eq. 5 to the multilayer case as:

$$z_{(i,m)} = \Big\|_{k=1 \dots K} \sigma\left(\sum_{(j,l) \in (\mathit{\Gamma}(i,m) \cup \mathit{\Psi}(i,m))} \alpha_{(i,m),(j,l)}^k W^k h_{(j,l)}\right) \tag{6}$$

We will refer to the proposed multilayer methods as $ML-GCN$ and $ML-GAT$.

Table 1. Performances (average and standard deviation over 10 runs) of the proposed methods and baseline over four real world datasets. Best results are highlighted in bold.

	ML − GAT		*ML − GCN*		*GAT*		*GCN*	
Dataset	*F1*	*MRR*	*F1*	*MRR*	*F1*	*MRR*	*F1*	*MRR*
CKM	94.46 ± 0.85	97.03 ± 0.49	85.8 ± 0.51	91.85 ± 0.28	**99.46 ± 0.00**	**99.73 ± 0.00**	98.17 ± 0.14	99.03 ± 0.07
Vickers	**96.81 ± 1.52**	**98.4 ± 0.76**	93.63 ± 1.0	96.81 ± 0.5	91.36 ± 3.6	95.68 ± 1.8	91.81 ± 0.6	95.9 ± 0.3
Congress	94.19 ± 0.85	97.1 ± 0.43	**95.87 ± 0.09**	**97.93 ± 0.04**	61.47 ± 0.00	80.73 ± 0.00	61.5 ± 0.03	80.75 ± 0.01
Balance	83.67 ± 1.0	90.84 ± 0.84	**86.82 ± 0.6**	**92.62 ± 0.41**	66.26 ± 0.11	81.82 ± 0.05	46.05 ± 0.00	71.71 ± 0.00

3 Preliminary experimental results

We tested the proposed methods in the context of a node classification task. We take into account four real world multilayer networks from different domains and showing different structural characteristics: Vickers [8], CKM Physicians Innovation [1], Congressional Voting [5] and Balance Scale [6]. We use as baselines the original *GAT* and *GCN* frameworks over the flattened networks (i.e., monoplex networks obtained by aggregating the relations over different layers, thus discarding the multilayer information). Table 1 reports the *F1-score* (F1) and *Mean Reciprocal Rank* (MRR), averaged over 10 runs. It can be noted how the multilayer methods obtain the best performance in 3 over 4 datasets. The improvement is more significant on networks which are bigger in size and with a higher number of layers (i.e., *Congress* and *Balance*), that also correspond to best results are obtained by $ML-GCN$. The fact that a baseline outperforms the multilayer approaches on *CKM* can be due to the fact that this network is composed by four isolated connected components, corresponding to different node labels: in this specific case, the monoplex flattening may be beneficial for the node classification task.

References

1. Coleman, J., Katz, E., Menzel, H.: The diffusion of an innovation among physicians. Sociometry 20(4), 253–270 (1957)
2. Kipf, T.N., Welling, M.: Variational graph auto-encoders. CoRR abs/1611.07308 (2016)
3. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th Int. Conf. on Learning Representations, ICLR 2017 (2017)
4. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. J. Complex Networks 2(3), 203–271 (2014)
5. Schlimmer, J.C.: Concept acquisition through representational adjustment (1987)
6. Siegler, R.S.: Three aspects of cognitive development. Cognitive psychology 8(4), 481–520 (1976)
7. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: 6th Int. Conf. on Learning Representations, ICLR 2018 (2018)
8. Vickers, M., Chan, S.: Representing classroom social structure. Victoria Institute of Secondary Education, Melbourne (1981)
9. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? In: 7th Int. Conf. on Learning Representations, ICLR 2019 (2019)

Multi-scale Anomaly Detection on Attributed Networks

Leonardo Gutiérrez-Gómez[1,2], Alexandre Bovet[2,3,*], and Jean-Charles Delvenne[2]

[1] Luxembourg Institute of Science and Technology (LIST), Esch-sur-Alzette, Luxembourg
[2] ICTEAM, Université catholique de Louvain, Belgium
[3] Mathematical Institute, University of Oxford, UK
alexandre.bovet@maths.ox.ac.uk,
WWW home page: http://alexbovet.github.io

1 Introduction

Many social and economic systems can be represented as attributed networks encoding the relations between entities who are themselves described by different node attributes. Finding anomalies in these systems is crucial for detecting abuses such as credit card frauds, web spams or network intrusions [1]. Intuitively, anomalous nodes are defined as nodes whose attributes differ starkly from the attributes of a certain set of nodes of reference, called the *context* of the anomaly. While some methods have proposed to spot anomalies locally, globally or within a community context, the problem remain challenging due to the multi-scale composition of real networks [2] and the heterogeneity of node metadata.

Here, we propose a principled way to uncover outlier nodes simultaneously with the context with respect to which they are anomalous, at *all relevant scales* of the network [3]. We characterize anomalous nodes in terms of the concentration retained for each node after smoothing specific signals localized on the vertices of the graph. Besides, we introduce a graph signal processing formulation of the Markov stability framework used in community detection, in order to find the context of anomalies. The performance of our method is assessed on synthetic and real-world attributed networks and shows superior results concerning state of the art algorithms. Finally, we show the scalability of our approach in large networks employing Chebychev polynomial approximations.

2 Results

Considering a graph $G = (V, E)$ composed by a set V with $|V| = N$ nodes or vertices, and a set of edges E. For simplicity, we will restrict this work to undirected, connected, simple graphs. Considering multidimensional node attributes, we associate to each vertex $u \in V$ a d-dimensional vector $\boldsymbol{f}(u) = \langle \boldsymbol{f}_1(u), \ldots, \boldsymbol{f}_d(u)\rangle$ where $\boldsymbol{f}_k(u) \in \mathbb{R}$ represents the $k-$th attribute of the vertex u.

In order to discover anomalous nodes and the context with respect to which they are anomalous, we introduce weights to the edges of the graph, expressing the similarity of the nodes' attributes linked by the edge. Here we use the Gaussian weighting function:

$$w(u,v) = \begin{cases} \exp\left(-\frac{\|\boldsymbol{f}(u)-\boldsymbol{f}(v)\|^2}{2\sigma^2}\right) & \text{if } (u,v) \in E \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where σ is a scaling parameter. This creates a weighted adjacency matrix $\boldsymbol{W}$ over the network, or similarity matrix, where two nodes are similar if they are neighbours with close values of the attributes. This adjacency matrix allows us to define a heat equation for any graph signal $\boldsymbol{x} \in \mathbb{R}^N$, i.e. any function $V \rightarrow \mathbb{R}$ assigning a scalar to every node. The heat equation is expressed as

$$\dot{\boldsymbol{x}}(t) = (\boldsymbol{W} - \boldsymbol{D})\boldsymbol{x}(t) = -\boldsymbol{L}\boldsymbol{x}(t),$$

where $\boldsymbol{D}$ is the diagonal matrix of node strengths, defined by $D_{uu} = d_u = \sum_v w(u,v)$. The matrix $\boldsymbol{L}$ is called the Laplacian of the weighted graph and the equation is solved by $\boldsymbol{x}(t) = e^{-t\boldsymbol{L}}\boldsymbol{x}(0)$. In terms of signal processing [4], the heat kernel $e^{-t\boldsymbol{L}}$ is often seen as a smoothing filter acting on the initial signal $\boldsymbol{x}(0)$, parametrized by the time t. It replaces every entry of $\boldsymbol{x}(0)$ with a weighted average of other nodes' signals (with a greater emphasis on neighboring nodes).

The *concentration* of a node $u \in V$ at scale t is defined by the L_2-norm of the filtered Kronecker delta $\boldsymbol{\delta}_u$ signal with a heat kernel at time t: $c_u(t) = \|e^{-t\boldsymbol{L}}\boldsymbol{\delta}_u\|_2$. A high-concentration node indicates a node that is very dissimilar in its attributes with its neighbors. We use the concentration as way to quantify the degree of deviation of a given node with respect to its context at a given time scale and provides a scoring for ranking potential outliers. Let $c(t) = [c_1(t), \ldots, c_N(t)]$ be the overall graph concentration. A node u is considered as *anomalous* at a time scale t if the following thresholding condition holds: $c_u(t) \geq \bar{c}(t) + 2s(c(t))$, with $\bar{c}(t)$ as the average concentration and $s(c(t))$ its standard deviation across nodes.

The choice of t selects the context with respect to which a node is anomalous. In a large t limit, a delta signal $\boldsymbol{\delta}_u$ will be smoothed to a constant over the whole graph. A node that stands as an outlier for a large t is an outlier globally, with respect to the whole graph. Conversely, a small t only allows heat diffusion with immediate neighbors, and thus a small t outlier is anomalously dissimilar to a small context. We consider that a set of nodes S is a suitable context for all potentially anomalous nodes lying in it, if this set is relatively poorly connected to the rest of the network, in a manner akin to community detection. Consider $\boldsymbol{h}_S$ the characteristic signal of S, i.e. the node signal taking unit values in S and zero values outside. The initial total energy of S is $|\boldsymbol{h}_S|_1$, the number of nodes in S. After smoothing, the energy remaining in S in excess of the energy $\frac{1}{N}|\boldsymbol{h}_S|_1$ that will remain at $t = \infty$ is: $r(t;\boldsymbol{h}_S) = \boldsymbol{h}_S^T e^{-t\boldsymbol{L}}\boldsymbol{h}_S - \frac{1}{N}|\boldsymbol{h}_S|_1$, which we want to be as high as possible. As we want to be able to provide a context for potentially each node of the graph, we look for a partition of the nodes, encoded by its $N \times K$ characteristic matrix $\boldsymbol{H}$, where K is the number of sets and every column $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K$ is the characteristic vector of a set of nodes. An optimal partition into contexts is given by the matrix $\boldsymbol{H}$ maximizing: $r(t;\boldsymbol{H}) = \sum_i \left(\boldsymbol{h}_i^T e^{-t\boldsymbol{L}}\boldsymbol{h}_i - \frac{1}{N}|\boldsymbol{h}_i|_1\right)$. This is essentially a particular case of the Markov stability [5]. It has the expected behaviour to provide a fine partition of small sets of nodes for low t, and a few large sets for large t.

See Figure 1 for an example of application to a real world dataset of co-purchases on Amazon. More details as well as synthetic and real world benchmarks showing the high performance of our method can be found in Ref. [3].

(2)

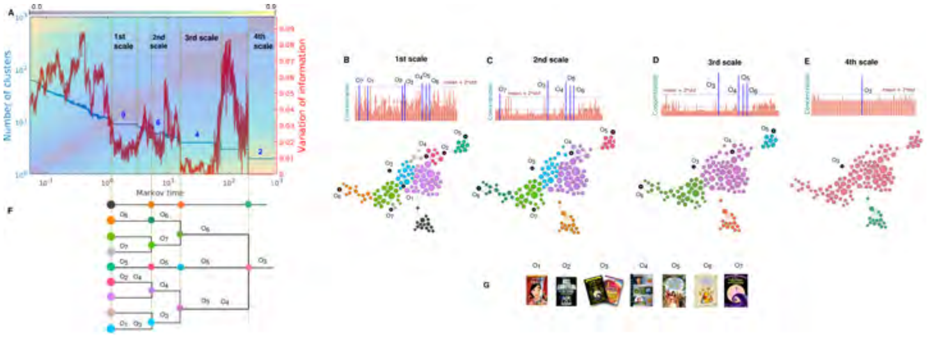


Fig. 1. **(A)** Number of clusters (blue curve) found by our algorithm, at each time step, the variation of information $VI(t)$ (red curve) between the ensemble of optimal partitions at each time and the variation of information $(VI(t,t'))$ between optimal partitions across times (background contour plot). Relevant partitions are determined by dips of $VI(t)$ and extended plateaus of $V(t,t')$. Visualization of four robust partitions and the node heat concentration (bar plots), indicating outlier nodes (blue bars) when evaluating the node concentration at **(B)** $t = 1.18$, **(C)** $t = 7.25$, **(D)** $t = 14.89$ and **(E)** $t = 204.8$. In all bar plots, the upper horizontal line (blue) indicates the detection threshold. **(F)** Dendrogram showing the hierarchy of clusters at each time with the contextual outliers. **(G)** Anomalous Disney DVD movies.

Summary.

- We propose a novel algorithm called MADAN (**M**ulti-scale **A**nomaly **D**etection in **A**ttributed **N**etworks) providing a principled mechanism to rank and localize outlier nodes within their *context* at *all scales* in a network [3].
- We conduct experiments on synthetic and real world benchmarks showing that our method allows to not only recover and rank the so called ground truth anomalies, but also to discover new anomalies jointly with their contexts.
- Finally, we show that our method is parallelizable and scales to large networks thanks to the Chebychev approximations of the exponential of the graph Laplacian. As side benefit, it also provides a faster methodology for the continuous-time Markov stability framework [5] for community detection.

References

1. Aggarwal, C. C. 2013. *Outlier Analysis*. Springer.
2. Boccaletti, S.; Latora, V.; Moreno, Y. *et al.* 2006. Complex networks: Structure and dynamics. *Phys. Rep.* 424(4-5):175–308.
3. Gutiérrez-Gomez, L. and Bovet, A. and Delvenne, J.-C. 2020. Multi-scale Anomaly Detection on Attributed Networks. *Proc. 34th AAAI Conf. Arti. Intel.*.
4. Shuman, D. I.; Narang, S. K.; Frossard, P.; Ortega, A.; and Vandergheynst, P. 2013. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine* 30(3):83–98.
5. Lambiotte, R.; Delvenne, J.; and Barahona, M. 2014. Random walks, markov processes and the multiscale modular organization of complex networks. *IEEE Trans. Net. Sci. Engi.* 1(2):76–90.

Calibrating Network Models for Real Networks Using graph2vec Embedding

Romain Lesauvage[1][2], Marcell Nagy[1], and Roland Molontay[1][3]

[1] Dept. of Stochastics, Budapest University of Technology and Economics, Hungary
[2] ENSAE Paris, France
[3] MTA-BME Stochastics Research Group, Hungary
romain.lesauvage@ensae.fr, {marcessz, molontay}@math.bme.hu

1 Introduction

The emergence of machine learning and data-driven solutions has opened up new research areas in network science. A related research question is how to fine-tune the parameters of network models to mimic real-world network as closely as possible. The traditional approach is to manually select a few graph metrics and calibrate the parameters of the network models in order to produce synthetic graphs that accurately match the properties of the original networks [1–3]. However, finding the most appropriate set of hand-crafted features that should be considered in the model-fitting is rather challenging [4].

In this work, we study an alternative framework where the model calibration is carried out in the embedding space learned by the graph2vec method. Graph2vec is an explicit graph embedding approach, which maps the graphs to real vectors such that similar graphs are mapped onto close points $\varphi : \mathbb{G} \to \mathbb{R}^{\delta}$ [5]. It is an unsupervised method that can capture the generic task-agnostic representations of arbitrarily sized graphs. Here we focus on four frequently used network models that we calibrate for 412 real-world networks from various domains (cheminformatics, social, food, brain, web). The four models that we consider are: the clustering version of the Barabási–Albert (CBA) model [6], the forest-fire (FF) model [7], the 2K model [8], which captures the joint-degree distribution of a graph, and the (degree-corrected microcanonical) stochastic block model (SBM) [9], which creates graphs with community structure. We study which models describe real-world networks the most accurately in each domain. Moreover, we compare the embedding based network calibration method to the feature-based method in terms of accuracy, time complexity, and interpretability. We can conclude that while embedding based network calibration is promising since it captures a lot of information, it can be cumbersome for large networks due to the running time, moreover it is more difficult to interpret than manually selected features.

2 Fitting the models

Our aim is to choose the parameters of the model in such a way that the ”distance” between the real network and the realizations of the network model is minimal. Formally, let G_R denote a real network, and let $M(\theta)$ be a network model with parameter vector θ, then the model calibration can be expressed as:

$$\theta^* = \arg\max_{\theta} \frac{1}{n}\sum_{i=1}^{n} s\left(G_{M(\theta)}^{(i)}, G_R\right) = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} d\left(G_{M(\theta)}^{(i)}, G_R\right), \tag{1}$$

where n is the number of the independent identically generated graphs $G^{(1)}_{M(\theta)}, \ldots, G^{(n)}_{M(\theta)}$. In this work, the d distance function is defined as the Euclidean distance between the PCA-projections of the graph2vec vector representations of the networks. The PCA is used to avoid the curse of dimensionality since the dimension δ is usually set to a high number ($\delta > 100$). The distance minimization is carried out with grid search. Since both the models themselves and the embedding procedures are incorporated with randomness, we propose to calibrate the parameter of the models using a weighted kNN regression on the parameters of the k closest model-generated graphs (we set $k = 5$). This procedure is reasonable if the distance between the model generated graphs in the embedded space varies ”continuously” with the parameters, that is what we check in Figure 1.

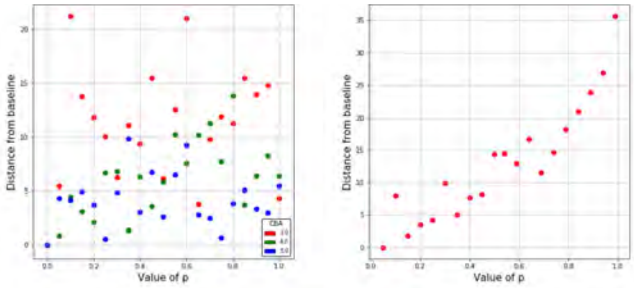


Fig. 1: The left figure shows the ”continuity” of the CBA model, and the right figure shows that of the FF model. The baseline graphs are generated with $p = 0$ parameter setting.

3 Finding the best models

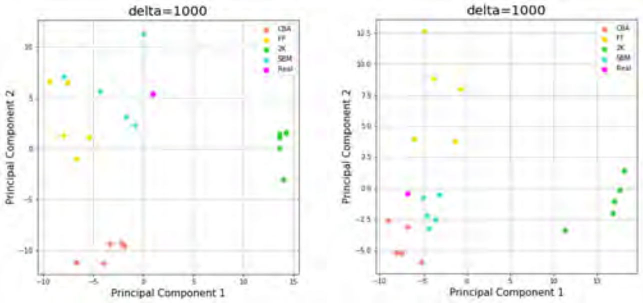


Fig. 2: The figures show the PCA projections of the embedding of the real and the calibrated models. The left figure shows the results for a brain network and the right shows the results for a social network.

After the parameters of the models are calibrated to mimic the given real-world network as closely as possible (i.e. being close in the embedded space), we study which network models can describe the networks the most accurately. We generate five instances with each model using the calibrated parameters and consider the graph2vec vector representations of the 20 generated graphs together with the real network. A couple of examples are depicted in Figure 2 using a 2-dimensional PCA projection of the $\delta = 1000$ dimensional embedded space. (We also experimented with other values of δ from 10 to 1500. We ran the entire procedure in each case and we compared the results. Ee aimed for well separated, robust and continuous results in the embedded space. Small δ values provided suboptimal results and from $\delta = 1000$, the results did not improve much.)

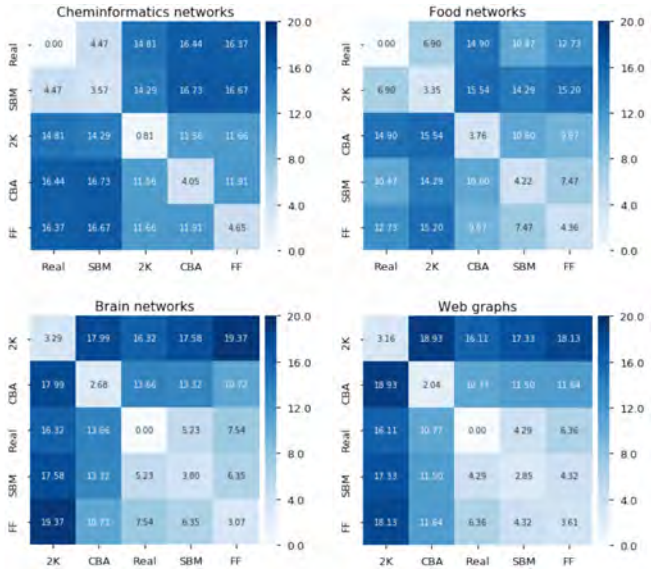


Fig. 3: The heatmaps of the mean distances between and within the model-generated and real networks for the four domains. The darker shades of blue indicate larger distances.

Figure 3 shows the heatmaps of the mean distances within and between the embedded coordinates of different models for each domain separately. We can observe that the best model in the cheminformatics, web, and brain domain is the SBM, while in the food domain FF model seems to be the best one. Table 1 shows that computation time must be taken into account because it can be indeed cumbersome for large networks.

Models	Brain	Food	Cheminformatics	Web
Average size	*946*	*118*	*55*	*4488*
CBA	23 min.	1 min. 20 sec.	20 sec.	7h 10 min.
FF	5 min.	30 sec.	10 sec.	1h 5 min.
2K	1 min 15 sec.	12 sec.	7 sec.	5 min. 10 sec.
SBM	2 min.	25 sec.	11 sec.	1h

Table 1: Average running time of computation for one network. The CBA, FF and 2K models have been computed on a computer with an i5-7300HQ processor and 8GB of RAM. SBM models have been computed with Google Colab.

References

1. I. Barnett, N. Malik, M. L. Kuijjer, P. J. Mucha, and J.-P. Onnela, "Feature-based classification of networks," *arXiv preprint arXiv:1610.05868*, 2016.
2. K. Ikehara and A. Clauset, "Characterizing the structural diversity of complex networks across domains," *arXiv preprint arXiv:1710.11304*, 2017.
3. N. Attar and S. Aliakbary, "Classification of complex networks based on similarity of topological network features," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 9, p. 091102, 2017.
4. M. Nagy and R. Molontay, "Data-driven analysis of complex networks and their model-generated counterparts," *arXiv preprint arXiv:1810.08498*, 2018.
5. A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal, "graph2vec: Learning distributed representations of graphs," *arXiv preprint arXiv:1707.05005*, 2017.
6. P. Holme and B. J. Kim, "Growing scale-free networks with tunable clustering," *Physical Review E*, vol. 65, no. 2, p. 026107, 2002.
7. J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proc. 11th ACM SIGKDD Int. Conf. on Knowledge Discovery in Data Mining*, pp. 177–187, 2005.
8. M. Gjoka, B. Tillman, and A. Markopoulou, "Construction of simple graphs with a target joint degree matrix and beyond," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 1553–1561, IEEE, 2015.
9. T. P. Peixoto, "Nonparametric Bayesian inference of the microcanonical stochastic block model," *Physical Review E*, vol. 95, no. 1, p. 012317, 2017.

Evaluating network embedding by community separability

Aldo Acevedo[1], Claudio Durán[1], Alessandro Muscoloni[1,2] and Carlo Vittorio Cannistraci[1,2]

[1]Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Cluster of Excellence Physics of Life (PoL), Department of Physics, Technische Universität Dresden, Germany. [2]Center for Complex Network Intelligence (CCNI) at the Tsinghua Laboratory of Brain and Intelligence (THBI), Department of Bioengineering, Tsinghua University, China

1 Introduction

Network embedding is a vibrant research topic of recent years, and the study of appropriate ways to evaluate its performance is currently on the verge. The most employed evaluation strategies are based on link prediction and community (cluster) detection in the geometrical space [1], [2]. Here, we introduce a new concept that is named community projection separability (CPS). It is the network science declination of the more general data science notion termed as projection separability [3], which is valid for any type of data embedded in a geometrical space [3]. After embedding, in order to assess the separability of nodes in two different communities, we project them on the separability line [3], which is the line that connects the centroids of the two node clusters in the geometrical space [3]. Then, the CPS strategy consists in applying any type of bi-class separability measure (such as the Matthews correlation coefficient (MCC), F-score, etc.) directly on the separability line in order to measure the two-group separability.

2 Results

For networks with multi-community structure, the average CPS among all pairs of clusters is considered as overall measure. Using CPS - with MCC as selected separability measure - in Fig. 1A, we investigate the performance of 4 baseline methods for network embedding (HOPE, ProNE (SMF), ProNE, Node2vec) in 8 social networks of different sizes (till thousands of nodes) whose metadata for community organization is known. Considering the mean and minimum performance (Fig. 1B) across all networks, ProNE (SMF) and ProNE offer the highest CPS and therefore seem to offer the highest community separability.

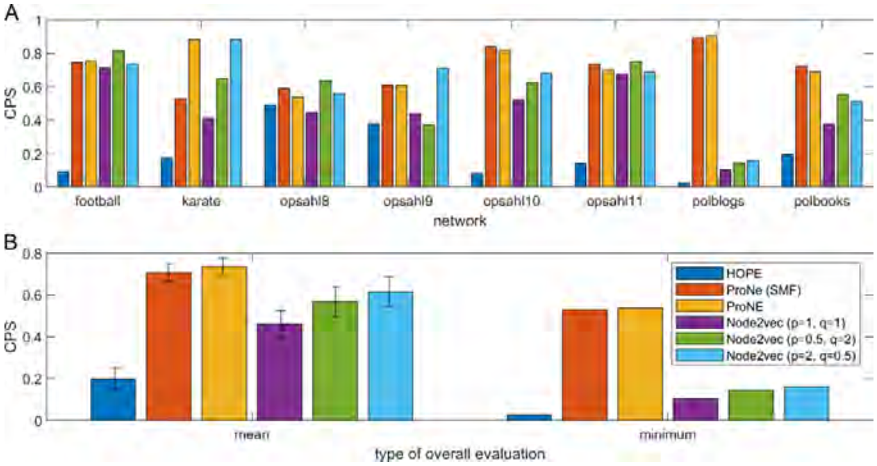


Figure. 1. CPS evaluation on real networks. (A) The barplots report the CPS (MCC based) performance of 6 network embedding methods for each of the 8 social networks. (B) The two barplots report the mean (left) and minimum (right) CPS performance of 6 network embedding methods over the 8 social networks. In case of the mean, the standard error is also shown. The legend indicates the list of the 6 network embedding methods: HOPE, ProNE (SMF), ProNE and Node2vec with 3 different parameter settings.

In order to visually inspect the reliability of CPS evaluation we focus on the Karate network, whose embedding performance pattern across the 4 methods (Fig. 1A) is close to the mean (Fig. 1B). Table 1 reports the performance of CPS versus 7 baseline cluster validity indexes [3] which are traditionally used in data analysis to evaluate clustering performance. We define as consensus (Table 1 last row) when at least 2/8 measures agree on the fact that an embedding method is the best. In Karate network ProNE has consensus 6/8, Node2vec (p=2, q=0.5) has 2/8 as well as Node2vec (p=0.5, q=2).

	Embedding methods					
Validity indices	**HOPE**	**ProNE (SMF)**	**ProNE**	**Node2vec (p=1, q=1)**	**Node2vec (p=0.5, q=2)**	**Node2vec (p=2, q=0.5)**
Community Separability Index (CPS, MCC based) [3]	0.18	0.53	**0.88**	0.41	0.65	**0.88**
Dunn Index [4]	**0.06**	0.01	**0.06**	0.01	0.07	0.02
Davies-Bouldin Index [5]	0.17	0.48	**0.66**	0.30	0.36	0.40
Generalized Dunn Index [6]	0.61	0.78	2.29	0.71	**0.80**	0.76
Calinski and Harabasz Index [7]	1.26	23.50	**123.98**	5.67	9.28	12.87
Silhouette Index [8]	0.00	0.40	**0.83**	0.23	0.32	0.31
Geometric Separability Index [9]	0.85	0.76	0.88	0.88	**0.91**	**0.91**
Cluster Validity Density-involved Distance [10]	4.88	0.58	**51.02**	3.63	17.22	10.13
Consensus	—	—	6/8	—	2/8	2/8

Table 1. Comparison of methods for community-based evaluation of network embedding in Karate network. The table reports the evaluation of performance in network embedding (6 methods considered) on the Karate network for CPS and 7 cluster validity indices. For each evaluator, the best performing embedding method (or methods) is highlighted in bold. The best results for the CPS index are also highlighted in red.

The last row reports the consensus, when at least 2/8 indices agree on the fact that an embedding method is the best.

Fig. 2 displays the embeddings of these 3 methods to visually investigate whether their community separability seems in agreement with the 8 evaluation measures. Comparing the values in Table 1 and the pattern in Fig. 2 emerges that CPS is the only method to detect the same level (just one node is misallocated) of community separability between ProNE (Fig. 2A) and Node2vec (p=2, q=0.5) (Fig. 2B) regardless of the geometrical shape of their embedding. Other evaluation methods wrongly assessed Node2vec (p=0.5, q=2) as best performance (Table 1), whereas CPS correctly evaluated its embedding as of lower quality (Fig. 2C). In conclusion, our results seem promising and offer evidence at support of CPS as a valid strategy for evaluation of network embedding.

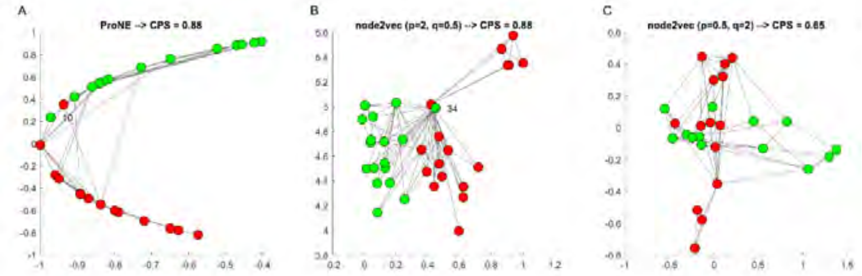


Figure. 2. Evaluation of Network embedding in Karate network. The panels represent the embedding in the 2-dimensional space of the Karate network using as embedding methods: (A) ProNE, (B) Node2vec with p=2, q=0.5 and (C) Node2vec with p=0.5, q=2. The title of each panel indicates the embedding method and the CPS (MCC based) performance. The nodes of the Karate network are colored according to their membership to the two communities. ProNE and Node2vec (p=2, q=0.5) offer the same level of community separability (CPS = 0.88) regardless of the geometrical shape of their embedding. In panels (A) and (B) we show the ID of the only node that has been misallocated by the two embedding methods, highlighting that the mistake is not on the same node (node ID 10 versus 34).

References

[1] H. Cai, V. W. Zheng, and K. C. C. Chang, "A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications," *IEEE Trans. Knowl. Data Eng.*, 2018, doi: 10.1109/TKDE.2018.2807452.

[2] P. Goyal and E. Ferrara, "Knowledge-Based Systems Graph embedding techniques, applications, and performance: A survey," *Knowledge-Based Syst.*, 2018, doi: 10.1016/j.knosys.2018.03.022.

[3] A. Acevedo, S. Ciucci, M. J. Kuo, C. Duran, and C. V Cannistraci, "Measuring group-separability in geometrical space for evaluation of pattern recognition and embedding algorithms," *arXiv:1912.12418 [cs.LG]*, 2019.

[4] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, 1973, doi: 10.1080/01969727308546046.

[5] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[6] J. C. Bezdek and N. R. Pal, "Cluster validation with generalized Dunn's indices," in *Proceedings - 1995 2nd New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems, ANNES 1995*, 1995, pp. 190–193, doi: 10.1109/ANNES.1995.499469.

[7] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat.*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610917408548446.

[8] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[9] C. Thornton, "Separability is a Learner's Best Friend," Springer, London, 1998, pp. 40–46.

[10] L. Hu and C. Zhong, "An internal validity index based on density-involved distance," *IEEE Access*, vol. 7, pp. 40038–40051, 2019, doi: 10.1109/ACCESS.2019.2906949.

Part VIII

Modeling Human Behavior

Impact of individual actions on the collective response of social systems

S. Martin-Gutierrez, J. C. Losada, and R. M. Benito

Grupo de Sistemas Complejos, Escuela Técnica Superior de Ingeniería Agronómica,
Alimentaria y de Biosistemas, Universidad Politécnica de Madrid, Av. Puerta de Hierro, 2,
28040 Madrid, Spain.

While we as individuals have the ability to influence those that are close to us (friends, family, colleagues...), the structure of our social networks can propagate that influence far beyond our personal social circle. This topic has received considerable attention and has been studied in different but closely related disciplines [1,5]. For example, in the field of diffusion on networked systems, which studies the spread of diseases or information and the emergence of cascading phenomena [6,7]. Other works focus on the Influence Maximization problem, taking advantage of the diffusion mechanisms to find a set of individuals whose actions would maximize the response [2].

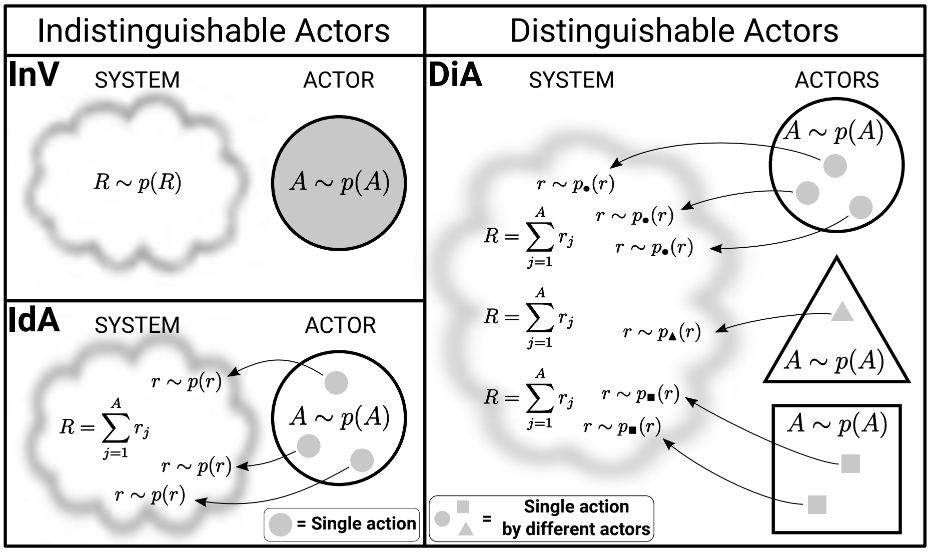


Fig. 1. Diagram that summarizes the main characteristics of the three models: Independent Variables (InV), Identical Actors (IdA) and Distinguishable Actors (DiA).

In this work [3] we present three mathematical models to explain how the actions performed by an actor, individual or agent (her activity) influence the collective reaction (or response) of the social system she is embedded in. The developed models consider different levels of dependence between response and activity. In the first model, called Independent Variables (InV) model, we consider activity and response to be completely independent, while in the second and third models the individual activity influences the response. The main difference between these last two models is the distinguishability of the actors. In the Identical Actors (IdA) model, the system is agnostic with respect to the individual that performs the actions, while in the Distinguishable Actors (DiA) model, the dependence between activity and response is determined by the features of the actor that performs the action. In Figure 1 we present a diagram that summarizes the main ideas behind each model.

We use the models to obtain the distribution of the efficiency metric η, defined as the ratio between the number of reactions R triggered by an actor on other members of the system and the number of actions A of the actor:

$$\eta = \frac{R}{A} \tag{1}$$

This metric is a generalization of the user efficiency, first introduced in the context of Twitter by some of the authors [4]. The simplicity of the relationship between A and R makes this metric optimal for analytical treatment; hence, we have chosen it over other standard metrics like the h-index, which is highly non-linear. The models are tested with 29 datasets from three systems of different nature: Twitter conversations, where A is the number of published tweets by a user and R the number retweets the user obtains; the scientific citations network, where A is the number of publications by an author and R the number of citations; and the Wikipedia collaboration environment, where A is the number of editions and R the number of received messages. In all the systems the efficiency distribution presents a universal shape, like the one shown in Figure 2, with small but relevant differences between systems.

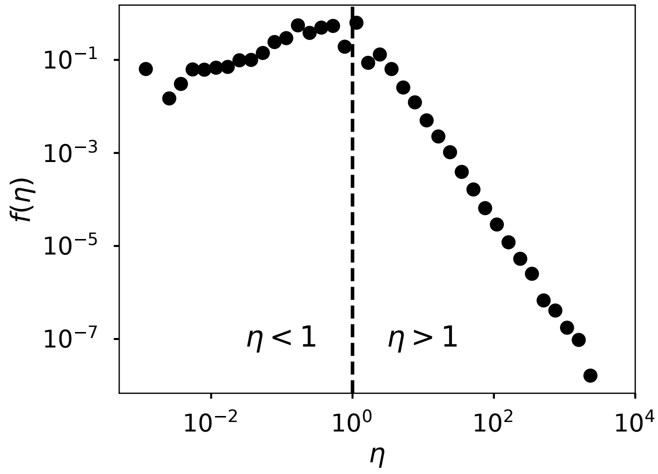


Fig. 2. Empirical probability density function of efficiency corresponding to a Twitter conversation.

The Independent Variables model is able to explain two fundamental characteristics of the efficiency distribution for which there was previously only empirical evidence: its universal shape and its independence with respect to changes in the activity distribution. Additionally, it reproduces the efficiency distribution for the scientific citations network appropriately. However, there are small discrepancies between the InV model and the data for Twitter and Wikipedia. We find the cause for the discrepancies and take it into account to develop the Identical Actors model, which improves the results for both systems. In this model, the correlations between individual actions and collective response emerge naturally from the hypothesis of the model. The theoretical correlations are comparable to those found in the empirical data. When it comes to the efficiency distribution, the model reproduces adequately the right tail ($\eta > 1$) for Twitter and Wikipedia. We again study the small discrepancies between the IdA model and the data and from this analysis we develop the Distinguishable Actors model, which fit the Twitter data remarkably well for the whole range of efficiencies. Moreover, the DiA model improves the concordance between theoretical and empirical activity-response correlations. However, the complexity and lack of comprehensive metadata for Wikipedia and the citations network makes the application of the DiA model to these systems too cumbersome.

To summarize, in this work we show the universality of the shape of the efficiency distribution in three systems of different nature. We explain how this universal shape emerges with a parsimonious model and develop two more sophisticated models to get a thorough understanding of the particularities of the efficiency distribution in each system considered. The three models have clear and intuitive interpretations. Furthermore, the models pave the way for more elaborated and domain-specific theories and can be used as null models or baselines for those.

References

1. Juul, J.S., Porter, M.A.: Hipsters on networks: How a minority group of individuals can lead to an antiestablishment majority. Phys. Rev. E 99, 022313 (Feb 2019), https://link.aps.org/doi/10.1103/PhysRevE.99.022313
2. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 137–146. ACM (2003)
3. Martin-Gutierrez, S., Losada, J.C., Benito, R.M.: Impact of individual actions on the collective response of social systems. Scientific Reports 10(1), 12126 (Jul 2020), https://doi.org/10.1038/s41598-020-69005-y
4. Morales, A.J., Borondo, J., Losada, J.C., Benito, R.M.: Efficiency of human activity on information spreading on twitter. Social Networks 39, 1–11 (2014), https://doi.org/10.1016/j.socnet.2014.03.007
5. Muchnik, L., Pei, S., Parra, L.C., Reis, S.D.S., Andrade Jr, J.S., Havlin, S., Makse, H.A.: Origins of power-law degree distribution in the heterogeneity of human activity in social networks. Scientific Reports 3, 1783 EP – (May 2013), https://doi.org/10.1038/srep01783, article
6. Pastor-Satorras, R., Castellano, C., Van Mieghem, P., Vespignani, A.: Epidemic processes in complex networks. Rev. Mod. Phys. 87, 925–979 (Aug 2015), https://link.aps.org/doi/10.1103/RevModPhys.87.925
7. Weng, L., Menczer, F., Ahn, Y.Y.: Virality prediction and community structure in social networks. Scientific Reports 3, 2522 EP – (Aug 2013), https://doi.org/10.1038/srep02522, article

Revealing semantic and emotional structure of suicide notes with cognitive network science

Andreia Sofia Teixeira[1,2,3], Szymon Talaga[4], Trevor James Swanson[5], and Massimo Stella[6]

[1] Center for Social and Biomedical Complexity, School of Informatics, Computing, & Engineering, Indiana University, Bloomington IN, USA
[2] Indiana Network Science Institute, Indiana University, Bloomington IN, USA
[3] INESC-ID, Lisboa, Portugal
[4] University of Warsaw, The Robert Zajonc Institute for Social Studies, Warszawa, Poland
[5] University of Kansas, Department of Psychology, Lawrence KS, USA
[6] Complex Science Consulting, Lecce, Italy

1 Introduction

Suicide notes represent the last remaining trace of mental states of people who committed suicide [4]. Investigating these notes is key to understanding how these individuals perceived their own choice, its context and its framing. To this aim, we build on the framework of cognitive network science [2, 10] for quantitatively extracting key ideas and emotions present in 139 genuine suicide notes [8].

This work relies on cognitive network science, a field merging data science, psychology and complex networks in order to understand cognition through large-scale knowledge graphs and language processing [2, 6, 11]. Unraveling these cognitive networks of concepts becomes key for opening a window onto people's minds and understanding them through data-driven analyses. This is also the main motivation of our approach, which overlaps with previous studies in psychology relying on human coding for understanding the language of suicide notes [1]. In comparison to these approaches, our key innovation lies in using automatic tools for quantitative reconstructing the semantic frames [3] and emotional perceptions [7] attributed by authors to key aspects of their suicidal ideation.

2 Methods

We conducted three studies. Study 1 investigated the "emotional syntax" of suicide notes, analysing whether the connectivity and assembly of words is somehow related to their valence. We used *structural balance theory* [5] to assess the degree of balance in the empirical network of syntactic dependencies as extracted from all texts. We used conceptual valence in order to classify words in three categories: positive, negative and neutral concepts (as defined by psycholinguistic norms, cf. [7, 9]). Structural balance investigated how positive and negative concepts were organised among neighbouring words. Study 2 delved more into the centrality of concepts in suicide notes, such as "love", of which semantic frames were then investigated and reconstructed. Study 3 combined semantic frames with emotional data in order to describe and quantify typical emotions associated with different concepts in suicide notes.

3 Results and Discussion

Our multidisciplinary approach enables a detailed cognitive map outlining key features of the narrative of suicide notes, see Figure 1 (left). Structural balance analysis indicated that, in comparison to randomised null models (see Figure 1, top right), suicide notes are found to display a high level of positive structural balance, where positively perceived concepts are prominently central and are found to cluster together, reducing contrast with more peripheral and negative concepts.

Language processing and network centrality measures highlighted that suicide notes display a psychological narrative crucially revolving around basic social aspects of life, like family relationships, love and personal identifiers (e.g. “I”, “he/she”, “you”). Semantic neighbourhoods and their network structure are highlighted in Figure 1 (bottom right). Notice that the concept of “love” was framed along both spiritual and interpersonal dynamics and it was perceived with a combination of trust, joy and sadness. This emotional profiling indicates a quantitative pattern of resignation, in contrast with the positive perception of “love” in mainstream language. Similar results hold also for “life”, “want” and “help”.

Our “text-as-data” approach highlights cognitive framings aimed at creating soothing narratives that minimize affective conflicts. This framework unearths more nuanced framings of central concepts in suicide notes, like “love”, and how their usage may differ from common language, quantitatively addressing debates on contextual valence shifting occurring within suicide ideation [4]. The evidence reported here indicates that future analyses of suicide notes should not interpret words as if they were used in commonly used language. Complex networks and conceptual associations can better capture how suicidal ideation alters the framing of concepts themselves. Cognitive networks can thus capture relevant contextual information that would be unavailable when considering words in isolation.

Our results reveal features relevant for next-generation cognitive modeling capable of accounting for contextual shifting and of relevance for AI-powered text analyses aimed at understanding suicidal risk [12].

References

1. Al-Mosaiwi, M., Johnstone, T.: In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical Psychological Science 6(4), 529–542 (2018)
2. Castro, N., Siew, C.S.: Contributions of modern network science to the cognitive sciences: Revisiting research spirals of representation and process. Proceedings of the Royal Society A 476(2238), 20190825 (2020)
3. Fillmore, C.J., et al.: Frame semantics. Cognitive linguistics: Basic readings 34, 373–400 (2006)
4. Galasinski, D.: Discourses of men’s suicide notes: A qualitative analysis. Bloomsbury Publishing (2017)

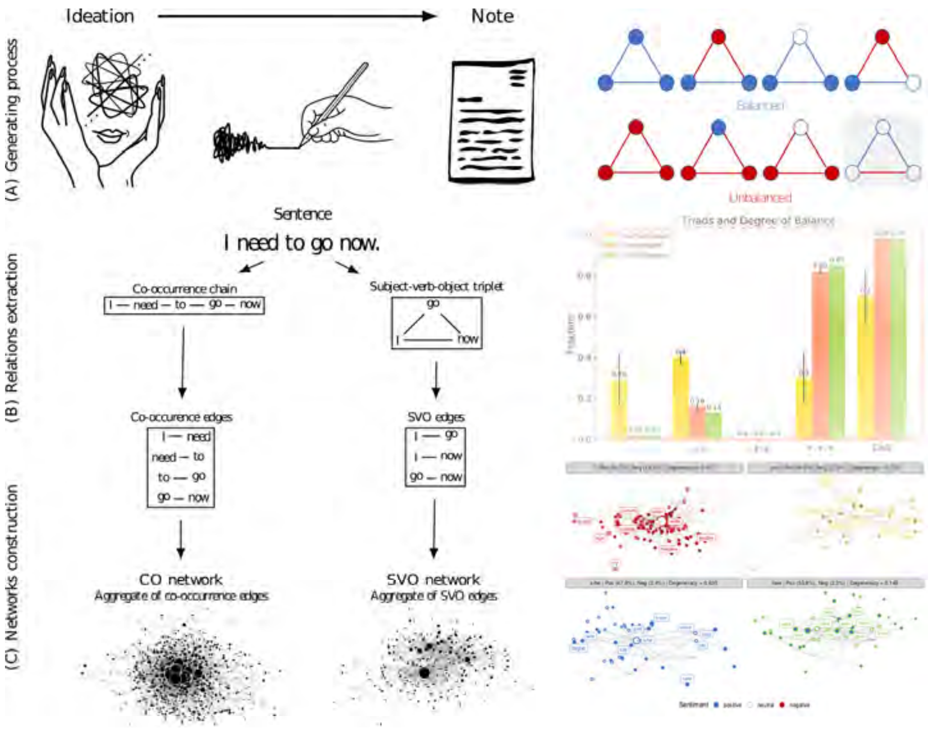


Fig. 1. Network construction methods (left). Structural balance analysis results (top right). Main concept clusters in the SVO network (bottom right).

5. Heider, F.: Attitudes and cognitive organization. Journal of Psychology 21, 107–112 (1946Journal of Psychology)
6. Kenett, Y.N., Levi, E., Anaki, D., Faust, M.: The semantic distance task: Quantifying semantic distance with semantic network path length. Journal of Experimental Psychology: Learning, Memory, and Cognition 43(9), 1470 (2017)
7. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word–emotion association lexicon. Computational Intelligence 29(3), 436–465 (2013)
8. Schoene, A.M., Dethlefs, N.: Automatic identification of suicide notes from linguistic and sentiment features. In: Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 128–133 (2016)
9. Stella, M.: Text-mining forma mentis networks reconstruct public perception of the stem gender gap in social media. arXiv preprint arXiv:2003.08835 (2020)
10. Stella, M., Beckage, N.M., Brede, M.: Multiplex lexical networks reveal patterns in early word acquisition in children. Scientific reports 7, 46730 (2017)
11. Stella, M., Restocchi, V., De Deyne, S.: #lockdown: Network-enhanced emotional profiling in the time of covid-19. Big Data and Cognitive Computing 4(2), 14 (2020)
12. Teixeira, A.S., Talaga, S., Swanson, T.J., Stella, M.: Revealing semantic and emotional structure of suicide notes with cognitive network science (2020)

Segregation dynamics with reinforcement learning and agent based modeling

Egemen Sert[1,2], Yaneer Bar-Yam[1], and Alfredo Morales[1,3]

[1] New England Complex Systems Institute, Cambridge, MA,
[2] Middle East Technical University, Ankara, Turkey egemen.sert@metu.edu.tr
[3] MIT Media Lab, Cambridge, MA

Societies are complex and their properties emerge from the interplay and weaving of individual actions. Rewards are key to understand people's choices and decisions. For instance, individual preferences of where to live may lead to the emergence of social segregation. In this paper, we combine Reinforcement Learning (RL) with Agent Based Modeling (ABM) in order to address the self-organizing dynamics of social segregation and explore the space of possibilities that emerge from considering different types of incentives. Our model promotes the creation of interdependencies and interactions among multiple agents of two different kinds that segregate from each other. For this purpose, agents use Deep Q-Networks (DNN) to make decisions inspired on the rules of the Schelling Segregation model and rewards for interactions. Despite the segregation reward, our experiments show that spatial integration can be achieved by establishing interdependencies among agents of different kinds. They also reveal that segregated areas are more probable to host older people than diverse areas, which attract younger ones. Through this work, we show that the combination of RL and ABM can create an artificial environment for policy makers to observe potential and existing behaviors associated to rules of interactions and rewards [1].

Our experiments consist of two types of agents in a 50x50 grid where they can move around and interact with other agents. Agents evaluate the number of agents per kind in an observation window centered around them and decide whether to move away or not and in which direction. We set up different values of incentives and observe the emergent collective behavior. During simulations, agents explore the space of possibilities and inform which behaviors are promoted under certain rewards and environmental rules. As a result, we create an artificial environment for testing hypotheses and obtaining information through simulations hard to anticipate given the complexity of the space of possibilities.

The rewards are scalar values that we individually provide to agents at each simulation time after evaluating their current state. A Segregation reward (SR) promotes agents' segregation, in the form: $SR = s - \alpha d$, where s is the number of agents of similar kind within the agent's observation window, d is the number of agents of different kind within the observation window and $\alpha \in [0,1]$ is a parameter we use to control the intolerance of agents to be close to those that are different. The segregation parameter α is equivalent to the threshold used in the original Schelling model [2]. An Interdependence reward (IR) promotes interactions among agents of different kind. The interdependencies among agents of different kinds is inspired by the dynamics of population models where agents need to interact with each other in order to extend their lifetime [3]. When an agent meets another agent of different kind, we choose a winner and a loser of the interaction. The winner receives a positive reward and an extension of its lifetime by one iteration. The loser ceases to exist. We use the IR as a parameter we can vary $IR \in [0, 100]$ in order to promote interactions among agents of different kind.

Figure 1 A shows the emergent collective behavior for multiple values of α (rows) at multiple times of the simulation (columns) without considering yet the effects of IR. Rows represent outcomes associated to different values of the segregation parameter (α). Columns show the state of the system at different points of the simulation. Panels show the average type occupation per location. Red regions denote biased occupation of one type of agents and blue regions denote biased occupation the other type. White areas indicate the average pattern. Lower values of α yield mostly white spaces, indicating a mixed population. As we increase α the segregation of agents begins. With high levels of α the segregation is pronounced and blue and red segregated clusters emerge. Similarly to the original Schelling segregation model, segregation occurs for smaller values of α in the long run (see $\alpha = 0.5$). The white regions for lower values of α indicate mixing, while the white regions of higher values of alpha are characterized for being emptier.

We provide rewards to create interactions and interdependencies among both populations. For this purpose, we combine the segregation dynamics shown in Fig. 1 A with the interdependence reward (IR). We explore multiple combinations of the segregation parameter α and the interdependence reward (IR)[4]. The resulting segregation of those simulations is visualized in Figure 1 B. The x-axis represents the segregation parameter α and the y-axis represents the interdependence reward (IR). The figure shows a contour plot of the expected amount of segregation in the system during the last 1000 iterations. We calculate segregation using entropy (see [1]). Red regions indicate high segregation and green regions show lower segregation. Segregation is high (red) when promoted (high α) and interdependencies are not rewarded. As interdependencies increase, the agents mix and the spatial segregation is significantly reduced (blue), even for high values of α. Therefore, high levels of interdependencies seem to counter the rewards for segregation. The resulting mixing for high levels of interdependencies are comparable to very low levels of α.

In summary, we created an artificial environment for testing rules of interactions and rewards by observing the behaviors that emerge when applied to multi-agent populations. We combine agent based modeling (ABM) with artificial intelligence (RL) in order to explore the space of solutions associated to promoted rewards. RL provides ABM the information processing capabilities that enables the exploration of strategies that satisfy the conditions imposed by the interaction rules. In turn, ABM provide RL with access to models of collective behavior that achieve emergence and complexity. While ABMs provide access to the complexity of the problem space, RL facilitates

[4] We share videos of segregation experiments at the following links:
(α=0) https://youtu.be/1qfbg4NLp8w,
(α=0.25) https://youtu.be/8nqll-jh9Ds,
(α=0.50) https://youtu.be/LXAKN3GrzEo,
(α=0.75) https://youtu.be/doNt7UJBqbg,
(α=1.00) https://youtu.be/YP0FGUo4tH4,

the exploration of the solution space. Our methodology opens a new avenue for policy makers to design and test incentives in artificial environments.

References

1. Sert, E., Bar-Yam, Y., & Morales, A. J. Segregation dynamics with reinforcement learning and agent based modeling. Scientific Reports, 10 (1), 1-12. (2020).
2. Schelling, T. C. (1971). Dynamic models of segregation. Journal of mathematical sociology, 1(2), 143-186. ISO 690
3. Lanchester, F. W. (1956). Mathematics in warfare. The world of mathematics, 4(Part XX), 2138-2157. ISO 690

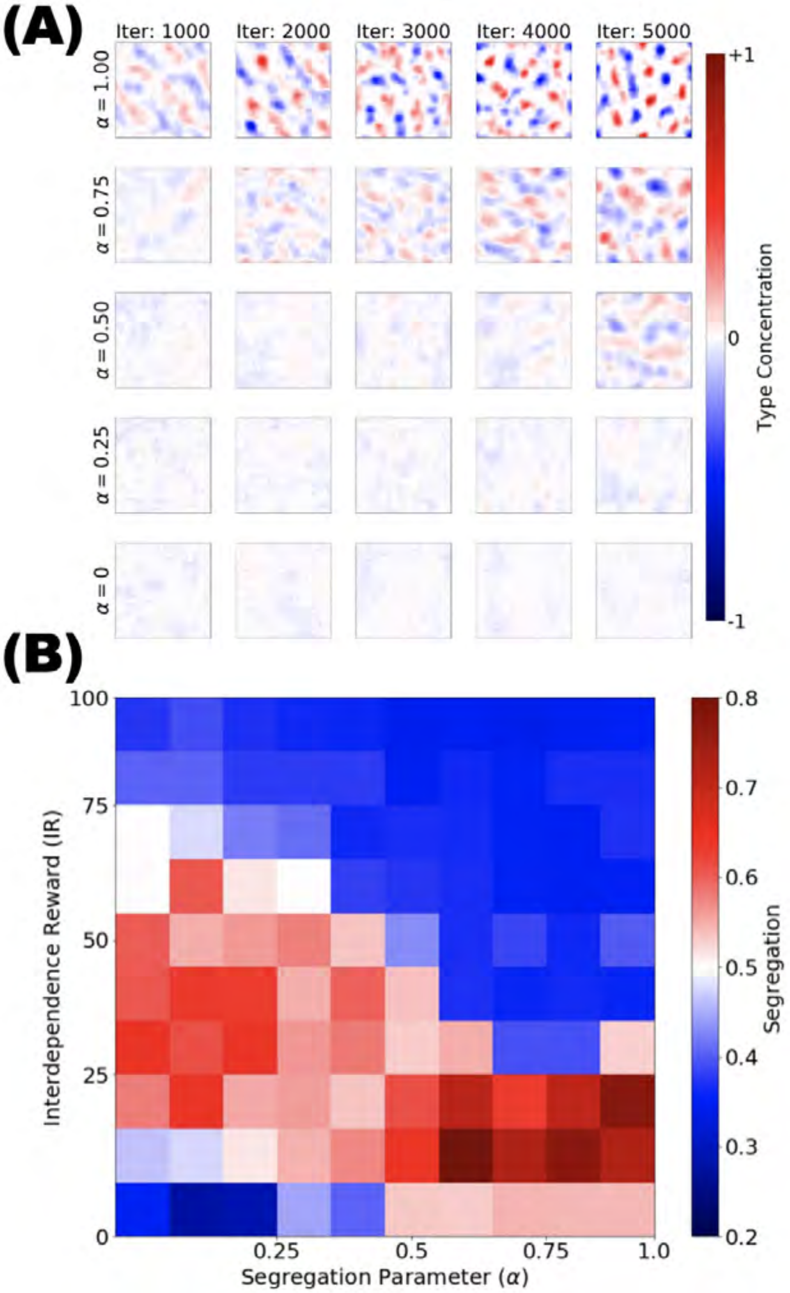


Fig. 1. Panel A: Agents collective behavior for multiple values of segregation reward α (rows) at multiple times (columns). Rows represent outcomes associated to different values of segregation reward (α). Columns show the state of the system at different points of the simulation. Colors indicate the concentration of both types of agents (blue and red). Panel B: Segregation values for multiple values of segregation reward (α) and interdependence reward (IR). Colors correspond to amount of segregation measured in the last 1000 iterations of the simulation. Scales in figure.

Behavioral gender differences are reinforced during the COVID-19 crisis

Tobias Reisch[1,2], Georg Heiler[2,3] Jan Hurt[2], Peter Klimek[1,2], Allan Hanbury[2,3], and Stefan Thurner[1,2]

[1] Institute for Complex Systems, Medcal University of Vienna,
Spitalgasse 23, 1090 Wien, Austria
[2] Complexity Science Hub Vienna, Josefstädter Straße 39, 1080 Wien, Austria
[3] Institute of Information Systems Engineering, Technical University Vienna,
Favoritenstraße 9-11, 1040 Wien, Austria

1 Introduction

Gender differences exist in a wide range of human activities including communication and mobility. Women and men are also differently affected by pandemics[6] such as the ongoing COVID-19 crisis. Studying the collective response to crisis is essential for first aid coordinators [8] and policy makers concerned with a population's health and safety [7] and, on the other hand, reveals human qualities that only surface when facing different kinds of perceived danger [2, 10, 4].

At the end of 2019 the SARS-Cov-2 virus emerged in China and subsequently caused an – at the moment of writing this work – ongoing, world-wide pandemic of the novel coronavirus disease. On March 15th Austria introduced a widespread lock-down in response to the COVID-19 pandemic. Right from the start, this lead to the apprehension that women could be affected worse by the lock-down due to additional childcare duties [3], domestic violence [11], employment in high exposure jobs and simultaneously higher unemployment[6]. In fact, women were more strongly affected by unemployment and partial layoffs [1], surveys registered an increase in domestic violence [9] and female scientist posted less pre-prints and started less projects [12, 5].

Here we quantify the gender specific response of behavior in mobility, communication and circadian activity under crisis conditions by examining the natural experiment of the COVID-19 pandemic in Austria. We use a large scale anonymized longitudinal dataset of 1.2 million cellphones, covering approximately 15% of the Austrian population. Our analyses compare various periods of time before, during and after the lock-down. Each period is based on the introduction or easing of one or more non pharmaceutical interventions.

2 Results

Typically, total call duration is high on weekdays and low on weekends, with women having 20% longer calls on weekdays and 30% longer calls weekends. After the first announcement of COVID-19 restrictions on March 10th, we observe a drastic rise in the median total call time per user T_i, with a peak increase of 200% on March 16th. The

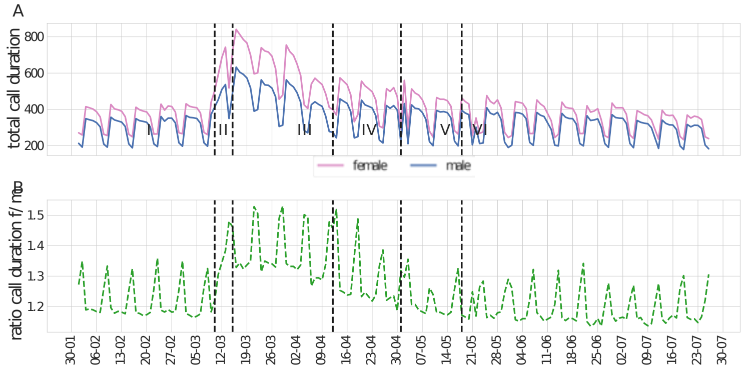


Fig. 1. Total call duration before, during and after the COVID-19 lock-down. (**A**) The total call duration T_i for men and women and (**B**) the gender ratio female/male for T_i. During the lock-down, both genders drastically increase their T_i with a monotonic decrease to normal levels after mid May. The gender ratio shifts towards women having longer T_i, with a similar decrease towards pre-crisis levels in mid May.

increase in in T_i is larger for women, resulting in the gender ratio T_i^f/T_i^m increasingly shifted towards females. The shift in gender ratio persists during the lock-down and fades out only in May after the lifting of the restrictions.

We calculate the median radius of gyration R_G, a measure for human mobility, for every day, from February 1st until the end of the study period, June 29th. Figure 2A shows the decline and return in median R_G over the study period for all of Austria, highlighting a drop in 40% of R_G in the first phase of the lock-down and a subsequently steady increase of the level of movement thereafter.

The female population is moving less than males for the whole period covered in our data set. We see that the difference of the ratio r_{R_G} on an average pre-crisis (phase I) weekday is 78% and 88% on weekends (see Figure 2B). After a brief transition period II the weekday ratio drops to around 73% during the lock-down phase III, but remains at the previous level on weekends.

As women restrict their movement more strongly than men, the gender ratio decreases. In the phase of easing the lock-down restrictions (phase IV), R_G for males begins to rise more strongly than R_G for females, hence decreasing the mobility ratio further to 67%. The ratio starts to recover towards pre-crisis levels after a further easing of public restrictions (such as the opening of restaurants in phase V).

Summary. Gender differences are increased during the COVID-19 lock-down. We find shifts in the gender-ratio of measures characterizing communication and mobility starting with the announcement of the COVID-19 lock-down and persisting until and even after the easing of restrictions. The COVID-19 pandemic provides a unique opportunity to learn about behavioral response to large scale crisis. Our findings imply that women intensify their social contacts stronger than men and react more strongly to the lockdown. Hence, careful investigations into permanent consequences are necessary.

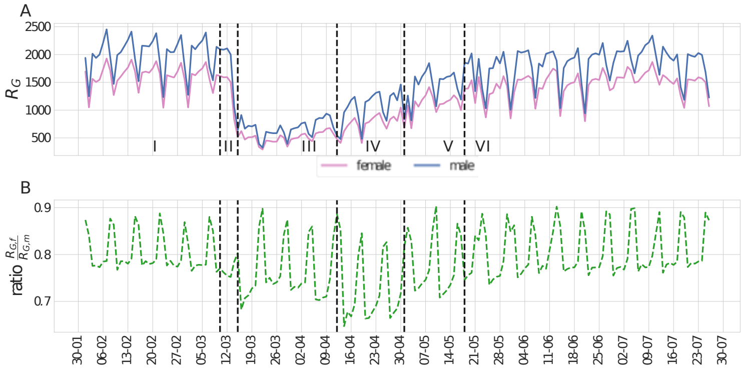


Fig. 2. Changes in mobility during the COVID-19 pandemic. (**A**) Radius of gyration R_G for both genders and (**B**) gender ratio female/male of R_G. During the lock-down mobility is drastically restricted during the lock-down and the gender ratio is shifted towards women moving less. After the first easing of the measures on Easter, the gender ratio drops further, indicating men increasing their mobility stronger.

References

1. Arbeitsmarktservice Austria: Arbeitsmarktdaten - Berichte und Auswertungen (2020), https://www.ams.at/arbeitsmarktdaten-und-medien/arbeitsmarkt-daten-und-arbeitsmarkt-forschung/berichte-und-auswertungen
2. Bagrow, J.P., Wang, D., Barabasi, A.L.: Collective response of human populations to large-scale emergencies. PloS one 6(3), e17680 (2011)
3. IFES and SORA: Home-Office: Positive Resonanz, aber mehr Stress (2020), https://www.ifes.at/arbeitsklima-index-2020-home-office
4. Matud, M.P.: Gender differences in stress and coping styles. Personality and individual differences 37(7), 1401–1415 (2004)
5. Myers, K.R., Tham, W.Y., Yin, Y., Cohodes, N., Thursby, J.G., Thursby, M.C., Schiffer, P., Walsh, J.T., Lakhani, K.R., Wang, D.: Unequal effects of the covid-19 pandemic on scientists. Nature Human Behaviour pp. 1–4 (2020)
6. OECD: Women at the core of fight against Covid-19 crisis (2020), https://www.oecd.org/coronavirus/policy-responses/women-at-the-core-of-the-fight-against-covid-19-crisis-553a8269/
7. Oliver, N., Letouzé, E., Sterly, H., Delataille, S., Nadai, M.D., Lepri, B., Lambiotte, R., Benjamins, R., Cattuto, C., Colizza, V., de Cordes, N., Fraiberger, S.P., Koebe, T., Lehmann, S., Murillo, J., Pentland, A., Pham, P.N., Pivetta, F., Salah, A.A., Saramäki, J., Scarpino, S.V., Tizzoni, M., Verhulst, S., Vinck, P.: Mobile phone data and covid-19: Missing an opportunity? (2020)
8. Petrescu-Prahova, M., Butts, C.T.: Emergent coordination in the world trade center disaster. Institute for mathematical behavioral sciences pp. 1–23 (2005)
9. Steiner, J., Ebert, C.: The impact of covid-19 on violence against women and children in germany. Preprint (June 2020), https://www.hfp.tum.de/globalhealth/forschung/covid-19-and-domestic-violence/

10. Taylor, S.E., Klein, L.C., Lewis, B.P., Gruenewald, T.L., Gurung, R.A., Updegraff, J.A.: Biobehavioral responses to stress in females: tend-and-befriend, not fight-or-flight. Psychological review 107(3), 411 (2000)
11. United Nations: UN supporting 'trapped' domestic violence victims during COVID-19 pandemic (2020), https://www.un.org/en/coronavirus/un-supporting-%E2%80%98trapped%E2%80%99-domestic-violence-victims-during-covid-19-pandemic
12. Viglione, G.: Are women publishing less during the pandemic? Here's what the data say (2020)

Dynamical patterns of user activity during Spanish electoral campaigns and debates in Twitter

L. Pérez-Sienes, J. Atienza-Barthelemy, S. Martín-Gutiérrez, J. C. Losada,
and R. M. Benito

Grupo de Sistemas Complejos, ETS Ingeniería Agronómica, Alimentaria y de Biosistemas,
Universidad Politécnica de Madrid, Avda. Complutense s/n 28040 Madrid, Spain ,
http://www.gsc.upm.es

1 Introduction

Twitter appears to be one of the most used social networks regarding topics of public interest [7]. On the one hand, there are not only anonymous people but politicians, political parties, reporters, newspapers, radio and TV that use the space to publish content and interact with other users. On the other hand, it is possible to download thousands of tweets per second and filter them by keywords and/or spatial geo-localization. This way, one can easily build a database containing a collection of tweets of a given topic of study. For these reasons, Twitter is an extremely useful data source for researching user behavior in electoral settings [3].

In the last years a lot of studies appeared in relation to this topic. In particular, special attention has been payed to electoral campaigns [1] using different approaches. For example, counting the number of retweets and mentions to the accounts of political parties and candidates, inferring collective opinion through the interaction network [6], etc. We focus in the analysis of the dynamical patterns of the time series of users activity during the last four elections that have taken place in Spain. These events have provided us with a unique opportunity to analyze behavioral trends and identify activity patterns in several consecutive electoral contexts. In particular, we have analyzed user behavior in the Spanish electoral campaigns of 2015, 2016 [5], April 2019 and November 2019.

To collect the data, we elaborated lists of words with terms related to the topic under study. The list for each scenario contains names of political candidates, parties, slogans, etc. Furthermore, we did not restrict our search to any language or national territory, filtering the tweets only by their content. The datasets were obtained by using the Twitter API.

2 Results

We have computed time series of the number of tweets and users per unit of time for different temporal scales for each electoral campaign. We have found that user activity diverges from the average activity profile during debates but also on the Election Day. We observe that during those days user activity rises, reaching a peak around five times greater than the general trend during regular campaign days (see Figure 1). Moreover, a common pattern appears: everyday activity begins with an exponential growth that

	α		r^2	
	2015	2016	2015	2016
Tweets	1.20 ± 0.06	1.04 ± 0.11	0.96	0.85
Retweets	1.25 ± 0.08	1.02 ± 0.12	0.94	0.80
Mentions	1.22 ± 0.10	0.9 ± 0.2	0.89	0.66

Table 1. Values of the exponents (α) and correlation coefficients (r^2) of the linear regressions of the log-log plots of tweets, retweets, and mentions with respect to the number of unique users per day shown in Figure 2 for both electoral campaigns.

leads to a plateau, continues with a peak of activity at around 22h-00h followed by an exponential decay [2]. We have also found that there exists a relationship of power

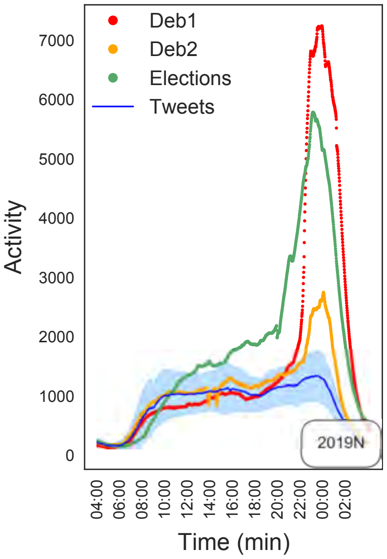


Fig. 1. Activity time series during two electoral debates (red and orange lines), Election Day (green line) and activity average (blue line) with its standard deviation for the rest of the campaign days.

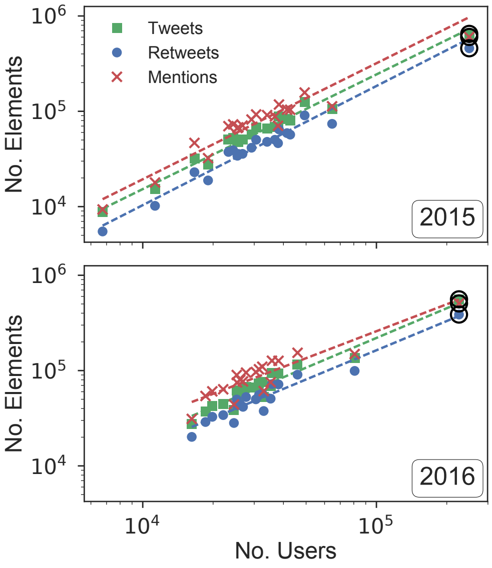


Fig. 2. Total number of tweets, retweets, and mentions per day as a function of the number of daily unique users for the 2015 campaign (top) and the 2016 campaign (bottom). Note that the data corresponding to the day of the elections (marked with a circle) were not included in the fits.

law, $y = x^{\alpha}$, between the total number of tweets, retweets and mentions per day (x) and the number of unique users (y) that have participated in the political conversation in Twitter for each day of the campaign. In Figure 2 we show least squares fits of the (logarithmically scaled) data to the previous expression, and in Table 1 the values of α and r^2 for the fits are presented. To understand this behavior we have followed the work by Leskovec et al. [4], where it is shown that when real-world networks evolve through time, the number of links $E(t)$ scale with the number of nodes $N(t)$ as $E(t) \propto N(t)^a$.

To get further insight into the evolution of user interest in electoral contexts we also analyzed the electoral debates. Nowadays it is very common for users to establish a parallel online conversation while watching TV at the same time. From the activity levels we find that when the debate starts on the TV users are already tweeting about it. Traditionally, candidates use the last minutes of the debate to address the audience directly asking for their votes; this is known as the "golden minute". The golden minute was thought to be the moment when most viewers were paying attention to the debate, so the candidates take advantage of the catchy slogans and political propaganda to gain votes. However, activity does not reflect this behavioral feature, as it peaks around mid-debate. After midnight there is an exponential decay showing the time at which users stop interacting. The Election Day is also the day when users post most of the tweets. In this case, activity rises faster than during debates and the peak is reached around the time when the results are published, quickly followed by another exponential decay.

Summarizing, in this work we have shown dynamical patterns of user behavior encoded in user activity time series during electoral events. We have used them to characterize the public response to offline events like the Election Day and the electoral debates.

References

1. Borondo, J., Morales, A.J., Losada, J.C., Benito, R.M.: Characterizing and modeling an electoral campaign in the context of twitter: 2011 spanish presidential election as a case study. Chaos: An Interdisciplinary Journal of Nonlinear Science 22(2), 023138 (2012), http://aip.scitation.org/doi/abs/10.1063/1.4729139
2. Candia, C., Jara-Figueroa, C., Rodriguez-Sickert, C., Barabási, A.L., Hidalgo, C.A.: The universal decay of collective memory and attention. Nature Human Behaviour 3(1), 82–91 (2019), https://doi.org/10.1038/s41562-018-0474-5
3. Jungherr, A.: Twitter use in election campaigns: A systematic literature review. Journal of Information Technology & Politics 13(1), 72–91 (2016), https://doi.org/10.1080/19331681.2015.1132401
4. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp. 177–187. ACM (2005)
5. Martin-Gutierrez, S., Losada, J.C., Benito, R.M.: Recurrent patterns of user behavior in different electoral campaigns: A twitter analysis of the Spanish general elections of 2015 and 2016. Complexity 2018, 2413481 (2018), https://doi.org/10.1155/2018/2413481
6. Olivares, G., Cárdenas, J.P., Losada, J.C., Borondo, J.: Opinion polarization during a dichotomous electoral process. Complexity p. 5854037 (2019)
7. Parmelee, J.H.: Political journalists and twitter: Influences on norms and practices. Journal of Media Practice 14(4), 291–305 (2013), http://www.tandfonline.com/doi/abs/10.1386/jmpr.14.4.291_1

Inequality is rising where social network segregation interacts with urban topology

Gergő Tóth[1,2], Johannes Wachs[3,4] Riccardo Di Clemente[5,6], Ákos Jakobi[7], Bence Ságvári[8], János Kertész[9], and Balázs Lengyel[1,10,11]

[1] ANET Lab, Centre for Economic and Regional Studies, Hungary
toth.gergo@krtk.mta.hu,
home page: anet.krtk.mta.hu
[2] Spatial Dynamics Lab, University College Dublin, Ireland
[3] Institute for Information Business, Vienna University of Economics and Business, Austria
[4] Complexity Science Hub Vienna, Austria
[5] Department of Computer Science, University of Exeter, United Kingdom
[6] Centre for Advanced Spatial Analysis, University College London, United Kingdom
[7] Department of Regional Science, Eötvös Loránd University, Hungary
[8] Centre for Social Sciences, Hungary
[9] Central European University, Department of Network and Data Science, Austria
[10] Corvinus University of Budapest, Institute of Advanced Studies, Hungary
[11] International Business School Budapest, Hungary

1 Introduction

Social networks amplify inequalities by fundamental mechanisms of social tie formation such as homophily and triadic closure. Together these forces sharpen social segregation reflected in network fragmentation. Geographical impediments such as distance and physical or administrative boundaries are also known sharpen social segregation. Yet, little is known about the joint relation between physical geography, social network structure, and inequalities.

Structural perspectives on inequality are at the core of sociology, which emphasises that social relations provide individuals with essential access to economic opportunities [1]. Social networks are claimed to maintain and even amplify inequalities when economic status plays a role in how social relations are established [2]. For example, a major micro-level mechanism for social-tie formation is homophily, the tendency for similar individuals to become friends [3]. Triadic closure, the phenomenon that friends of friends are more likely become friends [4], compounds the effect of homophily in tie formation [5]. Since wealth is the one of the most significant boundaries to social relations in most societies, these micro-level mechanisms can result in social segregation at the macro scale: groups with different socioeconomic status are separated from each other in social networks [6]. This kind of macro-scale network topology can lead to divergence of economic potentials between groups if access to resources or information runs through the network [2].

Towns in which amenities are spatially concentrated are also more socially segregated. To better understand how the structure of built environment relates to income inequalities through social relations, we use open source geographic data to develop three measures of urban segregation of Hungarian towns: 1. the average residential distance from the town center, 2. the extent of spatial concentration of amenities in towns, and 3. the degree to which physical barriers divide residential areas. Each of these indicators are significantly related to social network fragmentation. Using a machine learning approach, we find that these geographic indicators are better predictors of social network fragmentation than other social indicators of segregation. Finally, we deploy the urban indices as instrumental variables for social network fragmentation in a regression model predicting economic inequality. Model statistics show that our urban topology indicators have a significant relationship with economic inequality via their relationship with social network fragmentation.

2 Results

In this paper we analyze a Hungarian online social network and find that the fragmentation of social networks is significantly higher in towns in which residential neighborhoods are divided by physical barriers such as rivers and railroads and are relatively distant from the center of town. Our empirical analysis has established that the fragmentation of social network structure is positively associated with income inequality in cities and towns. Moreover, we have found that the relationship is dynamic - the interaction of fragmentation and existing inequalities predicts a significant growth in inequality in the future. The physical arrangement of a city's residential areas, the loci of social interactions, is also connected to social network fragmentation. We observe a tendency: if the urban fabric contains significant distances, physical barriers, or spatially concentrated amenities, social networks tend to be more fragmented. The relationship between geographic division and inequality manifests in this fragmentation.

Figure 1E presents the relationship between social network fragmentation and the change of town income inequality between 2011 and 2016 and plots β by levels of income inequality. We find that the interaction between inequality in 2011 and fragmentation has a positive and statistically significant relationship with inequality in 2016. However, the marginal effect of fragmentation informs us that social network fragmentation is positively related to future levels of income inequality only if the initial levels of inequalities are high. This result provides empirical support to the theory that social networks can increase inequalities when individuals sort based on their initial endowments [2]. Our analysis suggests how and why urban planning can be an effective tool to moderate inequalities in the long-run.

While it has long been known that segregation is often an implicit goal of urban planning, for example Detroit's Eight Mile Wall or the barrier that long separated suburban New Haven from Hamden in Connecticut [7] or the phenomenon of gated communities [8], our work suggests that even innocent design choices can lead to bad outcomes. Conversely, certain policies may facilitate mixing and block the compounding of inequality by fragmentation.

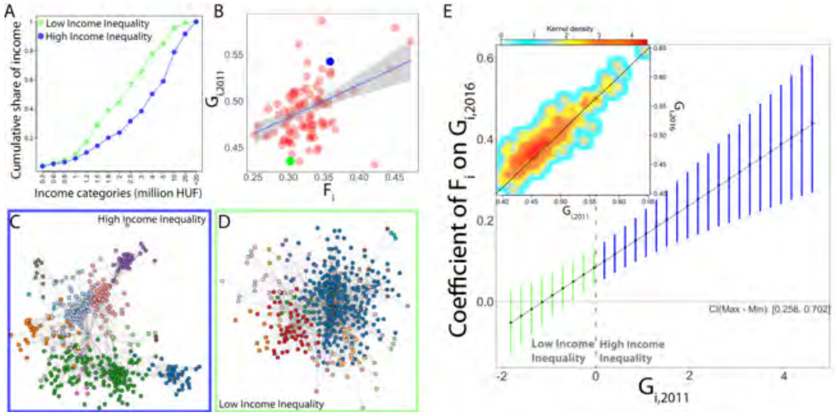


Fig. 1. Income inequality (G_i) correlates with network fragmentation (F_i) in towns. (A) Cumulative distribution of income in a relatively equal town (Ajka, green line) and a relatively unequal one (Gödöllő, blue line). **(B)** Income inequality measured by the Gini index ($G_{i,2011}$) for towns larger than 15,000 population correlates with the fragmentation (F_i) of social network within the town (Pearson's ρ = 0.44). Green dot: Ajka; blue dot: Gödöllő; red dots: all other towns. **(C)** The filtered social network structure in Gödöllő, the sample town that has high income inequality ($G_{i,2011} = 0.54, F_i = 0.36$). Node colors represent communities. **(D)** The filtered social network structure in Ajka, the sample town that has low income inequality ($G_{i,2011} = 0.43, F_i = 0.3$). Node colors represent communities. **(E)** Network fragmentation (F_i) intensifies income inequality stronger in those towns where initial inequality is high. **Inset**: We plot the correlation between town Gini scores in 2011 and 2016 ($G_{i,2011}$ and $G_{i,2016}$.)

References

1. Mark Granovetter. Economic action and social structure: The problem of embeddedness. *American journal of sociology*, 91(3):481–510, 1985.
2. Paul DiMaggio and Filiz Garip. Network effects and social inequality. *Annual review of sociology*, 38:93–118, 2012.
3. Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.
4. Ginestra Bianconi. Triadic closure as a basic generating mechanism of communities in complex networks. *Physical Review E*, 90(4), 2014.
5. Gueorgi Kossinets and Duncan J Watts. Origins of homophily in an evolving social network. *American journal of sociology*, 115(2):405–450, 2009.
6. Christoph Stadtfeld. The micro–macro link in social networks. *Emerging Trends in the Social and Behavioral Sciences*, pages 1–15, 2018.
7. Sarah B Schindler. Architectural exclusion: discrimination and segregation through physical design of the built environment. *Yale LJ*, 124:1934, 2014.
8. Setha M Low. The edge and the center: Gated communities and the discourse of urban fear. *American anthropologist*, 103(1):45–58, 2001.

Global Network of Hidden Military Support: its Structure and Evolution

Weiran Cai[1], Belgin San-Akca[2], Jordan Snyder[13], Grayson Gordon[1], Zeev Maoz[4],Raissa M. D'Souza[15]

[1] Department of Computer Science, UC Davis, Davis, CA 95616
[2] Department of International Relations, Koç University, Istanbul 34450, Turkey
[3] Department of Mathematics, UC Davis
[4] Department of Political Sciences, UC Davis
[5] Sante Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501

1 Introduction

Military confrontation is one of the key factors that has shaped human history. While warfare in the post-World War era have decreased, internal wars and low-intensity conflicts carried out by nonstate armed groups (NAGs) against nation states have become increasingly common [1, 2]. NAGs can include rebel, insurgent, guerrilla, or terrorist groups that engage in violent activity targeting the government, citizens, or institutions of nation-states. Unlike the warfare that has received extensive analyses, the hidden relations of NAGs and their supporting host states (HSs) are yet to be revealed. We study the bipartite network of NAGs and the HSs and its evolution over a large timespan by employing the latest pattern detection algorithm. We discover that the network has experienced a dynamical process that involves both highly biased attachment and detachment. Analogous to what is typically observed in ecological mutualistic and parasitic relations, a peculiar architecture showing both nested and modular patterns is persistently shown in a large portion of the timespan. Quantifying the supporting network provides an important perspective towards the comprehension of the origin, evolution and termination of this global hidden power.

2 Assembly and Disassembly

We focus on the relations between NAGs and external state supporters, extracted from the Dangerous Companions Database covering the time period between 1946 and 2010 [1]. We consider only violent NAGs (whose operations resulted in 25 or more deaths per year of operation). A HS is a country that provides any type of material support to a given NAG, including safe haven, training, or military or financial aid.

We examine the assembling and disassembling processes that unfold on it and find that both attachment and detachment of actors and links occur preferably at incumbent counterpart actors of higher degrees (approximately its fitness). While the positive dependence in the attachment can be understood as a generalized preferential attachment (PA), the generalized preferential detachment (PD) occurring at the leave of the actors from the network is somewhat counter-intuitive, which suggests that the network tends to disassemble at higher-fitness nodes.

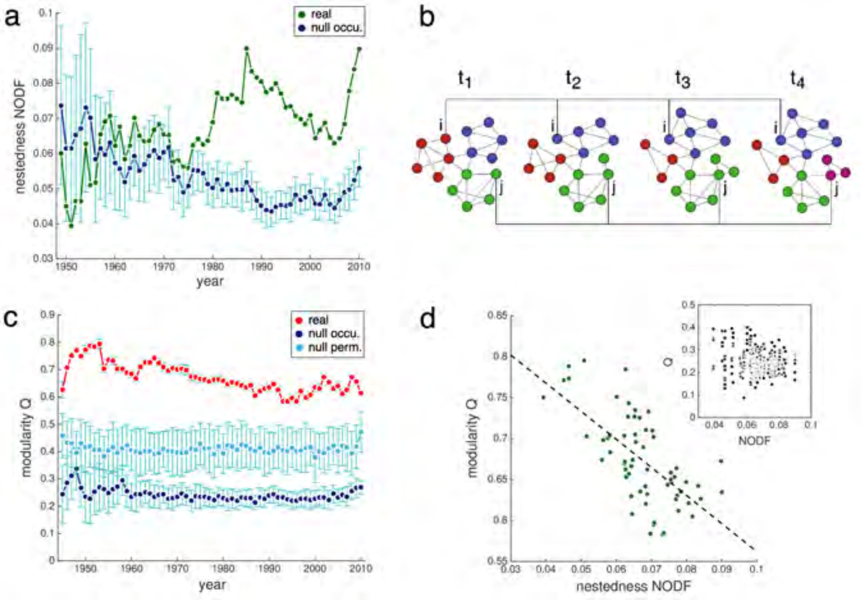


Fig. 1. Evolving pattern of NAG-HS network. **a, c** Temporal nestedness $NODF(t)$ and modularity $Q(t)$. **b** Dynamical module detection algorithm. **d** Negative correlation between $NODF$ and Q.

3 Peculiar Structure and its Evolution

Despite significant variations in the number of actors and links, this bipartite network exhibits a robust nested and modular structure over a large timespan (Fig. 1a and 1c). These patterns have been commonly observed in ecological plant-pollinator networks and in designer-contractor partnerships in the New York garment industry. Ecological organizing principles, originally discovered in natural systems, turn out to be shared by this military network [3]. The roles of the actors can then be determined based on the discovered network structure. The significantly high nestedness (from 1975 on) implies that a relatively more specialist-like node would interact predominantly with proper subsets of the partners of the more generalist-like nodes. Heuristically, the formation of the nested architecture follows the principle of cumulative advantages.

The network also shows a nontrivial modular structure, in which actors are more densely connected to partners within the same module than across modules. For this temporal network, we use a dynamical module detection algorithm based on the Markov stability, proposed by Mucha, et al [4] (Fig. 1b), which has the merit of allowing to identify the same module in consecutive time slices and thus to track the evolution of modules. The detected modularity value $Q(t)$ is significantly higher compared to that of the null models. Based on a representative partition, the temporal network over the 66 years is partitioned into 33 modules, including 9 major modules and 24 modules comprised of separate nodes or tiny groups (Fig. 2). The contrast is sharp: while the major modules can last for significantly long timespans, the rest ones are merely temporary. While the 24 temporary modules are almost constant in their membership, all the major

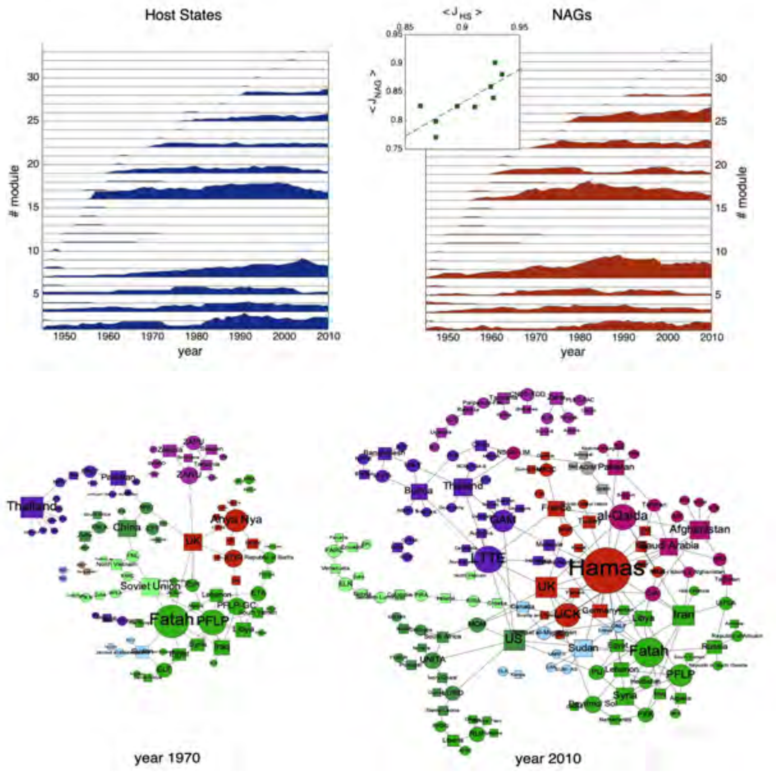


Fig. 2. Temporal modular structure. **a** Modules in HS and NAG guilds. **b** Partitioned bipartite networks for representative years 1970 and 2000.

9 modules have experienced large fluctuations in the content. This is consistent with the general tendency found in the unipartite coauthorship and the phone-call networks [5]: small modules survive when they are stationary in the membership, while large modules survive if they experience sufficient changes. In all, quantifying the state-based supporting network provides a unique and important perspective towards the comprehension of the origin, evolution and termination of NAGs, as the global hidden power.

References

1. San-Akca, B. States in Disguise: Causes of State Support for Rebel (Oxford, 2016).
2. Maoz, Z. and San-Akca, B. Rivalry and state support of non-state armed groups (NAGs), 1946–2001. Int. Stud. Q. 56, 720–734 (2012).
3. Olesen, J. M. and Bascompte, J. and Dupont, Y. L. and Jordano, P: The modularity of pollination networks. Proc. Natl Acad. Sci. USA 104, 19891–19896 (2007)
4. Mucha, P., Richardson, T., Macon, K., Porter, M. and Onnela, J. Community structure in time-dependent, multiscale, and multiplex networks. Science 328, 876–878 (2010).
5. Palla, G., Barabási, A. and Vicsek, T. Quantifying social group evolution. Nature 446, 664–667 (2007).

Detecting People Interested in Non-Suicidal Self-Injury on Social Media

Zaihan Yang and Dmitry Zinoviev

Suffolk University, Boston, MA 02108, USA
zyang13@suffolk.edu, dzinoviev@suffolk.edu

1 Introduction

Non-Suicidal Self-Injury (NSSI) is the intentional destruction of body tissue without the intent to commit suicide [1]. It is particularly prevalent among adolescents and young adults as a means of emotional control and release. Typical NSSI activities include skin cutting, banging or hitting oneself, and burns.

Recent prevalence estimates suggest that 14%–21% of adolescents and 17%–25% of young adults have engaged in NSSI at some point in their lives. NSSI is repeatedly found to be associated with significant emotional and behavioral dysfunction (such as eating disorders and suicide). This relationship highlights the urgency of providing early detection of people with NSSI engagement and prevention of their behaviors.

However, global provisions and services for detecting, supporting, and treating NSSI people have long been insufficient. There is no reliable laboratory test for diagnosing NSSI. Diagnostic largely depends on patients' self-reports or observations reported by relatives or friends. Yet, NSSI people often conceal their practices, which prevents detecting their engagement.

Early research work on NSSI people detection was primarily conducted within psychology, psychiatry, and medicine domains [1–3]. With the proliferation of social media, people are increasingly using online platforms to share their thoughts and opinions. Postings on these sites are made in natural settings and provide a means for capturing people's real thoughts, opinions, and moods. Researchers from Computer and Data Science fields have started to explore social media content to study people with NSSI engagement, their interests [4], the influence of social media on their behaviors [6], and their posted images [5]. However, to the best of our knowledge, no work has been done to provide an automatic learning system that can detect people with NSSI engagement.

We treat the detection of people interested in NSSI as a binary classification problem. We have collected data from LiveJournal.com, a social-blogging networking platform, and built Naïve Bayes and Logistic Regression classifiers based on the features extracted from users' self-declared interests. Experimental evaluation demonstrates that we can achieve 73% accuracy, 77% precision, 67% recall, and 71% F_1 score to detect people interested in NSSI and identify the most discriminating features.

2 Model

We have collected our data by crawling user profiles (demographics, self-declared interests, and friendship relationships) and NSSI-related thematic community profiles (membership, posts, and comments).

We assume that any LiveJournal user who is a member of one of the 139 manually selected NSSI-related thematic communities or contributes to such community by posting or commenting, is interested in NSSI. We designate such users as "harmers" for brevity, acknowledging that some may not practice self-injury. Following the harmers' friendship network, we further collect some of their immediate friends and friends-of-friends (the "non-harmers"), chosen randomly to match the size and age of the harmers.

The dataset has 11,972 harmers, 12,600 friends, 11,672 friends-of-friends, and 1,264 distinct self-declared single- or multi-word interests (on average, 26.2 interests per user). These self-declared interests serve as virtual profiles for each user. We regard each interest as a feature and represent the feature vectors in the following three ways:

Simple-Count-Vector: The feature value of each interest is represented by their occurrence frequency in user profiles.

TF-IDF-Vector: Each feature is represented by its TF-IDF value.

Topic-Distribution-Vector: We apply the Latent Dirichlet Allocation (LDA) model to learn a topic distribution over each user profile. We use the distribution as the feature vector. We choose the number of topics to be 10.

3 Results

We treat the harmers as positive samples and the non-harmers as negative samples. Since the original data set is unbalanced (with 1/3 positive and 2/3 negative samples), we pick 11,972 negative samples uniformly at random to form a balanced data set, and only consider those interests that are declared by more than 100 but fewer than 16,760 (70% of all users in the balanced data set) users to construct the three feature vectors mentioned above. We repeat the random sampling process five times and apply 5-fold cross-validation to evaluate the model's performance. Table 1 shows the classification results evaluated by accuracy, precision, recall, and F_1 score using the three feature vector representations and Naïve Bayes (NB) and Logistic Regression (LR) classifiers.

Table 1: Classification results
(Highest value of each metric is highlighted in bold)

Feature Vector (Classifier)	Accuracy	Precision	Recall	F_1 Score
Simple-Count (NB)	0.70	0.76	0.59	0.66
TF-IDF (NB)	0.71	0.74	0.64	0.68
LDA-Topic-Distribution (NB)	0.65	0.67	0.60	0.63
Simple-Count (LR)	0.72	**0.77**	0.62	0.69
TF-IDF (LR)	**0.73**	0.75	**0.67**	**0.71**
LDA-Topic-Distribution (LR)	0.67	0.72	0.56	0.63

We achieved up to 73% accuracy, 77% precision, 67% recall, and 71% F_1 score. For both Naïve Bayes and Logistic Regression classifiers, using TF-IDF feature vectors increases recall and F_1 score but slightly decreases precision as compared to Simple-Count vectors. Since our project aims to detect concealed behavior, higher recall and false positives are more desirable than a higher precision and false negatives.

We compute and sort each feature by its odds-ratio value learned from Naïve Bayes and report the top 20 positive and negative features. Table 2 shows the top 20 positive interests prevalent among the harmers (such as cutting, burning, self-injury, and self-harm) and the bottom 20 interests that the harmers tend to neglect (such as programming, linux, architecture, and chess). The harmers' interests are undoubtedly specific to NSSI, related disorders, and mental health in general. The non-harmers interests seem to be aligned with an adolescent/young adult's interests in the early XXIst century.

Table 2: Discriminating features for Simple-Count and TF-IDF

	Top positive	Top negative
Simple-Count	abuse, ednos, ocd	robots, technology, they might be giants, tom waits, firefly
TF-IDF	mental health, scars, mental illness	smallville, travel, naruto, american eagle, torchwood
Both	anxiety, pills, bulmima, suicide, bleeding, hurt, borderline personality disorder, razorblades, bipolar, cut, razors, burning, cutting, self mutilation, self injury, self harm, si	programming, battlestar galactica, linux, d&d, fanfic, farscape, science fiction, architecture, fishing, animation, tolkien, gaming, chess, rpg, doctor who

Summary. In this paper, we propose a supervised learning approach to detect people interested in NSSI. Experimental evaluation on a real-world dataset—the LiveJournal social blogging networking platform—demonstrates our proposed model's effectiveness. For future work, in addition to the pure content-based features, we would like to integrate network, demographic, sentimental, contextual, and other features. We consider building and comparing classifiers based on different algorithms (including logistic regression, random forest, and artificial neural networks). We plan to apply our approach to other social media platforms such as Twitter or Facebook.

References

1. American Psychastic Association: Diagnostic and Statistical Manual of Mental Disorders (5th ed.). American Psychiatric Association, 2013
2. Favazza, A.R.: Bodies under Siege: Self-mutilation, Nonsuicidal Self-Injury, and Body Modification in Culture and Psychiatry (3rd ed.). Johns Hopkins University Press, 2011.
3. Andover, M.: Non-Suicidal Self-Injury Disorder in a Community Sample of Adults. Psychiatry Research. 219, 2014
4. Zinoviev, D., Stefanescu, D., Fireman, G., Swenson, L.: Semantic Networks of Interests in Online Non-Suicidal Self-Injury Communities. Digital Health. Vol.2, Page 1–14, 2016.
5. Xian, L., Vickers, S.D., Giordano, A.L., Lee, J., Kim, I.K., Ramaswamy, L.: #Selfharm on Instagram: Quantitative Analysis and Classification of Non-Suicidal Self-Injury. International Conference on Cognitive Machine Intelligence (CogMI). Page 61–70, 2019.
6. Diane, K., Hawton, K., Singaravelu, V., Stewart, A., Simkin, S., Montgomery, P.: The Power of the Web: A Systematic Review of Studies of the Influence of the Internet on Self-Harm and Suicide in Young People. PLOS. Volume 8, Issue 10, 2013.

Part IX

Network Analysis

Highly comparative graph analysis

Robert L. Peach[1], Alexis Arnaudon[1,2], Henry Palasciano[1], and Mauricio Barahona[1]

[1] Department of Mathematics, Imperial College London, London, UK,
[2] Blue Brain Project, EPFL, Geneva, Switzerland,

Graphs offer a natural and powerful mathematical framework to represent complex systems across all scientific domains, from molecular interactions to astronomical dynamics. The wide applicability of graphs across many scientific domains has driven the development of various network analysis techniques purposed for investigating and revealing properties and structure of networks. Each method provides a window into the workings of complex systems, each providing a new perspective on the structure, and each with differing applicability. It is exactly this large corpus of graph theoretic research and the differing applicabilities of methodologies that can make it difficult to identify the best tool to generate insights into complex systems. Despite the vast quantity of graph theoretical methodologies and algorithms, there currently exists no extensive organisation of graph analysis methods for the purpose of highly comparative analysis.

Here, we introduce Highly Comparative Graph Analysis (HCGA), a modular Python software package which allows researchers to perform massive graph feature extraction and statistical analysis of their system [1] [3]. Given a graph dataset, HCGA transforms each graph to a set of a few thousand features, depending on the graphs, that each encodes a different interpretable network property. The resulting feature matrix facilitates data-driven, statistically controlled analysis of the dataset, removing the otherwise time-consuming and subjective task of implementing individual graph analysis methods.

Our computational framework illustrated in Figure 1 begins with a set of complex systems modelled as a set of graph structures, representing for example, molecules, proteins, transportation routes, neurons or socials networks. Subsequently, HCGA efficiently computes many theoretic features for each individual graph and combines the resulting feature vectors into a matrix that describes the complete set of graphs. Methods for graph analysis can take a variety of forms, from simple summations to complex convolutions. In HCGA, we have implemented each such method as an algorithm: an operation that summarizes an input graph with a single real number. The real-valued summary values are collected into a feature vector representation of each graph in a collection of graphs and results in a feature matrix. Given the computational inefficiency of some network based algorithms, we have implemented time-outs and provided flags for users to compute features that meet a particular speed criteria HCGA.

For analysis, HCGA includes a suite of statistical tools for classification, regression or unsupervised learning. Crucially, we include an in-depth feature importance analysis using Shapley Additive Values (SHAP) to drive scientific insights [2].

To validate our computational approach, we first test HCGA on benchmark datasets for graph classification, achieving state-of-the-art performance against all existing methods including deep-learning approaches. The complete set of classification results are

[3] https://github.com/barahona-research-group/hcga

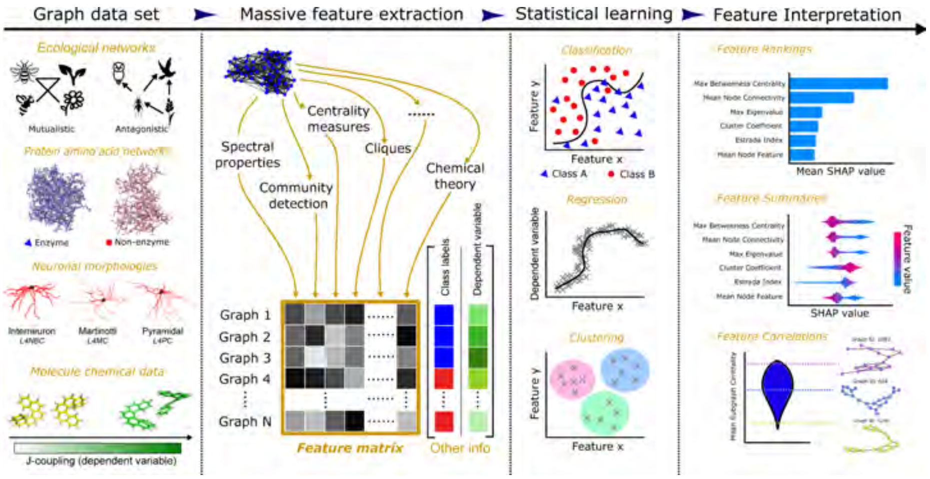


Fig. 1. The main workflow of HCGA. HCGA is compatible with binary classification, multi-class classification, regression or non-labelled data. Firstly, the user inputs a set of complex structures that are represented as graphs (the graph construction process is not included in HCGA since it is domain dependent). Subsequently, a massive collection of features ($n_f > 2000$) are computed for each graph and collected in a feature matrix. A suite of statistical learning algorithms are used to, for example, learn the optimal hyper-plane separating classes or identify the line of best fit for regression. The class label or dependent variable of new graphs can be predicted and the most important features can be identified and interpreted.

detailed in Tables 1 and 2 against popular deep-learning methodologies and Kernel algorithms using the results reported under Fair Comparison [3].

However, the power of HCGA lies in the scientific insights that can be derived from the importance of individual or groups of features. To demonstrate its versatility, we have used two novel datasets; (1) a neuronal morphologies dataset and (2) a crystal structure prediction dataset. In both datasets, we found that novel features can be derived that can be scientifically interpreted (results not shown here) [1].

In addition to the benchmark datasets and two phenotyping case studies, HCGA has general utility, including applications to functional or structural connectivity networks in brains, revealing properties of computer networks that make them weak to outside attacks and the differences in structure between varying ecological networks. Overall, HCGA is an intuitive tool to analyse any dataset that can be expressed as a collection of graphs, and is highly modulable with the option to restrict the computed features or add new ones.

References

1. R. L. Peach, A. Arnaudon, J. A. Schmidt, H. Palasciano, N. R. Bernier, K. Jelfs, S. Yaliraki, and M. Barahona, “hcga: Highly comparative graph analysis for network phenotyping,” *bioRxiv*, 2020.

2. S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, “From local explanations to global understanding with explainable ai for trees,” *Nature machine intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.
3. F. Errica, M. Podda, D. Bacciu, and A. Micheli, “A fair comparison of graph neural networks for graph classification,” *arXiv preprint arXiv:1912.09893*, 2019.

Method	Data Sets				
	ENZYMES	PROTEINS	D&D	NCI1	MUTAG
Multi-Hop	56.1 ± 9.1	76.7 ± 2.9		77.3 ± 1.7	89.8 ± 5.58
DGCNN	38.9 ± 5.7	72.9 ± 3.5	76.6 ± 4.3	76.4 ± 1.7	-
GIN	59.6 ± 4.5	73.3 ± 4.0	75.3 ± 2.9	**80.0 ± 1.4**	-
ECC	29.5 ± 8.2	72.3 ± 3.4	72.6 ± 4.1	76.2 ± 1.4	-
DiffPool	59.5 ± 5.6	73.7 ± 3.5	75.0 ± 3.5	76.9 ± 1.9	-
HCGA	**74.5 ± 4.5**	**78.4 ± 3.6**	**81.9 ± 3.5**	78.5 ± 2.4	**89.9 ± 6.1**

Table 1. Results on chemical benchmark datasets.

Method	Data Sets					
	COLLAB	IMDB-B	IMDB-M	REDDIT-B	REDDIT-5K	REDDIT-12K
Multi-Hop (RF)	78.2 ± 1.5	71.6 ± 4.4	45.2 ± 3.5	88.9 ± 2.2	51.3 ± 1.9	43.5 ± 1.0
DGCNN	71.2 ± 1.9	69.2 ± 3.0	45.6 ± 3.4	87.8 ± 2.5	49.2 ± 1.2	-
GIN	75.6 ± 2.3	71.2 ± 3.9	48.5 ± 3.3	89.9 ± 1.9	56.1 ± 1.7	-
ECC	-	67.7 ± 2.8	43.5 ± 3.1	-	-	-
DiffPool	68.9 ± 2.0	68.4 ± 3.3	45.6 ± 3.4	89.1 ± 1.6	53.8 ± 1.4	-
HCGA	**82.2 ± 2.3**	**74.4 ± 4.4**	**49.6 ± 3.4**	**93.5 ± 1.9**	**57.9 ± 1.5**	**49.6 ± 0.8**

Table 2. Results on social benchmark datasets.

Better weighted clustering coefficient: now continuous

Tanguy Fardet[1,2] and Anna Levina[1,2]

[1] University of Tübingen
[2] Max Planck Institute for Biological Cybernetics

1 Introduction

Network theory has provided an invaluable framework to study complex systems over the past decades, enabling analyses of interacting elements in various fields ranging from urban planning to genetics to social sciences and neuroscience.

Despite recent advances, many analyses in applied fields are carried on symmetrized adjacency matrices, ignoring orientation and weights. This choice simplifies procedures but leaves out crucial information regarding the influence of different edges on the network's dynamics. For networks inferred from data or direct measurements, noise and statistical biases can generate weak spurious connections that cannot easily be unambiguously separated from the original ones. We therefore propose a new *continuity* requirement: network measures should be a continuous function of the weights at zero. There should thus be no significant difference between the absence of an edge and its presence with an infinitesimal weight.

Nodes in real-world networks cluster into densely connected groups. This property is captured by the clustering coefficient, which was initially defined for unweighted symmetric networks. Several notable generalizations of the clustering coefficient have been proposed for weighted networks [1–3]. However, all of them violate the *continuity* condition. We propose a new consistent definition of the clustering coefficient that tackles this issue and satisfies previously formulated conditions [5].

2 A new definition for weighted clustering

Let W be a normalized weight matrix, $W^{[\alpha]} = \{w_{ij}^{\alpha}\}$ and $s_i^{[\alpha]} = \sum_j w_{ij}^{\alpha}$. For every node i the continuous local clustering coefficient is defined as:

$$C_i = \frac{\left(\left(W^{\left[\frac{2}{3}\right]}\right)^3\right)_{ii}}{\left(s_i^{\left[\frac{1}{2}\right]}\right)^2 - s_i}, \qquad (1)$$

To study more specific clustering motifs, we also adapted this definition for fan-in, fan-out, middleman, and circular cases as described in [4].

The new clustering coefficient fulfills the following conditions:

normalization: $C \in [0,1]$,

consistency: equivalent to binary definitions if all weights are equal,

continuity: zero-weight edges are equivalent to no-edge cases

linearity: if all weights in the neighborhood are scaled by a factor α, so is C.

3 Properties of our definition compared to previous proposals

Among established definitions, proposals from the teams of Barrat [1], Onnela [2], and Zhang [3] fulfill *normalization* and *consistency* conditions. They define triangle and triple intensities differently [5] and may be more or less suitable for specific problems, though Onnela and Barrat seem to be used most often. However, none of them satisfy the *continuity* condition, and Barrat also does not fulfill the *linearity* requirement.

The absence of *continuity* leads to edge-cases that were noted earlier in [5]. Unfortunately, the influence of these edge-cases does not disappear in larger networks, where the absence of *continuity* for Onnela and Barrat prevents them from capturing various structural properties in weighted networks. For example, in a weighted core-periphery graph, Barrat's method fails to capture any structure, while Onnela's does not distinguish proximal nodes from the weakly-connected periphery — Figure 1.A. However the continuous clustering differentiates all three types of nodes in the network.

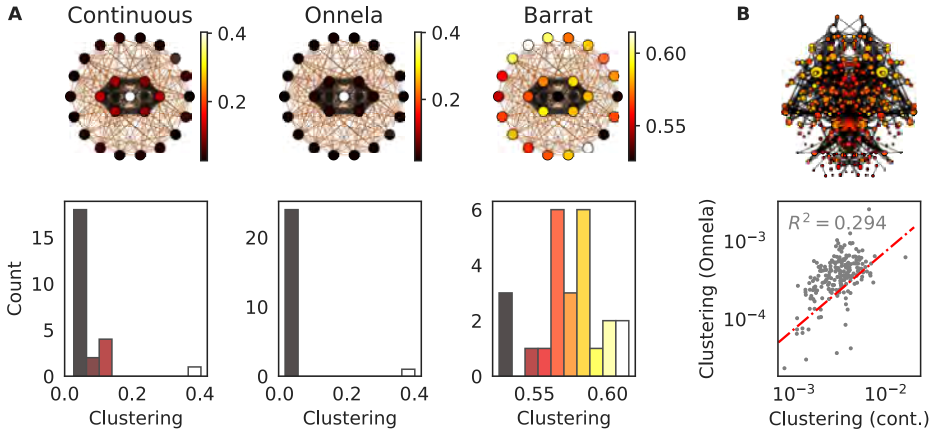


Fig. 1. Continuous clustering coefficient captures essential properties of networks with broad weight distribution. **A**. Clustering coefficients in a core-periphery network with 7 strongly connected core nodes (black edges) that interact with densely but weakly connected periphery nodes (light-brown edges). Top row: graphical view of the network; edge width gives the strength of the connection, node color indicates its clustering coefficient. Bottom row shows the distribution of clustering coefficients using the same colormap as above. **B**. Clustering of mouse brain areas. Top: continuous clustering is highest (yellow to white nodes) in regions of different strengths (size). Bottom: even the closest methods (continuous and Onnela) result in clustering coefficients that are only weakly correlated.

The absence of *continuity* in previous definitions becomes especially problematic in the case of networks inferred from data. For such graphs, the actual sparsely connected network may be drowned in a sea of weaker connections that arise from noisy or partial measurements, leading to imperfect inference. We investigate the robustness of the new definition to spurious, low-weight edges in an inferred network. To this end, we corrupt a Watts-Strogatz network with random, small-weights connections — Figure 2.A. The

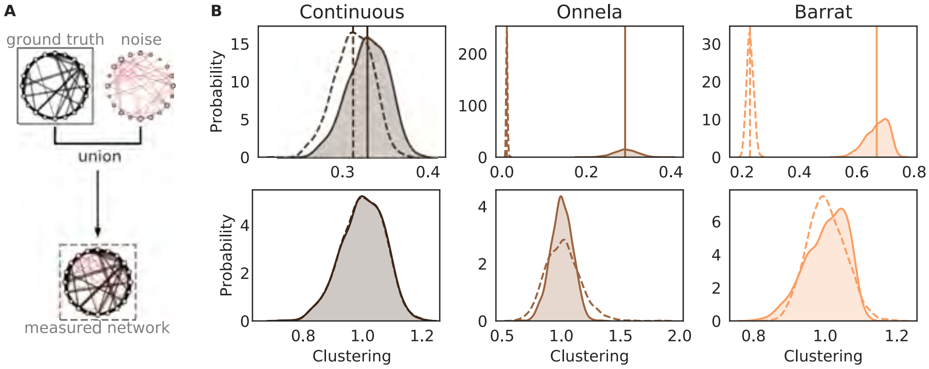


Fig. 2. Continuous clustering is robust to noise. **A.** A 1000-node Watts-Strogatz (WS, "ground truth") network is corrupted by spurious small-weight connections (noise) to generate the "measured network" . **B.** Top row: the clustering distribution of the original WS graph (full lines, shaded) and that of the "measured network" (dashed). Respective global clustering coefficients C_g are marked by the full and dashed vertical lines. Bottom row: the distributions of the normalized clustering C/C_g (shading and lines are the same as for the top raw).

continuous clustering coefficient is only mildly affected by the noise, and normalization with the total clustering coefficient recovers the correct distribution — Figure 2.B. On the other hand, Onnela and Barrat definitions lead to distributions that deviate strongly from the ground truth and cannot be recovered by normalization.

Finally, our new clustering coefficient gives notably different results on real-world networks — Figure 1.B. High-clustering nodes identified with continuous clustering in brain networks form strongly connected subnetworks that contain the strongest edges and display a higher number of reciprocal connections than groups of high clustering nodes obtained via other methods. Despite the absence of ground truth knowledge for mouse brain areas (data from [6]) and other real-world networks, the significant discrepancy between methods and substantial impact of noise on previously used method call for further investigation. We will discuss them in details in our presentation.

References

1. Barrat, A., Barthelemy, M., Pastor-Satorras, R. and Vespignani, A.: The Architecture of Complex Weighted Networks. PNAS 101, 3747–3752 (2004).
2. Onnela, J.-P., Saramäki, J., Kertész, J. and Kaski, K.: Intensity and Coherence of Motifs in Weighted Complex Networks. Phys. Rev. E 71, 065103 (2005).
3. Zhang, B. and Horvath, S.: A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology 4 (2005).
4. Fagiolo, G.: Clustering in Complex Directed Networks. Phys. Rev. E 76, 026107 (2007).
5. Saramäki, J., Kivelä, M., Onnela, J.-P., Kaski, K. and Kertész, J.: Generalizations of the Clustering Coefficient to Weighted Complex Networks. Phys. Rev. E 75, 027105 (2007).
6. Oh, S. W. *et al.*: A mesoscale connectome of the mouse brain. Nature 508, 207–214 (2014).

Query Oriented Temporal Active Community Search

Badhan Chandra Das[1] and Md Musfique Anwar[1] Md. Al-Amin Bhuiyan[2]

[1] Jahangirnagar University, Savar, Dhaka, Bangladesh
badhan0951@gmail.com, manwar@juniv.edu
[2] King Faisal University, Saudi Arabia
mb-huiyan@kfu.edu.sa

1 Introduction and Problem Statement

A considerable amount of research has been devoted towards the problem of community detection in online social networks (OSNs). More recently, a related but different problem is community search (aka local community) where the main objective is to ascertain the best potential meaningful community that contains the query node(s) and query attributes [1]. Huang et al. [2] presented a community search model for mining a community comprising multiple query nodes based on k-trusses. However, most of the existing investigations ignore the *topical activeness* of the community members and thus fail to determine the *active* communities for a given query. Moreover, these cited approaches did not contemplate how the users' interests for the given query attributes changes with time.

We propose a *temporal activity* biased weight model which gives higher weight to users' recent activities and also develop a framework of activity driven temporal active communities (ATAC) to search effective community for a given input query Q to find densely-connected community in which community members are temporally similar in terms of their activities related to the query attributes. An active online local community is considered as a connected induced subgraph in which each node has a degree of at least k(k-core) which indicates the structure cohesiveness of the desired community.

Activity: Each user u_i performs actions (e.g. posting tweets in Twitter, publishing research papers in coauthor network) known as activities at different time points (t_j) containing set of attributes ψ_{u_i}. An activity tuple $\langle u_i, \psi_{u_i}, t_j \rangle$ is used to represent an action.

Time-Based Forgetting Factor: Not all of a user's past activities are equally important and that the user's most recent activities can imply the most about her interests. A logarithmic time-decay function expressed in Equation 1 to assign lower importance (denoted as μ) to older activities, since they are less probable of corresponding to the user's recent interests.

$$\mu_{\langle u_i, \psi_{u_i}, t_j \rangle} = \frac{1}{1 + log_b(age_{\langle u_i, \psi_{u_i}, t_j \rangle} + 1)} \tag{1}$$

The base of the logarithm in Eq. 1, denoted by b, controls the speed of decay and $age_{\langle u_i, \psi_{u_i}, t_j \rangle}$ as the amount of time elapsed since it happened.

Activeness Score: The activeness score (denoted by σ) for each candidate community member $u_i \in U^Q$ is computed using Equations 3, where $\psi_{u_i} \in \mathscr{A}_q$ ($\mathscr{A}_q$ is set of query attributes within the input query $Q = (u_q, \mathscr{A}_q)$, u_q is a query node). This investigation deliberates two factors that are closely associated with the distinct activeness of a user

u_i. The first factor $f_1(u_i, \psi_{u_i})$ specifies the probability that u_i performs an activity related to Q.

$$f_1(u_i, \psi_{u_i}) = \frac{\sum \mu_{\langle u_i, \psi_{u_i}, t_j \rangle} \times |ACTS(u_i, \psi_{u_i})|}{|ACTS(u_i, *)|} \quad , \quad f_2(u_i, \psi_{u_i}) = \frac{\sum \mu_{\langle u_i, \psi_{u_i}, t_j \rangle} \times |ACTS(u_i, \psi_{u_i})|}{\sum_{u_z \in U^Q} |ACTS(u_z, \psi_{u_z})|} \tag{2}$$

where, $ACTS(u_i, \psi_{u_i})$ represents the set of activities comprising the set of attributes $\psi_{u_i} \subseteq \mathscr{A}_q$ performed by u_i and $ACTS(u_i, *)$ denotes the set of all the activities containing any attribute(s) performed by user u_i. The second factor $f_2(u_i, \psi_{u_i})$ designates the participation of user u_i compared to the total number of activities related to Q performed by U^Q. Then, the activeness (denoted as σ) of u_i related to Q is:

$$\lambda_{(u_i, \psi_{u_i})} = f_1(u_i, \psi_{u_i}) \times f_2(u_i, \psi_{u_i}) \quad \text{and} \quad \sigma_{(u_i, \psi_{u_i})} = \frac{\lambda_{(u_i, \psi_{u_i})}}{max_{u_z \in U^Q}\{\lambda_{(u_z, \psi_{u_z})}\}} \tag{3}$$

Problem Definition: Given an attributed graph *G = (U, E, $\mathscr{A}$)*, where *E* denotes set of edges, an input query $Q = \{u_q, \mathscr{A}_q\}$, two positive integers *h* (hop) and *k*, an attributed active local community $\mathscr{C}_q$ is an induced subgraph that meets the following constraints.

1. ***Connectivity.*** $\mathscr{C}_q \subset G$ is connected;
2. ***Structure cohesiveness.*** $\forall u \in \mathscr{C}_q, deg_{\mathscr{C}_q}(u) \geq k$;
3. ***Query cohesiveness.*** $\forall u \in \mathscr{C}_q$, activeness score of a user *u* is $\sigma_{(u,Q)} \geq \theta_a$ and $\theta_a \in [0, 1]$ is a threshold.

2 Experimental Evaluation

We conduct our experiment on a Twitter dataset named CRAWL where the input query set as {*social media*, *politics*, *entertainment*, *sports*}, an academic coauthor network dataset (DBLP) where the input query is set as {*data mining*, *natural language processing (NLP)*, *social network analysis (SNA)*, *machine learning*} and {*nature*, *festival*, *architecture*, *portrait*} set as input query for a flickr dataset.

We compare our proposed ATAC framework with two other methods. We select ACQ method, proposed by Yang et al. [1], for community search over attributed graphs based on *k*-cores. The key distinction with our work is that ACQ doesn't consider users' topical activeness as well as ignore the prospective temporality of users' interests. Finally, we consider a baseline solution (*k*-core) which forms communities based on only *k*-core i.e. focusing only the structural cohesiveness. We vary the length of query attributes $|\mathscr{A}_q|$ to $|\mathscr{A}_q|$= 1, 2, 3, 4 and use the following two quality evaluation metrics:

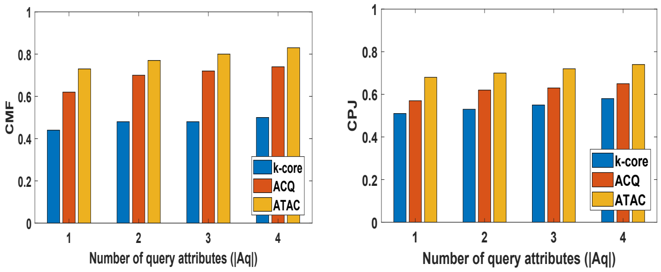
Community Member Frequency (CMF): The CMF measures the number of occurrences of query attributes in $\mathscr{C}_i$ to determine the degree of cohesiveness. Let $n_{i,p}$ be the number of nodes of $\mathscr{C}_i$ whose attribute sets contain the *p-th* attribute of $\mathscr{A}_q$. Then, $\frac{n_{i,p}}{|\mathscr{C} i|}$ is the relative occurrence frequency of this attribute in $\mathscr{C}_i$. The CMF is the average of this value in overall attributes in $\mathscr{A}_q$, and all communities in $N(\mathscr{C}_q)$:

$$CMF(N(\mathscr{C}_q)) = \frac{1}{\mathscr{L} \times |\mathscr{A}_q|} \sum_{i=1}^{\mathscr{L}} \sum_{p=1}^{|\mathscr{A}_q|} \frac{n_{i,p}}{|\mathscr{C} i|} \tag{4}$$

Community pairwise Jaccard (CPJ): Let $\mathscr{C}_{i,j}$ be the *j-th* node of $\mathscr{C}_i$. The CPJ is then the average similarity overall pairs of nodes of $\mathscr{C}_i$, and all communities of $n(\mathscr{C}_q)$:

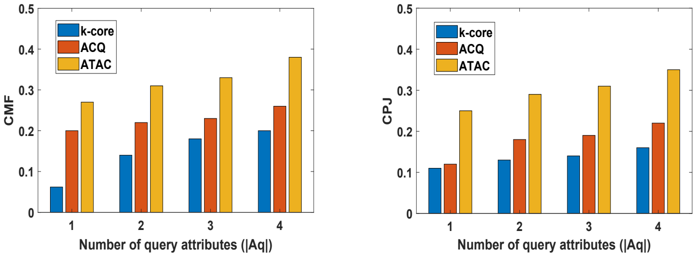
$$CPJ(N(\mathscr{C}_q)) = \frac{1}{\mathscr{L}} \sum_{i=1}^{\mathscr{L}} \frac{1}{|\mathscr{C}_i|^2} \left[\sum_{j=1}^{|\mathscr{C}_i|} \sum_{k=1}^{|C_i|} \frac{|\mathscr{A}_q(\mathscr{C}_{i,j})| \cap |\mathscr{A}_q(\mathscr{C}_{i,k})|}{|\mathscr{A}_q(\mathscr{C}_{i,j})| \cup |\mathscr{A}_q(\mathscr{C}_{i,k})|} \right] \tag{5}$$

CMF and CPJ ranges from 0 to 1. The larger value of those indicates the better cohesiveness of the community.



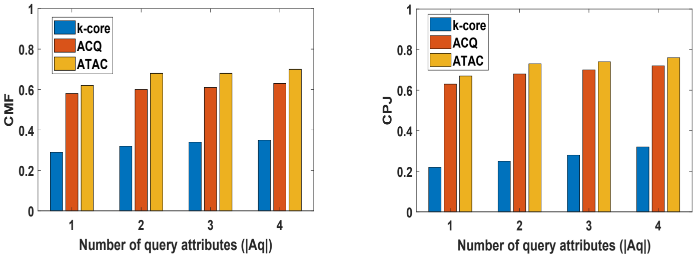
(a) CMF (T1) (b) CPJ (T1)

Fig. 1. Performance comparison on CRAWL dataset (in all cases, $k = 4$, $h = 3$, $\gamma = 150$, $\theta = 0.5$)



(a) CMF (b) CPJ

Fig. 2. Performance comparison on Flickr dataset (in all cases, $k = 4$, $h = 3$, $\gamma = 10$, $\theta = 0.5$)



(a) CMF (T1) (b) CPJ (T1)

Fig. 3. Performance comparison on DBLP dataset (in all cases, $k = 3$, $h = 3$, $\gamma = 3$, $\theta = 0.5$)

We see that ATAC always performs best in terms of CMF and CPJ (Figure 1(a), (b), Figure 2(a), (b) and Figure 3(a), (b)). The reason is that each community member has to perform certain number of activities related to $\mathscr{A}_q$ to become an active user and thus, most of them show their high degree of inclination towards multiple query topics. In case of ACQ, there are many low active community members who don't have interest in most of the query topics. So, the coverage of query topics within the communities are not that much better as in ATAC. On the other hand, the values of CMF and CPJ in *k*-core are very poor as it ignores users' association with the query topics.

References

1. Fang, Y., Cheng, R., Luo, S., Hu, J.: Effective Community Search for Large Attributed Graphs., VLDB, 1233 - 1244, (2016).
2. Huang, X., Lakshmanan, L. VS.: Attribute-driven community search. In: VLDB, 949 - 960, (2017)

Generalized optimal paths and weight distributions revealed through the large deviations of random walks on networks

Ricardo Gutiérrez[1] and Carlos Pérez-Espigares[2]

[1] Complex Systems Interdisciplinary Group (GISC), Department of Mathematics, Universidad Carlos III de Madrid, 28911 Leganés, Madrid, Spain,
rigutier@math.uc3m.es,
[2] Departamento de Electromagnetismo y Física de la Materia and Institute Carlos I for Theoretical and Computational Physics, Universidad de Granada, 18071 Granada, Spain

1 Introduction

Shortest paths are central in transportation networks, the internet and the human brain, among other complex systems, and also enter into the definition of network structural properties such as the closeness, efficiency or betweenness centrality [1]. The idea of a shortest path can be generalized to include obstacles or several targets from a given source, and a great deal of effort has been devoted to the solutions of such problems. But how can one find the path or paths that maximize the fluctuations of a given process; or the path that guarantees that a certain node is visited on average twice as frequently as another node; or what is the shortest path that a passenger or an information packet can take without saturating a given node or link? These are optimal paths in a generalized sense, as they ensure that the statistics of a given observable takes certain values or does not exceed certain bounds. Similarly, one could think of redistributing the weights of a network so as to ensure that a set of nodes is visited with some frequency in the resulting weighted network, or that no link carries more than a certain amount of flow.

We study such generalized optimal paths and weight distributions through a large-deviation study of ensembles of trajectories [2, 3]. By biasing the dynamics with time-integrated observables we obtain the stationary distribution that guarantees that such observables satisfy some statistical constraint, which may be related to its mean value, fluctuations, or higher-order cumulants. Moreover, the auxiliary process given by the Doob transform [4, 5] yields a new stochastic process whose transition rates give rise to the appropriate statistics for long times. By combining the biased stationary state with the Doob transform, we obtain generalized optimal paths. Furthermore, the Doob-transformed process itself yields an optimal distribution of link weights. We illustrate our approach by finding shortest paths in the presence of constrains through a large-deviation study of the maximal entropy random walk (MERW) [6]. We also study weight redistributions for maximal current of flows in the presence local constraints by means of the standard random walk on networks [7], which we do not discuss here for lack of space. This and several other examples will be discussed in Ref. [8] and in this presentation. The large-deviation approach to the study of processes on networks is a promising avenue for the exploration of structural properties of networks, as recently shown in Refs. [9, 10].

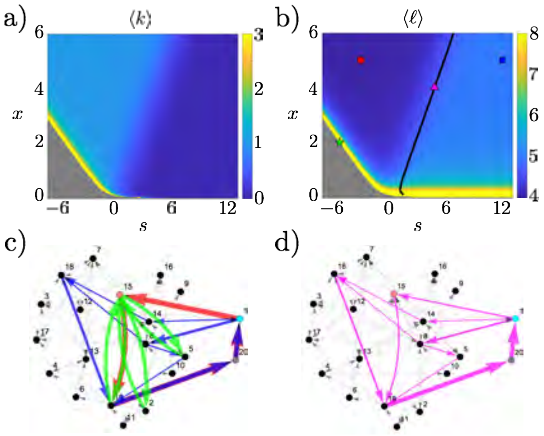


Fig. 1. Finding optimal paths in the presence of constraints. (a) Average activity $\langle k \rangle$ as a function of the tilting fields s and x in the graph shown in (c) and (d). (b) Average walk length $\langle \ell \rangle$ joining nodes 1 and 20. The solid black segment shows where $\langle k \rangle = 1/3$. (c) Trajectories obtained from the Doob transformed process corresponding to the red circle, the green star and the blue square shown in (b). (c) Trajectories for the magenta triangle in (b).

2 Results

Here we illustrate the idea of finding optimal paths in a small directed random graph of $N = 20$ nodes, see Fig. 1 (c) and (d). Much larger networks with Poissonian or power-law degree distributions, or networks arising from applications can be similarly studied [8]. A particle performs a MERW from the source node 1 and reaches the target node 20 after ℓ steps. We are interested in the walk length ℓ, but also on how frequently it goes through an "obstacle", node 15. For that purpose we use the activity k, which gives the number of times the particle goes through the obstacle before reaching the target. The large-deviation analysis allows us to access the average, fluctuations and higher cumulants of ℓ and k in a biased trajectory ensemble, with probability distribution

$$P_{sx}(k,\ell) = e^{-sMk-xM\ell} P(k,\ell)/Z(s,x) \qquad (1)$$

where $P(k,\ell) = P_{s=0,x=0}(k,\ell)$, s and x are tilting fields [3], $Z(s,x)$ is a normalizing factor, and $M \to \infty$. From the large-deviation functions of the operator governing the tilted dynamics we obtain the average activity $\langle k \rangle$ and path length $\langle \ell \rangle$, shown in Fig. 1 (a) and (b) respectively. The gray area corresponds to a region where the averages diverge. The transitions derived from the Doob transform for a point which lies close to the divergence [see the green star in Fig. 1 (b)] are shown as green arrows in panel (c). They clearly display a trajectory where the particle moves back and forth between the obstacle and its neighbors, without ever reaching the target (hence the divergence). Additionally, a crossover between a region where $\langle k \rangle \approx 0$ and a plateau where $\langle k \rangle \approx 1$ is observed in (a), which corresponds to a crossover from $\langle \ell \rangle$ larger than 4 to $\langle \ell \rangle \approx 4$ in (b). The trajectories for the point highlighted with a red disc in Fig. 1 (b), for which

$\langle \ell \rangle = 4$, are shown as red arrows in panel (c). They show a particle moving along the shortest path between the source and the target. As the obstacle lies on this path, we have $\langle k \rangle = 1$. On the other side of the crossover and for sufficiently large s, the walker chooses the shortest amongst the paths that avoid the obstacle, see the blue square in (b) and the blue arrows in (c). We next focus on the combination of paths that yield a frequency $\langle k \rangle = 1/3$ (the obstacle is only visited one in three times), while reaching the target node along the shortest possible route. The trajectory for the point highlighted with a pink triangle in Fig. 1 (b), is shown in panel (d). Apart from the shortest path, which occurs with a probability $1/3$ (as it should, given that contains the obstacle), the other 2/3 are equally split among the three shortest paths out of all five that do not cross the obstacle. Much more involved cases can be explored with as much ease, including paths that maximize the fluctuations of ℓ or other observables, see Ref. [8].

3 Conclusions

We illustrate how to find generalized optimal paths in a small network. In the presentation and in Ref. [8] less elementary cases, including optimal weight distributions adapted to flows, will be discussed. This approach can be employed in very large networks, as long as one can compute the largest eigenvalue of the appropriate tilted $N \times N$ operator. Paths and weights for specific network topologies, which can be directed, weighted, and also include spatial information, can be taylored to a given average or fluctuations of any time-integrated observable, so the approach is widely applicable. These are problems that are outside the reach of standard graph-theoretical algorithms.

References

1. Boccaletti,S. *et al.*: Complex networks: Structure and dynamics. Phys. Rep. 424(4–5), 175 – 308 (2006)
2. Touchette, H.: The large deviation approach to statistical mechanics. Phys. Rep. 478(1–3), 1 – 69 (2009)
3. Garrahan, J. P. *et al.*: First-order dynamical phase transition in models of glasses: an approach based on ensembles of histories. J. Phys. A: Math. Theor. 42(7), 075007 (2009)
4. Jack, R. L., Sollich, P.: Large deviations and ensembles of trajectories in stochastic models. Prog. Theor. Phys. Supp. 184, 304–317 (2010)
5. Chetrite, R., Touchette, H.: Nonequilibrium Markov processes conditioned on large deviations. Ann. Henri Poincaré 16(9), 2005–2057 (2015)
6. Burda, Z., Duda, J., Luck, J.-M., Waclaw, B.: Localization of the maximal entropy random walk. Phys. Rev. Lett. 102(16), 160602 (2019)
7. Noh, J. D., Rieger, H.: Random walks on complex network. Phys. Rev. Lett. 92(11), 118701 (2014)
8. Gutiérrez, R., Pérez-Espigares, C..: Generalized optimal paths and weight distributions revealed through the large deviations of random walks on networks. (In preparation)
9. De Bacco, C., Guggiola, A., Kühn, R., Paga, P.: Rare events statistics of random walks on networks: localisation and other dynamical phase transitions. J. Phys. A: Math. Theor. 49(18), 184003 (2016)
10. Coghi, F., Morand, J., Touchette, H.: Large deviations of random walks on random graphs. Phys. Rev. E 99(2), 022317 (2019)

Equality Measures in Complex Networks

Eva Barrena[1], Alicia De-los-Santos[2], M. Cruz López-de-los-Mozos[3], and Juan A. Mesa[3]

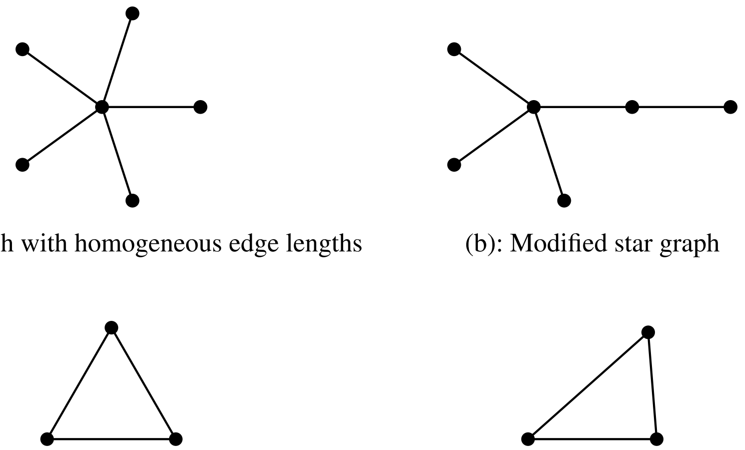
[1] Departamento de Economía, Métodos Cuantitativos e Historia Económica, Universidad Pablo de Olavide, Sevilla, Spain
[2] Departamento de Estadística e Investigación Operativa, Universidad de Córdoba, Córdoba, Spain
[3] Departamento de Matemática Aplicada I, Universidad de Sevilla, Sevilla, Spain
[4] Departamento of Matemática Aplicada II e Instituto de Matemáticas de la Universidad de Sevilla (IMUS), Universidad de Sevilla, Sevilla, Spain, jmesa@us.es, corresponding author

1 Introduction

Since their introduction in [1], the characteristic path length and the clustering coefficient have been used to characterize small-world networks as well as to evaluate several properties of complex networks. These measures have been upgraded and extended in order to cope with more general situations. In particular, these measures were substituted by the notions of global and local efficiency. In [2] these concepts were defined and extended to graphs endowed by a metric different than the topological one, that is to graphs where each edge has an associated weight (length, constructing cost, time to traverse, etc.). Also, the extension of the characteristic path length is straightforward, and the clustering coefficient has been extended to edge-weighted networks. During the last decade, the characteristic path length/clustering coefficient and the global/local efficiency have also been used to evaluate several properties of complex networks: efficiency, robustness (both against error or failures and intentional attacks), vulnerability, redundancy, and even resilience. Also, these measures are useful for comparing different networks and for optimal network design.

Some efforts have been done to measure the heterogeneity of the degree distribution, which is used in the definition of scale-free network. In [6] the sum of quadratic differences of the inverse square root of the degrees is defined as a network heterogeneity index, and in [7] some useful properties of this index are given.

However, each of the coefficients that characterize the small-world phenomenon synthesizes the network giving a unique (expected) associated value but, do not provide an idea of the dispersion of the nodes of the network regarding this measure, and therefore about the heterogeneity of the network. In fact, it is easy to find different (topological) graphs with the same number of vertices, and similar characteristic path length but different variance. This is the case of the star graph and modified star graph (Figures (a) and (b)), where the characteristic path lengths are $8/5$ and $9/5$, respectively but, their variance around the characteristic path length are $6/25 < 1$ and $71/50 > 1$. Moreover, for the global efficiency, which is the inverse of the harmonic mean, let us consider the triangle graphs with all the length of the edges equal to one, and the triangle graph with lengths: $1, 1/(1+\varepsilon), 1/(1-\varepsilon)$, where $0 < \varepsilon < 1$ (Figures (c) and (d).) Then, the global efficiency coincides but equality measures as the mean absolute deviation around the global efficiency give different values 0 and $\frac{2}{3}\varepsilon$, respectively. This is also the case when applying other equality measures as the variance, the sum of absolute differences, and the maximum absolute deviation around the global efficiency [3], [4], [5].



(a): Star graph with homogeneous edge lengths

(b): Modified star graph

(c): Triangular graph with edge lengths 1

(d): Edge lengths 1, $\frac{1}{1+\varepsilon}$, $\frac{1}{1-\varepsilon}$, for $\varepsilon = 0.2$

2 Results

Let $G = (V,E)$ be a graph where V is the vertex set, $|V| = N$, and E is the edge set. It is very common to have an edge-weighted graph, where each edge e has an associated value l_e. The edge-weight set induces the metric d given by the shortest path. Moreover, it is possible that vertices are also weighted. In what follows we provide some preliminary definitions of heterogeneous measures around the characteristic path length and global efficiency:

$$L(G) = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \qquad E(G) = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}}.$$

Thus, the variance around the characteristic path length and global efficiency are defined as:

$$V_L(G) = \frac{1}{N(N-1)} \sum_{i \neq j} (d_{ij} - L(G))^2 \qquad V_E(G) = \frac{1}{N(N-1)} \sum_{i \neq j} (\frac{1}{d_{ij}} - E(G))^2.$$

Analogously, the mean absolute deviation around the characteristic path length and global efficiency are defined as:

$$MAD_L(G) = \frac{1}{N(N-1)} \sum_{i \neq j} |d_{ij} - L(G)| \qquad MAD_E(G) = \frac{1}{N(N-1)} \sum_{i \neq j} |\frac{1}{d_{ij}} - E(G)|.$$

In this paper, we will properly define these measures and study their properties. Properties as monotonicity, normalization and membership to the interval $[0,1]$, scale invariance, modularity, and the relationship with vulnerability, among others, are explored in this paper. Moreover, we give algorithms for computing them, and their computational complexities that do not exceed that of computing the matrix distance. The main aim of this work is to propose and study these measures, and assess their usefulness when evaluating complex networks as well as when designing optimal networks from different viewpoints. Even though these measures provide useful information on the network, their importance takes a different value when they are accompanying other measures as the global efficiency. Another point of interest is their application to determine the efficiency, robustness, and other special properties of complex networks. Special attention is paid to the application to transportation networks with their special features. In rapid transit networks, each node can be considered as a station. If a node/station is not well connected, an accident in this station can block the line or even the network. Therefore, local connectivity measures can be used to determine the most vulnerable nodes in order to reinforce them and/or their surroundings. However, when comparing networks considering the mean of local connectivity measures, it may happen that two networks present the same mean but that one has more dispersion in the local measures, thus having more vulnerable nodes. Considering dispersion connectivity measures along with mean ones is a useful choice to globally detect these cases and therefore to prevent them in subsequent phases.

Acknowledgement: This work was in part supported by the Ministerio de Economía y Competitividad (Spain)/FEDER under grant MTM2015-67706-P, and by the Ministerio de Investigacón (Spain)/FEDER under grant PID2019-106205GB-I00.

References

1. Watts, D.J, Strogatz, S.H. : Nature 393, 440 (1998)
2. Latora, V., Marchiori, M.: Efficient Behavior of Small-World Networks, Phys. Rev. Lett. 87(19), 198701 (2001)
3. López-de-los-Mozos, M.C., Mesa, J.A.: The maximum absolute deviation measure in location problems on networks. Eur. J. Oper. Res. 135, 184–194 (2001)
4. López-de-los-Mozos, M.C., Mesa, J.A.: The sum of weighted absolute differences location problem. INFOR: Information Syst. and Oper. Res. 41, 195–210 (2003)
5. Mesa, J.A., Puerto, J., Tamir A.: Improved algorithms for several network location problems with equality measures. Discrete Applied Mathematics 130, 437–448 (2003)
6. Estrada, E.: Quantifying network heterogeneity, Physical review E, 82. 066102, (2010)
7. Estrada, E.: Degree heterogeneity of graphs and networks I. Interpretation and the heterogeneity paradox. Journal of Interdisciplinary Mathematics, 22, 503–529, (2019)

Analysis of Variable Stars via Visibility Graph Algorithm

Nayade Garcés[1] and Víctor Muñoz[2]

Universidad de Chile, Facultad de Ciencias, Las Palmeras 3425, Ñuñoa, Santiago, Chile
nayade.gh@gmail.com

1 Introduction

Nowadays Complex Networks are used in various areas and the number of topics in where they are being useful keeps growing. An interesting problem is how to build a complex network from a time series, which is a universal problem considering that time series is the main input that basic sciences receive from Nature.

Several ways to build these networks have been proposed [1], but there is one which leads to particularly interesting results, called the Visibility Graph algorithm [2] that takes a time series and maps it into a graph. In this graph, a node corresponds to a given datum in the time series, and two nodes are connected if visibility exists between the corresponding data, *i.e.* if there is a straight line that connects the data, provided that this "visibility line" is always above the data curve.

There are two main ways to build these visibility graphs. For the normal visibility graph (VG), the visibility line joins data points, which means lines can all have different slopes, depending on the relative height of the points. In the second way, data points are replaced by vertical bars, by joining them with the x-axis. Now, visibility lines are drawn parallel to the x-axis, starting at a data point, until it finds another datum's bar. This method is called the Horizontal Visibility Graph (HVG).

In Astronomy, light curves are time series of the luminosity of a star. In this work, we are interested in studying stars using the visibility graph algorithm, in particular variable stars.

Variable stars are stars that —seen from the Earth or satellites— change their luminosity in time. In many cases they do this periodically but in others they do not. Here we study one specific kind of variable stars known as pulsating stars, where sound waves travel across the stars interior making their radii change in time. When it gets bigger (smaller) the star gets colder (hotter). In turn, this change of temperature leads to luminosity change due to the Stefan-Boltzmann equation. The best known pulsating stars are the Cepheids, popular because of their period-luminosity relation [3]. These stars show regular changes in their light curves and their pulsating mechanism is fairly well understood. RR Lyrae is another type of pulsating star, also called "short-period Cepheids", due to their much smaller pulsating period compared to Cepheids. They also exhibit a period-luminosity relation. This relation is very important in Astronomy, as it can be used to infer our distance to the star, which is why Cepheids are popularly referred to as standard candles.

Besides the above, there is one particular type that has been difficult to understand: Delta Scuti stars. They are also Cepheids, because of their low mass they are also called dwarf Cepheids, and present period-luminosity relation. The difference between the classical Cepheids, RR Lyrae and Delta Scuti is their pulsation mode type. For the first two it is known that they pulsate only in radial modes, however Delta Scuti stars pulsate in both radial and non-radial modes, making their light curve difficult to understand. For this reason, several studies have been made to understand their pulsation mechanism [5, 6]. including its possible fractal behavior [4].

2 Results

We study these three types of Cepheids —classical, short period and dwarf— with the visibility graph algortihm using the OGLE-III catalog of variable stars [7]. We use both methods described earlier, VG and HVG, to map the light curves into graphs. Because of the existence of observational gaps present in almost all light curves, we build three types of visibility graphs. First, we made the (H)VGs using the complete light curve ignoring the existence of the gaps. Here it is important to mention that the normal visibility graph is affected by the spacing of the data, but the horizontal one is not. This means that the HVG is the same if the time series is evenly or unevenly spaced. Second, we build (H)VGs for each observation window. Finally, we build the (H)VGs using the phased light curves given by the OGLE catalog, that is, taking the star period given by the catalog, and putting all data points within a single period. Notice that the phased curves allow to ignore the effect of observational gaps, but need the star period as additional information.

For each Cepheid type, each graph model (VG, HVG), and each strategy to deal with the observational gaps, we calculate various network metrics: degree distribution, mean degree, clustering coefficient, transitivity quotient, average path length, and reversibility of the time series (via the Kullback-Leibler divergence [8]).

An interesting result is that the degree distribution for all stars present an exponential behavior (see Fig. 1(a)). On the other hands, some combinations of metrics seem to be able to discriminate between star types, as suggested in Fig. 1(b), which shows the average degree in the HVG versus the star period, for five star types.

Summary. We find that the visibility algorithm is a useful way to study the light curves of variable stars, showing interesting features, some of them universal, for all the stars studied, whereas others seem to discriminate between such types. It is also interesting that this method exhibits similar results for the three different ways to deal with the gaps, meaning that the —always difficult problem to resolve— gap existence is not too important for this construction, and that we can obtain valuable information from these light curves regardless of observational gaps.

References

1. Zhang, J. and Small, M.: Complex Network from Pseudoperiodic Time Series: Topology versus Dynamics. PRL 96, 238701-(1–4) (2006)
2. Lacasa, L. and Luque, B. and Ballesteros, F. and Luque, J. and Nuño, J. C. : From time series to complex networks: The visibility graph. PNAS 105 (13), 4972–4975 (2008)

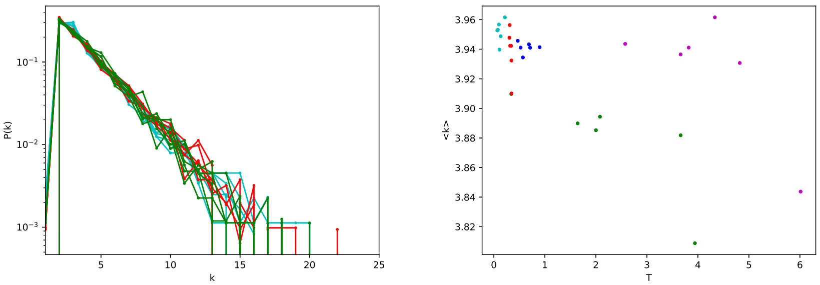


Fig. 1. Left panel: Degree distribution probability for HVG for Cepheids pulsating in their fundamental tone (cyan lines), delta Scuti stars (red lines), and Cepheids 10, *i.e.* Cepheids pulsating in their first overtone (green lines). Right panel: Average degree in HVG versus star period in days. Cyan points: delta Scuti; green points: Cepheids 10 ; blue points: RR Lyrae RRab; red points:RR Lyrae RRc; purple points: Cepheids pulsating in their fundamental tone.

3. Leavitt, H. : 1777 variables in the magellanic cloud. AHCO 60(4), 87–111 (1908)
4. de Franciscis, S. and Pascual-Granado, J. and Suárez, J. C. and García Hernández, .A and Garrido, R.: Fractal analysis applied to light curves of δ Scuti stars. MNRAS 481(4), 4637–4649 (2018)
5. Breger, M. and Lenz, P. and Antoci, V. and Guggenberger, E. and Shobbrook, R. R. and Handler, G. and Ngwato, B. and Rodler, F. and Rodriguez, E. and López de Coca, P. and Rolland, A. and Costa, V. : Detection of 75+ pulsation frequencies in the δ Scuti star FG Virginis. A&A 435(3), 955 –965 (2005)
6. Sánchez Arias, J. P. and Córsico A. H. and Althaus, L. G.: Astrosismología de estrellas variables Delta Scuti y Gamma Doradus. Boletín de la Asociación Argentina de Astronomía, 57, 99–101 (2014)
7. Soszyński, I. and Poleski, R. and Udalski, A. and Szymański, M. K. and Kubiak, M. and Pietrzyński, G. and Wyrzykowski, Ł. and Szewczyk, O. and Ulaczyk, K.: The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. I. Classical Cepheids in the Large Magellanic Cloud. AA, 58, 163–185 (2008)
8. Lacasa, L. and Nuñez, A. and Roldán, É., Parrondo, J. M. R. and Luque, B.: Time series irreversibility: a visibility graph approach. Eur. Phys. J. B, 85(217), 11 pages (2012)

Comparing Box-Covering Algorithms for Fractality of Complex Networks

Péter Tamás Kovács[1], Marcell Nagy[1], and Roland Molontay[12]

[1] Dept. of Stochastics, Budapest University of Technology and Economics, Hungary
[2] MTA-BME Stochastics Research Group
kptzeg@gmail.com, {marcessz, molontay}@math.bme.hu

1 The box-covering problem and fractal networks

The fractal nature of complex networks has received a lot of research interest recently since fractality has been verified in several real-world networks (WWW, actor collaboration networks, protein interaction networks) [1], moreover, fractality has been associated with many important features of networks such as robustness, modularity, and information contagion [2].

Heuristically, fractality of a network means that the network looks similar to itself on different scales: if one zooms in on a sub-network, one is expected to see the same qualitative behavior as in the whole network. A method to identify fractality of complex networks is similar to that of regular fractal objects: using the box-covering method. For a given network G, we partition the vertices into boxes of size l_B. A B box is a subgraph of G with $\mathrm{diam}(B) \leq l_B$. The goal of the box-covering algorithm is to find the **minimum number** of l_B-sized boxes needed to cover the entire network G, which is denoted by $N_B(l_B)$ [3]. The fractal or box-dimension d_B of a network can be defined via the relationship of $N_B(l_B)$ and l_B: $N_B(l_B) \sim l_B^{-d_B}$. If d_B **exists** and is finite, the network is **fractal**, otherwise it is non-fractal [4]. In other words, in fractal networks, the minimum number of boxes scales as a **power law** with the size of the boxes, i.e. $\log N_B$ scales linearly with $\log l_B$ with slope $-d_B$.

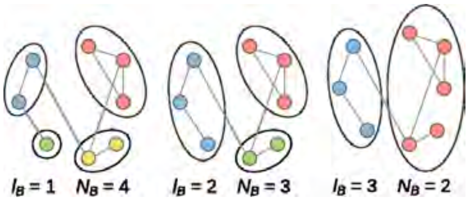


Fig. 1: The optimal box-covering for different box sizes.

Determining the minimum number of boxes needed to cover the entire network belongs to a family of NP-hard problems [3]. On the other hand, in practice, various algorithms are adopted to obtain an approximate solution; for an incomprehensive list see Table 1.

In this work, we provide a systematic review of the various box-covering algorithms proposed throughout the years. Furthermore, we compare the performance of the algorithms in terms of running time and approximation ability both using real-world networks and mathematical network models. We will also make our implementations of the algorithms publicly available as a Python module.

Classical box-covering algorithms	• RS (random sequential) [5] • Greedy coloring [3] • MA (merge algorithm) [6]	**Burning algorithms**	• CBB (compact-box-burning) algorithm [3] • MEMB (maximum-excluded -mass burning) [3] • MCWR (combines MEMB and RS) [7] • MVB (minimal value burning) [9]
Metaheuristic optimization algorithms	• SA (simulated annealing) [6] • Edge-covering with simulated annealing [8] • DEBC (differential evolution box-covering) [10] • PSOBC and MOPSOBC (single- and multi-objective discrete particle swarm optimization box-covering) [13, 14] • Max-Min ant-colony algorithm [16]	**Other algorithms**	• OBCA (overlapping box-covering algorithm) [11] • Fuzzy box-covering [12] • Sketch-based box-covering [15] • Sampling based box-covering [17]

Table 1: Approximating box-covering algorithms

2 Experimental results

We compare the performance of the algorithms on two networks. First, on the 5th iteration ($N = 684$) of the recursive fractal network model, the (u,v)-flower with $u = v = 2$ [2]. The (u,v)-flower has a ground-truth asymptotic box-dimension (for $v \geq u > 1$) given by $d_B = \frac{\ln(u+v)}{\ln u}$. Second, we compare the results on a real-world fractal network, the Tokyo metro network ($N = 248$). The results can be seen in Table 2 and Fig. 2.

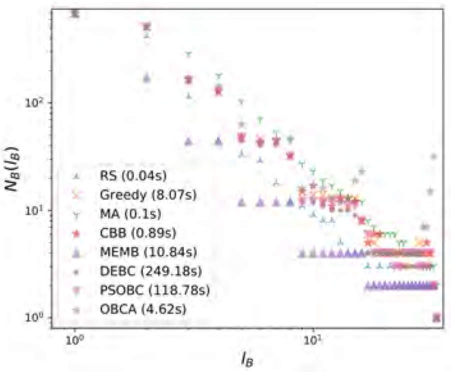
(2, 2)-flower

Algorithm	**Avg. rank**	**Avg. running time**	$\hat{d}_B$
RS	1.53	0.03 (1)	1.84 (5)
Greedy	5.84	8.16 (6)	1.82 (6)
MA	3.63	0.10 (2)	1.96 (1)
CBB	5.52	0.97 (3)	1.81 (7)
MEMB	1.00	10.97 (7)	1.62 (8)
DEBC	2.78	243.79 (9)	1.90 (3)
PSOBC	4.33	113.57 (8)	1.85 (4)
OBCA	3.31	4.59 (5)	1.62 (9)
Fuzzy	-	1.03 (4)	1.95 (2)

Tokyo metro network

Algorithm	**Avg. rank**	**Avg. running time**	$\hat{d}_B$
RS	1.57	0.01 (1)	1.76
Greedy	5.68	0.58 (5)	1.46
MA	4.78	0.02 (2)	1.56
CBB	5.57	0.11 (3)	1.47
MEMB	1.00	1.17 (7)	1.68
DEBC	3.00	27.06 (9)	1.48
PSOBC	4.00	10.56 (8)	1.49
OBCA	3.28	0.7 (6)	1.51
Fuzzy	-	0.22 (4)	2.02

Table 2: Comparison of different algorithms with respect to running time and approximation ability. Fuzzy has no avg. rank because it does not generate explicit boxes.



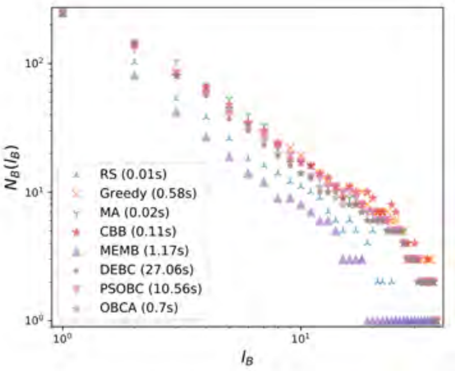


Fig. 2: The relationship between the number of boxes N_B and box size l_B.

The average rank is the lowest for the MEMB algorithm in both cases, thus in these examples, MEMB finds the fewest boxes on average (the ranks for each l_B value are averaged). The random sequential algorithm also ranked high in this respect. Moreover, it has the shortest running time for both networks. Another important feature of the boxing algorithms is the fractal exponent that they provide. On the other hand, the dimension that the algorithms yield does not solely depends on the approximation ability of the algorithm itself but also on the method how the $N_B(l_B) \sim l_B^{-d_B}$ relation is measured that needs further analysis. We also note that there might be significant differences in the asymptotic theoretical dimension of the network models and what we obtain on finite instances. Adjusting for this effect is currently under investigation.

References

1. E. Rosenberg, *A Survey of Fractal Dimensions of Networks*. Springer, 2018.
2. H. D. Rozenfeld, S. Havlin, and D. Ben-Avraham, “Fractal and transfractal recursive scale-free nets,” *New J. Phys*, vol. 9, no. 6, p. 175, 2007.
3. C. Song *et al.*, “How to calculate the fractal dimension of a complex network: the box covering algorithm,” *J. Stat. Mech.: Theory Exp*, vol. 2007, no. 03, p. P03006, 2007.
4. C. Song, S. Havlin, and H. A. Makse, “Self-similarity of complex networks,” *Nature*, vol. 433, no. 7024, p. 392, 2005.
5. J. S. Kim *et al.*, “Fractality and self-similarity in scale-free networks,” *New J. Phys*, vol. 9, no. 6, p. 177, 2007.
6. M. Locci *et al.*, “Three algorithms for analyzing fractal software networks,” *WSEAS Trans. Info. Sci. and App*, vol. 7, pp. 371–380, 2010.
7. H. Liao *et al.*, “Solving the speed and accuracy of box-covering problem in complex networks,” *Physica A*, vol. 523, pp. 954–963, 2019.
8. W.-X. Zhou, Z.-Q. Jiang, and D. Sornette, “Exploring self-similarity of complex cellular networks: The edge-covering method with simulated annealing and log-periodic sampling,” *Physica A*, vol. 375, no. 2, pp. 741–752, 2007.
9. C. M. Schneider *et al.*, “Box-covering algorithm for fractal dimension of complex networks,” *Phys. Rev. E*, vol. 86, no. 1, p. 016707, 2012.
10. L. Kuang *et al.*, “A differential evolution box-covering algorithm for fractal dimension on complex networks,” in *Congress on Evolutionary Computation*, pp. 693–699, IEEE, 2014.
11. Y. Sun, Y. Zhao, *et al.*, “Overlapping-box-covering method for the fractal dimension of complex networks,” *Phys. Rev. E*, vol. 89, no. 4, p. 042809, 2014.
12. H. Zhang *et al.*, “Fuzzy fractal dimension of complex networks,” *Appl. Soft Comput.*, vol. 25, pp. 514–518, 2014.
13. L. Kuang *et al.*, “A discrete particle swarm optimization box-covering algorithm for fractal dimension on complex networks,” in *Congress on Evolutionary Computation*, pp. 1396–1403, IEEE, 2015.
14. H. Wu *et al.*, “A multiobjective box-covering algorithm for fractal modularity on complex networks,” *Appl. Soft Comput.*, vol. 61, pp. 294–313, 2017.
15. T. Akiba, K. Nakamura, and T. Takaguchi, “Fractality of massive graphs: Scalable analysis with sketch-based box-covering algorithm,” in *16th Int. Conf. on Data Mining*, pp. 769–774, IEEE, 2016.
16. D. Li, X. Wang, and P. Huang, “A Max–Min ant colony algorithm for fractal dimension of complex networks,” *Int. J. Comput. Math*, vol. 95, no. 10, pp. 1927–1936, 2018.
17. Z.-W. Wei *et al.*, “Sampling-based box-covering algorithm for renormalization of networks,” *Chaos*, vol. 29, no. 6, p. 063122, 2019.

Revealing the complex comorbidity structure of internalizing disorders through hypergraph models

Marijn ten Thij[1], Lauren A. Rutter[2], Lorenzo Lorenzo-Luaces[2], and Johan Bollen[1,3]

[1] Center for Social and Biomedical Complexity, Indiana University Bloomington, Bloomington, IN, USA
[2] Department of Psychological and Brain Sciences, Indiana University Bloomington, Bloomington, IN, USA
[3] Centre for Urban Mental Health, University of Amsterdam, Amsterdam, the Netherlands

Internalizing disorders, such as anxiety and depression, frequently co-occur. In fact, major depressive disorder (MDD) and generalized anxiety disorder (GAD) overlap to such a degree that researchers have suggested that GAD is not independent of MDD because most people who develop GAD will eventually develop MDD (see [5]). The odds ratio of cross-sectional comorbidity of MDD and GAD is upwards of 8.2 [3]. While MDD and GAD are two of the most highly comorbid disorders, having any mood disorder substantially increases risk of another disorder with an adjusted odds ratio of 4.2 for any anxiety disorder, 2.0 for any drug use disorder, and 4.6 for any personality disorder [2]. Comorbidity is strongly associated with disorder severity, chronicity, and impairment [4]. Thus, a better understanding of the complex, time-evolving structure of comorbidity is crucial for prevention and treatment.

Here we use a hypergraph model to reveal the structure of comorbidity among internalizing disorders in a large naturalistic sample of online individuals by examining how groups of diagnoses and distinct disorders co-occur in reports of clinical diagnosis on social media. We specifically focus on two prevalent groups of internalizing disorders; one group centered around depression (Group 1, containing the following disorders; depression, dysthymia, seasonal affective disorder (SAD), and persistent depressive disorder (PDD)) and the other centered around anxiety and related disorders (which we will refer to as Group 2, containing the following disorders; agoraphobia, anxiety, GAD, obsessive compulsive disorder (OCD), panic disorder, and specific phobia).

We used the IUNI OSoMe [1], a service which provides searchable access to the Twitter "Gardenhose", a 10% sample of all daily tweets, to search for tweets posted in 2019, in which individuals explicitly state that they received a clinical diagnosis of one or more disorders. We do so by matching tweet content to the regular expressions "diagnos*" followed by a disorder name, e.g. "dysthymia". To ensure we are only including valid cases, 3 experts manually labelled whether the tweet contained an actual reference to a clinical diagnosis. We only retained "diagnosis" tweets for which all three labels were positive, thereby removing quotes, jokes, and external references. We found 4,049 individuals who thus expressed a clinical diagnosis of at least one of the ten considered disorders.

Individuals frequently report a diagnosis for multiple disorders, indicating they are comorbid for the individual. Hence, across all individuals, we can examine the structure of comorbidity through a hypergraph model in which each disorder is represented as a node and hyperedges correspond to individuals reporting comorbid diagnoses and

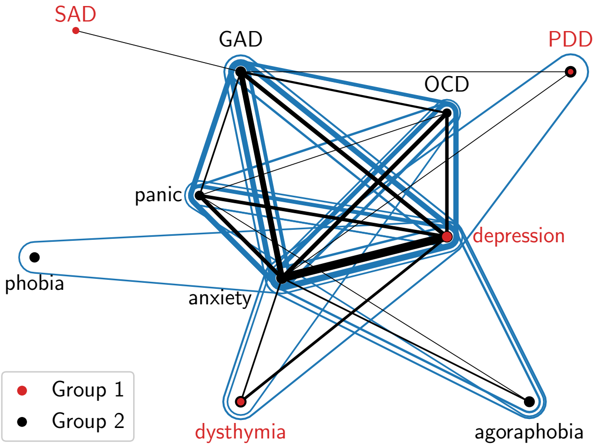


Fig. 1. Comorbidity network of internalizing disorders. Hyperedges indicate which combinations of individual clinical diagnoses were mentioned by our sample of N=4,049 individuals on Twitter. Hyperedge width scales with the number of individuals who reported a diagnosis of the particular set of disorders. Node sizes indicate the number of individuals that reported the diagnosis. Group membership for each disorder is indicated by the color of the node and the label.

hyperedge weights represent the number of individuals that reported this combination of diagnoses.

The network shown in Fig. 1 reveals a number of pertinent structural features of disorder comorbidity. First, as expected, the disorders in Group 1 are more strongly clustered within their own group than the disorders in Group 2. Second, seasonal affective disorder (SAD) is only connected to generalized anxiety disorder (GAD) in the graph, indicating it is comorbid only with this specific disorder. Given that SAD is a specifier for MDD, this reflect the comorbidity between depression and GAD. Third, if phobia is mentioned as a comorbid diagnosis, it is always mentioned in combination with both anxiety and depression, indicating a possible clinical connection between the three disorders.

The generated hypergraph representation of comorbidity from social media reports reveals a number of relevant features. We plan to extend this work to study the diachronic dynamics of comorbidity at high temporal resolutions by examining the order in which comorbidities, e.g. in a single tweet vs. at different moments in time for the same individual across specific geolocated cohorts. This could inform models of how internalizing disorders develop and evolve within the individual, possibly informing new treatment and prevention approaches. Ultimately, examining comorbidity allows for researchers to study causative genetic or neurobiological vulnerabilities, temperamental factors, or stressful life experiences involved in the development of internalizing disorders.

Disorder comorbidity as a complex phenomenon. Thus far, it has largely been studied in clinical samples, which allow for time-varying measurements but are not representative of the range of co-morbidity, or epidemiological samples, which are more representative but are usually cross-sectional, and their collection is limited by time and cost constraints. As a majority of the US population is now active on social media (over 70%), obtaining data from these platforms as a source allows for the rapid recruitment of large samples. Additionally, social media also provides new opportunities to include traditionally underrepresented groups, which could provide a deeper understanding of the social and clinical complexities of disorder comorbidity in non-treatment-seeking samples.

References

1. Davis, C.A., Ciampaglia, G.L., Aiello, L.M., Chung, K., Conover, M.D., Ferrara, E., Flammini, A., Fox, G.C., Gao, X., Gonçalves, B., et al.: OSoMe: the IUNI observatory on social media. PeerJ Computer Science 2, e87 (2016)
2. Hasin, D.S., Sarvet, A.L., Meyers, J.L., Saha, T.D., Ruan, W.J., Stohl, M., Grant, B.F.: Epidemiology of Adult DSM-5 Major Depressive Disorder and Its Specifiers in the United States. JAMA Psychiatry 75(4), 336–346 (04 2018), https://doi.org/10.1001/jamapsychiatry.2017.4602
3. Kessler, R.C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K.R., Rush, A.J., Walters, E.E., Wang, P.S.: The Epidemiology of Major Depressive DisorderResults From the National Comorbidity Survey Replication (NCS-R). JAMA 289(23), 3095–3105 (06 2003), https://doi.org/10.1001/jama.289.23.3095
4. Kessler, R.C., Stang, P.E., Wittchen, H.U., Ustun, T.B., Roy-Burne, P.P., Walters, E.E.: Lifetime Panic-Depression Comorbidity in the National Comorbidity Survey. Archives of General Psychiatry 55(9), 801–808 (09 1998), https://doi.org/10.1001/archpsyc.55.9.801
5. Moffitt, T.E., Harrington, H., Caspi, A., Kim-Cohen, J., Goldberg, D., Gregory, A.M., Poulton, R.: Depression and Generalized Anxiety Disorder: Cumulative and Sequential Comorbidity in a Birth Cohort Followed Prospectively to Age 32 Years. Archives of General Psychiatry 64(6), 651–660 (06 2007), https://doi.org/10.1001/archpsyc.64.6.651

Size agnostic change point detection framework for evolving networks

Hadar Miller and Osnat Mokryn[†]

Information Systems, University of Haifa, Israel

1 Introduction

The analysis of changes in dynamic social and complex networks in response to events, and the automatic detection of these points of change termed Change Point Detection (CPD) are of specific interest lately. Recent works identified changes in the community partitioning of the Enron email exchange immediately after the Californian blackouts [1], and a turtling-up of conversation networks between traders in response to significant stock price changes [2]. Understanding the network's reaction to unusual events provides improved abilities to analyze, understand and possibly take actions in a given system, infer its response to external shocks, and aid in predicting organizational and behavioral changes. Approaches for detecting changes range from finding a scalar values representing the longitudinal data [3], or probabilistic and model-based representations of the network [4, 1, 5]. However, the works mentioned here did not examine the complex network's structure as manifested through distributions.

The structural properties that are in the focus of our work here are the network's native statistical distribution, i.e., its degree distribution measure. Distribution functions are a measure of the division of resources within the network, and their relative positions, and are considered a fundamental tool in the understanding of complex systems [6]. Bhamidi *et al.* [7] further showed that degree-distribution measures reflect changes in the underlying structure better than the hyper-parameters of the corresponding network models. An additional valuable advantage of a degree distribution-based event detection is that it eliminates the need to know in advance the number of nodes in the network at each point in time and can work with as little information as the sequence of interactions for the periods under inspection. Thus, unlike other CPD schemes, the proposed solution here assumes no prior knowledge of the network, does not require pre-processing, and can be used in an online manner, where new network snapshots are generated on-the-fly.

In here we focus on interaction networks, such as phone calls, text messages, emails, and online social network postings. These networks, also termed temporal asynchronous human communication networks [8], can be characterized by the intertwining of the temporal topological structure and the interaction dynamics.

Method: For two consecutive graph snapshots $g_i, g_{i+1}, (i \in \{1,2,...\})$ we denote the two generated corresponding cumulative degree distribution functions by $S_i(x), S_{i+1}(x)$. Given the CDF degree distribution $S_j(x), j \in i, i+1$ for graph g_j: $S_j(x) = P_j(x \leq X)$

[†] Corresponding author: ossimo@gmail.com

we compute the KS statistic D, defined as the maximal difference between the two empirical distributions: $D(S_i, S_{i+1}) = \sup_x |S_i(x) - S_{i+1}(x)|$.

The KS null hypothesis is that the two samples were drawn from the same distribution. Namely,that the distance between the model graph distribution CDF, $S_i(x)$, and the consecutive graph's distribution CDF, $S_{i+1}(x)$, is typical for distances between distributions sampled from the same base model graph distribution.

Following Monte-Carlo bootstrap procedure [10] we generate $k = 1000 >> 1$ new samples from S_i and compute the distance $D(S_i, S_{i_k})$ between $S_i(x)$ and each of its bootstrap samples, $\{S_{i_k}(x), k \in 1..1000\}$. This results in a group of size $n = 1000$ of distances from S_i to its samples, the group of distances $\{d_{i_k}\}, k \in [1..n]$. Then, given $k \in [1..n]$: $\forall \alpha < 1$, $\exists\, d_{i,\alpha}$, s.t. $|\{D(S_i, S_{i_k}) < d_{i,\alpha}\}| = \alpha \cdot n$.
If $D(S_i, S_{i+1}) > d_{i,\alpha}$ then we can reject the null hypothesis with confidence α.

2 Results

We test our framework performance with two real world social organization datasets, Enron and AskUbuntu, and a series of synthetic datasets. We compare our framework performance[1] to that of [1]. We further conduct sensitivity tests with synthetic datasets and compared the performance of KS to that of two alternative distance metrics: Kullback-Leibler divergence (KL) [11] and Relaxed Hausdorff (RH) [9]. The synthetic datasets used were Erdös-Rényi (ER) and Caveman model networks of sizes $\sim N(\mu = 100, \sigma^2 = 10)$. The size and parameters were chosen as they correspond to the majority of the weekly temporal network snapshots of the real-world datasets we use in this research.

Table 1. CPD KS framework performance summary

Dataset	CPD framework		GHRG [1]	
	Prec.	Recall	Prec.	Recall
Enron	0.9	0.9	1.0	0.36
AskUbuntu	0.8	0.57		

	ER sensitivity				Caveman sensitivity			
	Precision		Recall		Precision		Recall	
Solution	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Framework	0.94	0.08	0.96	0.08	0.78	0.41	0.59	0.45
RH CPD	0.19	0.05	0.75	0.28	0.23	0.02	0.91	0.13
KL CPD	0.65	0.39	0.73	0.41	0.86	0.26	0.79	0.36

Table 1 Summarizes the framework's performance compared to the relevant competition for the real-wold dataset for which we had ground truth, and for the synthetic

[1] The Existing method requires that the size of the network is fixed, and hence can be run only for the Enron dataset. The number of participants in the AskUbuntu forum varies greatly with time, as as is the amount of interactions.

networks sensitivity tests. For the real dataset (Enron) our framework found 13 out of the existing 14 points of change in the data, and achieved recall and precision of 0.9, compared to the GHRG-based solution suggested in [1] that did not have false positives but failed to identify some of the change-points. In evaluating the distance metrics, for both random networks (ER) and Caveman-based networks of mid-sizes, our framework outperformed the competition. KL performs well and can detect rather small changes. RH is tailored for detecting changes between networks with a long-tail degree distribution and thus does not perform well for networks of small size. For the full set of experiments and results and a corresponding discussion please see the full paper [12].

Summary. A distribution-based framework like the one presented in this work enables a variable number of nodes at each window of time and hence is size agnostic. It also does not require historical information and can be used for online detection. This is in contrast to model-based frameworks, which are limited by definition to a stringent subset of traceable interacting players over time. The framework can be employed with different distance metrics. Our results demonstrate that for moderate-size networks, the KS distance metric yields good performance, better than KL and RH. It is also widely known to be fast [13].

References

1. Peel L, Clauset A. Detecting change points in the large-scale structure of evolving networks. 29th AAAI Conference on Artificial Intelligence (AAAI). 2015; p. 1–11.
2. Romero DM, Uzzi B, Kleinberg J. Social Networks under Stress: Specialized Team Roles and Their Communication Structure. ACM Transactions on the Web (TWEB). 2019;13(1):6.
3. McCulloh I, Carley K. Detecting change in longitudinal social networks. Journal of Social Structure. 2011;12:1–37.
4. Koutra D, Vogelstein JT, Faloutsos C. Deltacon: A principled massive-graph similarity function. In: Proceedings of the 2013 SIAM International Conference on Data Mining. SIAM; 2013. p. 162–170.
5. Wang Y, Chakrabarti A, Sivakoff D, Parthasarathy S. Fast Change Point Detection on Dynamic Social Networks. arXiv preprint arXiv:170507325. 2017.
6. Stumpf MP, Porter MA. Critical truths about power laws. Science. 2012;335(6069):665–666.
7. Bhamidi S, Jin J, Nobel A, et al. Change point detection in network models: Preferential attachment and long range dependence. The Annals of Applied Probability. 2018;28(1):35–78.
8. Lehmann S. In: Holme P, Saramäki J, editors. Fundamental Structures in Temporal Communication Networks. Cham: Springer International Publishing; 2019. p. 25–48. Available from: https://doi.org/10.1007/978-3-030-23495-9_2.
9. Aksoy SG, Nowak KE, Purvine E, Young SJ. Relative Hausdorff distance for network analysis. Applied Network Science. 2019;4(1):80.
10. Efron B, Tibshirani RJ. An introduction to the bootstrap. CRC press; 1994.
11. Kullback S, Leibler RA. On information and sufficiency. The annals of mathematical statistics. 1951;22(1):79–86.
12. Miller H, Mokryn O. Size agnostic change point detection framework for evolving networks. Plos one. 2020;15(4):e0231035.
13. Glazer A, Lindenbaum M, Markovitch S. Learning high-density regions for a generalized kolmogorov-smirnov test in high-dimensional data. In: Advances in neural information processing systems; 2012. p. 728–736.

Probabilistic Network Sparsification by Ego Betweenness

Amin Kaveh, Matteo Magnani, and Oskar Dahlin

InfoLab, Department of Information Technology, Uppsala University, 75105 Uppsala, Sweden
firstname.lastname@it.uu.se

1 Introduction

In this paper, we study the problem of probabilistic network sparsification. A probabilistic network is a network in which edges are associated with a probability of existence. Sparsification is the problem of removing a predefined percentage of edges from a network while a specific structural property is preserved [1,2]. Parchas el al. in [3] for the first time proposed a method to sparsify probabilistic networks while the expected degree of nodes have the least change. Since the number of edges in sparsified graphs is lower than the original one, the entropy[1] of the graph is normally lower. This is an advantageous result, as we can estimate measures with a smaller number of samples compared to what is needed in the original probabilistic networks. Networks sparsified by the method proposed in [3] can be used to estimate a number of measures like average shortest path length with low error. In this paper, we propose a new method in which ego betweenness is the structural property to be preserved in the sparsification process.

2 Problem Definition and Framework

A probabilistic network, $\mathcal{G}(V,E,P)$, is a network in which V is the set of nodes, E is the set of edges between the nodes and P is a function that assigns existence probability to each edge. Given a probabilistic network $\mathcal{G}(V,E,P)$ and sparsification ratio $0 < \alpha < 1$, *sparsification* is the process to extract a new probabilistic network $\mathcal{G}^{'}(V,E^{'},P^{'})$, where $E^{'} \subset E$ and $|E^{'}| = \alpha|E|$, while preserving a specific structural property and having lower entropy. In this paper, ego betweenness is the structural property that we aim to preserve. The reason is that in probabilistic networks, the probability of paths decreases as the number of constituent edges increases and as a result, the effect of lengthy shortest paths (SPs) is lower than the effect of short SPs in the calculation of betweenness. Hence, ego betweenness that just captures the betweenness of a node among its intermediate neighbors is an appropriate replacement, while also being faster to compute. However, the calculation of the ego betweenness is still computationally expensive because it needs averaging over all possible worlds. Ego betweenness of a node can be estimated in $O(L^2)$ [4], where L is the number of incident edges to that node.

Definition 1. (ego betweenness estimation) Given a probabilistic network $\mathcal{G}$ and a node u, ego betweenness of u can be estimated as:

$$B(u) = \sum_{\{v,w\} \subseteq N(u), v \neq w} p_{uv}\, p_{uw}\, (1 - p_{vw}) \tag{1}$$

where, p_{uv} is the probability of the edge between nodes u and v, and $N(u)$ is the set of nodes having an incident edge to u. We define the ego betweenness discrepancy of a node

[1] The entropy of a $\mathcal{G}$ is defined as: $H(\mathcal{G}) = \sum_{e \in E}(-p_e \log p_e) + (-q_e \log q_e)$, where $q_e = 1 - p_e$.

as $\delta_u = B_{\mathcal{G}}(u) - B_{\mathcal{G}'}(u)$, where $B_{\mathcal{G}}(u)$ and $B_{\mathcal{G}'}(u)$ are ego betweenness of u in $\mathcal{G}$ and $\mathcal{G}^{'}$.
Problem. (ego betweenness sparsification) Given a probabilistic network $\mathcal{G}(V,E,P)$ and sparsification ratio $0 < \alpha < 1$, extract a probabilistic network $\mathcal{G}^{'}(V,E^{'},P^{'})$, with $|E^{'}| = \alpha|E|$ for which the sum of discrepancies, $\sum_{u \in V} \delta_u$, is minimized.
Solution. Solving the aforementioned problem entails two phases: in the first phase, $\alpha|E|$ edges in the original network are selected (and the rest are removed). We perform this task with two different methods: iterative Maximum Spanning Tree (MST) and Monte Carlo sampling (MC). In the second phase, edges' probabilities are modified such that nodes have the least discrepancies. While performing the first phase is simple, finding the exact solution for the second phase is computationally prohibitive [3]. Therefore, we use the gradient descent algorithm (GDB) proposed in [3] to estimate edge probabilities, i.e. $P^{'}$. At each iteration, GDB selects one edge and optimizes its probability. The optimized probability is calculated based on discrepancy of nodes whose ego betweenness will be affected as the probability of the selected edge changes. In more detail, for edge e_{uv} in Figure 1, increasing its probability from p_{uv}^{old} to p_{uv}^{new}, will increase ego betweenness of its incident nodes, u and v, and decrease ego betweenness of common neighbors between u and v, which are w_1 and w_2. The following expressions show the change of discrepancies based on the change of probability p_{uv} at iteration $i+1$:

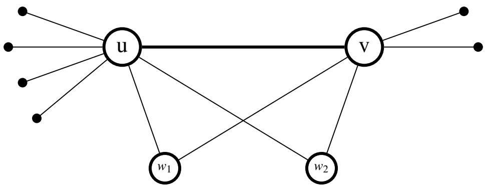


Fig. 1: The effect of changing p_{uv} on discrepancy of other nodes.

$$\delta_u^{i+1} = \delta_u^i - (p_{uv}^{i+1} - p_{uv}^i) \underbrace{\sum_{w \in W(u,v)} p_{uw}(1 - p_{vw}) + \sum_{x \in N(u)-W(u,v)} p_{ux}}_{C_u}$$

$$\delta_v^{i+1} = \delta_v^i - (p_{uv}^{i+1} - p_{uv}^i) \underbrace{\sum_{w \in W(u,v)} p_{vw}(1 - p_{uw}) + \sum_{x \in N(v)-W(u,v)} p_{vx}}_{C_v}$$

$$\delta_{w_j}^{i+1} = \delta_{w_j}^i - (p_{uv}^{i+1} - p_{uv}^i)\underbrace{(-p_{uw_j} p_{vw_j})}_{C(w_j)}$$

where, $N(u)$ is the set of nodes that are connected to node u, and $W(u,v) = N(u) \cap N(v)$ is the set of nodes that are connected to both nodes u and v. Hence, by calculating p_{uv}^{i+1} as follows, we will be assured that $\sum_{u \in V} \delta(u)$ will get one step closer to the minimum value:

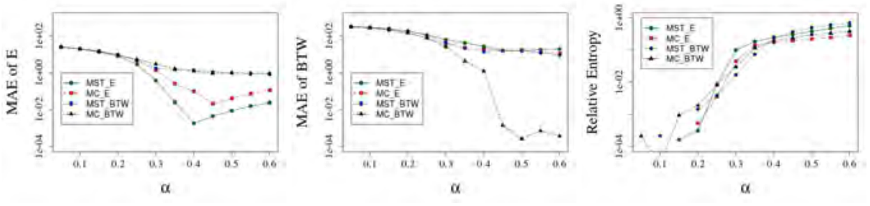
$$p_{uv}^{i+1} = p_{uv}^i + \Delta p \quad , \quad \Delta p = h \frac{\sum_{k \in \{u\} \cup \{v\} \cup W(u,v)} C_k \delta_k^i}{\sum_{k \in \{u\} \cup \{v\} \cup W(u,v)} {C_k}^2}$$

where, $0 < h \leq 1$ is gradient descent step size. We will include the mathematical proof in the main paper.

3 Results

Experiments on discrepancy and entropy Figure 2a shows the mean absolute error (MAE) of nodes expected degree in four sparsified graphs with regards to α: MST/MC states which algorithm has been used in the first phase and E/BTW specifies the structural property which has been preserved in the second phase (expected degree/ego betweenness). The MAE of the expected degree in graphs sparsified with our proposed method is almost 10 times higher than those sparsified with the method in [3]. However, MAE of ego betweenness in networks which are sparsified with the method in [3] is up to 1000 times higher than in sparsified network with our proposed method (specifically MC-BTW) for $\alpha > 0.3$, see Figure 2b. Figure 2c shows the relative entropy of sparsified graphs, $\frac{H(\mathcal{G}')}{H(\mathcal{G})}$. It shows that both methods have a similar trend in reducing the entropy of the probabilistic network. The entropy drops noticeably as α becomes lower that 0.3. That is because as the number of available edges in the sparsified graphs decreases, the probability of edges has to increase to compensate the shortage of expected degree/ego betweenness. As probability can not be higher than 1, then the compensation will be distributed among other involved nodes. As a result, the probability of many nodes will be 1 or close to 1 and the sparsified network becomes closer to a deterministic network with very low entropy.

Conclusion. In this work, we proposed a new probabilistic sparsification method in which nodes ego betweenness is the structural property to be preserved. Our experimental results show that the resulting sparsified graphs have preserved not only ego betweenness of the original graph, but also many other structural properties such as ranking and shortest path queries. All measures can be estimated with a lower number of samples, as the entropy of the sparsified graphs is lower that the entropy of the original graph.



(a) h = 0.5 (b) h = 0.5 (c) h = 0.5

Fig. 2: MAE and relative entropy of sparsified graphs of a probabilistic brain network with 116 nodes and 6670 edges. MST/E and MC/E are the methods proposed in [3] and MST/BTW and MC/BTW are the proposed methods in this paper.

References

1. Nagamochi, H., Ibaraki, T.: A linear-time algorithm for finding a sparse k-connected spanning subgraph of a k-connected graph. Algorithmica **7**(1-6) (1992) 583–596
2. Peleg, D., Schäffer, A.A.: Graph spanners. Journal of graph theory **13**(1) (1989) 99–116
3. Parchas, P., Papailiou, N., Papadias, D., Bonchi, F.: Uncertain graph sparsification. IEEE Transactions on Knowledge and Data Engineering **30**(12) (2018) 2435–2449
4. Kaveh, A.: Local measures for probabilistic networks. PhD thesis, Uppsala University (2019)

Vulnerability indexes in complex networks as a vulnerability component in disaster science

Giovanni G. Soares[1], Aurelienne A. S. Jorge[1], Jeferson F. Mendes[2], Tanishq Garg[3], Harshal Dupare[3], Kaushiki Dixit[3], Vander L. S. Freitas[4], and Leonardo B. L. Santos[5].

[1] National Institute for Space Research (INPE), Sao Jose dos Campos/SP, Brazil
giovanniguarnieri@id.uff.br, aurelienne.jorge@inpe.br,
[2] Sao Paulo State University (UNESP), Sao Jose Dos Campos/SP, Brazil
jeferson.feitosa8@gmail.com
[3] Indian Institue of Technology Kharagpur (IIT Kharagpur), India
gargtanishq@iitkgp.ac.in, harshal.dupare.iitkgp@gmail.com,
kaushikidixit21@gmail.com
[4] Federal University of Ouro Preto (UFOP), Ouro Preto/MG, Brazil
vandercomp@gmail.com
[5] National Center for Monitoring and Early Warning of Natural Disasters (Cemaden), Sao Jose dos Campos/SP, Brazil
santoslbl@gmail.com

1 Introduction

In a scenario of global change, extreme weather events are expected to increase in frequency and intensity and cause more social and economic impacts in several sectors, such as Critical Infrastructures, like Transportation systems. The Sendai Framework for Disaster Risk Reduction 2015-2030 is one of the most important documents in Disaster Science. Amidst its global targets, one refers exactly to "Substantially reduce disaster damage to critical infrastructure and disruption of basic services". For the Disaster Risk Reduction (DRR) scientific community, vulnerability is a key concept [1]. The measurement and mapping of vulnerability constitute a subject of global interest. The Complex Networks approach may offer a valuable perspective considering one type of vulnerability specially related to DRR on critical infrastructures: the topological vulnerability [2].

Geographical networks provide us a better viewpoint on the influence of physical world phenomena on network properties. In terms of the dynamic outlook, the complex network approach is apt for their analysis but, despite the observable interconnection between the network vulnerability and the complex networks, there still lies a scope for further discovery of the intricacies involved. The representation of street networks and other public utility networks using line and point features renders an efficient method to study a large number of fundamental properties, including shortest paths, connectivity and efficiency, which in turn assist in deducing significant conclusions.

In this work, we analyze two vulnerability indexes: the topological vulnerability index and the isolation index (a new one). In both cases, we consider a geographical representation for a street network, under a Disaster Risk Reduction point of view. Therefore, our work is related to two sustainable development goals: climate action, and industry, innovation and infrastructure.

Road transportation systems, such as highways and streets, are networks of routes and locations, represented as a framework of concrete hierarchical mesh structure along with physical spatial constraint, delimited and reserved for exclusive usage. Roads, bridges, tunnels, hill roads are highly vulnerable to disruptions due to inherent complexity and interdependent nature. Natural disasters cause impairment of physical infrastructures, shutdown and discontinuity in operations of road networks.

The vulnerability index, in complex network literature, is first introduced as a way to spot critical components of a critical infrastructure network. The index is based on the efficiency of the network and how taking away some nodes can change the efficiency.

The calculation of the vulnerability index V is as follows: one first determines the network's efficiency E, then removes all the edges connected to the node k and recalculates the efficiency E_k^* without it. Next, one compares the old efficiency E with the new one E_k^* as $V_k = (E - E_k^*)/E$, in which V_k is the vulnerability related to node k. The undirected network vulnerability is $\max(V_k)$, with $k = 1, \cdots, |N|$, for $|N|$ is the total number of nodes. There are several ways to compute the network's efficiency, each depending on the underlying problem. Yet, a classical definition is $E = (\sum_{i \neq j} e_{ij})/(|N|(|N|-1))$, in which e_{ij} is the efficiency in the communication between nodes i and j, defined as the inverse of the shortest path length between them.

The Isolation index serves as a useful measure to scrutinize the impact of isolated regions of a network. Given a undirected graph $G = (N, L)$, let L_k be the set of edges connected to a node k where $k \in N$. The isolation index of the node k is given by Q_k, which is the number of "infinite length paths" between any pair of nodes in the graph $H = (N, (L - L_k))$. This is equivalent to the number of disconnected pairs of nodes (i.e. there does not exist a path joining one node to another). The intuition behind the Isolation index is the following: it quantifies the inaccessibility in the network when a specific element (node/edge) is disconnected, like in a flooding or construction work.

While the V_k has the efficiency as the main ingredient, the Q_k is concerned with inaccessibility. Whenever removing the edges of a node, the former index quantifies the impact on existing routes, while the latter accounts for the number of isolated areas, making it easier to find bottlenecks in the network. Despite their differences, both capture structural vulnerabilities in complex networks, which have application in diverse critical infrastructures, ranging from mobility networks to power grids.

2 Results

Table 1 presents the vulnerability and isolation in five network topologies with seven nodes each.

Table 1. Vulnerability and isolation indexes applied to five network topologies.

Index	Isolation							Vulnerability						
node	0	1	2	3	4	5	6	0	1	2	3	4	5	6
tree	30.0	30.0	30.0	12.0	12.0	12.0	12.0	0.570	0.620	0.620	0.239	0.239	0.239	0.239
star	42.0	12.0	12.0	12.0	12.0	12.0	12.0	1.00	0.259	0.259	0.259	0.259	0.259	0.259
complete	12.0	12.0	12.0	12.0	12.0	12.0	12.0	0.286	0.286	0.286	0.286	0.286	0.286	0.286
line	12.0	22.0	28.0	30.0	28.0	22.0	12.0	0.220	0.425	0.522	0.522	0.522	0.425	0.220
ring	12.0	12.0	12.0	12.0	12.0	12.0	12.0	12.0	0.322	0.322	0.322	0.322	0.322	0.322

Figure 1 presents an example of the isolation index revealing the bottlenecks of a small-world network, while the vulnerability index labels the nodes more smoothly. Notice the differences in the outcomes of nodes 30 and 18 in both subfigures.

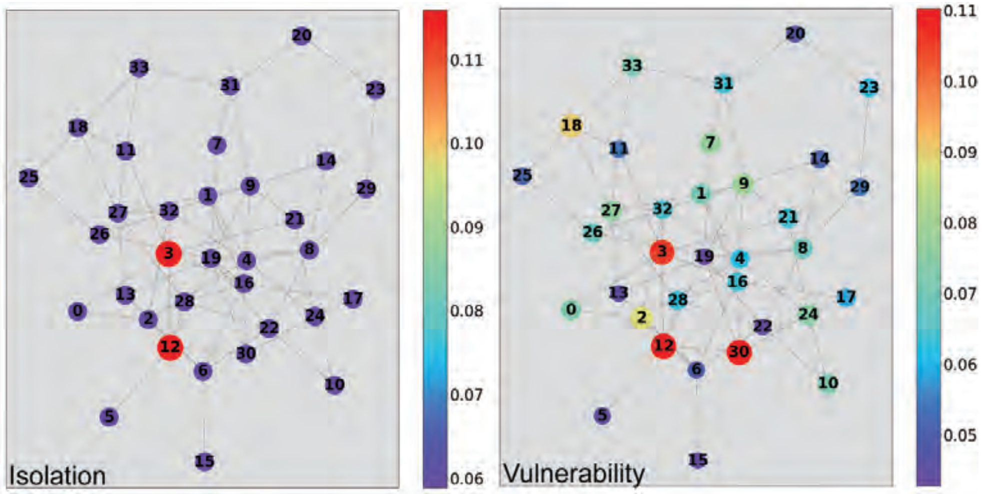


Fig. 1. Heat map for isolation and vulnerability in a Watts-Strogatz network produced with the igraph module in Python. Parameters: dimension= 1, size= 34, nei= 2, $p = 0.8$, and seed= 8666.

References

1. Wisner, B., Blaikie, P., Cannon, T., Davis, I.: At risk: natural hazards, people's vulnerability and disasters. 1st edition, 1994.

2. Santos, L. B. L., Londe, L. R., Carvalho, T., Menasche, D. S., Vega-Oliveros, D. A.: About Interfaces Between Machine Learning, Complex Networks, Survivability Analysis, and Disaster Risk Reduction. In Towards Mathematics, Computers and Environment: A Disasters Perspective, eISBN 978-3-030-21205-6, Springer-Nature, 2019.

Reconstructed potentials to characterize complex networks

Nicola Amoroso[1,2*], Loredana Bellantuono[3*], Saverio Pascazio[3,2], Alfonso Monaco[2], and Roberto Bellotti[3,2]

[1] Dipartimento di Farmacia-Scienze del Farmaco, Università degli studi di Bari “A. Moro”, I-70125 Bari, Italy
[2] Istituto Nazionale di Fisica Nucleare, Sezione di Bari, I-70126 Bari, Italy
[3] Dipartimento Interateneo di Fisica “M. Merlin”, Università degli studi di Bari “A. Moro”, I-70126 Bari, Italy
*Equal first author contribution.
loredana.bellantuono@ba.infn.it

1 Introduction

The investigation of complex network spectral properties represents a valuable tool to disclose hidden similarities and significant differences among graphs, especially in relation to their connectivity. We outline a novel quantum-inspired approach [1] to characterize complex networks based on the spectrum of the normalized network Laplacian, also known as the *graph spectrum*. We relate the network to a one-dimensional Schrödinger equation, such that its eigenvalues coincide with the graph spectrum. The potential that defines such equation, reconstructed by applying dressing transformations [2], provides a compact representation of the network properties, in particular those related to connectivity. To test the effectiveness of this new approach, we apply it to a well-known testbed in complex network theory, represented by Erdös and Rényi (ER) random networks [3]. At a critical value of the connection probability between pairs of nodes, ER networks are characterized by a phase transition, related to the emergence of a giant component. Our analysis tools, based on the reconstructed potentials, are able to capture the singular behavior of the network close to the transition. Specifically, we identify indicators of such criticality in the properties of the pointwise median potential, computed on several realizations of the ER network with the same size and connection probability. The main outcome is that the critical behavior of the network coincides with the emergence of a fractal structure in the median potential profile. We finally check the validity of the proposed approach by reconstructing the potentials corresponding to randomly subsampled real-world networks.

2 Results

A graph $\mathscr{G} = (\mathscr{N}, \mathscr{E})$ is defined through a set $\mathscr{N}$ of N nodes and a set $\mathscr{E}$ of edges that connect nodes to each other. Here, we focus on undirected and unweighted graphs and we also rule out the possibility of loops, namely edges connecting a node to itself. The spectrum of the normalized Laplacian $\mathscr{L}$ encodes information on the number and

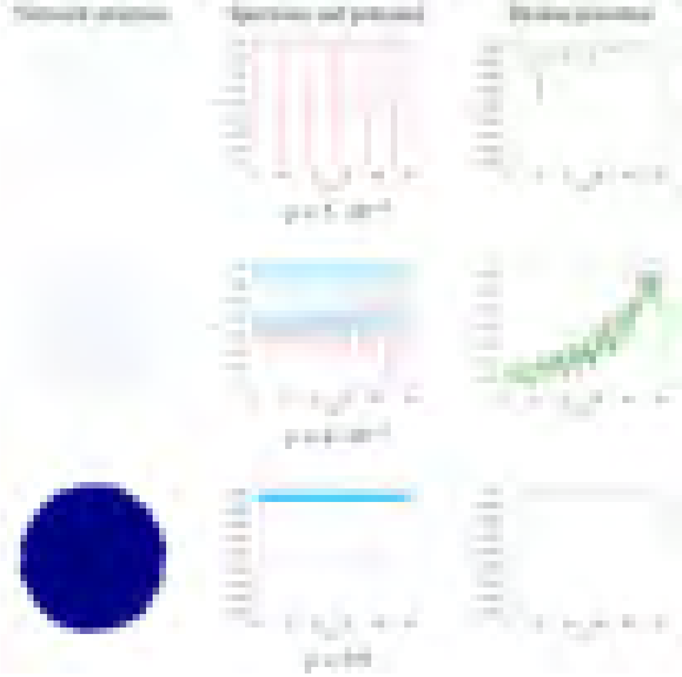


Fig. 1. Left column: representation of the network structure of a random ER network with $N = 250$ nodes and link probability p; central column: shifted graph spectrum (cyan) and reconstructed potential (red); right column: median potential, computed over an ensemble of 100 networks with the same N and p. The reported link probabilities p are below (top panels), equal to (central panels) and above (bottom panels) the critical value $p_c = 1/(N-1)$.

size of connected components and the presence of peculiar structures in the network. The methodology proposed in Ref. [1] introduces a compact and comprehensive representation of this information by associating a given network with a 1D Schrödinger equation. Specifically, we consider the spectrum of $\mathscr{L}$: $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N \leq 2$ and reconstruct the potential $V(x)$ with the constraint that the shifted graph eigenvalues $E_n = \lambda_n - \lambda_N$, belonging to $[-2,0]$, are energy levels of the Schrödinger equation $-\partial_x^2 \psi(x) + V(x)\psi(x) = E_n \psi(x)$ with $\hbar = 1, m = 1/2$. The potential is determined with an iterative N-step procedure, based on the application of dressing transformations [2].

The main testbed of our method is represented by the Erdös and Rényi (ER) network model [3]. Specifically, we focus on the formulation proposed by Gilbert [4], in which each pair of nodes can be linked with probability p irrespective of the connections in the rest of the network. Hence, the connectivity is quantified by the *average degree* $\langle k \rangle = p(N-1)$, where N is the network size. A phase transition is known to occur at the critical probability $p_c = 1/(N-1)$ (corresponding to $\langle k \rangle = 1$), related to the onset of percolation and to a change in the scaling behavior of the Largest Connected Component (LCC) size, which is logarithmic in N for $p < p_c$ and becomes linear in N for $p > p_c$ [3,4]. To understand the relation between the network structure, the graph spectrum and the reconstructed potential, we consider 100 realizations of the ER complex network with $N = 250$ nodes at three different values of p, and report the most relevant results in Fig. 1. Below the critical probability (top panels), the sampled networks are almost disconnected, typically characterized by sparse links, and the lowest shifted graph eigenvalue $E_1 = -2$ is highly degenerate, since its multiplicity coincides with the number of disjoint graph components; the potential reconstructed for a given network realization is characterized by evenly spaced wells. Far above the critical probability (bottom panels), the ER ensemble is mostly composed of single-component graphs, with a high density of links: in this case, the potentials are characterized by a single minimum and a rapid increase towards an almost constant value. Then, for both low and high p, the statistical variability of the reconstructed potentials is small, leading to a rather smooth profile of the function $V_m(x)$, defined as the pointwise median of the set of potentials $\{V_i(x)\}_{i=1,...,100}$ associated to all the network realizations. As one approaches $p = p_c$ (central panels) either from above or from below, the network structures and the associated spectra show a larger statistical variability: as a result, the profile of the median $V_m(x)$ becomes very irregular. We quantify the ruggedness of the median potential at different p through the Higuchi Fractal Dimension (HFD) [5], finding that a peak in this quantity is attained at the critical connection probability [1]. Such a result suggests the possibility to effectively employ our technique in the analysis of random networks, as an alternative probe of the percolation phase transition. Finally, we apply our methods to a real-world network, namely the US power grid [6]. In this framework the ensemble of graphs with a fixed average degree is obtained by randomly subsampling nodes from the original network and retaining links connecting them. Although no *bona fide* phase transition can be properly identified in this case, we show that the HFD of the median potential displays a behavior that is reminiscent of criticality, unlike standard indicators such as the size of the LCC [1].

Summary. We introduce a characterization of complex networks based on associating the potential of a Schrödinger equation to the graph spectrum. The reconstructed potential provides a compact representation of network properties. In particular, this technique is able to detect the percolation phase transition in Erdös-Rényi random networks in terms of the fractality of the ensemble median potential. The application to a randomly subsampled real-world network is finally performed to test the method.

References

1. Amoroso N., Bellantuono, L., Pascazio, S., Lombardi A., Monaco, A. Tangaro, S., Bellotti, R.: Potential energy of complex networks: a novel perspective, arXiv preprint arXiv:2002.04551 (2020).
2. Spiridonov, V.: Exactly solvable potentials and quantum algebras. Phys. Rev. Lett. 69(3), 398—401 (1992)
3. Erdös, P., Rényi, A.: On random graphs, I. Publ. Math. (Debrecen) 6, 290–297 (1959)
4. Gilbert, E.N.: Random graphs. The Annals Math. Stat. 30(4), 1141-–1144 (1959)
5. Higuchi, T.: Approach to an irregular time-series on the basis of the fractal theory. Physica D 31(2), 277-–283 (1988).

6. Kunegis, J.: KONECT – The Koblenz Network Collection. In Proc. Int. Conf. on World Wide Web Companion, p. 1343-–1350 (2013).

Part X

Network Models

Monotone Properties in Random Networks with Dependent Edges

András Faragó

The University of Texas at Dallas, Richardson, Texas 75080, USA,
farago@utdallas.edu,
WWW home page: http://www.utdallas.edu/~farago

Random networks occur in many practical scenarios. The oldest and most researched model of random networks is the Erdős-Rényi random graph $G_{n,p}$. This denotes a random graph on n nodes, such that each edge is added with probability p, and it is done independently for each edge. Many deep results are available about such random graphs, see expositions in the books [1,3,4]. However, the requirement that the edges are *independent* is often a severe restriction in modeling real-life networks. Therefore, numerous attempts have been made to develop models with various dependencies among the edges, see a survey in [2]. Here we consider a general form of edge dependency. We call a random graph with this type of dependency a *p-robust random graph.*

Definition 1. **(p-robust random graph)** *A random graph on n vertices is called p-*robust*, if every edge is present with probability at least p, regardless of the status (present or not) of other edges. Such a random graph is denoted by $\widetilde{G}_{n,p}$.*

Note that the classical Erdős-Rényi random graph is a special case of our model. However, we also allow (possibly messy) dependencies. For example, let $P(e)$ denote the probability that a given edge e is present in the graph, and let us condition on k, the number of other edges in the whole graph. For any fixed k, set $P(e|k) = 1 - k/n^2$. Since $k \leq n(n-1)/2$ always holds, therefore, $P(e|k) \geq 1 - \frac{n(n-1)}{2n^2} = 1 - \frac{n-1}{2n} \geq \frac{1}{2}$, for any k, implying $p(e) = \sum_{k=0}^{n(n-1)/2} p(e|k)p(k) \geq 1/2$. Thus, with $p = 1/2$, this random graph is p-robust. At the same time, the edges are not independent, since the probability that e is present depends on how many other edges are present.

Let Q be a set of graphs. We use it to represent a graph property: a graph G has property Q if and only if $G \in Q$. Therefore, we identify the property with Q. We are going to consider *monotone graph properties,* as defined below.

Definition 2. **(Monotone graph property)** *A graph property Q is called* monotone, *if it is closed with respect to adding new edges. That is, if $G \in Q$ and $G \subseteq G'$, then $G' \in Q$.*

Many important graph properties are monotone. Our result is that for any monotone graph property, and for any n, p, it always holds that $\widetilde{G}_{n,p}$ is more likely to have the property than $G_{n,p}$ (or at least as likely). This is very useful, as it allows the application of the rich treasury of results on Erdős-Rényi random graphs to the non-independent setting, as lower bounds on the probability of having a monotone property.

Theorem 1. *Let Q be a monotone graph property. Then the following holds:*

$$\Pr(G_{n,p} \in Q) \;\leq\; \Pr(\widetilde{G}_{n,p} \in Q).$$

Proof. We are going to generate $\widetilde{G}_{n,p}$ as the union of two random graphs, $G_{n,p}$ and G_2, both on the same vertex set V. $G_{n,p}$ is the usual Erdős-Rényi random graph, G_2 will be defined later. The union $G_{n,p} \cup G_2$ is meant with the understanding that if the same edge occurs in both graphs, then we merge them into a single edge. We plan to chose the edge probabilities in G_2, such that $G_{n,p} \cup G_2 \sim \widetilde{G}_{n,p}$, where the "$\sim$" relation between random graphs means that they have the same distribution, i.e., they are statistically indistinguishable. If this can be accomplished, then the claim will directly follow, since then a random graph distributed as $\widetilde{G}_{n,p}$ can be obtained by adding edges to $G_{n,p}$, which cannot destroy a monotone property, once $G_{n,p}$ has it. This will imply the claim.

We introduce some notations. Let $e_1,\dots,e_m$ denote the (potential) edges. For every i, let h_i be the indicator of the event that the edge e_i is included in $\widetilde{G}_{n,p}$. Further, let us use the abbreviation $h_i^m = (h_i,\dots,h_m)$. For any $a = (a_1,\dots,a_m) \in \{0,1\}^m$, the event $\{h_1^m = a\}$ means that $\widetilde{G}_{n,p}$ takes a realization in which edge e_i is included if and only $a_i = 1$. Similarly, $\{h_i^m = a_i^m\}$ means $\{h_i = a_i,\dots,h_m = a_m\}$. We also use the abbreviation $a_i^m = (a_i,\dots,a_m)$. Now let us generate the random graphs $G_{n,p}$ and G_2, as follows.

Step 1. Let $i = m$.
Step 2. If $i = m$, then let $q_m = \Pr(h_m = 1)$. If $i < m$, then set $q_i = \Pr(h_i = 1 \mid h_{i+1}^m = a_{i+1}^m)$, where a_{i+1}^m indicates the already generated edges of $G_{n,p} \cup G_2$.
Step 3. Compute

$$p_i' = \frac{p(1-q_i)}{1-p}. \tag{1}$$

Step 4. Put e_i into $G_{n,p}$ with probability p, and put e_i into G_2 with probability $q_i - p_i'$.
Step 5. If $i > 1$, then decrease i by one, and go to Step 2; else HALT.

First note that the value $q_i - p_i'$ in Step 4 can indeed be used as a probability. Clearly, $q_i - p_i' \le 1$ holds, as q_i is a probability and $p_i' \ge 0$. To show $q_i - p_i' \ge 0$, observe that $p_i' = \frac{p(1-q_i)}{1-p} \le q_i$, since the inequality can be rearranged into $p(1-q_i) \le q_i(1-p)$, which simplifies to $p \le q_i$. The latter is indeed true, due to $q_i = \Pr(h_i = 1 \mid h_{i+1}^m = a_{i+1}^m) \ge p$, which follows from the p-robust property.

Next we show that the algorithm generates the random graphs $G_{n,p}$ and G_2 in a way that they satisfy $G_{n,p} \cup G_2 \sim \widetilde{G}_{n,p}$. We prove it by induction, starting from $i = m$ and progressing downward to $i = 1$. For any i, let $G_{n,p}^i$, G_2^i denote the already generated parts of $G_{n,p}, G_2$, respectively, after executing Step 4 $m-i+1$ times, so they can only contain edges with index $\ge i$. Further, let $\widetilde{G}_{n,p}^i$ be the subgraph of $\widetilde{G}_{n,p}$ in which we only keep the edges with index $\ge i$, that is, $\widetilde{G}_{n,p}^i = \widetilde{G}_{n,p} - \{e_{i-1},\dots,e_1\}$. The inductive proof will show that $G_{n,p}^i \cup G_2^i \sim \widetilde{G}_{n,p}^i$ holds for every i. At the end of the induction, having reached $i = 1$, we are going to get $G_{n,p}^1 \cup G_2^1 \sim \widetilde{G}_{n,p}^1$, which is the same as $G_{n,p} \cup G_2 \sim \widetilde{G}_{n,p}$.

Let us consider first the base case $i = m$. Then we have $\Pr(e_m \in G_{n,p}) = \Pr(e_m \in G_{n,p}^m) = p$ by Step 4. Then in Step 4, edge e_m is put into G_2 with probability $q_m - p_m'$, yielding $\Pr(e_m \in G_2^m) = q_m - p_m'$. Now observe that the formula (1) is chosen such that p_i' is precisely the solution of the equation

$$p + q_i - p_i' - (q_i - p_i')p = q_i \tag{2}$$

for p'_i. For $i = m$ the equation becomes

$$p + q_m - p'_m - (q_m - p'_m)p = q_m, \tag{3}$$

and $p'_m = \frac{p(1-q_m)}{1-p}$ is the solution of this equation. Since by Step 4 we have $\Pr(e_m \in G^m_{n,p}) = p$ and $\Pr(e_m \in G^m_2) = q_m - p'_m$, therefore, we get that the left-hand side of (3) is precisely the probability of the event $\{e_m \in G^m_{n,p} \cup G^m_2\}$. By (3), this probability is equal to q_m, which is set to $q_m = \Pr(h_m = 1) = \Pr(e_m \in \widetilde{G}^m_{n,p})$ in Step 2. This means that $G^m_{n,p} \cup G^m_2 \sim \widetilde{G}^m_{n,p}$, as desired.

For the induction step, assume that the claim is true for $i+1$, i.e., $G^{i+1}_{n,p} \cup G^{i+1}_2 \sim \widetilde{G}^{i+1}_{n,p}$ holds. In Step 4, edge e_i is added to $G^{i+1}_{n,p}$ with probability p. It is also added to G^{i+1}_2 with probability $q_i - p'_i$. Therefore, just like in the base case, we get that $p + q_i - p'_i - (q_i - p'_i)p = \Pr(e_i \in G^i_{n,p} \cup G^i_2)$. We already know that p'_i satisfies the equation (2), so e_i is added to $\widetilde{G}^{i+1}_{n,p}$ with probability $q_i = \Pr(h_i = 1 \mid h^m_{i+1} = a^m_{i+1})$, given the already generated part, represented by a^m_{i+1}. By the inductive assumption, h^m_{i+1} is distributed as $\widetilde{G}^{i+1}_{n,p}$, which is the truncated version of $\widetilde{G}_{n,p}$, keeping only the $\geq i+1$ indexed edges. Hence, for h^m_{i+1}, we can write by the chain rule of conditional probabilities:

$$\Pr(h^m_{i+1} = a^m_{i+1}) = \Pr(h_m = a_m) \prod_{j=i+1}^{m-1} \Pr(h_j = a_j \mid h^m_{j+1} = a^m_{j+1}).$$

After processing e_i (i.e., adding it with probability q_i), we get

$$\begin{aligned}
\Pr(h^m_i = a^m_i) &= \Pr(h_i = a_i \mid h^m_{i+1} = a^m_{i+1}) \Pr(h^m_{i+1} = a^m_{i+1}) \\
&= \Pr(h_i = a_i \mid h^m_{i+1} = a^m_{i+1}) \Pr(h_m = a_m) \prod_{j=i+1}^{m-1} \Pr(h_j = a_j \mid h^m_{j+1} = a^m_{j+1}) \\
&= \Pr(h_m = a_m) \prod_{j=i}^{m-1} \Pr(h_j = a_j \mid h^m_{j+1} = a^m_{j+1}),
\end{aligned}$$

which, by the chain rule, is indeed the distribution of $\widetilde{G}^i_{n,p}$, completing the induction.

Thus, at the end, a realization $a = a^m_1 \in \{0,1\}^m$ of $\widetilde{G}_{n,p}$ is generated with probability $\Pr(h^m_1 = a) = \Pr(h_m = a_m) \prod_{j=1}^{m-1} \Pr(h_j = a_j \mid h^m_{j+1} = a^m_{j+1})$, indeed creating $\widetilde{G}_{n,p}$ with its correct probability. Therefore, we get $G_{n,p} \cup G_2 \sim \widetilde{G}_{n,p}$, so $\widetilde{G}_{n,p}$ arises by adding edges to $G_{n,p}$, which cannot destroy a monotone property. This implies the claim. ♠

References

1. B. Bollobás, *Random Graphs,* Cambridge University Press, 2001.
2. A. Faragó, "Network Topology Models for Multihop Wireless Networks," *ISRN Communications and Networking,* Vol. 2012, Article ID 362603, doi:10.5402/2012/362603
3. A. Frieze and M. Karoński, *Introduction to Random Graphs,* Cambridge Univ. Press, 2016.
4. S. Janson, T. Luczak, and A. Rucinski, *Random Graphs,* Wiley-Interscience, 2000.

Synchronization in complex networks with long-range interactions

Sarbendu Rakshit, Soumen Majhi, and Dibakar Ghosh

Physics and Applied Mathematics Unit, Indian Statistical Institute, 203 B. T. Road,
Kolkata-700108, India,
sarbendu.math@gmail.com

1 Introduction

In the work Ref. [1], we assume that each node in the network is associated with a d-dimensional dynamical system. Then the nodal dynamics of the i-th node in the network possessing long-range interaction can be described as follows,

$$\dot{\mathbf{x}}_i = f(\mathbf{x}_i) + \sum_{k=1}^{d_{max}} \epsilon_k \sum_{j=1}^{N} \mathscr{A}_{ij}^{[k]} \Gamma(\mathbf{x}_j - \mathbf{x}_i), \tag{1}$$

where $\mathbf{x}_i$ represents the d-dimensional state variable of the i-th node, $f : \mathbb{R}^d \to \mathbb{R}^d$ describes the dynamics of each isolated node, ϵ_k is the coupling strength between the i-th and j-th nodes if $d(i,j) = k$. $\mathscr{A}^{[k]}$ is the k-path adjacency matrix and Γ is the inner coupling matrix determining the state variables through which the nodes are interacting with each other.

When complete synchronization occurs in the dynamical network (1), then there exists a trajectory $\mathbf{x}_0 \in \mathbb{R}^d$, such that, for each $\epsilon > 0$ (however small) there exists $T > 0$ (however large), $\|\mathbf{x}_i(t) - \mathbf{x}_0(t)\| < \epsilon$ whenever $t \geq T$. We call the subset $\mathcal{S} = \left\{\mathbf{x}_0 \subset \mathbb{R}^d : \mathbf{x}_i = \mathbf{x}_0, \ \forall \ i = 1, 2, \dots, N\right\}$ as the complete synchronization manifold. Our aim of this paper is to determine the local and global stability of $\mathcal{S}$ in terms of the coupling and network parameters.

2 Local stability analysis

At first, we will analytically determine the local stability condition of the synchronization state for the coupled system (1) using the seminal MSF approach [2]. In this purpose, we will assume that, the individual nodal dynamics f is continuously differentiable with respect to its argument.

Theorem 1. *The parallel and transverse components along the synchronous solution respectively satisfy the system of equations,*

$$\dot{\eta}_P(t) = Jf(\mathbf{x}_0)\eta_P(t), \tag{2a}$$

$$\dot{\eta}_{T_i}(t) = \Big[Jf(\mathbf{x}_0) - \epsilon_1 {\gamma_i}^{[1]} \Gamma\Big]\eta_{T_i} - \sum_{k=2}^{d_{max}} \epsilon_k \sum_{j=1}^{N-1} {U_{ij}}^{[k]} \Gamma \eta_{T_j}, \tag{2b}$$

where, $i = 2, 3, \ldots, N$ and Jf denotes the Jacobian of the function f. $\eta_P(t) \in \mathbb{R}^d$ and $\eta_T(t) \in \mathbb{R}^{d(N-1)}$ are the state vectors which evolve parallel and transverse to the synchronization solution, respectively.

The above equation (2b) is the required transverse master stability equation (MSE) for the complete synchronization solution. It is $(N-1)d$-dimensional coupled equation. In general, this error system cannot be further reduced to low-dimensional form. Also, generally, it is not directly dependent on the eigenvalues of the k-path Laplacians.

We can measure the exponential contraction or expansion of the linearized variational equation by calculating its Lyapunov exponents. Among all the Lyapunov exponents, the maximum one (say Λ) plays a key role. If Λ is less than zero, the complete synchronization state turns out to be locally stable, while its positive value indicates the instability of the synchronization state. By adjusting the tuning parameters (coupling as well as network parameters), we can trace-out the synchronization region where the value of Λ is negative.

3 Emergence of complete synchrony: Numerical results

Without loss of generalization, we choose the chaotic Lorenz system [3] described as $\dot{x} = \sigma(y - x),\ \dot{y} = x(\rho - z) - y,\ \dot{z} = xy - \beta z$. We fix the system parameters at $\sigma = 10,\ \rho = 28$ and $\beta = 8/3$ for which the system remains in chaotic state. We also choose $\Gamma = diag[1, 1, 1]^{tr}$. If ϵ_k is the coupling strength between the nodes having shortest distance k ($k = 1, 2, \ldots, d_{max}$), then $\epsilon_k = \epsilon/k^{\alpha}$ where α is the power-law exponent governing the decay rate. Here we consider the underlying network as the ER random network architecture. Specifically, we choose the $G(N, p)$ graph model [4] with $N = 200$ as the number of nodes and p as the connection probability. We define the synchronization error as $E = \lim_{T\to\infty} \frac{1}{T} \int_t^{t+T} \sum_{i,j=1(i\neq j)}^{N} \frac{\|\mathbf{x}_i(t) - \mathbf{x}_j(t)\|}{N(N-1)}\, dt$. As conspicuous from its definition, in the state of complete synchronization, E necessarily becomes zero, and remains non-zero for the states of desynchrony.

We drawn the variation in the synchronization error E as a function of the interaction strength ϵ for different values of the power-law exponent α and connection probability p in Fig. 1(a) for various values of α. In Fig. 1(b), the transition from desynchrony to synchrony is characterized by the maximum transverse Lyapunov exponent Λ with respect to ϵ for the four different values of α and p as of Fig. 1(a). For these four cases, Λ crosses zero exactly at that point where the synchronization error becomes zero in Fig. 1(a), which indicates that our analytical local stability condition agrees well with our numerical simulations of the synchronization error plot.

4 Global stability analysis

It will strengthen the findings if the condition under which the synchronization solution is globally stable could be derived. The next theorem deals with the

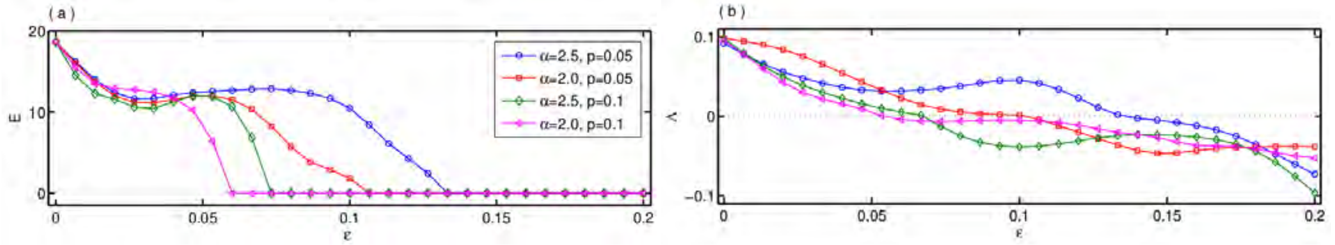


Fig. 1: Variation of (a) E and (b) Λ with respect to the coupling strength ϵ for different combinations of α and p. Here the pairs ($\alpha = 2.5,\ p = 0.05$), ($\alpha = 2.0,\ p = 0.05$), ($\alpha = 2.5,\ p = 0.1$) and ($\alpha = 2.0,\ p = 0.1$) are chosen and are shown by the blue circle, red square, green diamond, and magenta triangle lines, respectively.

global stability condition of the complete synchronization state with some mild assumptions.

Theorem 2. *Given a connected* 1*-path network of N nodes described by (1) with the following assumptions:*

1. *the isolate evolution function f satisfies the global Lipschitz condition. So there exists a non-negative constant M such that for any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\|f(\mathbf{x}) - f(\mathbf{y})\| \leq M \|\mathbf{x} - \mathbf{y}\|$,*
2. *the inner coupling matrix Γ is a symmetric positive definite matrix. That is, if $\{\mu_1, \mu_2, \ldots, \mu_d\}$ be the set of eigenvalues of Γ then $\mu_j > 0$ for all $j = 1, 2, \ldots, d$.*

Then, if $\sum\limits_{k=1}^{d_{max}} \epsilon_k \lambda_2[\mathscr{L}^{[k]} \otimes \Gamma] > M$, the complete synchronization state of the dynamical network (1) will be globally asymptotically stable.

Remark 1. For power-law decaying rate, the coupling strength for shortest-distance k is $\epsilon_k = \epsilon/k^\alpha$. Additionally, for the assumed inner coupling matrix Γ, all of its three eigenvalues are 1, so $\lambda_2[\mathscr{L}^{[k]} \otimes \Gamma] = \lambda_2[\mathscr{L}^{[k]}]$. Therefore, the global stability condition of Eq. (1) is $\epsilon > \frac{M}{\sum\limits_{k=1}^{d_{max}} \frac{\lambda_2[\mathscr{L}^{[k]}]}{k^\alpha}}$. Beyond this coupling strength, all the oscillators in the network converge toward the identical trajectory, irrespective of the initial conditions.

5 Conclusion

In this work, we have presented our results on the manifestation of complete synchronization in ER random network subject to long-range interactions. In order to describe the impact of long-range couplings, we have chosen the decaying interaction strength among the nodes with reference to the power-law, while considering the paradigmatic Lorenz systems for casting the nodes in the network. We have shown how the appearance of synchrony depends on the variation of the coupling strength and the power-law exponent. More importantly, we have provided comprehensive analysis on the stability of the obtained synchronization solution. Besides a thorough investigation of local stability based on the master stability function approach, we also presented global stability analysis for appropriate choice of Lyapunov function.

References

1. Rakshit S., Majhi S., & Ghosh D., J. Phys. A: Math. Theor. **53**, 154002 (2020).
2. Pecora, L. M., & Carroll, T. L. *Phys. Rev. Lett.* **80** 2109-2112 (1998).
3. Lorenz E. N., *J. Atmos. Sci.* **20** 130 (1963).
4. Erdös P., & Rényi A., *Publ. Math. Debrecen* **6** 290 (1959).

Statistical properties of edges and bredges in configuration model networks

Ofer Biham[1], Haggai Bonneau[1] Eytan Katzav[1], and Reimer Kühn[2]

[1] Racah Institute of Physics, The Hebrew University, Jerusalem 9190401, Israel
biham@phys.huji.ac.il,
http://cond-mat.phys.huji.ac.il/ofer/
[2] Mathematics Department, King's College London, Strand, London WC2R 2LS, UK

A bredge (bridge-edge) in a network is an edge whose deletion would split the network component on which it resides into two separate components. Since the integrity of most networks (particularly transportation and communication networks) is essential for their functionality, bredges are vulnerable links that play an important role in network collapse processes. Therefore, the abundance and properties of bredges affect the resilience of the network to both inadvertent failures and deliberate attacks. We present analytical results for the statistical properties of bredges in configuration model networks [1]. Using a generating function approach based on the cavity method, we calculate the probability $\widehat{P}(e \in \mathrm{B})$ that a random edge e in a configuration model network with degree distribution $P(k)$ is a bredge (B). We examine the distinct properties of bredges on the giant component (GC) and on the finite tree components (FC) of the network. On the finite components all the edges are bredges and there are no degree-degree correlations.

In Fig. 1 we present an Erdős-Rényi (ER) network of $N = 100$ nodes with mean degree $c = 1.7$. The giant component of this network coexists with many finite tree components. The non-bredge edges (solid lines) connect pairs of nodes that reside on the 2-core of the giant component [2]. The giant component is decorated by tree branches, on which all the edges are bredges. The bredge that connects each tree branch to the 2-core of the giant component is called root bredge (dashed line). The end-node of the root bredge that resides on the 2-core is called root end-node. All the other bredges (dotted lines) connect pairs of nodes that reside on the tree branches, which are not on the 2-core. The distinction between root bredges and all the other bredges on the giant component may be useful for optimized dismantling algorithms [3] and targeted attacks [4, 5]. This is due to the fact that the deletion of a root bredge disconnects the whole tree branch that is held by this bredge. In contrast, random deletion of bredges may require a large number of deletion steps in order to chop each tree branch from the 2-core of the giant component.

In Fig. 2 we present analytical reults for the probability $\widehat{P}(e \in \mathrm{B})$ (solid line) that a randomly sampled edge in an ER network is a bredge as a function of the mean degree c. The probability $\widehat{P}(e \in \mathrm{B})$ can be expressed as a sum of two components: the probability $\widehat{P}(e \in \mathrm{B}, \mathrm{GC})$ (dashed line) that a randomly sampled edge is a bredge that resides on the giant component, and the probability $\widehat{P}(e \in \mathrm{B}, \mathrm{FC})$ (dotted line) that a randomly sampled edge is a bredge that resides on one of the finite components. The analytical results (solid, dashed and dotted lines) are in excellent agreement with the results of

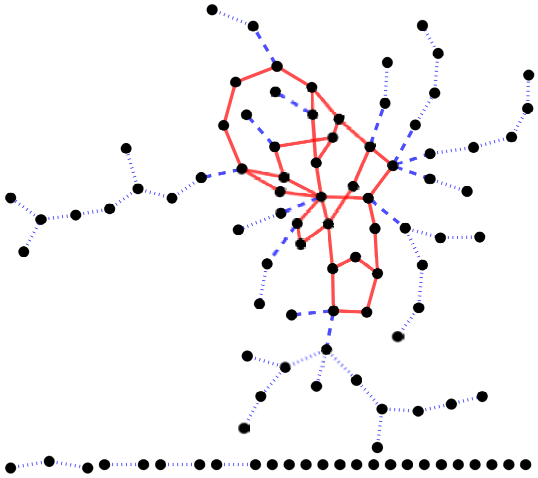


Fig. 1. The structure of an instance of an ER network of $N = 100$ nodes with mean degree $c = 1.7$, which exhibits a coexistence between a giant component and finite tree components. The non-bredge edges (solid lines) connect pairs of nodes that reside on the 2-core of the giant component. The 2-core exhibits a complex web of cycles. The root bredges (dashed lines) connect the tree branches on the giant component to the 2-core. All the other bredges (dotted lines) connect pairs of nodes on that reside on the tree branches of the giant component and pairs of nodes on the finite tree components.

computer simulations (circles), performed for an ensemble of ER networks of $N = 10^4$ nodes.

In order to analyze degree-degree correlations between the end-nodes of bredges, we calculate the joint degree distribution $\widehat{P}(k,k'|\mathrm{B})$ of the end-nodes i and i' of a random bredge. We also calculate the joint degree distribution $\widehat{P}(k,k'|\mathrm{B},\mathrm{GC})$ of the end-nodes of bredges and the joint degree distribution $\widehat{P}(k,k'|\mathrm{NB},\mathrm{GC})$ of the end-nodes of non-bredge (NB) edges on the giant component. Surprisingly, it is found that the degrees k and k' of the end-nodes of bredges are correlated, while the degrees of the end-nodes of non-bredge edges are uncorrelated. We thus conclude that all the degree-degree correlations on the giant component are concentrated on the bredges. We calculate the degree-degree correlation function $\Gamma(\mathrm{B},\mathrm{GC})$ between the end-nodes of bredges, also called the assortativity coefficient [6], and show it is negative, namely bredges tend to connect high degree nodes to low degree nodes. We apply this analysis to ensembles of configuration model networks with degree distributions that follow a Poisson distribution (ER networks), an exponential distribution and a power-law distribution (scale-free networks).

The properties of bredges in a wide range of real-world empirical networks were recently studied [7]. The fraction of bredges in each empirical network was calculated using an algorithm based on depth-first search. An ensemble of configuration model networks, whose degree distribution coincides with the degree sequence of the empirical network, was generated using degree-preserving randomization. The fraction of

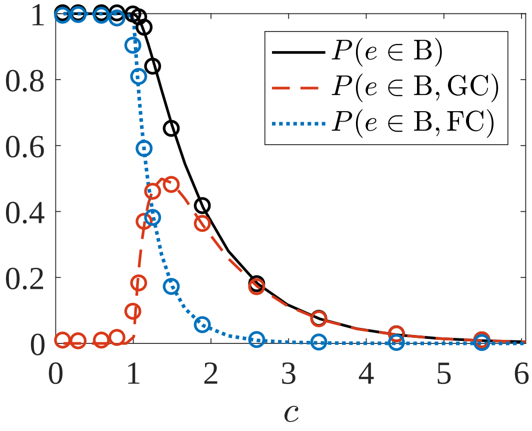


Fig. 2. The probability $\widehat{P}(e \in \mathrm{B})$ (solid line) that a randomly sampled edge in an ER network is a bredge, as a function of the mean degree c; The probability $\widehat{P}(e \in \mathrm{B})$ is equal to the sum of two components: the probability $\widehat{P}(e \in \mathrm{B}, \mathrm{GC})$ (dashed line) that a randomly sampled edge is a bredge that resides on the giant component, and the probability $\widehat{P}(e \in \mathrm{B}, \mathrm{FC})$ (dotted line) that a randomly sampled edge is a bredge that resides on one of the finite components. The analytical results (solid, dashed and dotted lines) are in excellent agreement with the results of computer simulations (circles), performed for an ensemble of ER networks of $N = 10^4$ nodes.

bredges in each ensemble was calculated both numerically and using a generating function formalism. It was found that the fraction of bredges in the randomized ensembles is very similar to their fraction in the corresponding empirical networks. This indicates that the information about the number of bredges is captured in the degree distribution. Thus, configuration model networks are likely to provide useful predictions for the statistical properties of bredges in empirical networks and their effect on the resilience or vulnerability of these networks to failures and attacks.

References

1. H. Bonneau, O. Biham, R. Kühn and E. Katzav, Statistical analysis of edges and bredges in configuration model networks, *Phys. Rev. E* **102**, 012314 (2020).
2. M.E.J. Newman and G. Ghoshal, Bicomponents and the robustness of networks to failure, *Phys. Rev. Lett.* **100**, 138701 (2008)
3. A. Braunstein, L. Dall'Asta, G. Semerjian and L. Zdeborová, Network dismantling. *Proc. Natl. Acad. Sci. USA* **113**, 12368 (2016).
4. R. Cohen, K. Erez, D. ben-Avraham and S. Havlin, Breakdown of the Internet under intentional attack, *Phys. Rev. Lett.* **86**, 3682 (2001).
5. X. Yuan, Y. Dai, H.E. Stanley and S. Havlin, *k*-core percolation on complex networks: Comparing random, localized, and targeted attacks, *Phys. Rev. E* **93**, 062302 (2016).
6. M.E. J.Newman, Assortative mixing in networks, *Phys. Rev. Lett.* **89**, 208701 (2002)
7. A.-K. Wu, L. Tian and Y.-Y. Liu, Bridges in complex networks, *Phys. Rev. E* **97**, 012307 (2018).

Latent Space Modelling of Hypergraph Data

Kathryn Turnbull [1], Simón Lunagómez [2], Christopher Nemeth [2],
Edoardo Airoldi [1]

[1] Fox School of Business, Temple University, Philadelphia, PA, USA
[2] Department of Mathematics and Statistics, Lancaster University, UK

1 Introduction

The ubiquity of network data describing interactions among a population has motivated a broad literature on statistical network analysis (see [Kolaczyk, 2009]). Whilst the existing literature is primarily concerned with pairwise interactions, there are many settings in which interactions occur between higher-order sets of the population. Data of this type are more appropriately represented as a hypergraph comprised of nodes, indexed by $V = \{1,2,\ldots,N\} = [N]$, and hyperedges $E = \{e \subseteq [N] |$ elements of e are unique$\}$. As a motivating example, consider a coauthorship network where nodes represent authors and an interaction occurs between authors when they have collaborated on an article together. Typically, more than two authors will contribute to an article and this can be naturally represented by a hyperedge.

Since its introduction in [Hoff et al., 2002], the latent space approach for network data has inspired a rich modelling literature. In this framework, the probability of an edge forming between each node pair is modelled as a function of low-dimensional latent coordinates associated with the nodes. The underlying geometry imposes desirable properties on the networks, such as transitive relationships, allows exploration of predictive distributions and provides a visualisation of the data. We consider extending this approach to the hypergraph setting, where we rely on tools from computational topology to develop a parsimonious model. Our approach models the hyperedges directly, and therefore avoids the loss of information associated with representing each hyperedge by a clique of pairwise relationships. In the context of coauthorship, this is made clear by noting that three authors writing a paper together is not equivalent to each pair of authors writing papers together. Throughout, we focus on the setting in which hyperedges of order $k \in \{2,3,\ldots,K\}$ are observed.

2 Proposed model

We propose the non-simplicial Random Geometric Hypergraph (nsRGH) model in which the probability of each hyperedge forming depends on the relative positions of latent coordinates $\boldsymbol{U} = \{u_i\}_{i=1}^N$, where $u_i \in \mathbb{R}^d$ is the coordinate for the i^{th} node. Our construction is based on the Cěch complex (S3.2 of [Edelsbrunner and Harer, 2010]) in which the order k hyperedge $e_k = \{i_1, i_2, \ldots, i_k\}$ is present if $\cap_{j \in e_k} B_r(u_j) \neq \emptyset$, where $B_r(u)$ denotes the ball of radius $r > 0$ and centre u.

The Cěch complex is a simplicial complex, meaning that the presence of the hyperedge e_k implies that all hyperedges described by subsets of e_k are also present. In our

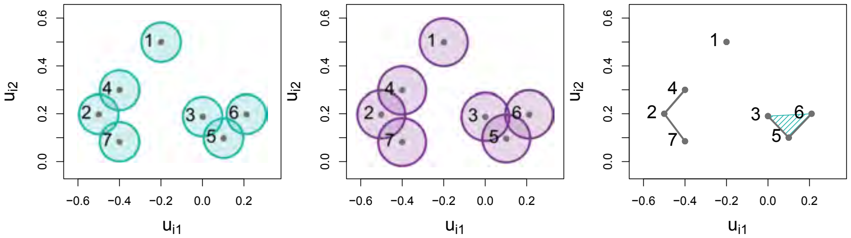


Fig. 1. Left: $\{B_{r_2}(u_i)\}_{i=1}^7$. Middle: $\{B_{r_3}(u_i)\}_{i=1}^7$ for $r_3 > r_2$. Right: nsRGH with no noise. The shaded region represents the hyperedge $\{3,5,6\}$.

motivating example of coauthorship, this property is typically not observed and so we extend this construction to the non-simplicial setting through the introduction of additional radii. More specifically, conditional on $\boldsymbol{U}$, we generate a hyperedges of order k according to the radii $r_k > 0$ and combine hyperedges of each order. Furthermore, we impose $r_k > r_{k-1}$ for $k \in \{3,4,\dots,K\}$ to ensure the hypergraph is non-simplicial. An example of this construction for $K = 3$ and $d = 2$ is given in Figure 1, where order 2 hyperedges from the left panel and order 3 hyperedges from the middle panel are combined in the right panel. Finally, to extend the support of the model and aid estimation, we allow the state of the order k hyperedges to be modified from present to absent, or vice versa, independently according to a small probability.

3 Identifiability, estimation and inference

Similarly to latent space network models, our nsRGH model exhibits non-identifiability from the distance-preserving transformations of $\boldsymbol{U}$. We propose to address this via Bookstein coordinates ([Bookstein, 1986], Section 2.3.3 of [Dryden and Mardia, 1998]), a concept from shape theory, in which a subset of the latent coordinates are specified as anchor points that remain fixed. This removes the need for post-processing to address non-identifiability, as typically applied in the latent space network literature (see [Hoff et al., 2002]).

Estimation is performed via an MCMC scheme and we rely on the GUDHI library ([The GUDHI Project, 2020]) to evaluate the Cěch complex, as required for calculation of the likelihood. Given posterior estimates of the model parameters, we are able to visualise the hypergraph through the latent coordinates and explore predictive distributions using the generative model. Finally, we consider the application of our model to real world data.

Summary. We extend the latent space approach of [Hoff et al., 2002] to the hypergraph setting by relying on tools from computational topology. Our model is both parsimonious and flexible, and avoids expensive likelihood calculations implied by a direct analogue of the latent space network model of [Hoff et al., 2002]. Estimation is carried out via MCMC, and our nsRGH model facilities visualisation of the hypergraph and exploration of predictive distributions.

References

[Bookstein, 1986] Bookstein, F. L. (1986). Size and shape spaces for landmark data in two dimensions. *Statist. Sci.*, 1(2):181–222.

[Dryden and Mardia, 1998] Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*. Wiley, Chichester.

[Edelsbrunner and Harer, 2010] Edelsbrunner, H. and Harer, J. (2010). *Computational Topology - an Introduction.* American Mathematical Society.

[Hoff et al., 2002] Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

[Kolaczyk, 2009] Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated, 1st edition.

[The GUDHI Project, 2020] The GUDHI Project (2020). *GUDHI User and Reference Manual*. GUDHI Editorial Board.

Distribution of the sizes of blackouts on electrical grids and some theoretical models

Gabriel Cwilich, Zvi Goldstein, and Sergey V. Buldyrev

Department of Physics, Yeshiva University, NewYork, New York 10033, USA
cwilich@yu.edu

It has been noted by several authors who studied the sizes of blackouts in real electrical grids that the sizes of those blackouts are distributed as power laws, and it was suggested that this is because they are driven to the self organized critical-state [1, 2]. On the other hand, recent studies [3, 4] point to the distribution of cascades being bimodal, as in first order phase transitions, with either very small or very large blackouts. The objective of this work is to reconcile these results, by looking at models of electric grids, and simpler models of betweenness centrality overload, like the Motter and Lai [5, 6] model. We investigate how the distribution of the sizes of the cascades depends on the parameters of the models, including the tolerance criteria for the lines and the dynamic rules of failure during the cascades.

The standard criterion of resiliency in electrical grids is the $N-1$ criterion [7]: the grid must safely operate in the event of the failure of any single line. This is quite different from the resilience criteria in many overload models where a uniform level of protection for all the lines is considered [6, 8]. Also, in previous models of the grid [1], it was assumed that if at any stage of the cascade one of the lines with loads exceeding the maximum values imposed by the $N-1$ condition fails, immediately all the currents in the grid are redistributed adjusting to the new network topology, and if a previously overloaded line falls below the threshold it can survive. So, this dynamics of "one-at-a-time" failure of the lines in the cascade slows down the cascade. In contrast the overload models considered that all the lines or nodes failing at one stage of the cascade are removed simultaneously before the currents are redistributed.

Here we reconcile the two approaches and show that the power law distribution of the sizes of the cascades emerges for high protection level, and also in cases where the network topology is close the percolation point. We also show that the “one-at-a-time” removal rule significantly reduces the sizes of large blackouts and their probabilities, replacing the bimodal distribution of blackouts by an approximate power-law for intermediate protection levels when the “all-at-once” rule still leads to a bimodal distribution. We show that these features are held for both the direct current model of a power grid and for the Motter and Lai model of overload, suggesting that the exact physical laws of flow on the network (Kirchhoff’s laws in real grid and their models,vs. minimal path rule in the Motter and Lai case) are not as relevant a factor as the protection level, network topology and the cascade dynamics rules.

For all the models of the power grid considered we observe the emergence of an approximate power law part in the distribution of the size of the small cascades when the one-by-one update rule is adopted. However, for each of the models studied (with the exception of one particular model, designated as RML in Figure 1, which is close to the percolation threshold), in addition to small blackouts characterized by an approximate power law distribution, we still find some large blackouts with a significant fraction of power loss. [Fig. 1(a)] . But when we consider the $N-1$ condition with the all-at-once update rule, the power law part of the blackout distribution disappears [Fig. 1(c)]. The distribution becomes strictly bimodal. Additionally, large blackouts become much larger with the all-at-once rule than with the one-by-one rule; the system is much more prone to failure.

The models presented in figure 1 correspond to approximately 12,000 nodes (models denoted as USWI and DADA) and 10,000 nodes (models denoted as RR and RML)

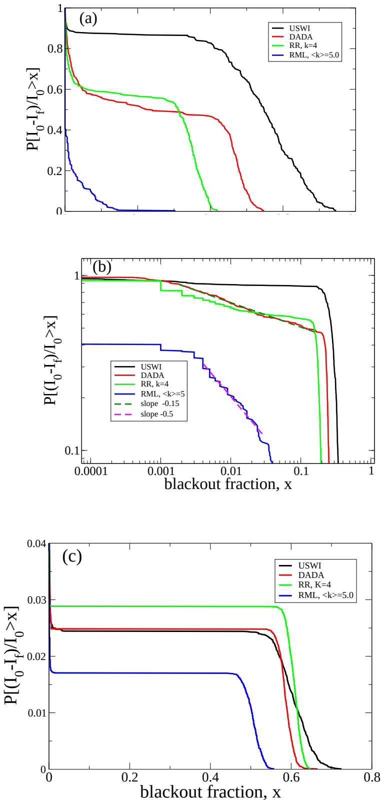


Fig. 1. Cumulative distributions of blackout sizes, measured as a function of the fraction of the consumed power lost in the blackout for four different grid topologies with the direct current approximation and the $N-1$ condition implemented. (a) One-by-one update rule. All distributions have a small range of a power law decay but eventually become bimodal except for the RML model with $\langle k \rangle = 5.0$. A clear power-law distribution with $\tau - 1 = 0.5$ emerges for the case of the RML model. (b) The same figure in a double logarithmic scale. (c) All-at-once update rule. All distributions remain bimodal.

In the case of the Motter and Lai model the $N-1$ condition is not necessary for the emergence of the power-law distribution of the cascades For example, if we take an ER network with $\langle k \rangle = 1.5$ we see the transition from a bimodal distribution of the cascades to a power-law distribution with an exponential cut-off [Fig. 2(a)] as the tolerance increases. This is somewhat similar to the effect observed for large protection and small fraction p of nodes that survive the initial attack in Ref. [8], where the bimodal distribution of the cascades ceases to exist. Replacing the all-at-once update rule by the one-by-one rule shifts the transition from bimodal to a power-law distribution at a smaller α [Fig. 2(b)], but both behaviors can still be seen.

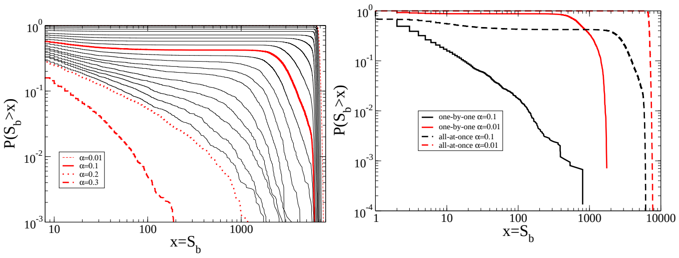


Fig. 2. (a) Cumulative distributions of blackout sizes, measured as the fraction of failed nodes in the Motter and Lai model with different values of the tolerance α and all-at-once update rule. Black lines indicate graphs for values of $\alpha = 0.02, 0.03, ...0.19$. (b) Comparison of all-at-once and one-by-one update rules for selected values of the tolerance. One can see the emergence of the power law distribution in the one-by-one removal case for much smaller values of tolerance .

References

1. B. A. Carreras, V. E. Lynch, I. Dobson, and D. E. Newman. "Critical points and transitions in an electric power transmission model for cascading failure blackouts." Chaos **12**, 985-994 (2002).
2. B. A. Carreras, V. E. Lynch, I. Dobson, and D. E. Newman. "Complex dynamics of blackouts in power transmission systems." Chaos **14** 3, 643-652 (2004).
3. R. Spiewak, S. Soltan, Y. Forman, S. V. Buldyrev, and G. Zussman, A study of cascading failures in real and synthetic power grid topologies, Network Science 6, 448468 (2018).
4. S. Pahwa, C. Scoglio, and A. Scala. (2014) "Abruptness of cascade failures in power grids." Sci. Rep. **4**, 3694.
5. A. E. Motter and Y.C. Lai. Phys. Rev. E, **66**, 065102 (2002).
6. A. E. Motter. Phys. Rev. Lett., **93**, 098701 (2004).
7. H. Ren, I. Dobson, and B. A. Carreras, IEEE transactions on power systems 23, 1217 (2008).
8. Y. Kornbluth, G. Barach, Y. Tuchman, B. Kadish, G. Cwilich, and S. V. Buldyrev, Phys.Rev. E 97, 052309 (2018).

A tail of two distributions: why negative binomial may be a better model than power law.

Jeremie Fish[1,2], Erik Bollt[1,2]

[1] Clarkson University, Department of Electrical and Computer Engineering
fishja@clarkson.edu,
bolltem@clarkson.edu
Potsdam NY, 13676
[2] Clarkson Center for Complex Systems Science (C3S2)

Abstract

Complex networks ahave becom a central structure of scientific analysis, from our social networks [1], to biological networks [2], networks of engineered systems [3], to gene networks [4]. Clearly these complex networks play a prominent role in our everyday lives, and understanding the processes running, on and inside of, these networks cannot be achieved without clear knowledge of these networks themselves.

Recently, thanks to both technological advances as well as the onset of "big data" it has become clear that real world networks are generally not simple random networks, such as the Erdős-Rényi networks with a Poisson degree distribution [5]. It was suggested in [5] that the heavy tails of the degree distributions of real world networks could be fit with a power law. While this suggestion has certainly been challenged, (see for example [6–8]) much of the network community has accepted the approximate power law degree distribution of real world networks as de facto. However, we find that most real world networks exhibit either a negative binomial, or more frequently, a scaled version of the negative binomial degree distribution. We will refer to both of these as negative binomial for the remainder of this abstract unless otherwise stated. Below we outline evidence for negative binomial degree distributions in real world networks.

We consider our case for negative binomial degree distributions. First we provide a simple generative model, to be presented in a forthcoming paper, which provides negative binomial or scaled negative binomial degree distributions. Second, we model the degree distributions from real world networks showing that commonly real world network degree distributions fit to the negative binomial distribution. Fitting to these real world networks is shown in Fig.1. The proposed network model produces one of three types of networks depending on the parameter regime, either power law, negative binomial or scaled negative binomial degree distributions. To provide further evidence we show that the Barabási-Albert (BA) networks have power law degree distribution, which is not fit by the negative binomial distributions, and furthermore that power law does not fit to the proposed network generative method.

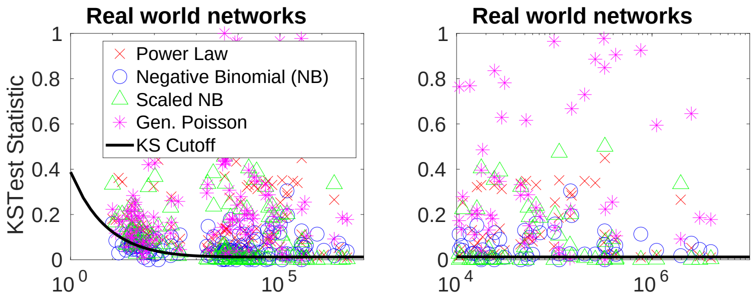


Fig. 1. Left: We show the Kolmogorov-Smirnov (KS) two sample test statistic (y-axis smaller indicates better fit) of real world networks fitted to one of four models. The black line indicates the cuttoff beneath where the null hypothesis of the two samples not being drawn from the same distribution would be rejected. Right: Networks of size 10,000 nodes or larger. We note that small networks often fit to multiple distributions for the KS test for large networks this overlap is less common. In this case we see that about 80 % of all networks (with greater than 10,000 nodes) fit to one of the 4 distributions, about 55 % fit to the scaled negative binomial, about 25 % fit to power law, about 19 % fit to negative binomial and about 4 % fit to Poisson.

The proposed generative model is a growth model, with nodes added successively one at a time. As a node joins the network, it ”chooses” to attach to existing nodes via a Krapivsky-Redner [9] type preferential attachment mechanism, with a redirection parameter (r), and a parameter that controls the ”memory” of previous attachments, (β), and the parameter m controls the average degree of the network. When a node joins the network, with probability $1 - r$ m nodes are chosen at random to connect to as in [10]. However, with probability r a node choses its attachments via weights w, which are given by

$$w = \sum_{k} e^{-\beta(t-\tau(j_k))} / \sum_{i=1}^{n} \sum_{l} e^{-\beta(t-\tau(i_l))}. \tag{1}$$

In Eq. 1, $t \in \{1, 2, ...n\}$ is an integer value representing the current node joining ”time”, and $\{\tau(j_k)\}$ is the set of times when links have been connected into node j. In the case of $\beta \to 0$ we recover the model presented in [10], which produces a power law distribution, and if $\beta = 0, r = 0.5$ the BA model is recovered [9]. The case of $\beta = 0$ is the infinite memory case, in which all previous connections play a role in the probability of attachment, while $\beta \to \infty$ represents no memory of previous attachment. So the degree of a node is only related to aging effects in the network, this produces a negative binomial (more specifically a left truncated and zero-inflated negative binomial) distribution. With $0 < \beta < \infty$ we find that as m grows, negative binomial fits poorly to the negative binomial, matching what we find in the real world networks.

In Fig. 2 and Fig. 3 we show that in all cases the best fitting model to simulated networks is the model we woud expect, for instance in the BA model the best

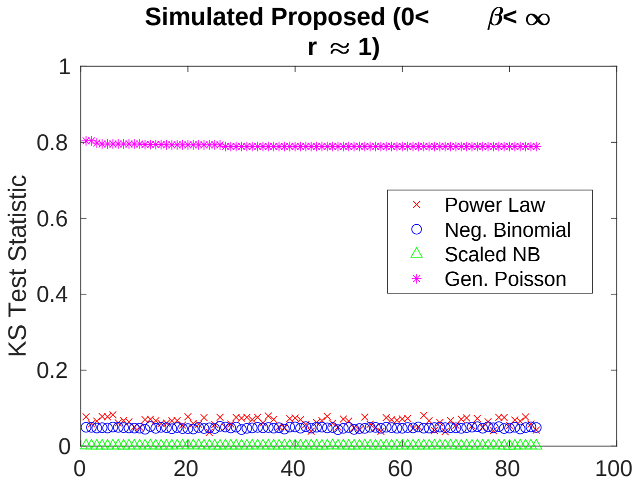


Fig. 2. We show the Kolmogorov-Smirnov (KS) two sample test statistic of various distributions when fitting to the proposed network model with 100,000 nodes under 4 different distributions; power law, negative binomial, scaled negative binomial and generalized poisson. It is clear that the scaled negative binomial gives a better fit than alternatives to this model.

fitting distribution in the KS test statistic is the power law, and in the situation where β is small but nonzero, the best fitting model is the truncated negative binomial model. Although we have not shown this case, we have also verified that when $\beta \to \infty$ the best fitting model is negative binomial as this situation (zero memory) leads to an exponential tail. We conclude from this that the KS two sample test can be used to differentiate between the various types of degree distribution listed above.

In conclusion, we have produced a generative model which produces scaled negative binomial, negative binomial or power law degree distributions. We have shown that real world networks tend to fit our model more closely to this scaled negative binomial, than the usual power law distribution. Finally we have shown that the KS two sample test statistic is valid for use in separating the various types of degree distribution, verifying our claims above.

References

1. M. Nekovee, Y. Moreno, G. Bianconi, M. Marili *Theory of rumor spreading in complex social networks* Physica A 374, 457-470 **(2007)**
2. L. K. Gallos, C. Song, S. Havlin, H. A. Makse *Scaling theory of transport in complex biological networks* PNAS 104, 7746-7751 **(2007)**
3. A. E. Motter, Y. C. Lai *ICascade-based attacks on complex networks* PRE 66, 065102 **(2002)**

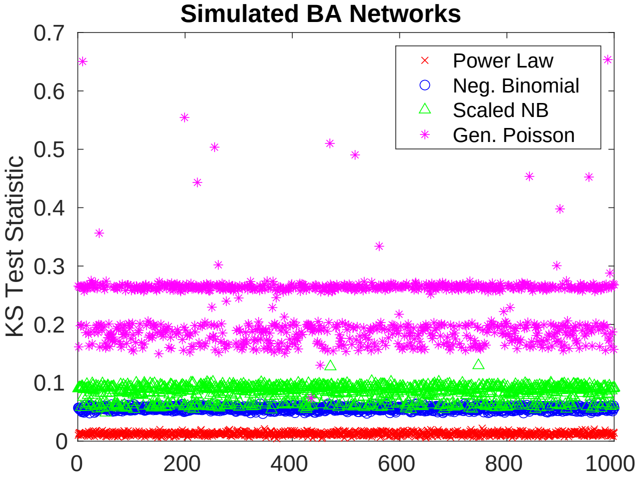


Fig. 3. We show the Kolmogorov-Smirnov (KS) two sample test statistic of various degree distributions on simulated BA Networks. It is clear that power law provides the best fit to these types of networks as would be expected, as it is known that BA Networks produce power law degree distributions.

4. R. Milo et. al. *Network motifs: simple building blocks of complex networks* Science 298 **(2002)**
5. A. Barabási, R. Albert *Emergence of scaling in random networks* Science 286, 509-512 **(1999)**
6. A. D. Broido, A. Clauset *Scale-free networks are rare* Nature Communications 10, 1-10 **(2019)**
7. I. Voitalov, P. van der Hoorn, R. van der Hofstad, D Krioukov *Scale-free networks well done* PRR 1, 033034 **(2019)**
8. P. Holme *Rare and everywhere: Perspectives on scale-free networks* Nature Communications 10, 1-3 **(2019)**
9. P. L. Krapivsky, S. Redner textitOrganization of growing random networks PRE 63, 066123 **(2001)**
10. H. D. Rozenfeld, D. ben-Avraham *Designer nets from local strategies* PRE 70, 056107 **(2004)**

Part XI

Networks in Finance and Economics

Market Interaction Structure Behind Price Heterogenity in a Monopolistic Market

Tamás Sebestyén[1] and Balázs Szabó[2]

[1] University of Pécs, Faculty of Business and Economics, Pécs, Hungary,
MTA-PTE Innovation and Economic Growth Research Group, Pécs, Hungary,
EconNet Research Group, Pécs, Hungary,
sebestyent@pte.hu
[2] University of Pécs, Faculty of Business and Economics, Pécs, Hungary,
EconNet Research Group, Pécs, Hungary,
Doctoral Prgramme in Business Administration, University of Pécs, Faculty of Business and Economics, Pécs, Hungary,
szabo.balazs@pte.hu

1 Introduction

While real life market interactions tend to be selective in the sense that not all suppliers are connected to all customers and vica versa, standard economic analysis frequently assume away this incompleteness and use models where consumers have access to all product varieties and producers supply the entire market. An interesting aspect of this incomplete connectedness is price heterogeneity: if supplier-buyer interactions are not complete in the above sense, consumers make decisions on different information bases with respect to prices and if rational suppliers take these differences into consideration then heterogeneous prices may arise from the specific network structure on which market interactions are based.

In this study we build a simple model of monopolistic competition where the network structure of supplier-buyer interactions is explicitly taken int account. In this model framework we then analyze the existence and properties of optimal prices. In particular, we are interested in the distribution of prices and the extent to which the selective nature of market interactions can give rise to price dispersion.

Standard literature on price dispersion seeks to explain heterogeneous prices either through some heterogeneity in the inherent characteristics of producers or consumers or assuming some incompleteness in price information on the side of consumers which arises from positive search costs [1, 2]. We join this line of research from the viewpoint of complex systems: taking the structure of information flows through market interactions as given, we infer on how this structure shapes price heterogeneity.

Recent literature supports this view by emphasizing the advantages of complex systems approach in economic analysis ([3, 4]), while the results of network science also indicate that the aggregate performance of a system is inherently related to the structure of connections among agents (see e.g. [5–7]).

While some studies have explicitly taken into account the relationship between network structure or network formation and the pricing behavior of firms, these contributions typically refer to network externalities where the connection structure is assumed to play a role between consumers [8–11], to situations where firms can engage in some kind of information exchange [12] or markets where networks (like telecommunication networks) compete with each other [13]. Also, a relatively recent line of research focuses on network structure in externalities, but these contributions consider the network of consumers (local network externalities) explicitly [11, 14–18]. Up to our knowledge, however, there is no attempt so far which examines the role of explicit network structures between buyers and sellers, and how this structure affects price dispersion.

The model is set up as follows. Its key element is the exogenous adjacency matrix $\mathbf{A}$ describing the connection structure between suppliers indexed by $i \in \{1,2,...,M\}$ and consumers indexed by $j \in \{1,2,...,N\}$. An exogenous connection structure can be relevant in this case as the pace of change in connections (changing consumption habits, information channels, etc.) is lower than the pace at which companies can set/reset their prices [19]. This adjacency matrix shows whether actor j consumes/buys from actor i ($a_{ij} = 1$) or not ($a_{ij} = 0$). Let $\mathbf{B}$ denote the column-standardized version of $\mathbf{A}$, with general element b_{ij}. Then, demand of actor j from the product of actor i is defined by the following demand function:

$$x_{ij} = b_{ij}\left[\gamma_j - \varepsilon\left(p_i - \overline{p}_j\right)\right], \ \forall i, j \tag{1}$$

where $\gamma_j > 0$ is a consumer-specific intercept of the demand function, $\varepsilon > 0$ is an elasticity parameter reflecting the substitutability of product varieties, p_i is the price charged by producer i and $\overline{p}_j$ is the price index perceived by consumer j. The latter is defined as

$$\overline{p}_j = \sum_k b_{kj} p_k, \ \forall j \tag{2}$$

which means that the perceived price index of consumer j is the average price of the producers it is connected to. The above demand structure explicitly depends on the network structure $\mathbf{A}$ between producers and consumers. Total demand is of producer i is the sum of the demands from all consumers j:

$$y_i(\mathbf{p}) = \sum_j x_{ij} = \sum_j b_{ij}\gamma_j - \varepsilon \sum_j b_{ij} p_i + \varepsilon \sum_j \sum_k b_{ij} b_{kj} p_k, \ \forall i \tag{3}$$

The price vector $\mathbf{p}$ in parentheses indicates that the demand of a single producer depends on the price of others through the connection structure. Producers then try to set their prices p_i to maximize their profit

$$\pi_i(\mathbf{p}) = p_i y_i(\mathbf{p}) - \omega \ell_i = y_i(\mathbf{p})\left(p_i - \frac{\omega}{\alpha}\right), \ \forall i, \tag{4}$$

taking into account the structure of market interactions in $\mathbf{A}$ and subject to the demand functions (3). In the profit function above, we assumed a linear production technology $y_i = \alpha \ell_i$, $\forall i$, where ℓ_i is labor use and α is labor productivity.

The model thus describes a landscape where consumers have established channels through which they compare and purchase product varieties and they possibly lack access to full information on prices. On the other side, firms take these channels as given and maximize profits by setting prices. The model thus allows decision makers to design policies targeting the redirection of these channels of information flows in order to adjust price heterogeneity.

2 Results

The results of this study are two-fold:

(*i*) We analytically prove that the optimal price vector exists under non-restrictive conditions of the market structure and
(ii) we show with simulations that incomplete market structure gives rise to price heterogeneity and this price heterogeneity hinges on the density of the interaction network while the asymmetry of the network plays a minor role.

The producer-wise first order conditions of the profit-maximizing problem can be used to derive the following formula for optimal prices:

$$\mathbf{p} = \frac{1}{\varepsilon}\mathbf{Q}^{-1}\left(\mathbf{BG1} + \frac{\varepsilon\omega}{\alpha}\mathbf{C1}\right). \tag{5}$$

where $\mathbf{G}$, $\widetilde{\mathbf{B}}$ and $\mathbf{C}$ are diagonal matrices with γ_i, $\sum_j b_{ij}$ and $\sum_j b_{ij}(1-b_{ij})$ on their main diagonals respectively, while $\mathbf{1}$ stands for column vectors of ones with adequate sizes. Finally, $\mathbf{Q} = \widetilde{\mathbf{B}} + \mathbf{C} - \mathbf{BB}^\mathsf{T}$ describes the strength of indirect connections between producers, the extent to which two producers share the same consumers *and* these consumers are exposed to them with a low number of alternative varieties at their reach. The existence of an optimal price vector then depends on whether $\mathbf{Q}^{-1}$ exists. We show that for this inverse to exists, it is enough that

$$\sum_j b_{ij}(1-b_{ij}) > 0, \ \forall i. \tag{6}$$

holds. This is true if there is at least one consumer j for every producer i for which $d_j \geq 2$. So every producer must be connected to at least one such consumer who is connected to at least one other producer. Put it differently, this condition rules out isolated producers ($b_{ij} = 0$ for all j) and producers without competition (monopolies).

With numeric simulations, we analyzed how network structure shapes price heterogeneity. First, we used the random network model of [20] to fill up the adjacency matrix $\mathbf{A}$. Although the network represented by $\mathbf{A}$ is bipartite, the algorithm is easily adjusted to generate networks where the degree distribution is Poisson for both types of the nodes. Fig 1 shows how the relative standard deviation of the resulting price vector is shaped by the size (N) and expected density (r) of the network. The results indicate that price dispersion is significant only for quite sparse networks ($r < 0.1$) or small ones ($N < 10$).

As the Erdős-Rényi random network gives symmetric structure with less deviation around average degree, we experimented with a network generation algorithm which is able to generate asymmetric degree distributions. The algorithm was proposed by [21] and [22], and it works along a weight parameter k which sets the asymmetry of the degree distribution: for $k = 0$ we have Poisson, and for $k = 1$ power law distribtion. In our setting, we let the degree distribution of suppliers (k^1) and that of consumers (k^2) to vary independently, therefore the bipartite nature of the generated network is handled.

Fig 2 show that the asymmetry in the degree distribution of suppliers affects price heterogeneity: the more skewed the degree distribution of the suppliers, the more dispersed prices are. However, compared to the extent to which sparsity can shape price

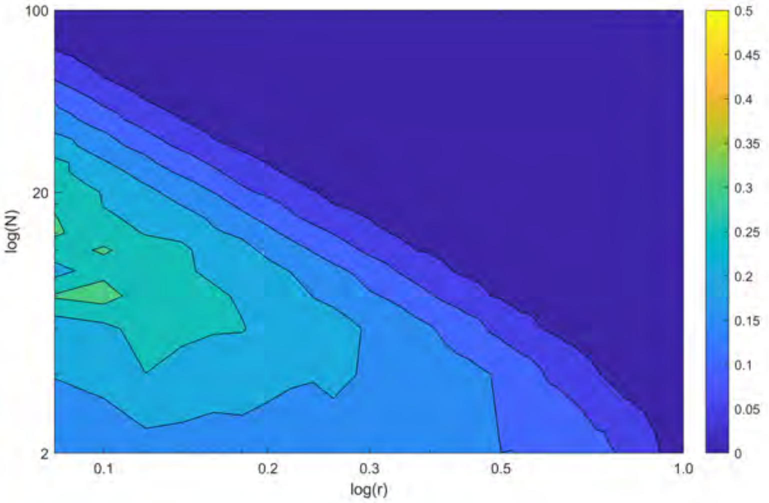


Fig. 1. Standard deviation of the optimal price vector **p** (indicated by the coloring) in function of expected network density (r) and network size (N). The figure was constructed with parametrization $\alpha = \gamma = \varepsilon = \omega = 1$ and for all combinations (r,N) 1000 independent simulations were run and the relative standard deviations then averaged.

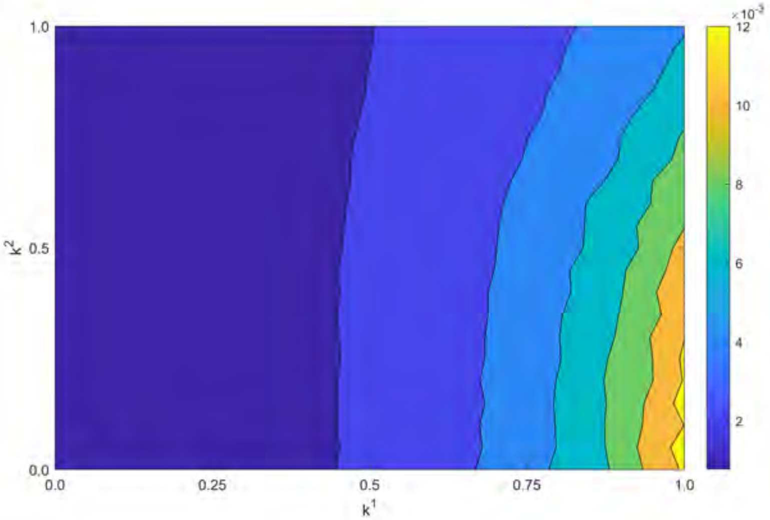


Fig. 2. Standard deviation of the optimal price vector **p** (indicated by the coloring) in function of the asymmetry of the degree distribution of suppliers (k^1) and consumers (k^2). The figure was constructed with parametrization $\alpha = \gamma = \varepsilon = \omega = 1$ and for all combinations (k^1, k^2) 1000 independent simulations were run and the relative standard deviations then averaged.

heterogeneity, the effect of asymmetric network structure remains only minor as shown by the range of the color scale. Moreover, asymmetry in the degree distribution of consumers seems to affect price heterogeneity even less.

Summary. In this study we have investigated the impact of incomplete market interaction networks on price dispersion. This attempt fits well into the literature on price heterogeneity and incomplete information concerning competitive market structures. We set up a model of monopolistic competition in which suppliers and buyers are interacting according to an exogenously given and possibly incomplete network. Within this framework, we show that the optimal price vector exists under realistic market conditions and that a slight deviation from the complete network results in heterogeneous prices. However, this heterogeneity becomes economically significant only under sparse or small networks. Sparsity as the main determinant of price heterogeneity dominates network asymmetry: relatively dense networks show minimal price dispersion even if its degree distribution follows a power law. As a result, we can say that price heterogeneity can be reasonably decreased through a more dense information/interaction structure rather than simply restructuring the existing connections.

References

1. Walsh, P.R., Whelan, C. (1999): Modeling price dispersion as an outcome of competition in the Irish grocery market. Journal of Industrial Economics, 47, 325–343.
2. Baye, M., Morgan, J., Scholten, P. (2006): Information, Search, and Price Dispersion. In: T. Hendershott (ed.) Handbook of Economics and Information Systems, ch 6. Elsevier Press.
3. Farmer, J.D., Foley, D. (2009): The economy needs agent-based modelling. Nature, 460, 685–686.
4. Farmer J.D., Gallegati, M., Hommes, C., Kirman, A., Ormerod, P., Cincotti, S., Sanchez, A., Helbing, D. (2012): A complex systems approach to constructing better models for managing financial markets and the economy. The European Physical Journal Special Topics, 214, 295–324.
5. Barabási, A.L. (2016): Network Science. Cambridge University Press.
6. Bala, V., Goyal, S. (2000): A Noncooperative Model of Network Formation. Econometrica, 68(5), 1181–1229.
7. Jackson, M.O., Wolinsky, A. (1996): A Strategic Model of Social and Economic Networks. Journal of Economic Theory, 71(1), 44–74.
8. Katz, M.L., Shapiro, C. (1985): Network Externalities, Competition, and Compatibility. The American Economic Review, 75(3), 424–440.
9. Laussel, D., Van Long, N., Resende, J. (2015): Network effects, aftermarkets and the Coase conjecture: A dynamic Markovian approach. International Journal of Industrial Organization, 41, 84–96.
10. Norbäck, P-J., Persson, L., Tåg, J. (2014): Acquisitions, entry, and innovation in oligopolistic network industries. International Journal of Industrial Organization, 37, 1–12.
11. Banerji, A., Dutta, B. (2009): Local network externalities and market segmentation. International Journal of Industrial Organization, 27, 605–614.
12. Billand, P., Bravard, C. (2004): Non-cooperative networks in oligopolies. International Journal of Industrial Organization, 22, 593–609.
13. Hoering, S. (2014): Competition between multiple asymmetric networks: Theory and applications. International Journal of Industrial Organization, 32, 57–69.

14. Chen, Y-J., Zenou, Y., Zhou, J. (2018): Competitive pricing strategies in social networks. RAND Journal of Economics, 49(3), 672–705.
15. Jullien, B. (2011): Competition in Multi-sided Markets: Divide and Conquer. American Economic Journal: Microeconomics, 3(4), 186–220.
16. Blume, L.E., Easley, D., Kleinberg, J., Tardos, É. (2009): Trading networks with price-setting agents. Games and Economic Behavior, 67(1), 36–50.
17. Bloch, F., Quérou, N. (2013): Pricing in social networks. Games and Economic Behavior, 80, 243–261.
18. Aoyagi, M. (2018): Bertrand competition under network externalities. Journal of Economic Theory, 178, 517–550.
19. Wilhite, A. (2006): Economic Activity on Fixed Networks. In Handbook of Computational Economics. Agent-Based Computational Economics. Ed. by L. Tesfatsion and K. Judd. North-Holland: Handbooks in Economics Series. p. 1013-1045.
20. Erdős, P., Rényi, A. (1959): On Random Graphs. I. Publicationes Mathematicae, 6, 290–297.
21. Goh, K.I., Kahng, B., Kim, D. (2001): Universal Behavior of Load Distribution in Scale-Free Networks. Phys. Rev. Lett., 87(27), 278701.
22. Catanzaro, M., pastor-Satorras, R. (2005): Analytic solution of a static scale-free network model. Eur. Phys. J. B, 44, 241–248.

Group Centrality Indices
in the International Trade Networks

Fuad Aleskerov[1,2], Alina Roman[1] and Viacheslav Yakuba[2]

[1] National Research University Higher School of Economics, Moscow, Russia
[2] V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences
Moscow, Russia
alesk@hse.ru

Keywords: Group influence in networks, Bundle index, Pivotal index, international trade network.

Extended Abstract

Although the classic centrality indices [1] are common, they consider only individual relations between nodes in the networks. Such classic centrality indices, for example, PageRank or In-degree centrality, disregard these important issues as group influence in the analysis of the influence of the nodes in the networks. The notion of group centrality of nodes can be consistently incorporated into a model evaluating the group influence in networks along with the parameters of the nodes themselves. For example, in a banking network the values on arcs can be volumes of the credit flow among banks, and the parameters of the nodes can reflect the assets of the banks themselves. So, if several creditors of the bank act simultaneously and the total value of the credit exceeds some critical value, which, in general, depends on the total credit activity of the bank or the banks' assets, then the stability of the financial institution might differ drastically comparing to that if the actors act asynchronously. In the international trade network the situation in which the import to the country is reduced from one exporting country alone, differs from the situation in which the import from several exporting country is reduced. The effect induced on the importing country also depends on the share of the amount of import the set of countries constitutes, does the amount exceeds some critical level, and are there countries in key positions in terms of the amount of trade. This kind of group influence in the networks can be modelled by the proposed group centrality indices, such as Bundle and Pivotal indices. The indices are illustrated by evaluating the centrality of the countries in the international trade network of 2015 year using the WITS Comtrade data [4,5].

The Bundle and Pivotal indices in a general form have been proposed in [3]. Let us define the indices. Let a country in the international trade network imports goods from several other countries, and let it be some critical level, say 10%, called quota, of the total import, to the country. If the total import from some subset of the countries exceeds the quota then such subset is called critical. The proposed Bundle index calculates the number of the critical sets for each country and then normalize it over

all countries. One can notice that some countries can exit the critical sets, and then the set ceases to be critical. These countries called pivotal ones. The proposed Pivotal index evaluates the number of pivotal countries aver all critical sets and then normalize the values among the countries. For the proposed indices, the restriction on the size of the coalition of the exporting countries can be imposed. Only coalitions of not more than 5 exporting countries are considered.
More formally, for the country i, the critical sets S for the country can be defined using the following function

$$BI_i(S) = \begin{cases} 1, if \sum_{j \in S} w_{ji} \geq q_i \,, \\ 0, else. \end{cases}$$

where w_{ji} – weights of the arcs, that are the volumes of the bilateral import to the country i from the country j, and the quota for the country i is denoted as q_i. Thus, the Bundle index for the country i just counts the number of such critical sets

$$BI_i = \sum_S BI_i(S),$$

the values of BI_i then normalized over all countries so that the sum be equal to 1.
The Pivotal index counts the number of the pivotal vertices in the critical sets, i.e. the vertices v_p, which satisfy both of the following conditions

$$\sum_{j \in S} w_{ji} \geq q_i, \text{ and } \sum_{j \in S \setminus \{v_p\}} w_{ji} < q_i,$$

where, as before, w_{ji} are the weights of the arcs directed to the country i, and q_i are the quota for this country. Having denoted the number of the pivotal vertices in the critical set S for the country i as $PI_i(S)$, the formula for the number of pivotal nodes has the following form

$$PI_i = \sum_S PI_i(S),$$

then the values PI_i are normalized over all countries.

For the international trade network, the Bundle and Pivotal indices are evaluated using the data on total import to the countries of the world for 2015. United Nations Statistics Division International Trade Statistics Database (UN Comtrade) is a database of export-import information, obtained using World Integrated Trade Solution (WITS) software. The database contains yearly export/import volumes between the pairs of the countries over an extensive structure of the categories of the goods. For our analysis, the total amount of the trade between each pair of the countries is taken into account. The quota is taken to be equal 10% of the total import to the country, and the maximum size of coalitions is set to 5. The Copland in-degree index is also calculated. The results show that the most influential countries are France, Spain, South Africa, Poland, Netherlands, Thailand, Germany, Korea, United Kingdom, and United States. In this year according to Bundle index the influence of France is around 2.1%, of Spain is about 1.9%, and within 1.7–1.8% for each of the rest of these top ten countries. According to the Pivotal index the most influential countries are France, Thailand, Slovakia, Korea, Poland, Singapore, Mexico, Czechia, New Zealand, and Canada. These countries have around 2% of Pivotal index influence each. Comparing to the classic Copeland in-degree index the United States, China, Germany, United Kingdom, France, Hong Kong, Japan, Netherlands, Italy, and Korea are among the top 10 countries. In contrast to the group influence indices there is around 14% for the USA, 9% for China, 7% for Germany, and about 4% for United Kingdom, France, and Hong Kong each. The difference of the group centrality comparing to classic in-degree centrality originates both from the fact that the group centrality indices use the share of the total import of the country contrary to the in-degree index which takes into account the import in absolute terms, and the fact that the new group centrality indices reveal so-called the structure of the import among the exporting countries not taken into account in the in-degree index.

References

1. Newman M.E.J. The Structure and Function of Complex Networks // SIAM Review, 45(2), 2003, pp. 167–256.
2. Aleskerov F., Meshcheryakova N., Shvydun S. Power in Network Structures. In: Kalyagin V., Nikolaev A., Pardalos P., Prokopyev O. (eds) Models, Algorithms, and Technologies for Network Analysis. NET 2016. Springer Proceedings in Mathematics & Statistics, v. 197. Springer, Cham. First Online: 24 June 2017.
3. Aleskerov F., Yakuba V. Matrix-vector approach to construct generalized centrality indices in networks. WP7/2020/01. Moscow: Higher School of Economics Publ. House, 2020, 21 p. SSRN: https://ssrn.com/abstract=3597948
4. UN Comtrade, International Trade Statistics database, https://comtrade.un.org/
5. WITS, World Integrated Trade Solution database, https://wits.worldbank.org/

A network analysis of personnel exchange and companies' relevant sector: the LinkedIn case study.

Tommaso Cavalieri[1], Andrea Fedele[1], Federica Guiducci[1], Valentina Olivotto[1], and Giulio Rossetti[2]

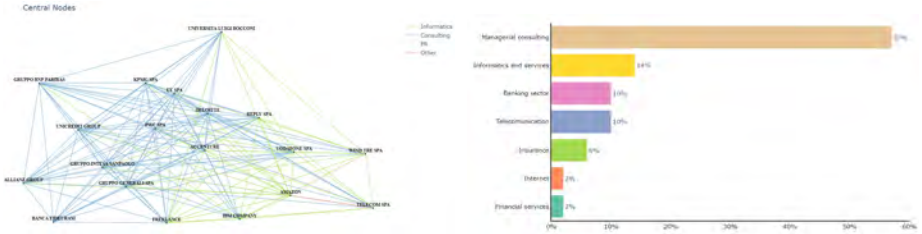
[1] University of Pisa, Italy,
cavalieri.tommaso@gmail.com, andrea-fedele@hotmail.it,
guifede3@gmail.com, valentina.olivotto19@gmail.com,
[2] KDD Lab, ISTI-CNR, Italy,
giulio.rossetti@isti.cnr.it

1 Introduction

In nowadays fast world, people tend to change workplace very often, carrying along with them the old company's know-how and, possibly, some strategic information. Analysing how companies exchange personnel is interesting in both social and economic terms and, moreover, it could be a useful tool for every company that does not want to lose its best employees. In fact, over the last few years, one of the major problems managers had to deal with was employee retention [3, 6]. To gain insight into this problem we built a professional network using data retrieved from the **LinkedIn** social network and we analysed how employees from different sectors moved between companies. The final goal of the study was to understand whether diverse sectors presented some peculiar characteristics in terms of turnover, with a follow-up focus on those companies exchanging a significant number of employees. A further objective was to investigate whether it was possible to infer companies' prevalent sector from employees' exchange information. A sample of 221,782 users was gathered filtering by location (Italy), industrial sector and the number of employees. Such sample was later exploited to obtain a network connecting various companies, which were represented as nodes in the graph. An edge between two companies was built when at least two persons have been employed in both of them at some point in time. Each of these links was weighted according to the total number of people that had worked in both firms and characterized by an attribute representing the industrial filter. The building of such edges did not consider the chronological order of the employee's exchanges, resulting in an undirected graph of 14,875 nodes and 43,932 weighted edges. During the scraping phase we selected 23 of the industrial sectors available on LinkedIn, that were later aggregated into 4 macro-sectors: **Consulting**, **Informatics (IT)**, **Public Relations (PR)** and **Others**.

2 Experimental Results

To investigate the correlations between companies through the exchanged personnel we analysed how sectors, communities and central nodes influenced the structure of the



(a) Central nodes by macro-sector (b) *Accenture*'s sectors percentage distribution

Fig. 1. Central nodes subgraph and *Accenture* sectors' distribution

graph. The communities analysed were discovered via the Louvain algorithm [1]. Our analysis underlines that each of the identified community is mainly composed by companies belonging to the same sector. Moreover, we observed the sectors composing the IT macro-sector to be uniformely distributed - being spread all over the communities - while the sectors composing the Consulting and the PR macro-sector were confined in fewer communities. However, these last two macro-sectors appeared to behave differently: in fact PR sectors covered most of the communities in which they were presented - revealing a weak connection between them - while the Consulting ones followed a better balance distribution.

We then carried out a sector distribution analysis on those nodes that were found to be central in the network (by means of classic centrality indexes). The first thing to notice was that these nodes represent well-known companies in their industry as shown in figure 1(a). A common pattern was indeed found in these central nodes' sector distribution: each of them was mainly composed by one sector with a value between 60% and 80% and several others with a value smaller than 10% separately. Figure 1(b) highlights this phenomenon for the *Accenture* node. Depending on the company, the major sector of its edges changes as expected: while in nodes representing well-known banks (eg. Gruppo BNP Paribas) it is the banking industrial sector, in IT companies (eg. Amazon) it is the IT sector and in Consulting firms (eg. Accenture) it is the consulting one. We observed interesting characteristics about the *Freelancer* node, which presented the most sparse sector's distribution. This might be due to the nature of the node that does not represent a specific company, but a temporary status (if not a life choice).

To examine the correlation between sectors we created a co-presence matrix that stores in each cell (i,j) the number of nodes that had both a link of sector i and a link of sector j attached to them. We observed, as expected, high co-presence values between pairs of sectors belonging to the same macro-sector. However, pairs composed of different macro-sectors were also observed: the highest value was reached by the couple *(Managerial consulting - Informatics and services)*, followed by *(Managerial consulting - Human resources)* and *(Financial services - Informatics and services)* reflecting what the authors concluded in [4]. Moreover, it emerged that *Managerial consulting* was the most present sector in such pairs and this result could be associated with the multidisciplinarity of the Consulting macro-sector [5], capturing people with various backgrounds and moving from different sectors.

Finally, we constructed a derivative dataset using both semantic and topological information characterising the network nodes. We then exploited such information as attributes in a Machine Learning classification problem, whose aim was to predict the industrial macro-sector of the links. The main goal of this task was to infer whether and to which extent the network's properties influenced such characteristic. The selected classifier to address such task was the Random Forest (RF). We achieved satisfactory results reaching an overall accuracy of 76% and a F1-score value of the Consulting, IT, PR and Others macro-sectors labels respectively of [0.83, 0.75, 0.70, 0.50]. It is important to notice that the lowest performance was obtained for those labels with lower support due to the unbalance in the dataset. Finally, to identify, explain and interpret some classification rules, we applied a local rule based explainator, namely LORE [2]. While eigenvector and pagerank centralities appear to be the features of uttermost importance for the RF classifier (considering their prediction weights), the local classification rule explainator highlighted something different. In fact, by analysing several random records with LORE, it emerges that another feature has the most discriminative power: community belonging. The classifier decides the classification label according to the community to which the nodes connected by the link belong. Moreover, the final prediction is set to the class of the most common industrial sector in the node community. These kind of LinkedIn networks' analysis could be exploited to study the knowledge exchange development between companies and discover what are the social phenomena underneath these changes. As a future work, we plan to collect a larger amount of data and store more information in the network such as users' education, location and job's role. A different network could also be explored taking into account the chronological order of employees' movements to better investigate people turnover, migration and knowledge exchange between companies.

Acknowledgements

This work is supported by the scheme 'INFRAIA-01-2018-2019: Research and Innovation action', Grant Agreement n. 871042 'SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics'.

References

1. Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008.10 (2008): P10008.
2. Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F. and Giannotti, F. 2018. Local rule-based explanations of black box decision systems. arXiv:1805.10820
3. James, L and Mathew, L. (2012). Employee Retention Strategies: IT Industry. SCMS Journal of Indian Management, 2012.
4. Miele, M.G. and Scognamiglio, A. *Indagine FinTech nel sistema finanziario italiano*. Divisione Editoria e stampa della Banca d'Italia. 2019
5. Pinza, A. and Zanchi, R. (April 2020). *La domanda di servizi di consulenza in Italia I risultati dell'indagine*. Confindustria, ASSOCONSULT
6. Singh, D. A Literature Review on Employee Retention with Focus on Recent Trends. International Journal of Scientific Research in Science and Technology (IJSRST)

Wealth distribution for agents with spending propensity, interacting over a network

Víctor Muñoz

Departamento de Física, Facultad de Ciencias, Universidad de Chile, Chile
vmunoz@fisica.ciencias.uchile.cl

1 Introduction

Agent-based models have been a useful strategy to model social systems, suggesting that complex collective behaviors can be described by sets of simple local interactions. One such problem has been the wealth distribution in societies, and various models have been proposed to understand its power-law behavior (Pareto's law). [3, 9, 8]

We have studied a simple model for wealth distribution, in which agents interact by sharing money, but with a certain spending propensity. [7] Specifically, a certain amount of money M is distributed uniformly across all agents, Then, at each iteration, two agents i and j are chosen randomly, and they exchange money according to:

$$\begin{aligned} x_{i,t+1} &= (1-\Lambda_i)x_{i,t} + (\Lambda_i x_{i,t} + \Lambda_j x_{j,t})\varepsilon_{i,j,t} \ , \\ x_{j,t+1} &= (1-\Lambda_j)x_{j,t} + (\Lambda_i x_{i,t} + \Lambda_j x_{j,t})(1-\varepsilon_{i,j,t}) \ , \end{aligned} \tag{1}$$

where $x_{i,t}$ is the wealth of agent i at iteration t, Λ_i is the spending propensity of agent i, and $\varepsilon_{i,j,t}$ is a random number between 0 and 1, taken from a uniform distribution. Notice that if $\Lambda_i = 1$, agents can exchange any portion of their wealth, and the equilibrium wealth distribution is exponential, as obtained by the earliest versions of these models, which did not consider spending or saving propensity [4].

In Ref. [7], it was shown that the model is able to realistically account for the wealth distribution in a society, not only in the high-end where the Pareto's law is valid. Also, we show that if the spending propensity is distributed according to a power-law, then the resulting wealth distribution is a power-law as well. Any Pareto index can be obtained by changing the power-law index of the spending propensity. As mentioned in Ref. [7], the power-law behavior of Λ should be validated from real data, and would require reliable information on the consumption habits in a given community.

As seen in Eq. (1), one of the assumptions in Ref. [7] is that all agents can exchange money with all others. However, in real societies, not all interactions are possible: one individual may buy at certain stores and not others, depending on her/his actual location, two arbitrary individuals may never meet, etc. Thus, it is worth considering the effect of this nonhomogeneity of economic interactions on the wealth distribution.

Various works have studied wealth distribution over a network, based on strategies such as discretized kinetic theory [1], the Bouchard and Mézard model [5, 6], the yard-sale model [10] or the Conservative Exchange Market Model [2]. Some of these works study the effects of network topology, such as having homogeneous instead of heterogeneous networks, or the correlation between wealth and connectivity of the agents.

In this work, we follow a similar approach, by taking agents as nodes in a complex network, but starting from the model proposed in Ref. [7]. Thus, we can study how the existence of a spending propensity affects wealth distribution, when economic interaction is only possible if those agents are connected.

2 Results

In order to investigate the effect of network topology on wealth interaction, we will consider three cases, representing various distributions of social contacts: a network where all pairs of nodes are connected (equivalent to the model in Ref. [7]), a random network with Poisson degree distribution, and a scale-free network.

Thus, in our modified model, Eq. (1) is used, but only if i and j are connected by a network vertex. The model is run for networks with 1000 nodes, with an initial wealth for all nodes equal to $x_0 = 10$, and iterated for 10^8 timesteps, which ensures convergence in all cases. Random networks were created with 5000 edges, whereas the scale-free networks were given the degree distribution $P(k) \propto k^{-\alpha}$, in the domain $k \in [1,20]$, and with $\alpha = 0.1, 1, 3, 10$. Notice that for the largest values of α, the maximum degree is less than $k = 20$, but this is considered only for illustration purposes, as the relevant parameter for the wealth distribution exponent is α. For each network topology, three choices for the spending propensity distributions were tried: (a) $\Lambda_i = 1$; (b) uniform distribution in the interval $[0,1]$; (c) scale-free distribution $\sim \Lambda^{-\beta}$, with $\beta = 0.1, 0.5, 1, 5, 10$.

Figure 1(a) shows the complementary cumulative distribution function ($F(x)$ is the cumulative distribution function of wealth x), for the three described topologies, in the case where all agents have maximum spending propensity, and Fig. 1(b) for the case where spending propensity follows a power-law distribution.

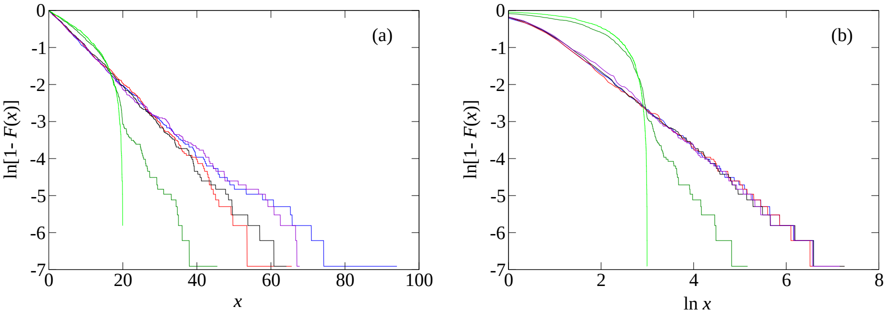


Fig. 1. Wealth distribution for the agent-based model for various network topologies: fully connected network (blue line), Poisson-distributed degree (red line), scale-free with $\alpha = 0.1$ (black line), $\alpha = 1$ (violet line), $\alpha = 3$ (dark green line), and $\alpha = 10$ (green line). (a) Maximum spending propensity. (b) Power-law distribution of spending propensity, with $\beta = 0.5$.

Notice that Fig. 1(a) is a semilog graph, whereas 1(b) is a log-log graph. For the reference model, that is, fully connected network, the distribution follows an exponential curve if agents have maximum spending propensity, and a power-law curve if the spending propensity follows a power-law distribution. This is consistent with Ref. [7]. The same wealth distribution is observed for random and scale-free networks ($\sim k^{-\alpha}$), for power-law exponents below $\alpha \simeq 1$. Thus, it is interesting that such topologies have no major incidence on the distribution of wealth, which may be related to the universality of Pareto's law, as observed in actual economic systems.

However, for larger values of α (green and dark green curves in Fig. 1), the wealth distribution clearly differs from the fully connected behavior, also losing its purely exponential or power-law behavior.

Summary. Wealth distribution in an economic system is studied by means of an agent model, where agents have a certain spending propensity, and they interact over a given network. When the network is random, or scale-free ($\sim k^{-\alpha}$) with α below 1, approximately, results are equivalent to having all agents allowed to interact with any other agent. However, values of $\alpha > 1$ affect both the wealth distribution, and the behavior at the tail. These results hold both in the absence of spending propensity, and when the spending propensity follows a power-law. Although these are preliminary results, and a larger diversity of network topologies should be investigated, as well as non-conservative models, they nevertheless suggest that Pareto's law is a very robust phenomenon with respect to the details of the connectivity of the agents, and that the ubiquity of Pareto's law in actual systems may have implications on the topological properties of the underlying networks of interaction.

References

1. Bertotti, M.L., Modanese, G.: Discretized kinetic theory on scale-free networks. Eur. Phys. J. Special Topics 225, 1879–1891 (2016)
2. Braunstein, L.A., Macri, P.A., Iglesias, J.R.: Study of a market model with conservative exchanges on complex networks. Physica A 392, 1788–1794 (2013)
3. Chatterjee, A., Chakrabarti, B.K., Manna, S.S.: Money in gas-like markets: Gibbs and Pareto laws. Phys. Scr. T106, 36–38 (2003)
4. Drăgulescu, A., Yakovenko, V.M.: Statistical mechanics of money. Eur. Phys. J. B 17, 723–729 (2000)
5. Garlaschelli, D., Loffredo, M.I.: Wealth dynamics on complex networks. Physica A 338, 113–118 (2004)
6. Ichinomiya, T.: Wealth distribution on complex networks. Phys. Rev. E 86, 066115 (2012)
7. Lammoglia, N., Muñoz, V., Rogan, J., Toledo, B., Zarama, R., Valdivia, J.A.: Quantitative description of realistic wealth distributions by kinetic trading models. Phys. Rev. E 78, 047103 (2008)
8. Mohanty, P.K.: Generic features of the wealth distribution in ideal-gas-like markets. Phys. Rev. E 74, 011117 (2006)
9. Patriarca, M., Chakraborti, A., Germano, G.: Influence of saving propensity on the power-law tail of the wealth distribution. Physica A 369, 723–736 (2006)
10. Vásquez-Montejo, J., Huerta-Quintanilla, R., Rodríguez-Achach, M.: Wealth condensation in a Barabasi-Albert network. Physics 389, 1464–1470 (2010)

Learning through the Grapevine: The Impact of Noise and the Breadth and Depth of Social Networks

Matthew O. Jackson[1] and Suraj Malladi[2] David McAdams[3]

[1] Department of Economics, Stanford University, and Santa Fe Institute,
jacksonm@stanford.edu
[2] Graduate School of Business, Stanford University,
surajm@stanford.edu
[3] Fuqua School of Business and Economics Department, Duke University,
david.mcadams@duke.edu

1 Introduction

While bias and errors arise in all forms of media and communication, policy-makers have been particularly concerned with misinformation and disinformation spreading through online channels. With little accountability on social media and messaging platforms, false rumors are easy to start. And due to the ease of forwarding and content modification, even true messages may get corrupted as they get relayed over long chains.[4] [5] While it may be difficult to measure the extent to which misinformation on online platforms alone has distorted decision-making, governments worldwide have been developing an array of tools to regulate these channels. However, governments involved in censorship may inject their own bias in the communication process by selectively filtering and promoting certain messages. Crowd sourced fact-checking can similarly reflect the prevalent bias among the majority of users. Meanwhile, platforms have been reluctant to police certain types of content and risk alienating users. What sort of policies, then, could curb misinformation and improve social learning on online platforms without relying on some authority or majority to decide what is true.

Motivated by these questions, we develop a framework to study more generally how people learn as a function of their networks when information is subject to mutation, deliberate manipulation, and content-based transmission failure. In particular, we characterize how learning depends on the depth and breadth of a person's network. Since information that travels a longer path is less likely to survive, and less likely to be accurate if it does survive, increasing network size does not necessarily improve learning. Both depth and breadth increase the size of a network, but they also increase the relative

[4] Like in the children's game of telephone, messages frequently mutate in online channels due to intentional distortions and noisy communication. In [1]'s study of online viral memes, one meme was reposted more than 470,000 times, with a mutation rate of around 11 percent and more than 100,000 variants. This was not an outlier in their analysis: 121 of the 123 most viral memes each had more than 100,000 variants.

[5] [3] found instances of Internet chain letters that traveled median distances of over one hundred links. [1] examined hundreds of millions of instances of thousands of memes and found chains with lengths in the hundreds and typical distances well into the dozens. [2] explain why the resulting trees can be much longer than they are wide.

amount of noise by increasing the relative number of sources at greater distances compared to those closer by. We examine how depth and breadth impact learning and how their effects relate to each other. Our analysis shows how limiting the network can improve the accuracy of overall content, without the need for censorship or for private or public monitoring of messages. This is important, given the many problems associated with censorship, even with benign intentions.

2 Model and Results

In our model, information is relayed from original sources via sequences of individuals to an eventual Bayesian receiver (or learner), who wishes to learn the state of the world.[6] The state of the world and the corresponding messages are "1" or "0" (e.g., the state can be "climate change is real" or "climate change is not real"; or less drastically, "no study shows that 5g radio waves are carcinogenic" or "5g radio waves *are not* carcinogenic"). With noiseless word-of-mouth communication and sufficiently many starting sources of conditionally independent information, the receiver learns the true state. However, along each chain, the message may mutate or be dropped – reducing the information content of the signals that reach the receiver.

First, if the receiver knows the mutation and dropping rates, we show that the receiver learns the state (with a probability approaching 1 as depth increases) if and only if the number of chains that they have access to (the breadth of their network) exceeds a threshold that grows exponentially in the depth of the network, as well as in the mutation and dropping rates. The increased noise from greater depth has to be countered by more total sources. This provides a precise relationship between breadth and depth: breadth has to exceed a threshold that increases with depth (or equivalently, depth has to be lower than a threshold that increases with breadth) in order for learning to be possible. The threshold is sharp in that full learning happens asymptotically above the threshold, but no learning happens below the threshold.

Next, we show that even small amounts of uncertainty about mutation rates precludes learning from any number of long chains. The intuition here is that over long chains most messages that survive have mutated, and in the limit, information content disappears. The ratio of 1's to 0's is slightly different depending on the starting state, but that difference vanishes as chains get longer. If there is any uncertainty about the relative likelihood of mutating from 1 to 0 or vice versa, then that uncertainty swamps the tiny differences that emerge from the starting state. Learning is completely precluded.

The key to overcoming this is to limit the depth of messages, or at least the relative ratio of messages that are coming from far away. This is the subject of some of our key results. By limiting the total distance that messages can travel, one limits the chances that messages are distorted. This allows for partial learning from nearby messages. People see fewer messages, but ones that are more likely to be informative - and this increases the overall signal to noise ratio. Thus, capping depth can be helpful, and we characterize an optimal cap. We also explore what happens when depth is not easily

[6] Learners on a platform may or may not be Bayesian in practice. But since our focus is on how debilitating noise can be to social learning, we do not introduce constraints to the receiver's ability to process information.

capped. For instance, some social media platforms do not track whether a message is new or somehow forwarded. However, it is easy for them to track how many people someone broadcasts a message to.[7] By capping the number of people a person can send a message to (or at least making it more difficult), one can control the breadth of the network.[8] This can also help, since decreasing breadth increases the *relative* number of nodes in a network that are close compared to farther away for any given depth. Thus, without being able to control depth, limiting breadth can also improve learning. Again, we provide bounds and explore optimal policies.

Interestingly, breadth-limits have been adopted by online messaging platforms. For instance, *WhatsApp* has capped the number of people that someone can message, for the express purpose of curbing the spread of false information.

Finally, we extend the model to study how well a receiver learns when dropping rates also depend on the message being relayed. For instance, a person may be more likely to pass along information that they find surprising, or that is in line with their prior beliefs. Even in the publication of scientific articles, reviewers may be more likely to agree to publish (i.e., pass along) statistically significant or surprising results than insignificant or expected ones. In this case, a receiver can learn from how many messages she receives. Hearing about very few studies, even though many were conducted, is informative: the studies most likely did not find 'exciting' results that prompted them to be discussed and forwarded.

We bound how much more likely a fully Bayesian agent – who updates based on both message survival and content – is to guess the state compared to someone who looks only at message content or only at message survival. We show that the Bayesian's advantage vanishes as distance to primary sources increases: for any parameters of the model, all the information is contained in either message frequency alone or in message content alone. This implies that full Bayesian learning is fully approximated in the limit by simple rules of thumb conditioned on just one dimension of the available information. Thus, learning need not involve sophisticated calculations but simple thresholds: believe 1 if and only if more than a certain number of messages are received, or if and only if the ratio of messages containing 1's is above some threshold.

We gratefully acknowledge support under NSF grants SES-1629446 and SES-2018554.

References

1. Adamic, Lada A., et al. ”Information evolution in social networks.” Proceedings of the ninth ACM international conference on web search and data mining. 2016.
2. Golub, Benjamin, and Matthew O. Jackson. ”Using selection bias to explain the observed structure of internet diffusions.” Proceedings of the National Academy of Sciences 107.24 (2010): 10833-10836.
3. Liben-Nowell, David, and Jon Kleinberg. ”Tracing information flow on a global scale using Internet chain-letter data.” Proceedings of the national academy of sciences 105.12 (2008): 4633-4638.

[7] Users might individually re-target messages, but this becomes much more burdensome.

[8] What matters is average in-degree, but what can be most easily capped is out-degree. However, note that average in-degree must equal average out-degree, and so capping one caps the other.

Foreign lockdown in supply networks: a cross-country analysis of economic independence

Erik Braun[1], Tibor Kiss[2], and Tamás Sebestyén[3]

[1] University of Pécs, Faculty of Business and Economics
EconNet Research Group
Pécs, Rákóczi út 80., Hungary,
[2] University of Pécs, Faculty of Business and Economics
Pécs, Rákóczi út 80., Hungary,
[3] University of Pécs, Faculty of Business and Economics
MTA-PTE Innovation and Economic Growth Research Group
EconNet Research Group
Pécs, Rákóczi út 80., Hungary,
sebestyent@ktk.pte.hu

1 Introduction

There has been an increasing interest in analyzing the structure of domestic and global supply chains in the past decade in parallel to which network theoretic approaches gained ground in economic literature. First, with the help of network analytic tools the central actors can be identified in production networks and it is possible to track down how shocks propagate between them (see e.g. [1–3]). Second, a multi-sector and multi-country dataset allows the examination of the structure of trade in value-added [4], detecting the community structure [5], and to measure the length of production chains both locally and globally [6]. Third, a complex systems approach to economic modelling reveals that the asymmetric structure of intersectoral transactions increase the volatility of aggregate output [7] for which the central actors are primarily responsible [8].

Among other things, the COVID-19 pandemic has also focused attention to the structure of global supply chains and its economic effects. A series of recent studies [9–12] have analyzed the macroeconomic impact of the pandemic, and several studies [13–16] have been conducted on how the GDP declines in response to economic lockdown in input-output network economies. Most closely related is [17], who showed how shocks propagate through global supply chains and what is the role of indirect effects in this propagation. If some countries execute economic lockdown, it can lead to economic decline in other countries through their production network. For example, [18] reported that three-quarters of companies in the US faced difficulties in sourcing inputs. This sheds light on the fact that countries are closely connected to each other on the global scale, resulting in a moderate ability to operate independently. Building on this literature, this paper presents a new method to measure the extent to which economies are independent from global supply chains, and are able to build on their domestic supply chains. The novelty in this measure is that it allows us to decompose a single independence score into two parts: the extent to which a given sector is dependent on foreign inputs in its operations and the extent to which it contributes to the domestic economy through its direct and indirect output towards other sectors.

2 Methods

In this study, we analyze data collected from WIOD, which contains sector-level multi-country input-output data. We consider a 2408x2408 inter-sectoral input-output matrix as a weighted adjacency matrix of a network where the nodes are the different country-sectors and the input-output links present the directed and weighted edges. In order to determine the value of an economy's independence, we use this database and apply the bootstrap percolation process as follows [2]:

1. step: Select a country and extract all sectors and the links between them which belong to this country. As a result, we get a new adjacency matrix representing within-country trade flows.
2. step: Extract the elements of the final demand vector which belong to the given country and subtract the value of imports and input flows from them. As a result, we get a modified final demand vector for the country.
3. step: Create the technical coefficients matrix using the new adjacency matrix and the modified final demand vector. Determine the Leontief-inverse from the technical coefficients matrix.
4. step: Using the diagonal elements of the Leontief-inverse, calculate the contribution of individual sectors to the domestic economy as a supplier.
5. step: Determine the ratio of domestic inputs to all inputs in the case of the selected country's sectors.
6. step: Based on the final demand, calculate the economic weight of the selected sectors.
7. step: Multiply both the value of contribution and the ratio of domestic inputs by these weights. The independence measure is obtained by multiplying these two factors.
8. step: Repeat this process for all countries.

One of the main advantages of this method is that we can decompose the independence measure into two parts: the first part gives information about the openness of a country from the production network perspective, and the second part can show how closely connected the sectors are to each other through the direct and indirect links. Another advantage is that we can analyze these properties on a sectoral level. For these reasons, we can better understand why some countries show to be more independent while others are not.

This method clearly focuses on the demand side effects. Although it is just part of the total effects of lockdowns, the method described above can be easily extended to supply-side analysis where the Leontief inverse is substituted for the Gosh index (based on the transpose of the adjacency matrix), and augmented with the export shares instead of import shares. This approach reflects the extent to which sectors are served by (embedded in) domestic partners and sell their output towards the local economy. Due to size limitations, in this abstract we only refer to the demand side calculations. As a result, the results shown here apply only to demand side shock propagation. A limitation of our method is that it does not take into account the feedbacks from shocks to the structure of the input-output relations (coefficients).

3 Results

The results of this study are summarized in Figure 1. The dots represent all countries in the WIOD database. The horizontal axis shows the average contribution of a country's sectors to other domestic sectors, while the vertical axis shows the importance of foreign inputs in the operations of the country's sectors. The coloring of the nodes represent the overall level of economic independence of a country. The two black lines denote the average values along the two dimensions.

The results show that the most independent country is China because the Chinese sectors are closely connected as suppliers in their production network. It is interesting to note that the US economy is on a similar level with respect to the ratio of domestic inputs, however, due to the structure of the production networks, especially in the case of manufacturing industries, the Chinese economy is less dependent on the global supply chains. Furthermore, the results reveal that the countries are very heterogeneous in terms of the two factors and the extent of independence. One of the most exposed economies to the global supply chains is Hungary. It is visible from the picture that this country performs badly along all dimensions as its key sectors (the automotive and electronic industries) use mostly foreign inputs, and as suppliers, sell their output to other Hungarian sectors only to a small extent.

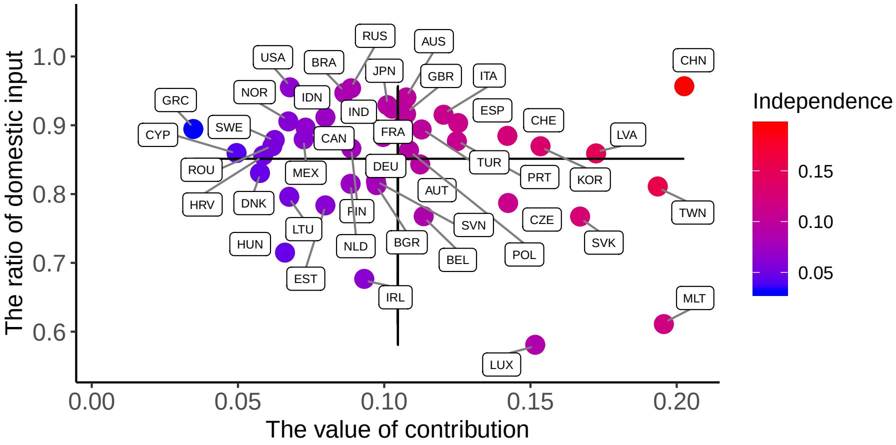


Fig. 1.

Summary. Due to the COVID-19 pandemic, economists are increasingly focusing on the structure of global supply chains and the economic effects of the virus. A series of recent studies have indicated that economic lockdown in a country can significantly influence the performance of other countries indirectly, through the production networks. With this perspective in mind, this research presented a method to measure the level of a country's independence of global supply chains which can be used to decompose this independence into two factors: the share of domestic inputs and the dominance of domestic sales.

References

1. Blöchl, F., Theis, F. J., Vega-Rendondo, F. Fisher, E. O. N.: Vertex Centralities in Input-Output Networks Reveal the Structure of Modern Economies. Phys. Rev. E 83(4) 046127 (2011)
2. Fan, Y., Ren, S., Cai, H., Cui, X.: The State's Role and Position in International Trade: A Complex Network Perspective. Economic Modelling, 39 71–81 (2014)
3. Cingolani, I., Panzarasa, P., Tajoli, L.: Countries' Positions in the International Global Value Networks: Centrality and Economic Performance. Applied Network Science, 2:21 (2017)
4. Amador, J., Cabral, S.: Networks of Value-Added Trade. The World Economy, 40(7) 1291–1313 (2017)
5. Cerina, F., Zhu, Z., Chessa, A., Riccaboni, M.: World Input-Output Network. PloS One 10(7) e0134025 (2015)
6. Wang, Z., Wei, S. J., Yu, X., Zhu, K.: Characterizing Global Value Chains: Production Length and Upstreamness. NBER Working Paper 23261 (2017)
7. Acemoglu, D., Carvalho, V. M., Ozdaglar, A., Tahbaz-Salehi, A.: The Network Origins of Aggregate Fluctuations. econonetrica, 80(5) 1977–2016 (2012)
8. Contreras, M.G. - Fagiolo, G.: Propogation of Economic Shocks in Input-Output Networks: A Cross-Country Analysis. Phys. Rev. E 90(6) 062812 (2014)
9. Acemoglu, D., Chernozhukov, V., Werning, I., Whinston, M. D.: A Multi-Risk SIR Model with Optimally Targeted Lockdonw. NBER Working Paper 27102 (2020)
10. Alvarez, F. E., Argente, D., Lippi, F.: A Simple Planning Problem for COVID-19 Lockdown. NBER Working Paper 26981 (2020)
11. Eichenbaum, M. S., Rebelo, S., Trabandt, M.: The Macroeconomics of Epidemics. NBER Working Paper 26882 (2020)
12. Krueger, D., Uhlig, H., Xie, T.: Macroeconomic Dynamics and Reallocation in an Epidemic. NBER Working Paper 27047 (2020)
13. Baqaee, D. R. Farhi, E.: Nonlinear Production Networks with an Application to the COVID-19 Crisis. NBER Working Paper 27281 (2020)
14. Barrot, J.-N., Grassi, B., Sauvagnat, J.: Sectoral Effects of Social Distancing. HEC Paris Research Paper No. FIN-2020-1371 (2020)
15. Barthélémy, B., Huo, Z., Lechenko, A. A., Pandalai-Nayar, N.: Global Supply Chains in the Pandemic. NBER Working Paper 27224 (2020)
16. Giammetti, R., Papi, L. Teobaldelli, D., Ticchi, D.: The ITalian Value Chain in the Pandemic: The Input-Output Impact of COVID-19 Lockdown. Journal of Industrial and Business Economics 47, 483–497 (2020)
17. Guan, D., Wand, D., Hallegatte, S., Davis, S. J., Hue, H. Li, S., Bai, Y., Lei, T., Xue, Q., Coffman, D. Cheng, D. Chen, P., Lian, X., Xu, B., Lu, X., Wang, S., Hubacek, K., Gong, P.: Global Supply-Chain Effects of COVID-19 Control measures. Nature Human Behaviour 4, 577–587 (2020)
18. Institute for Supply Management: COVID-19 Survey: Impacts on Global Supply Chains. https://www.instituteforsupplymanagement.org. (2020)

Prioritizing investments in critical facility access during and following natural hazard events using geospatial data and network perturbation models

AE Schweikert[1], GF L'Her, and MR Deinert

Colorado School of Mines, Golden CO 80401, USA,
aschweikert@mines.edu

1 Introduction and background

The value of critical infrastructure systems is the services they provide. A key vulnerability of these systems is the damage to, or failure of, system components during and following natural hazard events. In 2017, global economic losses related to natural hazard events are estimated at $337 billion, with less than half insured [1]. These direct impacts are expensive and result in part from the failure of critical infrastructure systems including roads, railways, transmission infrastructure, buildings and other assets. Further, indirect impacts such as loss of income, disruption of supply chains and injuries and morbidity, among other factors, increase these costs to society. Particularly in lower income economies and vulnerable regions such as small island states, these impacts are exacerbated [2].

Ongoing work from multiple disciplines is focused on ways to reduce vulnerability, manage risk, and increase resilience to natural hazards. A 2019 article by Koks et al. [3] looked at multi-hazard risks to transport infrastructure (road and rail). The authors found expected global annual damages of $3.1-22 billion USD from direct impacts of natural hazards, the majority coming from flooding. Network vulnerability of infrastructure to hazards has been investigated for seismic risk to a historic transmission grid in the US [4], to gas networks in Europe [5] as well as a small set of water and power infrastructure [6]. Vulnerability as a function of the dependency between systems is highlighted as an important consideration in [5], [6]. The impact of storm surge and sea level rise on roadways and accessibility of regions was investigated using simulated risks for Spain [7]. Critical road segments for agricultural transport to market in [8] account for factors such as flood risk, climate change, poverty and other factors.

The focus of the present work is to quantify the impacts that road and transmission network disruptions have on critical facilities in the Commonwealth of Dominica, a small island state in the Caribbean that is exposed to hurricanes, flooding, landslides and other hazards. The case study results below examines road network disruption on household access to hospitals resulting from combined pluvial, fluvial and storm surge at 0.02 annual ("50-year flood") probability. The algorithm developed for this purpose is flexible and can be applied to any network (e.g. rail, electricity distribution, water, and others), any set of origin-destination pairs, and flexible inputs to perturbance (including natural hazards), failures and costs.

2 Methods

Two scenarios were developed to evaluate the impact that road failure would have on critical facilities. First, a drop-link analysis, where every network segment is dropped individually and an analysis of travel cost between origin(s) and destination(s) is calculated. Second, a Monte Carlo analysis where multiple networked infrastructure segments are dropped based upon failure probabilities. Both scenarios rely on user-defined origin and destination pairings (such as travel from houses (“origins”) to hospitals (“destinations”) on roads (“network”)). Costs are assessed using the Dijkstra algorithm, which finds the shortest paths between nodes in a graph [9]. We assign a cost (time) to the vertices in the network and, for each pair of origin nodes and destination nodes, it finds the optimal path according to the aggregate defined cost.

Failure probabilities for road segments can be user-defined, or, where data is available, informed by hazard probability and infrastructure fragility data. For example, some portion of the road network is exposed to a flood. The probability of a network segment failure in the example below is calculated using engineering-based fragility curves to estimate the likelihood of damage to the network based upon flood depth, road surface and classification (see [10] for full background and source data).

These two scenarios used here produce three quantitative metrics. First, *Segment Drop-Link Criticality*, which is the impact of dropping a segment on travel cost between all origin and destination pairs (every segment is dropped once, and travel cost is calculated over entire network routing). The impact is defined by the mathematical distance between ‘normal’ and ‘perturbed’ network matrices of travel cost between all origin and destination pairs, calculated over the entire drop-link segment analysis. Second, *Impact on Service Delivery/Supply Chain* which defines the importance of each segment, defined as the increase in the cost of delivery of good/person from origin to destination relative to all other network segments. Third, the *Impact on Facility Access*, which is the impact to destinations, based upon their increase or decrease in expected goods received based on least-cost travel routes if a link or set of links is lost relative to the overall network. The outcome of this analysis is the number of persons, goods, or other flow on a network that are rerouted to a different destination based on segment disruption.

Future work will include congestion (traffic, in the context of road networks) resulting from the perturbed network scenarios. This consideration is especially important in hazard scenarios or on infrastructure such as power systems where overloading can result in cascading network failures. A limitation here is the available data, although methods for estimation may be useful.

3 Results

Figure 1A details all of the origins (households), the three destinations (hospital/medical facility) and the road network. The road network in 1A shows the criticality of every road segment from drop-link analysis (Segment Drop-Link criticality). Figure 1B shows the impact that road segment failure has on each destination in terms of the change of flow (households) (Impact on Service Delivery/Supply Chain). The impact on facility access is measured as the number of households re-routing to a different (closest) hospital based on road network perturbances from the flooding scenario.

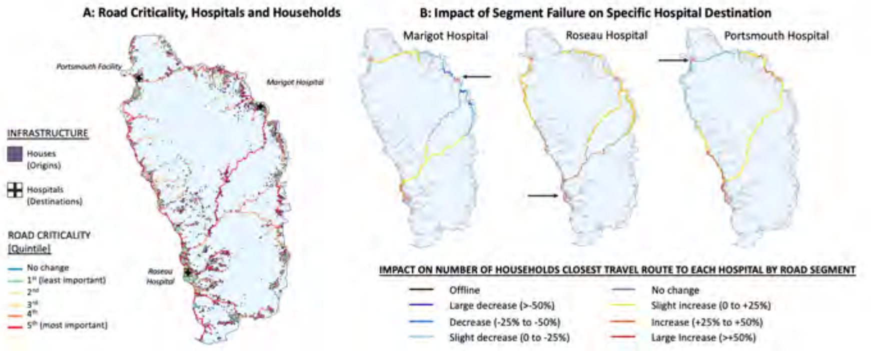


Fig. 1. Dominica road network segments impact on hospitals resulting from flooding. Figure 1A details all origins (households), destinations (hospitals) and road segment criticality. Figure 1B details the impact that segment failure has on each specific destination.

Additional analysis was run to determine the impact on each facility in terms of increase or decrease in households (Impact on Facility Access) from 1,000 Monte Carlo runs. While the majority of runs show only small increases/decreases in overall change, some runs show increases of nearly half the total households rerouting to Portsmouth or Marigot hospital. This shows that, in terms of stocking supplies, Roseau is likely the least necessary facility to stock (for this specific hazard). Figure 1 shows that some road segments are very important to the overall re-routing and, particularly in the southwest part of the island, failures of these segments contribute to large increases in hospital destination change.

Summary. Because of reality-constrained budgets for upgrading infrastructure and hazard preparedness the focus of this work is to provide cost-benefit analyses that allow for comparative assessments of investments in different sectors (such as health or transport). The analysis method and metrics presented in this study have specific application to decision-making for limited resources. In the example of Dominica illustrated, tradeoffs between allocating resources to reduce vulnerable road sections (to avoid damage) or to hospitals (to increase capacity) is an inter-disciplinary policy question. This methodology allows for multiple scenario analysis and outputs quantitative data. The developed algorithm has been applied to fuel supply chains to power plants (origin: ports, destination: power plant, network: roads), water distribution networks, and power transmission. Ongoing work is focused on incorporating additional constraints such as congestion, multiple hazards and the interconnectedness of multiple networked systems.

References

1. Swiss Re, “At USD 144 billion, global insured losses from disaster events in 2017 were the highest ever, sigma study says,” Swiss Re, Zurich, Apr. 10, 2018.
2. S. Hallegatte, J. Rentschler, and J. Rozenberg, Lifelines: The Resilient Infrastructure Opportunity, vol. English Overview. Washington, DC: The World Bank Group, 2019.
3. E. E. Koks et al., “A global multi-hazard risk analysis of road and railway infrastructure assets,” Nat. Commun., vol. 10, no. 1, pp. 1–11, Jun. 2019, doi: 10.1038/s41467-019-10442-3.
4. F. Cavalieri, P. Franchin, J. A. M. B. Cortés, and S. Tesfamariam, “Models for Seismic Vulnerability Analysis of Power Networks: Comparative Assessment,” Comput.-Aided Civ. Infrastruct. Eng., vol. 29, no. 8, pp. 590–607, 2014, doi: 10.1111/mice.12064.
5. K. Poljanšek, F. Bono, and E. Gutiérrez, “Seismic risk assessment of interdependent critical infrastructure systems: The case of European gas and electricity networks,” Earthq. Eng. Struct. Dyn., vol. 41, no. 1, pp. 61–79, 2012, doi: 10.1002/eqe.1118.
6. L. Dueñas-Osorio, J. I. Craig, and B. J. Goodno, “Seismic response of critical interdependent networks,” Earthq. Eng. Struct. Dyn., vol. 36, no. 2, pp. 285–306, 2007, doi: 10.1002/eqe.626.
7. H. Demirel, M. Kompil, and F. Nemry, “A framework to analyze the vulnerability of European road networks due to Sea-Level Rise (SLR) and sea storm surges,” Transp. Res. Part Policy Pract., vol. 81, pp. 62–76, Nov. 2015, doi: 10.1016/j.tra.2015.05.002.
8. X. Espinet, J. Rozenberg, K. Singh Rao, and S. Ogita, “Piloting the use of network analysis and decision-making under uncertainty in transport operations,” World Bank Group, Policy Research Working Paper WPS8490, Jun. 2018.
9. E. W. Dijkstra, “A note on two problems in connexion with graphs,” Numer. Math., vol. 1, no. 1, pp. 269–271, Dec. 1959, doi: 10.1007/BF01386390.
10. A. Schweikert, G. L’her, and M. Deinert, “Resilience in the Caribbean: Natural Hazards Exposure Assessment and Areas for Future Work,” The World Bank, forthcoming.

A network model of World Trade inequality and how to mitigate it

Javier García-Algarra[1], Mary Luz Mouronte-López[2], Javier Galeano[3], and Gonzalo Gómez-Bengoechea[4]

[1] Engineering Department, Centro Universitario de Tecnología y Arte Digital, Spain,
javier.algarra@u-tad.com
[2] Higher Polytechnic School, Universidad Francisco de Vitoria, Spain,
[3] Complex Systems Group, Universidad Politécnica de Madrid, Spain,
[4] Economics Department, Universidad Pontificia Comillas, Spain

1 Introduction

The Global Trade Network (WTN) is a network of exchange flows among countries whose topological and statistical properties are a valuable source of information. Degree and strength are key magnitudes to understand its structure. We have built a stochastic generative model that yields synthetic networks that closely mimic the properties of annual empirical data[6]. Agreement between empirical and synthetic networks is checked using the available series from 1962 to 2017. One of the properties of the WTN is inequality. This model can explain how it is a self-sustained property of the newtork and suggests possible strategies to tackle it. We apply a mild mitigation method and assess the quantitative impact on the Gini index.

2 Results

The study of the statistical and topological properties of the global trade as a network has been an active research topic in recent years[1, 2]. In these studies, degree and strength are two key connectivity properties of each node. The WTN is modeled as a bipartite network, and each country plays a double role, as exporter or importer[3]. Probability distributions of degree and strength shed light on the network properties. If the network is scale-free, these distributions follow a power-law. Trying to fit empirical data to a power-law has been a common procedure in network analysis, despite the fact that this behavior is not universal[4]. The log-normal distribution is a second possible choice in many fields, but specially in Economics[5]. The discussion about the best model has run for a long time, because both may fit the long tail of some empirical series. Our approach takes into account two observations: dominant countries, both as exporters or importers, attract with high probability new trade chances and global yearly trade distributions are approximately log-normal. In this paper, we describe a generative stochastic synthetic model of trade that mixes two processes, preferential attachment and proportional effect, to mimic the properties of the weighted matrixes of WTN historical data series. The first one works during a very short period of simulation time and produces a Pareto-like heterogeneous distribution of trade strength for exporter and importer guilds. The second, acts on a much larger time scale and by its multiplicative effect makes the yearly trade distributions log-normal.

As a result of both processes, we obtain the synthetic trade matrix, whose statistical properties may be compared to those of the empirical one (see Fig.1). For each year, only three parameters of the empirical network set the final configuration of the synthetic matrix: number of nodes of each guild and number of links of the network. The Gini coefficient of yearly traded volume is above 0.8, an evidence of the inequality of this network.

The force that drives the birth of new links through mutual benefit is proportional to the product of the strengths of both nodes. As new trade opportunities tend to join the main Exporter-Importer pairs, dominant nodes attract more and more volume through a positive feedback loop. The exponents of the degree-strength distribution show that the share of global trade these countries is quite above their share on GDP, while smaller countries strive to play their role in the network.

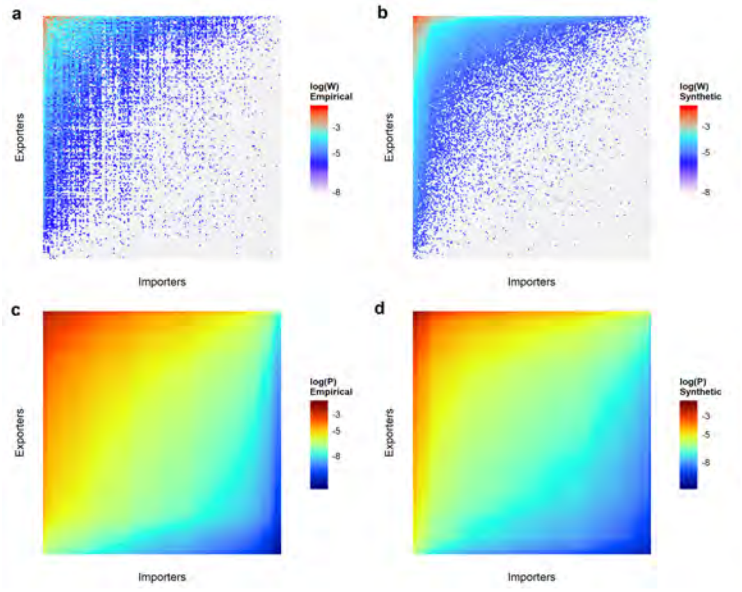


Fig. 1. Empirical and synthetic trade volume and probability matrices for year 2015

To assess trade inequality reduction policies, we simulate the impact of agreements modifying three parameters: improved trade percentage, boost percentage, and scope[7]. With the first one, we select those nodes that are eligible for the improvement policy. For instance, the value of 1% discards all probability matrix cells in the upper 99% of volume distribution. The boost percentage ranges from 25% to 200% and is the increase in the probability of that particular matrix cell. Finally, the scope may restrict trade improvements to countries that belong to the same World Bank region or allow free connections among any pair of countries.

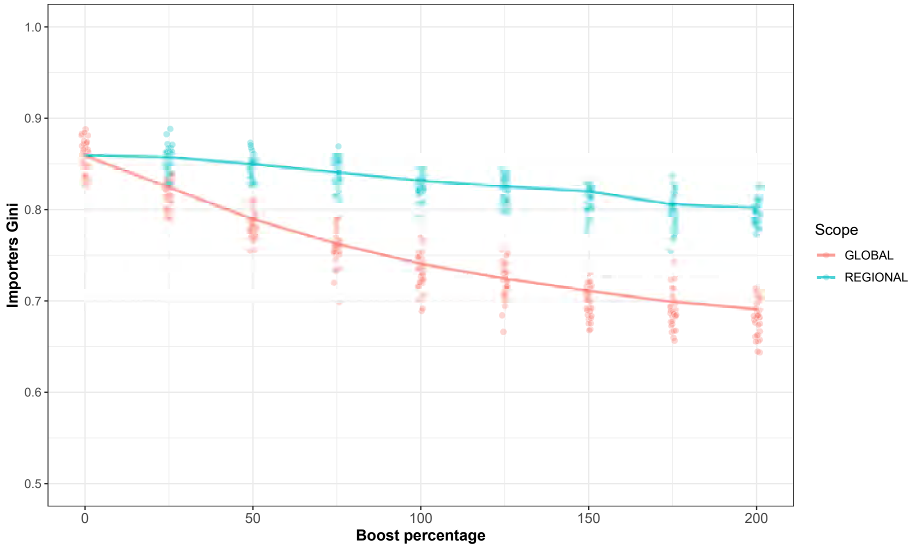


Fig. 2. Effect of a improvement policy for year 2005. The policy is applied to the lower 2% of global trade. For each boost percentage value there are 30 dots, one for experiment, the solid line joins the average values. Gini index lowers following a quadratic law if the policy is global. If trade boosting is restricted to countries of the same region the effect is weaker.

The results of the numerical experiments show that trade agreements between underrepresented countries with no regional limitations may have a sensible larger impact on reducing inequality of the international trade network (Fig.2). Improvement ratio, without regional restrictions, follows a quadratic pattern an so, small actions may be quite effective if they are focused on the group of extremely poor nodes. Results also confirm one of the network's most relevant properties: self-fulfilling structural inequality.

References

1. Fagiolo, G., Squartini, T. and Garlaschelli. J. Econ. Interact. Coord. 8, 75–107, (2013).
2. Garlaschelli, D. and Loffredo, M. I. , Phys. A: Stat. Mech. its Appl. 355, 138–144, (2005).
3. Bhattacharya, K., Mukherjee, G., Saramäki, J., Kaski, K. and Manna, S. S. J. Stat. Mech.: Theory Exp. 2008, P02002 (2008).
4. Broido, A. D. and Clauset, A. Scale-free networks are rare. Nat. Commun. 10 (2019).
5. Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. Internet mathematics 1, 226–251, (2004).
6. Garcia-Algarra, J., Mouronte-Lopez, M.L., and Galeano, J. Scientific Reports, 9(1), 1–10, (2019)
7. Garcia-Algarra, J., Gómez-Bengoechea, G., Mouronte-Lopez, M.L. Complexity, 2020, (2020)

Shock propagation channels behind the global economic contagion network

Zita Iloskics[1] and Tamás Sebestyén[2]

[1] University of Pécs, Faculty of Business and Economics
EconNet Research Group
Pécs, Rákóczi út 80., Hungary,
[2] University of Pécs, Faculty of Business and Economics
MTA-PTE Innovation and Economic Growth Research Group
EconNet Research Group
Pécs, Rákóczi út 80., Hungary,
sebestyent@ktk.pte.hu

1 Introduction

Examining the spread of macroeconomic phenomena between countries has become increasingly popular after the 2008 economic crisis, but it became an even more relevant issue today, due to the COVID-19 pandemic. This study aims to examine how economic linkages between countries such as trade in goods and services or foreign direct investment shapes shock transmission on a global scale.

The study of international trade networks gained special attention in the past decades ([1],[2],[3],[4],[5],[6]), showing that structural changes in this network can be conclusive about the development paths of countries ([4],[6]) and that overall instability in the global economic system is a result of increased globalization and complexity of the underlying networks ([7],[8],[9]).

Apart from investigating trade networks, attention is also directed towards the extent to which economic activity is synchronized across countries. Most of this research like [10], [11], [12], [13], [14], [15] focus on the contemporaneous correlation between some macroeconomic variables, while some others like [16], [17] and [18] use Granger causality tests to identify cross-country macroeconomic effects.

The relationship between business cycle co-movement and trade have also been examined for a while. According to the empirical studies, there is a moderate positive significant relationship between trade interdependence and cyclical co-movement of macroeconomic indicators ([19]). Although other studies have confirmed the relationship between business cycle synchronization and commercial relations, they argue that it is less pronounced than in previous studies [22]. Also, the relationship between bilateral trade and the co-movement of business cycles differs between subsets of countries [15]: the link is stronger within OECD countries than within non-OECD countries and between OECD and non-OECD countries.

In this study, we build on previous work by [16] which captures the spread of shocks by estimating causal relationships between the business cycles of countries. The analysis is based on trade and output data for the period between 1996 and 2018 and for the OECD countries except Turkey, plus Bulgaria and Romania. Within this time frame we construct time windows of 52 quarters and estimate Granger causality between the cyclical components of country-level GDP series to obtain a network of shock contagion. The nodes of this directed binary network are the countries and the links between them are present if the GDP-cycle of a country affects that of another country. Using the rolling time windows, we are able to draw this contagion network for all years separately through the sample.

Contrary to previous studies which examine the link between trade and business cycle synchronization ([19],[14],[20],[21],[15]), our approach is novel in the sense that instead of establishing synchronization on the basis of cross-correlation, we employ a causality approach and examine whether trade linkages contribute to the extent to which one country's cyclical behavior affects that of another country. While cross-correlation may falsely identify co-movement if business cycles are driven by unobserved common factors, using causality tests we can focus on events when the economic situation in one country truly affects that in an other country.

More precisely, we extract the cyclical component $\hat{c}_{t,i}$ of GDP series for all country i and period t from a sample of 42 countries, using HP-filtering. Then we run pairwise Granger causality tests for all pairs of filtered GDP series, by estimating the following two regression models for every pair of countries i and j, with time lag L:

$$\hat{c}_{t,i} = \beta_0^1 + \sum_{l=1}^{L} \beta_l^1 \hat{c}_{t-l,i} + \varepsilon_{t,i} \tag{1}$$

$$\hat{c}_{t,i} = \beta_0^2 + \sum_{l=1}^{L} \beta_l^2 \hat{c}_{t-l,i} + \sum_{l=1}^{L} \gamma_l \hat{c}_{t-l,j} + \mu_{t,i} \tag{2}$$

Having these models estimated, we calculate the F-statistic for all country-pairs as $F_{i,j} = (RSS_1 - RSS_2)/RSS_2(N-L)/L$, where RSS_1 and RSS_2 are the residual sum of squares of models Eq (1) and Eq (2) respectively, while N is the sample size. Given these test statistics we calculate the probability of the F-values ($P(F_{i,j})$), and establish a contagion link from country i to j as

$$a_{i,j} = \begin{cases} 1, & \text{for } P(F_{i,j}) < f_{i,j}, i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

More details of this method can be found in [16]. Given this network drawn, in this paper we examine whether the existence of these contagion links can be explained by bilateral trade volumes. The latter data is extracted from the UN Comtrade database. To bring the trade data in line with the contagion networks calculated for the 52 quarter rolling time windows, trade data were averaged for the same periods by country pairs. Data on exports and imports of goods for all periods and country pairs are fully available between 1996 and 2018.

2 Results

Using the data described previously, we employ logistic panel regression to estimate the relationship between trade connections and shock contagion. This type of estimation allows to take into account the cross-sectional and time dimension of the data. The independent variable in the regression is relative trade ($T_{c,t}$) volume between countries which is calculated as follows:

$$T_{c,t} = \frac{EX_{c,t} + IM_{c,t}}{Y_{i,t} + Y_{j,t}} \tag{4}$$

where $c = i \rightarrow j$ labels a directed country pair, $EX_{c,t}$ is the average export and $IM_{c,t}$ is the average import of goods between the country pair c in time period t, while $Y_{i,t}$ is the average GDP of country i in time period t. In order to have a more detailed picture, we split total trade between goods and services, and use the same formula Eq 4 to calculate relative trade volumes.

The dependent variable is binary and measures if the cyclical component of the GDP series in country i causes the cyclical component of country j over time period t, as shown previously. For further methodological details of these calculations, see [16]. We use the difference in the level of development between countries as a control variable, measured by the absolute difference in GDP per capita. If this value is positive, it means that the origin country in the shock propagation or trade network is more developed than the receiving country. If it is negative, then shock/trade originates from an economically less developed country and absorbed by a more developed one. In the random effect panel models we also include the geographic distance of country pairs, the data on which were collected from the CEPII gravity database. In the fixed effects model we can not include distances separately as they do not change over time, so the effect of distance is merged intot the estimated fixed effect coefficients. The sample is strongly balanced, all observations are available for country-pairs and time periods. Due to the binary nature of the dependent variable, we use logit panel regression in the form:

$$a_{c,t} = \alpha_0 + \alpha_1 T_{c,t} + \alpha_2 DevDiff_{c,t} + \alpha_3 Dist_c + \mu_c + \gamma_t + \varepsilon_{c,t} \tag{5}$$

where μ_c is a unit-specific constant and γ_t is a time-specific constant. Applying multivariate regression to panel data, the problem of omitted variables can be reduced, as unobserved heterogeneity can be captured by unit-specific coefficients μ_c. We run four different models, which differ in whether trades in goods or services are included as independent variable ($T_{c,t}$) and whether fixed or random effects model is estimated. The random effect model can give a more efficient estimation, while the fixed effect model is less effective but consistent in all cases. To decide which estimation method to use, the Hausman test was applied ([23]) which shows that the fixed effects estimate is consistent. However, we present the results of the random effects models as well, because in these cases we can include geographic distance as a separate explanatory variable.

The results in Table 1 show that trade in either goods or services has a positive significant effect on shock contagion: the more countries trade, the more likely that their economic fluctuations affect each other. This result is robust across the different

	Model 1 (RE)	Model 2 (FE)	Model 3 (RE)	Model 4 (FE)
Const.	-2.8722***	-	-2.2435***	-
Goods	0.0002***	0.0005***		
Services			0.0001***	0.0002***
Dev.diff.	0.000001	-0.00003***	-0.00001**	-0.00004***
Dist.	-0.00007***	-0.0001***		
Wald χ^2	85.31	210.34	42.11	66.92

Table 1. Results of logit panel regression. Dependent variable: shock contagion as in [16]. Goods: relative trade volume as in Eq 4, calculated for goods, Services: relative trade volume as in Eq 4, calculated for services, Devdiff.: difference between GDP per capita, Dist.: geographical distance.

estimation methods. Looking at the consistent fixed effect estimations, the difference in development level also seems to affect contagion. The larger the gap between the per capita GDP of the origin and receiver country, the less likely it is that shocks spread between them. This means that shocks spread from less developed countries towards more developed ones and not vica versa. Finally, geographic distance have a negative effect – in line with intuition: countries farther away are less likely to transmit shock to each other.

Summary. In this study we further explore the relationship between trade connections and the contagion of economic shocks across countries. In contrast to previous studies, we use a causality approach to identify synchronization and as a result we can infer on how trade affects the spread of economic shocks between countries. The results show a positive significant relationship between trade and shock contagion: trade channels thus seem to be conducive in spreading business cycle fluctuations across countries in general. The difference between the level of development also has an effect on shock contagion: shocks are less likely to spread from more developed towards less developed countries.

References

1. Fagiolo G. Reyes J, Schiavo S. World-trade web: Topological properties, dynamics, and evolution. Physical Review E. 2009;79(3):036115
2. Serrano Á, Boguñá M. Topology of the world trade web. Physical Review E. 2003;68(1):015101.
3. Askari M, Shirazi H, Samani KA. Dynamics of financial crises in the world trade network. Physica A: Statistical Mechanics and its Applications. 2018;501:164-169.
4. Cristelli M, Tacchella A, Pietronero L. The Heterogeneous Dynamics of Economic Complexity. PLoS ONE 10(2): e0117174. doi:10.1371/journal.pone.0117174 (2015)
5. Saracco F, Di Clemente R, Gabrielli A, Squartini T. Detecting early signs of the 2007-2008 crisis in the world trade. Scientific Reports 2016;6:30286.
6. Straka MJ, Caldarelli G, Saracco F. Grand canonical validation of the bipartite international trade network. Phys. Rev. E. 2017;96:022306.
7. Sheng A. Globalization and Growth Implications for a Post-Crisis World. In: Spence M, Leipziger D, editors. Financial crisis and global governance: A network analysis. Washington DC: World Bank; 2010. p. 69–93

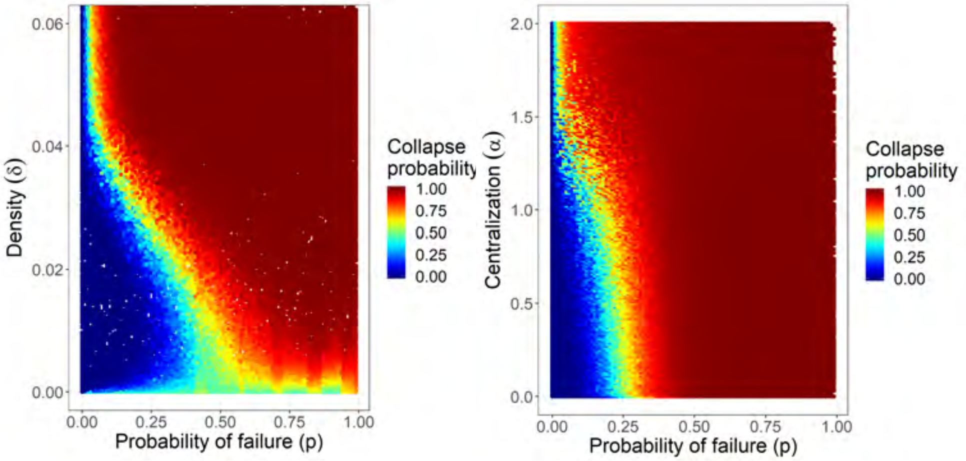
8. Schweitzer F, Fagiolo G, Sornette D, Vega-Redondo F, Vespignani A, White RD. Economic Networks: The New Challenges. Science. 2009;325:422–425.
9. He J, Deem MW. Structure and Response in the World Trade Network. Phisical Review Letters. 2010;105:198701.
10. Blonigen BA, Piger J, Sly N. Comovement in GDP trends and cycles among trading patners. Journal of International Economics. 2014;94(2):239-247.
11. Kose MA, Prasad ES, Terrones ME. How Does Globalization Affect the Synchronization of Business Cycles? The American Economic Review. 2003;93(2):57-62.
12. Doyle, BM, Faust, J. Breaks in the Variability and Comovement of G-7 Economic Growth. The Review of Economics and Statistics. 2005;87(4):721-740.
13. Shin K, Wang Y. The Impact of Trade Integration on Business Cycle Co-Movements in Europe. Rev. World Econ. 2005;141:104.
14. Inklaar, R., Jong-A-Pin, R., De Haan, J.: Trade and business cycle synchronization in OECD countries—A re-examination. European Economic Review, 52(4), 646-666 (2008).
15. Di Giovanni, J., Levchenko, A. A.: Putting the parts together: trade, vertical linkages, and business cycle comovement. American Economic Journal: Macroeconomics, 2(2), 95-124, (2010).
16. Sebestyén T, Iloskics Z. Do economic shocks spread randomly?: A topological study of the global contagion network. PLoS ONE 15(9): e0238626. 2020
17. Selover DD. International co-movements and business cycle transmission between Korea and Japan. Journal of the Japanese and International Economies. 2004;18(1):57-83.
18. Sander H, Kleimeier S. Contagion and causality: an empirical investigation of four Asian crisis episodes. Journal of International Financial Markets, Institutions and Money. 2003;13(2):171-186.
19. Canova, F., Dellas, H.: Trade interdependence and the international business cycle. Journal of international economics, 34(1-2), 23-47 (1993)
20. Frankel, J.A., Rose, A.K. Is EMU more justifiable ex post than ex ante? European Economic Review, 41(3-5), 753-760. 1997.
21. Frankel, J.A., Rose, A.K. The endogenity of the optimum currency area criteria. The Economic Journal, 108(449), 1009-1025. 1998.
22. Kose, M. A., Yi, K. M.: Can the standard international business cycle model explain the relation between trade and comovement? Journal of international Economics, 68(2), 267-295 (2006)
23. Wooldridge, J, M.:Introductory Econometrics: A modern approach. Canada: South (2009).

The hidden cost of interdependencies: Collapse of complex economic systems and network structure

Aymeric Vié Alfredo J. Morales
New England Complex Systems Institute
MIT Media Lab
Mathematical Institute, University of Oxford
Corresponding: alfredom@mit.edu

Economic interdependencies have become increasingly present in globalized production, financial and trade systems. As highly complex and networked systems, the properties of economies are characterized by the behavior and interdependencies of their components (Hidalgo et al., 2007). Whether they arise from investments, trade or supply chains, interdependencies are increasingly important in contemporary economic systems, and fundamental for risk assessment and evaluation (Schewitzer et al., 2009). Interconnections enable the diversification of outputs, improve efficiency of economies, and increase the growth of economic complexity (Hidalgo et al., 2007). However, at the same time, they also introduce paths for risk contagion and generate large-scale vulnerabilities to systemic failure (Balsa-Barreiro et al., 2020). Given the current context of increasing international trade, financialization and globalization in economies, it is crucial to understand the effects of connectivity on networked economies and its relationship to economic collapse (Vié and Morales, 2020). In this analysis, we consider collapse as incapacity of a given network structure to attain production levels of a network-free production system.

In (Vié and Morales, 2020), we characterized how the structure of the interdependencies shape the macroeconomic variables of the system. In the continuity of this work, we discuss how the structure of interconnections among economic agents increases the fragility of economic systems despite an apparent improvement of their production complexity. We model the spread of failure in economic systems and explore multiple ways in which systems can be interconnected. We consider network density -the number of connections that are drawn among agents independently-, centralization, which refers to the emergence of highly connected nodes that bridge across



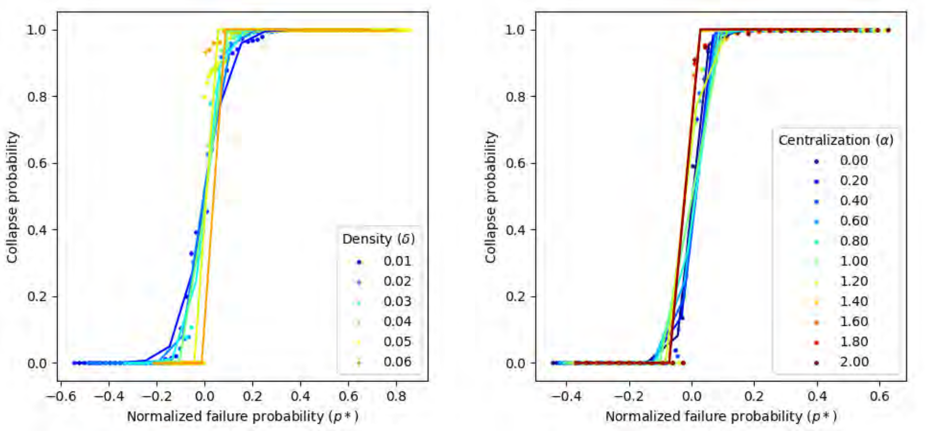
(a) Density model

(b) Centralized model

Figure 1: Probability of systemic collapse as a function of model parameters. Color indicates the collapse probability. The left panel shows the outcomes of the density model. The right panel shows the outcomes of the centralized model. The x-axis represents the probability of individual failure in both panels. The y-axis represents the network density (left panel) or centralization (right panel).

large parts of the network, and multilayer models of supply chain networks. Finally, we have applied our models to empirical international trade and supply chain networks. The results on realistic simulations and real data are consistent with the more generalized network framework.

We show that the transition to collapse is universal and independent of a specific network structure (see Figures 1 and 2), in artificial and empirical networks. The amount to which network structure -measured by density and centralization- affects collapse probability tends to follow a sigmoidal curve. Sparser networks may have a lower productivity but are more resilient to the spread of failure. Previous research has investigated the risks of creating agents ”too big to fail”, or more recently ”too central to fail” (Battiston et al. 2012). In the continuity of this observation, our model emphasizes that without further hypotheses, economic agents may in some situations become too interconnected to thrive. Risk diversification improves global robustness only in an interval of individual risk of failure, the system becoming too sensitive to individual failure if density and centralization are too high.



(a) Density model

(b) Centralized model

Figure 2: Universal behavior of collapse probability. The left panel shows the results for the density model. The right panel shows the results for the centralized model. Dots represent the resulting collapse probability (y-axis) of model simulations. The solid lines show the fit to the sigmoid function.

References

1. Balsa-Barreiro, J., Vié, A., Morales, A.J. et al. Deglobalization in a hyper-connected world. Nature Palgrave Commun 6, 28 (2020). https://doi.org/10.1057/s41599-020-0403-x
2. Battiston, S., Puliga, M., Kaushik, R., Tasca, P., and Caldarelli, G. (2012b). Debtrank: Too central to fail? financial networks, the fed and systemic risk. *Scientific reports*, 2, 541.
3. Hidalgo, C. A., Klinger, B., Barabási, A. L., and Hausmann, R. (2007). The product space conditions the development of nations. *Science*, 317(5837), 482-487.
4. Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., and White, D. R. (2009). Economic networks: The new challenges. *Science*, 325(5939), 422-425.
5. Vié, A., Morales, A.J. How Connected is Too Connected? Impact of Network Topology on Systemic Risk and Collapse of Complex Economic Systems. Comput Econ (2020). https://doi.org/10.1007/s10614-020-10021-5

Economic Integration Index Evaluated from the Loop Flow Component in Global Value-Added Network

Sotaro Sada and Yuichi Ikeda

Graduate School of Advanced Integrated Studies in Human Survivability,
Kyoto University, Kyoto 606-8303, Japan,
sada.sotaro.87a@st.kyoto-u.ac.jp

1 Introduction

Nations worldwide are connected through trade, producing to meet the demands of their own countries and the demand in other countries. How much of the GDP generated in a country is induced by foreign demand? The extent to which sectors worldwide are affected by demand in other countries is represented in a complex network. We created a network in which the value-added induced in the sector by demand in other countries is the weight of the links and analyzed in terms of flows. In the network of trade values, the weights of the links from downstream sectors are overestimated because most products of a sector include value added by other sectors. However, the production share of value-added, which is also used in the calculation of GDP, caused by demand in other countries, as a link, gives a more realistic picture of how the world economy depends on avoiding overestimates.

One of this study's goals is to identify the communities of global value-added activities. The second is to propose an index of economic integration based on the circulation degree in the global economy. Lastly, we investigate how the world economy's integration changed before and after the economic crisis by analyzing the more detailed process of change in the value-added network.

2 Data and Method

We had constructed a Global Value-Added Network (GVAN) from World Input-Output Database (WIOD) [1] and conducted a community analysis. WIOD contains data for 56 sectors in 43 countries for the period 2000–2014 including developing countries. Therefore, in each year, the network consists of 2408 nodes. In our study, the value-added was calculated based on the Trade in Value Added (TiVA) calculation method [2], a method of Input-Output analysis. We used international TiVA which means value-added induced by foreign demand as weights of the directed network (the node is a sector in a country). Then, a flow-based community analysis, the map equation [3], was used to identify community change over 15 years, and the Helmholtz-Hodge decomposition (HHD) was applied to the complex value-added flows within the community [4]. By applying HHD, the value-added flow F_{ij} from node i to node j can be separated into the circular flow F_{ij}^{c} and gradient flow F_{ij}^{g}: $F_{ij} = F_{ij}^{c} + F_{ij}^{g}$. Here the gradient flow F_{ij}^{g} is given by $F_{ij}^{g} = w_{ij}(\phi_i - \phi_j)$ where ϕ_i is the Helmholtz-Hodge potential and w_{ij} is a positive weight for link between node i and j. The circular flow satisfies $\sum_j F^c_{ij} = 0$ in which incoming flow and outgoing flow are balanced in each node. F^g_{ij} represents the difference in potentials between the nodes, i.e. the hierarchical relationship. On the other hand, F^c_{ij} represents a value circulation in the network.

These data and methods allowed us to determine economic integrated areas objectively by the characteristics of the data (by minimizing the map equation). Moreover, we could evaluate the extent of value circulation within the economic areas using circular flows, which illustrated economic integration of the countries/sectors of the areas quantitatively.

3 Results

The community analysis of GVAN using the map equation method for 15 years revealed the communities in Europe and the Pacific Rim including the US between 2004 and 2011 (Fig. 1). From 2000 to 2003 and 2012 to 2014, orange component: nodes in the Pacific Rim, and green component: nodes in Europe had been in the same giant community. However, these two regions had been virtually divided, and Europe had been fragmented into small communities for eight years. The regional characteristics of the communities were different compared to industrial communities observed in international network [5]. Moreover, the structural changes 2003–2004 and 2011–2012 coincided with the changes of the inclination of the ratio of international trade and world GDP.

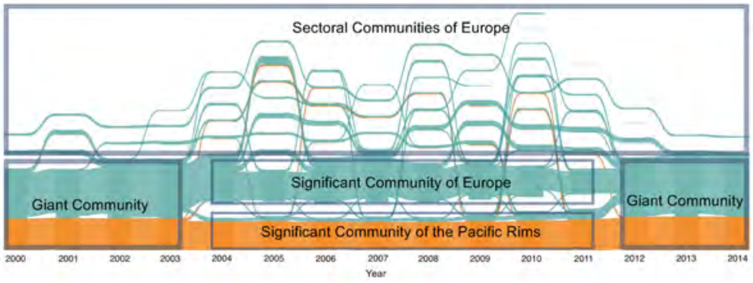
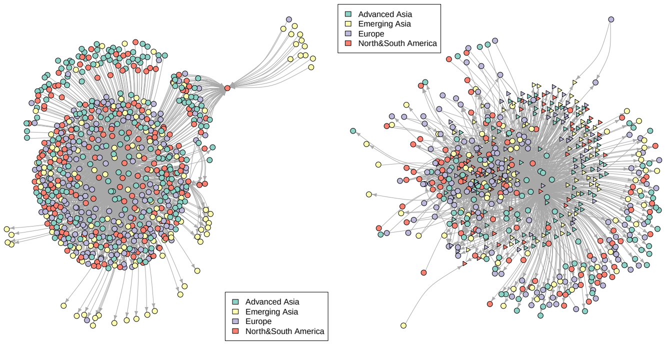


Fig. 1. Alluvial diagram of annual community changes in Global Value-Added Network for 15 years. Orange and green mean nodes of the Pacific Rim and Europe respectively. Isolated nodes are not represented. These two regions were in almost different community 2004–2011.

We applied HHD to investigate how value circulated in two significant regional communities, and Fig.2 showed the result. We confirmed that the value-added was rotating in GVAN as the Strongly Connected Component nodes, which were reachable to each other as shown as triangular nodes in Fig.2. The results indicated GVAN has not only Global Value Chain, but also Regional Value Circulation. Furthermore, we calculated the degree of the circulation within the communities as the Economic Integration Index shown in Fig.3. This is a preliminary estimation calculated by aggregate amount of loop flow component divided by the total weight of the links. The change of the index showed unstable circulation in Europe from 2004 to 2011.



(a) The potential flow. (b) The circular flow.

Fig. 2. The potential and circular flows in the Pacific Rim community of Global Value-Added Network in 2011. The links were cut with 120 million dollars, and nodes ware colored in regions. Triangular nodes represent Strongly Connected Components in the network.

This presentation will show how these flows changed before and after the economic crisis and quantitatively assesses the community's annual economic integration development. In the future research, we will apply this index for understanding what roles regional/sectoral economic integration played in international trade and why the inclination of international trade's ratio to world GDP changed after 2011.

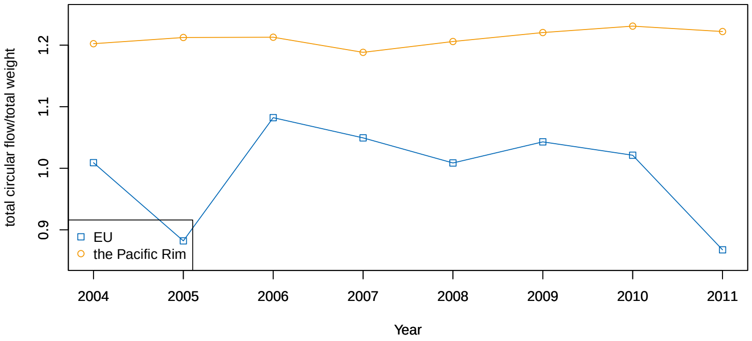


Fig. 3. Preliminary estimation of the Economic Integration Index of the two significant regional communities 2004–2011.

References

1. Timmer, M.P., Dietzenbacher, E., Los, B., Stehrer, R. and de Vries, G.J.: User Guide to World Input–Output Database. Review of International Economics, 23: 575–605. (2015)
2. Stehrer, R.: Trade in Value Added and the Value Added in Trade, WIOD Working Paper 8, 1–19 (2012)
3. Rosvall, M., Axelsson, D., & Bergstrom, C. T.: The map equation. European Physical Journal: Special Topics, 178(1), 13-23. (2009)
4. Kichikawa, Y., Iyetomi, H., Iino, T., & Inoue, H.: Community structure based on circular flow in a large-scale transaction network. Applied Network Science, 4(1), 92. (2019)
5. Ikeda, Y., Aoyama, H., Iyetomi, H., Mizuno, T., Ohnishi, T., Sakamoto, Y., & Watanabe, T.: RIETI Discussion Paper Series 16-E-026 Econophysics Point of View of Trade Liberalization : Community dynamics , synchronization ,. RIETI Discussion Paper Series. (2016)

Shock propagation in supply and demand constrained input-output economies

Anton Pichler[1] and J. Doyne Farmer[1]

Institute for New Economic Thinking at the Oxford Martin School, University of Oxford,
United Kingdom, anton.pichler@maths.ox.ac.uk

1 Introduction

Social distancing measures adopted to combat the COVID-19 pandemic have created severe disruptions to economic output. During lockdown firms are required to shut down or substantially reduce their economic activity if they cannot comply with social distancing rules and are located in non-essential industries. Another source of negative direct shocks arises from changed consumption behavior of individuals to avoid infectious exposure. The shocks to the economy are highly industry-specific and therefore affect firms in heterogeneous ways [1]. Since firms are embedded in production networks, these direct shocks will propagate upstream (due to reduced demand to suppliers) and downstream (due to reduced supply for customers) [2][3].

Input-output (IO) analysis provides the traditional toolbox to quantify overall economic impacts in networked economies arising from exogenous supply or demand shocks [4]. However, these models are not able to incorporate supply and demand shocks simultaneously. This makes them hardly applicable to the current health crises which is characterized by both, substantial supply and demand shocks.

We show that standard IO models which allow for binding demand and supply constraints yield infeasible solutions when applied to empirical data from the United Kingdom. We then introduce a mathematical optimization procedure which is able to determine optimal and feasible market allocations, giving a lower bound on total shock propagation. We find that even in this best-case scenario network effects substantially amplify the initial shocks. To obtain more realistic model predictions, we study the propagation of shocks out of equilibrium by imposing different rationing rules on firms if they are not able to satisfy incoming demand. Our results show that overall economic impacts depend strongly on the emergence of input bottlenecks, making the rationing assumption a key variable in economic predictions.

2 Results

Let us consider the basic national accounting identity

$$x = Z\mathbf{1} + c, \tag{1}$$

where x is a vector of total output per industry, Z_{ij} the intermediate consumption of product i by industry j and c the vector of final consumption. Assuming fixed production technologies for every industry yields the classical Leontief framework,

$$x = Ax + c = Lc, \tag{2}$$

where A is the technical coefficient matrix and L the Leontief inverse.

The Leontief model is demand-driven. To incorporate demand and supply shocks simultaneously, the mixed endogenous/exogenous model (MEEM) has been suggested [5]. Yet when initializing the MEEM with shocks derived for the current pandemic [1], we find infeasible solutions for almost half of all 55 UK industries in our dataset. Specifically, the MEEM solution requires negative consumption and violates further binding economic constraints.

To quantify how initial shocks are propagating through the production network, we introduce a simple optimization procedure which globally maximizes total gross output or final consumption under supply and demand constraints imposed by the pandemic. Maximizing gross output under binding shock constraints can be formulated as linear programming problem of the form

$$\max_{c\in[0,c^{\max}]} \quad \mathbf{1}^\top(\mathbb{I}-A)^{-1}c, \qquad (3)$$
$$\text{subject to} \quad (\mathbb{I}-A)^{-1}c \in [0, x^{\max}],$$

and similarly for maximizing final consumption. Fig. 1 shows industry-specific results of the optimization procedures. While these results indicate best-case scenarios of shock propagation, we find that total economic impacts are nevertheless substantial, and strongly vary between industries. Interestingly, output and consumption maximization yield very different market allocations for industries depending on their upstreamness/downstreamness location in the economic network.

To gain a better understanding of more realistic shock propagation mechanisms, we implement various rationing algorithms. Rationing describes the decision of a firm in case it faces larger demand than it can produce. In contrast to conventional IO models, this allows us to investigate the out-of-equilibrium dynamics triggered by economic shocks. Fig. 2 shows the overall economic impact for three different rationing rules: proportional, priority and random rationing. We see for all three schemes that overall

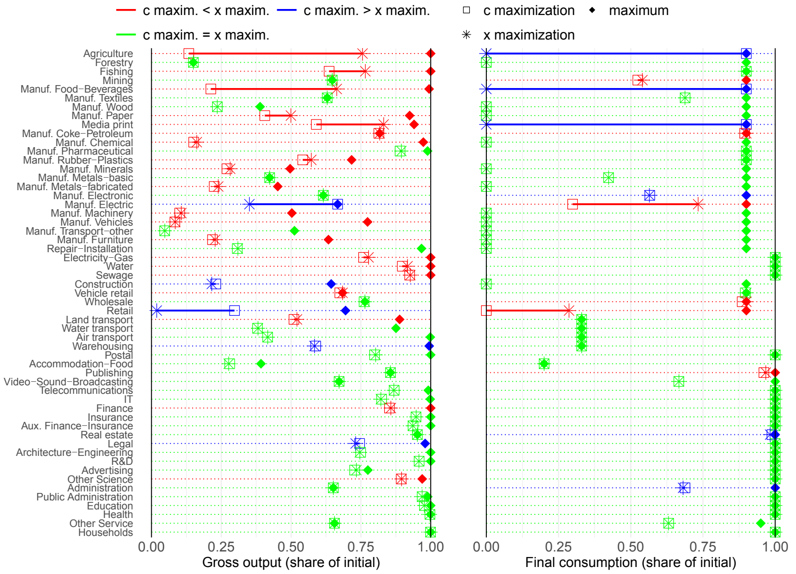


Fig. 1. Market allocations from linear optimization. The y-axis depicts industries ordered by ISIC codes. The x-axis of the left panel shows gross output values as share of initial pre-shock levels and the x-axis of the right panel shows the same for final consumption. Diamonds show the maximum output and consumption levels as obtained from the direct shocks. Squares and stars show values obtained from the consumption and output maximizations procedures, respectively. A green color indicates that output and consumption maximization yield the same outcomes. A red color indicates that the output maximization gives a higher value than consumption maximization and a blue color the opposite.

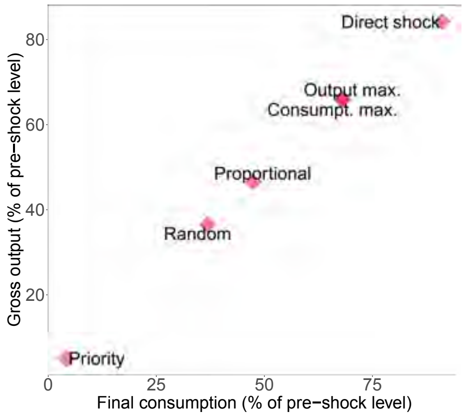


Fig. 2. Comparison of different shock propagation mechanisms.

economic impacts are much larger than what the optimization methods predict. Interestingly, priority rationing – where largest customers are served first – yields the largest economic downturns. We show in further analysis that the ordering of rationing mechanisms is not generic, but strongly depends on the underling production network topology.

Summary. We have shown that conventional IO models are not able to incorporate the large supply and demand shocks of the current COVID-19 pandemic. We derive lower bounds of economic impacts in Leontief economies by linear programming. We find even for this best-case scenario substantial total impacts. When adopting more realistic shock propagation models, we find that total economic impact estimates are highly sensitive regarding the assumed rationing mechanism. In particular, we find interaction effects between the rationing mechanisms and the network topology.

References

1. del Rio-Chanona, RM., Mealy, P., Pichler, A., Lafond, F., Farmer JD.: Supply and demand shocks in the COVID-19 pandemic: An industry and occupation perspective. Covid Economics 6. 65–103 (2020)
2. Pichler, A., Pangallo, M., del Rio-Chanona, RM., Lafond, F., Farmer, JD.: Production networks and epidemic spreading: How to restart the UK economy? Covid Economics 23, 79–151 (2020)
3. Guan, D., Wang, D., Hallegatte, S., Davies, SJ., et al.: Global supply-chain effects of COVID-19 control measures. N Human Behaviour 4, 577–587 (2020)
4. Miller, RE., Blair, PD. Input-output analysis: Foundations and extensions, Cambridge University Press (2009)
5. Dietzenbacher, E., and Miller, RE. Reflections on the inoperability input–output model. Economic Systems Research 27(4), 478–486 (2015)

Socially Responsible Investing in the Global Ownership Network and its implications for International Security

Takayuki Mizuno[1], Shohei Doi[1 2], Takahiro Tsuchiya[3], and Shuhei Kurizaki[4]

[1] National Institute of Informatics
[2] Hokkaido University
[3] Kyoto University of Advanced Science
[4] Waseda University

1 Introduction

We connect the corporate ownership network to the network of financial instruments that inject money into the ownership network. The ownership network in our data consists of 66 million nodes (i.e., companies and their shareholders) and over 90 million ownership links among them. Recent studies on the ownership network often analyze networks with more than 10 million nodes and estimate the structure of corporate control in the network based on the flow of equity stakes and associated control through a sequence of subsidiaries [1–3]. The previous studies on the network of corporate ownership and control emphasize the importance of roles played by banks and other financial institutions in the network. However, it is left unanswered where the financial institutions collect the capital from so that they can inject cash into the capital market (i.e., corporate ownership network). We therefore look into ETFs and mutual funds as the financial instruments through which the investors (individual and institutional alike) supply money in the shareholding network. In this way, we extend the existing research on corporate control in the ownership network to the network of investment funds purchased by investors (such as central banks) from asset managers and other the institutional investors.

As the shareholding network has rapidly expanded at the global scale in the recent, the international society has just begun to notice its impact on our society. Impacts of globalized ownership network, however, are hard to measure because the structure of this network has become increasingly complex with the rise of a passive investment strategy and because diversified portfolio investing strategies call for complex financial instruments such as a fund of funds and exchange-traded funds (ETF). Yet, the awareness of Environment, Social, and Governance (ESG) investing makes it more important than ever before to understand the structure of the global shareholding network and how capital flows therein to shape the distribution of corporate control and the associated risks and responsibilities.

2 Data and Method

We use three kinds of databases. (1) We obtain global corporate network data, including 66 million shareholders, from the Bureau van Dijk's Orbis database. (2) We use

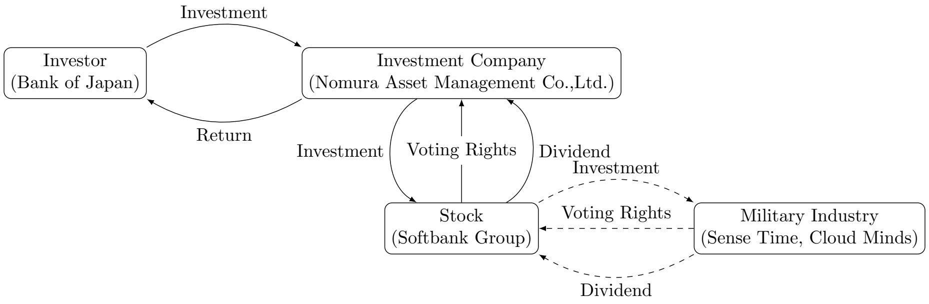


Fig. 1. Example of Investment Funds and Ownerhisp Network

the data on ETFs and mutual funds that are traded in the US and Japan from Lipper database from Refinitiv. These are about 800 thousand financial instruments in our database that supply money into the shareholding network through the asset management nodes (companies). A classic example that goes against the spirit of ESG is investment in military companies. (3) We have collected 3,793 companies globally whose company profiles in the Orbis database indicate their corporate activities involve military productions. Of those, the Chinese companies with military ties turn out to be designated by the US Department of Commerce as the companies that contribute to the Chinese government's procurement of commodities and technologies for military end-use. As an example, we investigate how investment money in the form of financial instruments such as ETFs flows from U.S. investors to Chinese munition companies through the global shareholding network.

We estimate three quantities for our analysis. First, we measure the flow of investment from the investors into the military industry through the global shareholding network. As the second quantity of interest, we measure returns for investment in the form of dividends from military industry to shareholders and in the form of profit distribution from the fund managers to the investors. Here we use the page rank. Third, we use indicators of corporate control held by the shareholders based on the voting rights. We use Network Power Index [4] rather than Network Control [1]. This is because ETFs and mutual funds imply equity ownership is dispersed and voting rights are therefore fragmented among widely spread shareholders. The indicator for corporate control must adequately account for how asset managers and other institutional investors might consolidate fragmented voting rights to establish corporate control. Further, if and when investments made by ESG investors with preferences for SRI end up injecting capital to vice industries, that would not be a result of their direct investment but as unintended consequences through the indirect ownership paths. As we demonstrate elsewhere [4], indirect corporate ownership can be established either through consolidation of dispersed ownership or through a sequence of transitive ownership, and NPI accounts for both while Network Control accounts only for the latter (i.e., transitivity).

3 The Result

The analysis of the shortest (not fastest) path from asset management companies to a Chinese munition company reveals that no asset manager either in the US or Japan directly invests in any Chinese munition companies with one link of ownership. That is, no Chinese munitions company's stock was included in the ETFs or mutual funds listed in the US and Japanese markets. However, almost all the asset managers have at least one path through which their equity stakes eventually reach a Chinese munitions company with the second or further apart links in the global ownership network. This includes ETFs purchased by the Bank of Japan from Nomura Asset Management Co. Ltd., which comprise of, among others, shares of Softbank Group that in turn invests in the military-related companies in China such as Sense Time and Cloud Minds (Fig. 1). In this particular example, Japanese citizens' tax money is invested in the Chinese military complex that poses security threat to the Japanese citizens themselves.

This result points to an important social consequence, namely causing the disparity between the power of corporate control and the stewardship responsibility. In the ideal world where the investors who strive for socially responsible investing should be empowered by their own equity stakes to make positive impacts on corporate activities related to the ESG issues. However, our analysis suggests that two obstacles encourage decoupling of equity stakes and social responsibilities so that socially responsible investors become incapable of making positive impacts with their investing strategy. The first obstacle is the fact that ETFs and other similar financial instruments separate capital and corporate control. Using our example in the figure, as Bank of Japan purchases ETFs to inject cash into the capital market, it incurs the cost of investment but Nomura Asset Management company obtains all the voting rights attached to the ownership share. If Bank of Japan invested by itself rather than purchasing ETFs, it could have obtained the power to control managerial decisions of the investing targets. The second obstacle is the complexity of ownership network itself. While Nomura Asset Management company may have the potential to control companies like Cloud Minds, the complexity of the ownership network is likely to prevent the Nomura from knowing its own potential.

References

1. Vitali, S., Glattfelder, J.B., Battiston, S.: The network of global corporate control. PloS ONE **6**(10) (2011) e25995
2. Brancaccio, E., Giammetti, R., Lopreite, M., Puliga, M.: Centralization of capital and financial crisis: A global network analysis of corporate control. Structural Change and Economic Dynamics **45** (2018) 94–104
3. Fichtner, J., Heemskerk, E.M., Garcia-Bernardo, J.: Hidden power of the big three? passive index funds, re-concentration of corporate ownership, and new financial risk. Business and Politics **19**(2) (2017) 298–326
4. Mizuno, T., Doi, S., Kurizaki, S.: The power of corporate control in the global ownership network. PloS one **15**(8) (2020) e0237862

Part XII

Social Networks

Scientific collaboration of researchers and organizations: A two-level blockmodeling approach

Marjan Cugmas, Franc Mali, and Aleš Žiberna

Faculty of Social Sciences, University of Ljubljana, Kardeljeva ploščad 5, SI-1000 Ljubljana

1 Introduction

The development and successful implementation of R&D policies depend on understanding patterns of scientific collaboration (SC). Existing studies on SC typically focus on the individual level [1–4], despite SC occurring on many interdependent social levels [5, 6]. Therefore, this research aims to provide a simultaneous insight into SC patterns among researchers (individual level) and organizations (organizational level) from the field of social sciences in Slovenia.

SC on the individual level is operationalized by co-authorship of a scientific paper, whereas two organizations are said to collaborate if they collaborate on a joint research project. The data on 788 researchers from the field of social sciences and the data about their research papers (3,367 research papers, i.e. original scientific article, review article and short scientific article) and corresponding research organizations (64 distinct organizations) were retrieved from Slovenian national information systems for the period 2006 – 2015. The 64 research organizations worked on 3,367 national and international research projects in the defined period.

Based on these data, the two-level collaboration networks were formed. These networks were analyzed by using k-means-based blockmodeling approach for linked networks [7]. A blockmodeling is an approach for reducing large and complex networks to a smaller and more interpretable structure. The nodes in a blockmodel are clusters of equivalent (regarding their structure of links) nodes from the studied network. The term block refers to a submatrix showing the link between two clusters [8].

The applied k-means based blockmodeling approach looks for homogenous blocks in terms of tie values (i.e., row/column densities) and it is considerably faster compared to the generalized blockmodeling for multilevel networks. One of the characteristics of this blockmodeling approach is that the levels with a much higher number of nodes can have a disproportionally high impact on the blockmodeling solution. Therefore, the weighting was used to ensure that both network levels have a comparable influence on the results. The number of clusters was chosen by observing the values of the criterion function [8].

2 Results

The obtained blockmodel is visualized in Fig. 1. The chosen number of clusters on an individual level is 15 (blue coloured nodes) while the number of clusters on an organizational level is 6 (red coloured nodes). Nodes' sizes for individuals are proportional to the number of researchers in each cluster, while nodes' sizes of organizations are proportional to the total number of researchers from organizations belonging to a given cluster. Widths and colour-intensity of the edges within individuals and within organizations are proportional to the blocks' density which connect two given clusters of individuals or organizations. The widths and colour-intensity of the edges between clusters from different levels are arranged to represent a share of individuals from a given cluster of individuals who affiliate with a given cluster of organizations.

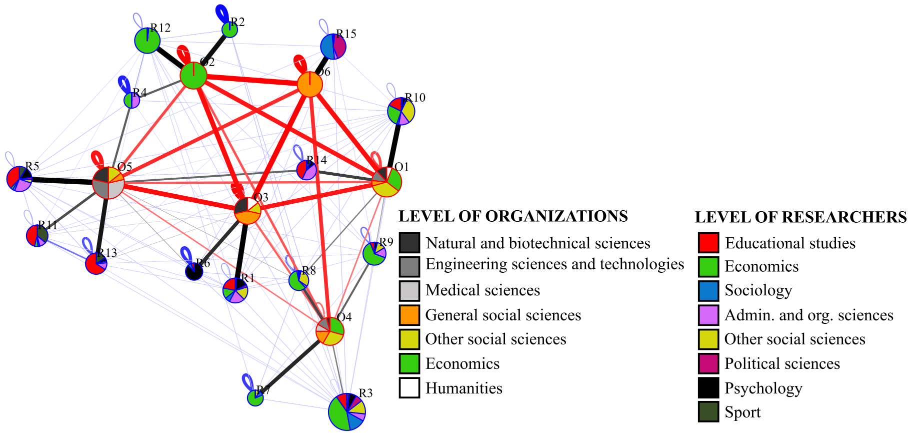


Fig. 1. Visualization of the blockmodeling solution in graph form. Nodes marked with letter "O" (red-colored nodes) represent clusters of organizations (their sizes are proportional to the total number of researchers from organizations belonging to a given cluster). Nodes marked with letter "R" (blue-colored nodes) represent clusters of researchers (their sizes are proportional to the number of researchers in each cluster). The legend corresponds to different considered scientific fields and disciplines.

It can be seen that a global network structure of researchers consists of many internally more or less linked groups of researchers. Also, the clusters of organizations are well-linked internally as well as to each other. Furthermore, SC between organizations is generally not manifested in co-authored publications between researchers from these organizations. The latter can be seen by only a few links between the clusters of researchers who are employed at organizations from otherwise linked clusters of organizations.

By looking at the obtained results, the two clusters (clusters O2 and O6 on Fig. 1) of organizations with only one organization in each (i.e., School of Economics and Bussines, University of Ljubljana and Faculty of Social Sciences, University of Ljubljana) can be identified. While all researchers from the Faculty of Social Sciences are in a single cluster, the researchers from the School of Economics and Bussines are in three distinct clusters with a low level of SC among them. The other cental organizations from the field of Social sciences in Slovenia are clustered in a single cluster (O1). Some of the researchers from this cluster collaborate with the researchers that belong to another cluster of organizations (cluster O5) from the scientific disciplines Administrative and organizational sciences, Educational studies, Sport and Psychology. Next, a cluster of organizations exists, which are geographically positioned at the coastal region of Slovenia or run research activities in this region (cluster O3). The organizations in the last cluster (cluster O4) are different faculties and institutes. These organizations are less linked to each other, and the corresponding researchers are from different scientific disciplines.

Summary. Formal incentives for SC between researchers from different organizations might not be reflected in SC on an individual level. There is a high level of inter-organization SC. Organizational SC is often not manifested in co-authored scientific papers. Their organizational affiliations highly determine the clusters of researchers.

References

1. Karlovčec M., Mladenić D.: Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. Scientometrics. 1(102), 433 – 454 (2015)
2. Kronegger L., Mali F., Ferligoj A., Doreian P.: Collaboration structures in Slovenian scientific communities. Scientometrics. 90(1) 631 – 647 (2012)
3. Mali F., Kronegger L., Ferligoj A.: Co-authorship trends and collaboration patterns in the Slovenian sociological community. Corvinus J Sociol Soc Policy CJSSP. 2(1), 29 – 50 (2010)
4. Cugmas M., Ferligoj A., Kronegger L.: The stability of co-authorship structures. Scientometrics. 106(1), 163 – 186 (2016)
5. Barbillon P., Donnet S., Lazega E., Bar-Hen A.: Stochastic block models for multiplex networks: an application to a multilevel network of researchers. J R Stat Soc Ser A Stat Soc. 180(1), 295 – 314 (2016)
6. Lazega E., Jourda M.T., Mounier L., Stofer R.: Catching up with big fish in the big pond? Multilevel network analysis through linked design. Soc Netw. 30(1), 159 – 176 (2008)
7. Žiberna A.: k-means-based algorithm for blockmodeling linked networks. Soc Netw. 60(1), 153 – 169 (2020)
8. Doreian P., Batagelj V., Ferligoj A.: Generalized blockmodeling. Cambridge University Press (2005)

Sockpuppet Detection: a Telegram case study

Gabriele Pisciotta[1], Miriana Somenzi[1], Elisa Barisani[1], and Giulio Rossetti[4]

[1] University of Pisa, Italy,
{g.pisciotta1, m.somenzi, e.barisani}@studenti.unipi.it
[2] KDD Lab, ISTI-CNR, Italy
giulio.rossetti@isti.cnr.it

1 Introduction

In Online Social Networks (OSN) numerous are the cases in which users create multiple accounts that publicly seem to belong to different people but are actually fake identities of the same person. These fictitious characters can be exploited to carry out abusive behaviors such as manipulating opinions, spreading fake news and disturbing other users [2, 4]. In literature this problem is known as the ”Sockpuppet problem”.
In our work we focus on Telegram, a wide-spread instant messaging application, often known for its exploitation by members of organized crime and terrorism, and more in general for its high presence of people who have offensive behaviors. In Italy, for example, it stepped into the spotlight because of a revenge porn case in early 2020. In this OSN users can chat both privately and in groups, and its peculiarity is that they can interact anonymously, because the service allows you to display only your nickname of choice, keeping any other personal information hidden. This feature facilitates the emergence of various fake accounts, exploiting also free VoIP numbers to create a variety of sockpuppet accounts.
The detection of sockpuppet accounts is a challenging task and the approaches [5] that have been tried during these years involve many different platforms and OSN; furthermore these techniques can be tailored or applied in a more general way. The majority of them often relies on the use of Natural Language Processing (NLP) methods, OSN-dependent features and mined behavioral patterns.
Thus we decided to tackle the problem in an innovative way to try and have more interesting results, that is by involving Network Science as a tool to match different virtual users that are actually the same person offline. To do so we took several Italian public groups on Telegram, choosing them because of the similarity in their academic contents and the fact that these groups have some users in common (they actually belong to the same virtual network of self-declared ”friends group”). We scraped these groups' messages and exploited the replies between users, all this in a completely undetectable way from the inside thanks to the fact that it is not necessary to join the groups to retrieve the informations.
We then created a directed weighted network of user interactions that connects the users who write a message, the source, to the users to whom they reply, the target. We took 17747 users as nodes and 191526 explicit replies to messages as edges, weighted by the number of replies.

2 Preliminar Results

We decided to investigate the power of Network Science applied on Sockpuppet Detection in Telegram. More in depth, we saw this problem as an instance of the link prediction task based on similarity measures, where the link has the meaning of ”same-as”. In agreement to this the whole problem can be seen as an instance of the data linking problem that, according to the particular scenario taken into account, can be named as deduplication (w.r.t. databases), instance matching (w.r.t knowledge graphs) and entity resolution: we decided to to treat it in a similar way, trying to find cluster of similar instances.

We propose a scalable method that exploits neighbours as features. In order to assess the similarity between the nodes we should compute the similarity matrix according to a certain function and filter that need to be the closest possible to a given threshold. However this operation is very costly, both in memory and time, being O(n^2) (where n is the number of nodes). To tackle this scalability issue we involved a family of algorithms known as Locality-Sensitive Hashing (LSH) [1] that create blocks in which similar entities are stored based on an hash function that takes each entity's feature as input: similar entities should lead to similar hash, so they should belong to the same cluster. Instead of computing the similarity between each pair of nodes, using LSH algorithms we drop from a complexity of O(n^2) to O(n), allowing us to scale w.r.t the number of nodes. In specific we use `SimHash` [3], to approximate the hamming distance between the the nodes and get the clusters.

Before computing the similarity between nodes, we preprocessed the data in this way:

- we normalized neighbours' weight according to each user
- we dropped links having weight less than 0.5 to only take into account the more meaningful interactions

For each node we wanted to limit the number of similar ones to have few candidates: to do this, we set the fingerprint size to 128 and the max distance to 20.

Using a real life scraped dataset we were not able to create a full ground truth; moreover we only had a partial knowledge of all the users in it. Considering this, we acted as an oracle and we tested our method on a small number of users that we knew for sure being accounts of the same person.

With only this topological information (up to now we had just used the neighbours similarity) we have found both bidirectional exact match and one-to-many matches. For the last case, we found also users that are not related, based on our knowledge. Varying the value of the signature size and the max distance, we're able to both find new correct match and to lose them. We suggest to find a trade-off between these values, according to the precision and the recall metrics.

Unfortunately the downside we encountered using this feature was the discovery of a lot of false positives belonging to people that we know to be different from the main sockpuppet account we were looking for.

Network Science can provide a very effective framework to tackle the search of sockpuppet accounts in OSN, and it can be an interesting research direction because of

the possibility of exploiting the structure of interactions that emerges with time. Furthermore this is something that is not OSN-dependent making it possible to hypothetically create an interaction graph from every OSN.
As future works we also plan to extend the experimentation involving labeled datasets and trying to combine other features mined from the graph. We have also started to explore the usage of lexical features that can be extracted from sent messages, using them in combination with the previously described ones, in order to investigate the influence of structural informations and linguistic ones. The first results are really promising and we are continuing to improve our research on the subject.

Acknowledgements

This work is supported by the scheme 'INFRAIA-01-2018-2019: Research and Innovation action', Grant Agreement n. 871042 'SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics'.

References

1. Indyk, P., Motwani, R.: Approximate nearest neighbors: Towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing. STOC '98, New York, NY, USA, Association for Computing Machinery (1998) 604–613
2. Li, J., Yuan, C., Zhou, W., Wang, J., Hu, S.: Who are controlled by the same user? multiple identities deception detection via social interaction activity (student abstract). Proceedings of the AAAI Conference on Artificial Intelligence **34**(10) (apr 2020) 13853–13854
3. Manku, G.S., Jain, A., Das Sarma, A.: Detecting near-duplicates for web crawling. In: Proceedings of the 16th International Conference on World Wide Web. WWW '07, New York, NY, USA, Association for Computing Machinery (2007) 141–150
4. Solorio, T., Hasan, R., Mizan, M.: Sockpuppet detection in wikipedia: A corpus of real-world deceptive writing for linking identities. CoRR **abs/1310.6772** (2013)
5. Yamak, Z., Saunier, J., Vercouter, L.: Sockscatch: Automatic detection and grouping of sockpuppets in social media. Knowledge-Based Systems **149** (2018) 124 – 142

Evolution of Political Polarization on Slovenian Twitter

Bojan Evkoski[1,2], Igor Mozetič[2], Nikola Ljubešić[2], and Petra Kralj Novak[2]

[1] Jozef Stefan International Postgraduate School, Ljubljana, Slovenia,
[2] Jozef Stefan Institute, Ljubljana, Slovenia,
{Bojan.Evkoski, Igor.Mozetic, Nikola.Ljubesic,
Petra.Kralj.Novak}@ijs.si

We analyze the evolution of Twitter activities in Slovenia in recent years. We construct networks, with Twitter users as nodes, and retweet relations as edges. We detect communities and influential users in them, and track how they evolve during times of political changes and start of the Covid-19 pandemic. We observe the following: Most of the influential users around which communities emerge are related to politics, the political polarization is increasing, and the right leaning Twitter users are considerably more active.

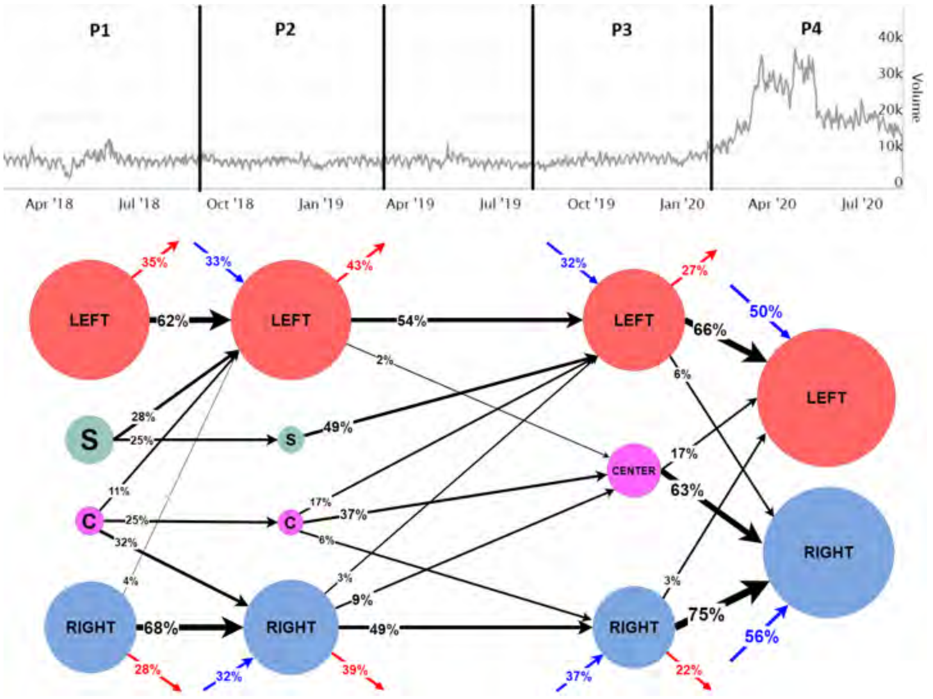


Fig. 1. Volume of tweets (top) and evolution of the top communities (bottom) across the four periods analyzed, P1–P4. Considered are only communities with more than 2% of the Twitter users in each period (S stand for the SPORTS, and C for the CENTER community). Size of a community corresponds to the number of its users. Black arrows show the flow of users between the communities, and percents refer to fractions of the source community. Red arrows indicate users leaving a community, and blue arrows indicate new users (or users from smaller communities) joining a community (shown just for the LEFT and RIGHT communities).

It turns out that the retweet communities very well reflect the actual political alignments. We already demonstrated that political parties and nationality of the members of the European Parliament can be reconstructed solely from their retweet activities [3]. We also showed that there is a correspondence between the co-voting and retweeting in the European Parliament, while higher Twitter activity was observed for the right-wing parties [2]. Also, in the case of Brexit, the Leave proponents showed much higher activity and influence on Twitter than the Remain proponents [4].

We collected most of the tweets from the Slovenian users in recent years with the TweetCat tool [6], built specifically for acquisition of Twitter data for "smaller" languages. For the current work, the collected tweets are split into four 6-months periods corresponding to major political events:

- P1 (Mar. 2018 - Aug. 2018) - government resignation and snap parliamentary elections (on 14 Mar. and 3 Jun. 2018, respectively),
- P2 (Sep. 2018 - Feb. 2019) - left-wing government formation (on 13 Sep. 2018),
- P3 (Aug. 2019 - Jan. 2020) - left-wing government resignation (on 27 Jan. 2020),
- P4 (Feb. 2020 - July 2020) - right-wing government formation (on 13 Mar. 2020) and emergence of the Covid-19 pandemic in Slovenia.

For each period, P1–P4, we detect communities by the Louvain method [1] which maximizes modularity. We identify influential users in terms of the Hirsch index (*h-index*) [5], adapted to Twitter [4]. A user with an index of h has posted h tweets and each of them was retweeted (RT) at least h times: $h\text{-}index(RT) = \max_i \min(RT(i), i)$.

Table 1 shows basic network statistics for the four time periods. If we ignore smaller communities which contain less than 2% of the users, the largest four communities comprise more than 92% of all the users. The communities are labeled as LEFT, SPORTS, CENTER, and RIGHT by their most influential members.

Period	P1	P2	P3	P4
Twitter users (nodes)	8,334	7,952	7,315	9,760
retweeted tweets	155,730	146,806	165,733	410,206
retweets (weighted edges)	448,962	412,434	424,729	1,648,807
communities (> 1%)	6 (95%)	5 (95%)	5 (96%)	2 (96%)
communities (> 2%)	4 (92%)	4 (93%)	3 (93%)	2 (96%)
modularity	0.40	0.38	0.35	0.32
Average *h-index* of the top 20 influencers				
LEFT	15	14	13	27
SPORTS	7	5	/	/
CENTER	12	11	18	/
RIGHT	48	46	43	66

Table 1. The Slovenian retweet networks during periods P1–P4. Size of the networks, the number of communities with more that 1% or 2% of the users with corresponding fractions of all the users covered (top), and average influence of the top 20 users in the largest communities (bottom).

Fig. 1 shows the transitions of the users between the communities across the time periods P1–P4. We observe the dominance of the LEFT and RIGHT communities, and how they eventually absorb the smaller communities (SPORTS is largely absorbed by

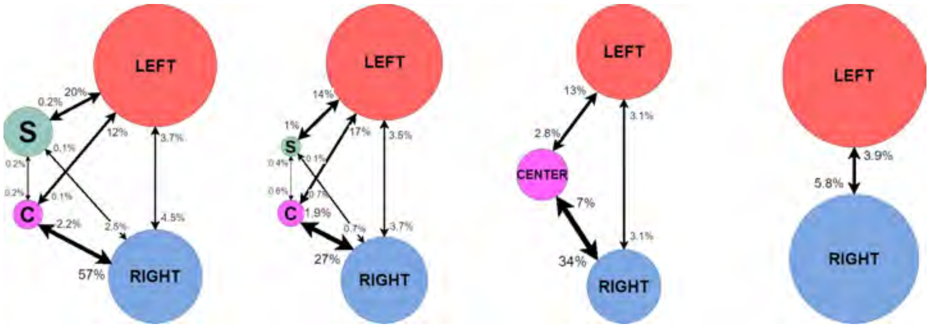


Fig. 2. Inter-community retweeting between the top communities during each of the four periods, P1–P4. Size of a community corresponds to the number of its users, and arrows correspond to fractions of retweets. For example, in P3, 34% of all the retweets by CENTER are from the RIGHT community, and 13% are from LEFT. The rest (53%, not shown) are retweets from the CENTER community itself, i.e., defining it as a community.

LEFT, and CENTER by RIGHT). There is a relatively large fraction of the Twitter users leaving the dominant communities or joint them anew (indicated by the red and blue arrows, respectively). We did not yet exhaustively check the robustness of the Louvain community detection algorithm on this data, but so far there are strong indications that the cores of the communities, in terms of the influential users, remain stable.

Fig. 2 shows the retweeting activity between different communities. CENTER is retweeting more from the RIGHT and is eventually absorbed by the RIGHT. However, CENTER also acts as a link between LEFT and RIGHT (most pronounced in period P3), and this link disappears in period P4. The last period, P4, is characterized not only by an increase in Twitter activities, due to the Covid-19 pandemic, but also by increased political polarization. Our preliminary experiments also indicate that the amount of inappropriate, offensive and violent hate speech is increasing in this recent period.

Acknowledgments. The authors acknowledge financial support from the Slovenian Research Agency (research core funding no. P2-103 and P6-0411), and the European Union's Rights, Equality and Citizenship Programme (2014-2020) project IMSyPP (grant no. 875263).

References

1. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.
2. D. Cherepnalkoski, A. Karpf, I. Mozetič, and M. Grčar. Cohesion and coalition formation in the European Parliament: Roll-call votes and Twitter activities. *PLoS ONE*, 11(11):e0166586, 2016.
3. D. Cherepnalkoski and I. Mozetič. Retweet networks of the European Parliament: Evaluation of the community structure. *Applied Network Science*, 1(1):2, 2016.
4. M. Grčar, D. Cherepnalkoski, I. Mozetič, and P. Kralj Novak. Stance and influence of Twitter users regarding the Brexit referendum. *Computational Social Networks*, 4(1):6, 2017.
5. J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 16569–16572, 2005.
6. N. Ljubešić, D. Fišer, and T. Erjavec. TweetCaT: a tool for building Twitter corpora of smaller languages. *Proc. LREC'14*, 2279–2283, 2014.

Clustering Active Users in Twitter Based on Top-*k* Trending Topics

Tanjim Taharat Aurpa, Md Shoaib Ahmed, and Md Musfique Anwar

Department of Compurter Science and Engineering, Jahangirnagar University, Bangladesh
(taurpa22,shoaibmehrab011)@gmail.com, manwar@juniv.edu

1 Introduction

Discovering meaningful topical clusters in online social networks (OSNs) has recently occupied an overwhelming research interest owing to its diverse applications. Most existing approaches focus on the contents generated by the social users and link structure of the underlying social network [1,2]. However, the degree of users' *temporal topical activeness* has not been thoroughly studied to identify its effect on the formation of topical clusters. As a result, the resulting clusters may contain mix of high and low active users as well as may contain users who have no inclination towards the query attributes. This research investigates on how the users' behaviors and topical activeness vary with time and how these parameters can be employed in order to improve the quality of the detected topical clusters for top-*k* trending topics at different time intervals. Our observation is that users have different degrees of topical activeness which vary widely over time. The proposed approach is commenced on measuring the degree of activeness for each candidate cluster member with respect to the given query attributes to enhance the quality of the detected topical clusters. Instead of giving the query topics manually, our system identify the top-*k* trending topics by taking into account the number of mentions on that topics and the coverage of that topics in OSN.

2 Problem Statement and Proposed Framework

We introduce some relevant concepts before defining the problem statement.
Activity: Activity refers to an action that a user performs at a time point. For example, a user u in Twitter posts a tweet (message) containing a specific topic T_i at time t_j. This activity is recorded as an activity tuple $\langle u, T_i, t_j \rangle$.
Query: An input query $Q = \{\mathscr{T}_q\}$ consisting top-*k* trending Topics $\mathscr{T}_q = \{T_i, T_{i+1} ..., T_n\}$ at a particular time interval.
Topical Interest Score: For each user $u_i \in U$, we compute her topical interest score (denoted by Ω) to measure the involvement of u_i towards the given query attributes $\mathscr{T}_q$ of Q, using Equation (where $Q_{u_i} \in Q$) here below:

$$\Omega_{I_m}(u_i, Q_{u_i}) = \frac{|ACTS(u_i, Q_{u_i})|}{\Lambda_{(Q,U^Q_{I_m})}} \quad \text{,where} \quad \Lambda_{(Q,U^Q_{I_m})} = \frac{\sum_{u_i \in U^Q_{I_m}} |ACTS(u_i, Q_{u_i})|}{|U^Q_{I_m}|} \tag{1}$$

where, $ACTS(u_i, Q_{u_i})$ indicates the set of activities related to Q performed by u_i and $\Lambda_{(Q,U^Q_{I_m})}$ denotes the *average* number of activities related to Q performed by $U^Q_{I_m}$ in attributed graph G, where $U_{I_m}^{Q}$ indicates only those users who posted tweets related to Q at time interval I_m. Then, the activeness (denoted as σ) of u related to Q is

$$\sigma_{(u_i,Q_{u_i})} = \frac{\Omega_{I_m}(u_i,Q_{u_i})}{max_{u_z \in U_{I_m}^{Q}} \{\Omega_{I_m}(u_z,Q_{u_z})\}} \tag{2}$$

Problem Definition: Given an attributed graph G = *(U, E, $\mathscr{T}$)*, an input query $Q = \{\mathscr{T}_q\}$, a positive integer k, and a threshold value of θ, we want to group users into three different clusters (namely $\mathscr{C}_{\mathscr{H}}$, $\mathscr{C}_{\mathscr{M}}$ and $\mathscr{C}_{\mathscr{L}}$ as high, medium and low active groups respectively) based on their topical interest scores. We consider a threshold $\theta \in [0,1]$ so that each user has to show her inclination to at least $|Q|\ .\ \theta$ topics.

In our proposed model, we set the value of the query Q at each time interval I_m as the top-k trending (busty) topics at that I_m. We define trending score ($\eta_{(T_z,I_m)}$) for each topic T_z according to equation mentioned below::

$$\eta_{(T_z,I_m)} = \alpha\ .\ \left|ACTS(*,T_z)\right| + (1-\alpha)\ .\ U_{T_z,I_m} \tag{3}$$

where $\left|ACTS(*,T_z)\right|$ indicates the total number of activities related to topic T_z and U_{T_z,I_m} represents the number of users who showed their interests on T_z at time interval I_m. The weighting parameter $\alpha \in [0,1]$ balances the above two factors.

3 Experimental Evaluation

We use a Twitter dataset from SNAP [1] data collection repository and randomly choose 4,00,000 users and consider their tweets from June 16, 2009 to June 30, 2009. The interest scores of *active users* (having at-least 10 activities related to Q at I_m) ranges are greater than 0.75, between 0.41 to 0.75 and between 0.25 to 0.4 for $\mathscr{C}_{\mathscr{H}}$ (high), $\mathscr{C}_{\mathscr{M}}$ (medium) and $\mathscr{C}_{\mathscr{L}}$ (low) clusters, respectively. We use two measures of *entropy* and *cluster topical expertise level* to evaluate the quality of the detected clusters.

$$entropy(\{\mathscr{C}j\}_{j=1}^{r}) = \sum_{j}^{r} \frac{|U(\mathscr{C}_j)|}{|U|} entropy(\mathscr{C}_j) \quad \text{,where} \quad entropy(\mathscr{C}_j) = -\sum_{i=1}^{n} p_{ij} log_2 p_{ij} \tag{4}$$

Here $\frac{|U(\mathscr{C}_j)|}{|U|}$ is the weighted probability of a cluster's user and p_{ij} is the percentage of users in cluster $\mathscr{C}_j$ which are active on query topic T_i.

The term *entropy* $(\{\mathscr{C}_j\}_{j=1}^{r})$ measures the weighted entropy considering all the query topics over all the (r) clusters. Entropy indicates the randomness of topics discussed in clusters. Next, we measure the semantic cohesion related to Q in each cluster considering the main topic of interest of each user u_i according to Equation given below.

$$\psi_{(u_i,I_m)} = freqmax_Q ACTS(u_i,Q_{u_i}) \tag{5}$$

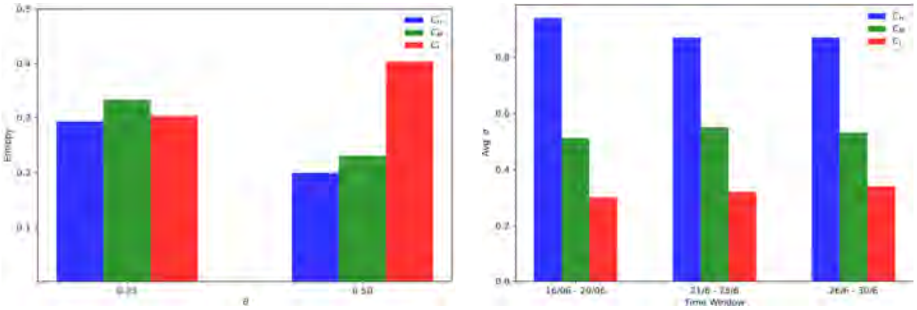
Similarly, Equation **??** defines the most frequent topic in a cluster $\mathscr{C}_j$ at I_m.

$$\lambda_{(\mathscr{C}_j,I_m)} = freqmax_Q \psi_{(u_i,I_m)} \tag{6}$$

Finally, we measure the expertise level of a cluster (denoted as $\rho_{(\mathscr{C}_j,I_m)}$) for a particular topic T_z at time interval I_m (mentioned in Equation 7). Generally, a good topical cluster should have low entropy value and high semantic cohesion towards the given Q.

$$\rho_{(\mathscr{C}_j,I_m)} = \frac{\#\{u_i \in \mathscr{C}_j,\ \ \psi_{(u_i,I_m)} = \lambda_{(\mathscr{C}_j,I_m)}\}}{|\mathscr{C}_j|} \tag{7}$$

[1] http://snap.stanford.edu/data/twitter7.html



(a) Entropy (b) Average interest score

Fig. 1. (a) Entropy at time interval (26/06 - 30/06), (b) Average interest score at different time intervals at each topical cluster for top-*k* trending topics (in all cases, $k = 4$, $\alpha = 0.5$, $\theta = 0.5$))

Fig. 1 (a) shows the entropy values at a particular time interval (26/06-30/06) where the trending topics set Q is $\{$*Jobs, News, Father's day and Politics*$\}$. Threshold θ values of 0.25 and 0.5 indicate that users have to *active* at-least 1 and 2 topics related to Q respectively. In both cases, we see that entropy values are higher in $\mathscr{C}_{\mathscr{M}}$ and $\mathscr{C}_{\mathscr{L}}$ as not all the users show their inclination to all the topics related to Q. On the other hand, as the members in $\mathscr{C}_{\mathscr{H}}$ have high degree of activeness towards Q, so most of them have to pay attention to all the query topics. Fig. 1 (b) shows that the members in high cluster ($\mathscr{C}_{\mathscr{H}}$) has higher average score than other clusters at different time intervals.

Table 1. Semantic cohesion ($\rho_{(\mathscr{C}_j, I_m)}$) for Top-*k* Trending Topics

Time Window	θ **= 0.25**	θ **= 0.50**
I_1 (16 / 06 - 20 / 06)	$\mathscr{C}_{\mathscr{H}}$ = 0.750 (Business)	$\mathscr{C}_{\mathscr{H}}$ = 0.692 (Business)
	$\mathscr{C}_{\mathscr{M}}$ = 0.406 (Business)	$\mathscr{C}_{\mathscr{M}}$ = 0.611 (Business)
	$\mathscr{C}_{\mathscr{L}}$= 0.667 (Business)	$\mathscr{C}_{\mathscr{L}}$ = 0.633 (Business)
I_2 (21 / 06 - 25 / 06)	$\mathscr{C}_{\mathscr{H}}$ = 0.50 (News)	$\mathscr{C}_{\mathscr{H}}$ = 0.538 (News)
	$\mathscr{C}_{\mathscr{M}}$ = 0.414 (News)	$\mathscr{C}_{\mathscr{M}}$ = 0.324 (Politics)
	$\mathscr{C}_{\mathscr{L}}$ = 0.412 (Politics)	$\mathscr{C}_{\mathscr{L}}$ = 0.333 (Politics)

Table 1 shows the expertise level ($\rho_{(\mathscr{C}_j, I_m)}$) of each cluster for the most frequent topic of that cluster at different time intervals. We find that *Business* is the most frequent topic in all the clusters at time interval I_1 for θ values of 0.25 and 0.5, respectively. On the other hand, *News* become most frequent topic at time interval I_2 in most cases. In all the above cases, $\mathscr{C}_{\mathscr{H}}$ is found as most coherent cluster.

4 Conclusion

Our observation is that the users' individual activeness vary widely for different attributes. This research outlined an activeness score function for the social users and developed methods to effectively cluster them for top-*k* trending topics.

References

1. M. Michelson and S. A. Macskassy. Discovering Users' Topics of Interest on Twitter: A First Look.In CIKM, pp. 73-80 (2010)
2. G. Qi, C. C. Aggarwal, and T. Huang. Community detection with edge content in social media networks. In Proc. ICDE, pp. 534–545 (2012)

The Italian Twittersphere discussion on migration: a network analysis

Tommaso Radicioni[1,2], Fabio Saracco[2], and Tiziano Squartini[2]

[1] Scuola Normale Superiore, P.zza dei Cavalieri 7, 56126 Pisa (Italy)
[2] IMT School for Advanced Studies, P.zza S. Francesco 19, 55100 Lucca (Italy)

1 Introduction

The advent of social media and microblogging platforms over the last decade has brought fundamental changes to the way people access and share information, communicate, etc. According to Eurobarometer, the percentage of Europeans making *daily* use of social networks to access news has increased from 18% in 2010 to 42% in 2017 [1]. So far, however, researchers have mainly focused on the *behaviour* of social media users, individuating several stylized facts. For example, it has been observed that online users are more likely to select information adhering to their system of beliefs and joining groups (the so-called "echo chambers") supporting it; these groups, in turn, are often found to be polarized, hence promoting an approach to discussions which is strongly negatively biased [2]. Relatively less attention has, instead, been paid to the *semantic* aspect of online conversations and its nexus with the relationships amongst users participating to them.

Our contribution aims at overcoming the limitations of present studies [3] by exploring the structural properties of both the networks of *actors* and the networks of *topics*. In other words, our approach allows us to combine both the *behavioral* and the *semantic* aspect of online political debates to unveil the communication strategies and the backbone of the narratives developed by different political groups. Our data set consisted of approximately 5 millions of tweets induced by the Italian debate about migration and posted by 306.894 users across the period May-November 2019. For the present analysis, we have focused on Twitter, a choice driven by the evidence that it is massively used during political debates [4], by the vast majority of public figures (e.g. political leaders, journalists, official media accounts), to provide visibility to their statements. Amongst all types of interaction modes characterizing the Twitter platform, the current study grounds on *retweets*, i.e. a relational mechanism which is particularly insightful when studying collective political identities [5].

2 Results

Following the approach of [6, 7], we have considered the bipartite networks of *verified users* retweeted by *non-verified users* at a monthly time scale and proceeded in a two-step fashion. First, we have obtained the corresponding set of (monopartite) projections on the layer of verified accounts, by comparing the empirical number of times any couple of verified users has been retweeted by the same non-verified users with the outcome

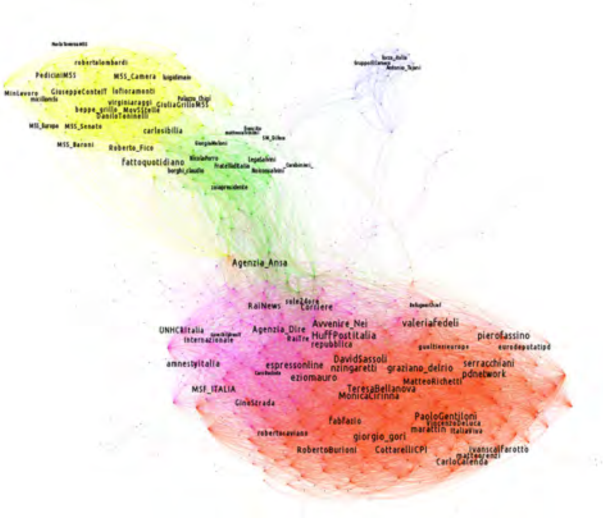


Fig. 1. The monopartite projection on the layer of verified users "summing up" the entire observation period (i.e. May-November 2019). This network clearly spots out five largest communities corresponding to the (members and supporters of) the major Italian political parties.

of a properly-defined benchmark model (in our case, the so-called *Bipartite Configuration Model* [8]); then, we have run the Louvain community detection algorithm on the obtained projections.

Remarkably, the detected groups - named *discursive communities* and interpreted as clusters of users sharing similar contents, in turn able to trigger a discussion - can be identified with (members and supporters of) the main Italian political parties. As reported in Figure 1, the projection "summing up" the entire observation period (and obtained by properly combining the monthly projections) is characterized by the presence of five largest communities, i.e. the supporters of "Movimento 5 Stelle" (*Five Star Movement*) (yellow), the supporters of far-right parties (green), the supporters of center-left parties (red), a community of media and NGO accounts (purple) and a community of politicians and supporters of the center-right party "Forza Italia" (*Go Italy*). Moreover, at a finer grained level (i.e., the monthly time scale), all the discursive communities follow the evolution of the corresponding political coalitions. For instance, our analysis clearly spots out the effects of the 2019 Italian government crisis, leading the center-left leaning community to split into two sub-communities.

In order to provide a deeper insight into the online discussions characterizing each discursive community, we have also analyzed the structure of the corresponding monthly *semantic networks* which are obtained by projecting each bipartite user-hashtag network on the layer of hashtags (as hashtags play a central role in the Twitter environment, we have defined the nodes of our semantic networks to be precisely the hashtags extracted from the text of the collected tweets). Out of the four different benchmarks employed in our work to obtain the semantic networks, here we consider the so-called *naïve* one (prescribing to link any two hashtags sharing a non-negative number of users) and the

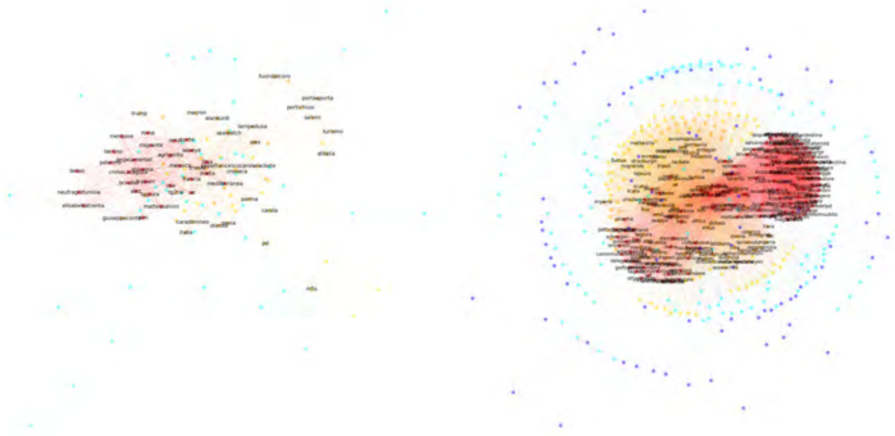


Fig. 2. The July 2019 semantic network of the far-right leaning community obtained by employing two different filtering procedures, i.e. the *naïve* one, prescribing to link any two hashtags sharing a non-negative number of users (right panel) and the one defined by the *Bipartite Partial Configuration Model* (left panel) [8].

one defined by the *Bipartite Partial Configuration Model* [8]. The projection obtained by employing the latter lets the most persistent (groups of) hashtags emerge: as shown in Figure 2, the bulk of the discussion characterizing the far-right leaning community, in July 2019, is represented by keywords such as *#portichiusi*, *#fuoridalcoro*, *#salvini*.

Interestingly, the *k-shell decomposition* of the semantic networks obtained by applying the naïve recipe has revealed the presence of a *core-periphery* structure (i.e. a bunch of very well-connected vertices, linked to a group of low-degree, loosely inter-linked nodes) characterizing the monthly activity of each discursive community.

References

1. Eurobarometer. Standard eurobarometer 88 “Media Use in the European Union” Report. European Commission - Public Opinion, (2017). Accessed July 30, 2020.
2. Del Vicario, M. et al. *Sci. Rep.* 7, 40391 (2017). https://doi.org/10.1038/srep40391
3. Cherepnalkoski, D. and Mozetič, I. *Appl. Netw. Sci.* 1, 2 (2016). https://doi.org/10.1007
4. Jungherr, A. *J. Inf. Technol. Politics* 13, 1 (2016). https://doi.org/10.1080/19331681.2015.1132401
5. Conover, M. D. et al. 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, 2011, pp. 192-199. https://doi.org/10.1109/PASSAT/SocialCom.2011.34.
6. Becatti, C. et al. *Palgrave Comm.* 5, 91 (2019). https://doi.org/10.1057/s41599-019-0300-3
7. Caldarelli, G. et al. *Commun. Phys.* 3, 81 (2020). https://doi.org/10.1038/s42005-020-0340-4
8. Saracco, F. et al. *New J. Phys.* 19, 053022 (2017) https://doi.org/10.1088/1367-2630/aa6b38

Paternal-maternal surname networks reveal the population structure of Santiago, Chile

Naim Bro[1] and Marcelo Mendoza[1][2]

[1] Millennium Institute Foundational Research on Data, Santiago, Chile,
naim.bro@imfd.cl,
http://imfd.cl/en/investigador/naim-bro/
[2] Department of Informatics, Universidad Técnica Federico Santa María, Santiago, Chile,
marcelo.mendoza@usm.cl,
https://www.inf.utfsm.cl/mmendoza/

1 Introduction

Insofar surnames are associated with ancestry, they contain social information. Previous research has used last names to uncover the genetic [1,2], ethnic [3,4], and linguistic [5] composition of populations. Importantly, surnames can be a source of relational data. Previous research constructs surname networks from geographic proximity; if two surnames tend to co-occur in space, then they are made to connect [6]. In countries where individuals hold paternal and maternal surnames, last names are a direct source of relational data. For example, if Elena's paternal last name is González and her maternal last name is Muñoz, this is likely to mean that, at some point, a González and a Muñoz — her parents — had a relationship.[3]

This article exploits the social and relational properties of surnames to produce a synthetic view of the population of Santiago, Chile. From administrative records of more than four million names, it creates a network of paternal-maternal last names and associates them to an approximate socioeconomic index. It identifies the clusters that emerge and describes them. The hypothesis is that socially similar surnames form clusters.

2 Data and methods

The Chilean electoral registry of 2012 contains the full name, the unique identifying number,[4] and the address of all individuals eligible to vote for political authorities in Chile.[5] Only residents of Santiago were included in this analysis, totaling 4,652,933 individuals. An initial surname network was made using all paternal and maternal surname pairs and weighting them after their occurrence. For example, if the electoral registry contains 100 individuals named González Muñoz, then the González-Muñoz

[3] On some occasions this is not the case. For example, when people change their last names, or when the identity of the father is not known.

[4] Registro Único Tributario: RUT, in Chilean administrative parlance

[5] Persons over 18 years of age, including Chilean citizens and foreigners that have resided in Chile more than five years

pair is given an initial weight of 100.

Ties that do not seem to express affinity are dropped. The González and Muñoz of Santiago may possess multiple connections; however, if their occurrence (n_{ss}) is less than expected given their large sizes, then their tie is not indicative of affinity. In contrast, two rare surnames — say Yarur and Manzur — may have fewer connections, but still more than expected given their reduced numbers; if so, this connection is preserved. Following [7], we remove surname pairs if their n_{ss} falls below a threshold defined by $n_{ss} > \frac{kn_{s1}n_{s2}}{N}$, where K is a constant, n_{s1} is the occurrence of the first surname, n_{s2} is the occurrence of the second surname, and N is the total number of individuals. Constant k establishes a threshold for tie unexpectedness, and is determined by the researcher. We chose a $k = 25$. Finally, if n_{ss} is above 1, then the pair is kept. In a second filtering step, all nodes (surnames) that fall below a network coreness threshold equal to 4 are dropped. The resulting graph contains 3728 nodes/surnames. Clusters are identified using the Louvain community detection algorithm [8].

The second source of data is the Índice de Bienestar Territorial (IBT) of 2012 [9], which indexes the mean socioeconomic level of every census administrative unit down to the block level. The data building phase involved geocoding every address contained in the electoral registry using the Google Maps API, which yields four types of definitions: approximate, geometric center, range interpolated, rooftop. Only addresses geocoded with rooftop- and range interpolated-level precision were kept in the analysis, leaving 3,720,431 registers. Then each address was matched with a census block, and the socioeconomic status of individuals was imputed based on the mean socioeconomic level of the block they live in.

3 The clusters

The community structure of the resulting graph contains 10 clusters. Three of these clusters facilitate an intuitive interpretation. In Table 1, clusters 0 and 5 possess a high socioeconomic level, and cluster 2 possesses low socioeconomic status. The latter is formed by Mapuche surnames such as Collio, Painen, Curiqueo, and Curín (see Appendix A for a detail of representative surnames per cluster). The most representative surnames of cluster 5 are Larraín, Vial, and Errázuriz, which contextual familiarity with Chilean surnames suggests represent the traditional upper class. The most representative surnames of cluster 0 are Jadue, Nazar, Awad, and Manzur, all Palestinian surnames. The upper-class and Palestinian clusters possess similar socioeconomic levels, but, historically, the former has been dominant in politics. Figure 4 represents the proportion of parliamentarians holding surnames belonging to each cluster; we see that cluster 5 has been dominant throughout two centuries, especially in the nineteenth century.[6]

The clusters also differ in their geographical distribution. Figure 1 through Figure 3 visualize where in Santiago individuals holding the most representative surnames of

[6] The list of Chilean parliamentarians from 1834 to 2018 was extracted from the site of the Chilean congress at www.bcn.cl (accessed on 20 May 2020).

each cluster live. Both the aristocratic and Palestinian clusters are concentrated in the north-east quarter of the city.[7] The Mapuche cluster is overrepresented in the lower-income neighborhoods of Cerro Navia in the north-west, and La Pintana in the south part of Santiago.

4 Conclusion

This paper builds a network of paternal-maternal surname pairs to reveal the community structure of the population of Santiago, Chile. It reveals 10 clusters, three of which are intuitively interpretable, those representing the Mapuche people, the descendants of Palestinians, and the traditional Chilean upper class. Individuals holding Mapuche surnames were considerably more deprived economically than the Palestinian and upper-class clusters. The two latter possess similar socioeconomic status and live in the same part of Santiago, but the upper-class cluster is better represented in politics.

References

1. Roguljić, D., Rudan, I., Rudan, P. Estimation of Inbreeding, Kinship, Genetic Distances, and Population Structure from Surnames: The Island of Hvar, Croatia, American Journal of Human Biology, 1997. 9(5):595–607.
2. Manni, F., Toupance, B., Sabbagh, A., Heyer, E. New Method for Surname Studies of Ancient Patrilineal Population Structures, and Possible Application to Improvement of Y-Chromosome Sampling, American Journal of Physical Anthropology, 2005. 126(2):214-228.
3. Scapoli, C., Mamolini, E., Carrieri, A., Rodriguez-Larralde, A., Barrai, I. Surnames in Western Europe: A Comparison of the Subcontinental Populations through Isonymy, Theoretical Population Biology, 2007. 71(1):37–48.
4. Cheshire, J., Mateos, P., Longley, P. Delineating Europe's Cultural Regions: Population Structure and Surname Clustering, Human Biology, 2011. 83(5):573–598.
5. Scapoli, C., Goebl, H., Sobota, S., Mamolini, E., Rodriguez-Larralde, A., Barrai, I. Surnames and Dialects in France: Population Structure and Cultural Evolution, Journal of Theoretical Biology, 2005. 237(1):75–86.
6. Novotný, J., Cheshire, J. The Surname Space of the Czech Republic: Examining Population Structure by Network Analysis of Spatial Co-Occurrence of Surnames, PLoS ONE, 2012. 7(10), PMID: 23119060.
7. Mateos, P., Longley, P., O'Sullivan, D. Ethnicity and Population Structure in Personal Naming Networks, PLoS ONE, 2011. 6(9), PMID: 21909399.
8. Blondel, V., Guillaume, J., Lambiotte, R., Lefebvre, E. Fast unfolding of communities in large networks, Journal of Statistical Mechanics, 2008. P10008.
9. CIT (Centro de Inteligencia Territorial). Índice de Bienestar Territorial 2012. Santiago: Universidad Adolfo Ibáñez, 2012.

Appendix A: figures and tables

[7] Palestinians are also overrepresented in Patronato, a lower-income neighborhood specialized in commerce. Historically, this is where Palestinians migrating to Santiago arrived.

Table 1. Descriptive statistics per cluster

	N	Percent of population	Mean NSE	Mapuche	Palestinian	Jewish
0	345	2.4	77.9	0.9	15.7	2.9
1	482	8.4	61.3	3.1	0.0	0.6
2	769	3.8	41.6	86.2	0.0	0.3
3	56	0.8	67.6	1.8	0.0	5.4
4	486	9.0	62.4	4.1	0.0	3.3
5	874	6.8	77.3	0.8	0.2	3.0
6	315	5.4	61.9	5.1	0.3	1.6
7	79	0.7	63.4	2.5	0.0	0.0
8	147	1.7	60.7	5.4	0.0	0.0
9	175	1.9	69.3	2.3	0.0	16.6

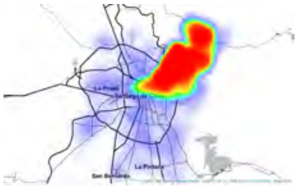


Fig. 1. Geographical distribution of aristocratic cluster (representative surnames)

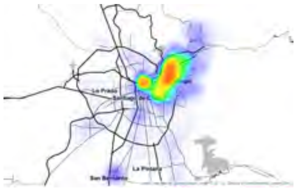


Fig. 2. Geographical distribution of Palestinian cluster (representative surnames)

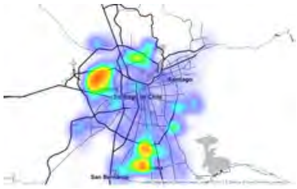


Fig. 3. Geographical distribution of Mapuche cluster (representative surnames)

Fig. 4. Political representation of clusters

Towards Mesoscopic Structural Analysis of the Fediverse of Decentralized Social Networks

Lucio La Cava, Lucas E. Ruffo, and Andrea Tagarelli

DIMES, University of Calabria, Rende (CS), Italy
tagarelli@dimes.unical.it

1 Introduction

Open-source, distributed decentralized online social networks (DOSNs) are emerging as alternatives to the popular though centralized platforms – i.e., hosted and controlled by a single company – like Facebook or Twitter. In DOSNs, everyone is allowed to get the code and set up their own server. A particularly relevant model is the *federated* one, whereby each server can communicate with the others using the same protocol, meaning that if a user is signed up for a certain server, s/he is still able to interact with users on another server – in a similar fashion as for the email service. In addition, any platform that implements the common protocol, such as ActivityPub, becomes part of a massive social network, called *Fediverse*, thus enabling individuals to use their accounts on a platform to follow users on other platforms without needing an account there.

The Fediverse currently provides several services, such as Mastodon and Friendica for microblogging, PeerTube and Funkwhale for video hosting, PixelFed for image hosting. To our knowledge, Mastodon is the only platform in the Fediverse that has received relative attention from the research community [1–6]. From a network science perspective, the studies by Zignani et al. [3, 5] are particularly relevant: they are the first to analyze a large portion of the Mastodon user-network, focusing on degree distribution, triadic closure, and assortativity aspects, and comparing such characteristics to those in Twitter [3]; also, in [5], they investigate how the decentralization process affects relationships between users, unveiling that instances show individual footprints (based on degree distribution and clustering coefficient statistics observed on the top-10 instances in Mastodon) that influence relationships. However, several open questions still remain to address on Mastodon, which motivated us to focus on this platform in this work. Our major contribution is to fill a gap in the understanding of the mesoscopic structure of Mastodon: in fact, unlike existing works, our study builds upon a network model over the instances and exploits it to provide insights into connections among instances, and hence their users, over the detected modularity-based community structure as well as core-decomposition. Given the versatility of our analysis methodology, we believe our study can serve as a starting point to analyze the entire Fediverse of DOSNs.

2 Methodology and Results

To conduct our study, we referred to the Mastodon data provided in [3], which contains about 6.5 million *following* relations for 566 000 users, covering 4 015 instances – which is more than half of the Mastodon instances created to date.[1] Upon this data,

[1] Source: https://instances.social/.

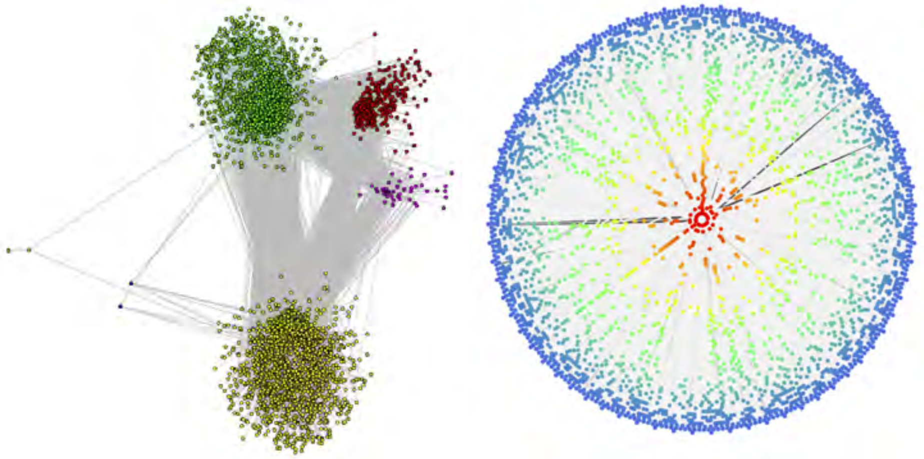


Fig. 1. *(On the left)* Community structure of the *instances* network detected by the directed Louvain method. *(On the right)* Core decomposition of the *instances* network, based on vertex in-degrees: vertices having the same core-index are assigned the same color (inner-most, resp. outer-most core correspond to red, resp. blue); edge width is drawn proportionally to the edge weight

we derived the *instances* network as a directed, weighted network with set of nodes as instances, set of edges (95 221) each modelling a follow relation from a user in an instance to a user in another instance, and edge weights each storing the count of users in a source instance following users in a target instance.

Basic structural statistics on this network, such as the high values of clustering coefficient (0.848) and reciprocal edges (70.9%), already provide hints at the presence of a federation structure. Nonetheless, our goal is to gain insights into the mesoscopic structure of the *instances* network to unveil whether and to what extent the federated mechanism actually holds in Mastodon. To this aim, we leveraged on two main exploratory tools, namely community detection and core decomposition.

We used the well-known modularity-maximization-based Louvain algorithm for discovering communities in the *instances* network, also employing its directed variant.[2] Figure 1 shows a graphical illustration of the 6 communities found, where two macro-communities stand out (colored in green and yellow in the figure), containing about 89% of the instances. This also couples with a modularity of 0.397, which hints at dense connections existing between instances within the same community, though the top-3 Mastodon instances by relevance separately lay on these macro-communities. Like previous studies have shown, communities tend to be geographically distinct, as a result of different cultures and languages of their users. However, a deeper analysis of the instances in the communities allowed us to draw more interesting remarks.

The decentralized mechanism enables the creation of instances based on specific topics. While this allows users to select which instance best fits her/his interest, a "sec-

[2] https://github.com/nicolasdugue/DirectedLouvain.

torization" bias of the instances network could occur. However, we found out that the community structure of the *instances* network is *not* strictly topically-induced: in fact, in every community, the description of most instances share Not-specified or Generic as main "topics" (followed by actual topics, e.g., Technology, Gaming). We believe such looseness of topics should also positively be interpreted together with the adoption by most instances of the content-prohibition feature concerning a variety of undesired contents (e.g., spam, advertising), which works out to ensure reciprocal respect between users of the same or different instances. As a side yet relevant remark, the *instances* network also exhibits a degree assortativity significantly negative (-0.291): this is another distinctive trait of Mastodon w.r.t. most of the centralized social networks, which might be ascribed to the opportunity given to the users of following and interacting with users in different instances, regardless of the popularity of their home-instances.

Our analysis of core decomposition the *instances* network revealed more peculiarities of this network which further strengthens our hypothesis of federation. First, the *instances* network has a degeneracy of 141 (resp., 69 and 70) w.r.t. full-degree (resp., in-degree and out-degree). One major finding is a significant presence of instances not only in the periphery but also in the inner cores of the network. In particular, the full degree (resp., in-degree and out-degree) based inner-most core contains about 120 (resp., 123 and 115) instances. Also, besides the expected high density (0.76-0.79) and clustering coefficient (0.87-0.88), we observed an extremely high reciprocity (above 90%).

Further interesting findings are drawn from the observation of Fig. 1, whose display of the cores unveils evidence of solid radial lines, which means there is a significant amount of connections between users of instances in the inner-most core and users of peripheral instances. This relates to the negative degree assortativity we previously discussed, and it represents quite a novel pattern in social networks, which do not actually show direct links of any type between core-users and periphery users.

Summary. Our study based on mesoscopic structural analysis has confirmed a clearly federative status characterizing Mastodon, which is today the most important platform in the Fediverse. Nonetheless, our study also paves the way for future research on the understanding of relevant features that might be shared among the various platforms of the Fediverse of DOSNs, unveiling new social patterns expressed by DOSN users.

References

1. Cerisara, C., Jafaritazehjani, S., Oluokun, A., Le, H.T.: Multi-task dialog act and sentiment recognition on Mastodon. In: Proc. COLING Conf. (2018) 745–754
2. Trienes, J., Cano, A.T., Hiemstra, D.: Recommending users: Whom to follow on federated social networks. CoRR **abs/1811.09292** (2018)
3. Zignani, M., Gaito, S., Rossi, G.P.: Follow the "mastodon": Structure and evolution of a decentralized online social network. In: Proc. ICWSM Conf. (2018) 541–551
4. Raman, A., Joglekar, S., Cristofaro, E.D., Sastry, N., Tyson, G.: Challenges in the Decentralised Web: The Mastodon Case. In: Proc. ACM IMC Conf. (2019) 217–229
5. Zignani, M., Quadri, C., Gaito, S., Cherifi, H., Rossi, G.P.: The Footprints of a "Mastodon": How a Decentralized Architecture Influences Online Social Relationships. In: Proc. IEEE INFOCOM Workshops. (2019) 472–477
6. Zulli, D., Liu, M., Gehl, R.: Rethinking the Social in Social Media: Insights into Topology, Abstraction, and Scale on the Mastodon Social Network. New Media & Society **22**(7) (2020) 1188–1205

Digital Sousveillance: A Network Analysis of US Surveillance Organizations

Colin Burke

University of California, San Diego, La Jolla CA 92093, USA,
cmburke@ucsd.edu

1 Introduction

While it is a given that private entities play a crucial role in allowing the government to engage in mass surveillance, whether through "backdoors" [1] or "revolving doors" [2], we know far less about how this vast assemblage of public and private organizations actually operates. Because of the tight-lipped nature of the US government, as well as the private corporations involved in surveillance activities, the study of what some have termed the "surveillance-industrial complex" has struggled to "unmask" the actors involved.

This study introduces a new methodological approach to the study of surveillance that I call *digital sousveillance* – the co-optation of digital data and the use of computational methods and techniques to resituate technologies of control and surveillance on individuals to instead observe the organizational observer. To illustrate the potential of this methodological approach, I employ quantitative network analytic methods to trace the changes and development of the vast network of public and private organizations involved in surveillance operations in the United States – what I term the *US surveillant assemblage* – from the 1970s to the 2000s. The results of the network analyses suggest that the US surveillant assemblage is becoming increasingly privatized and that the line between "public" and "private" is becoming blurred as private organizations are, at an increasing rate, partnering with the US government to engage in mass surveillance.

2 Data and Methods

As noted above, this study uses network analytic techniques to map the historical development of the US surveillant assemblage, and its public-private linkages, across four decades spanning from the 1970 to the 2000s. The primary sources of network data for this article are drawn from the Transparency Toolkit's "ICWatch" database. The data employed here are scraped from individual profiles from LinkedIn.com. The scraper collected public profile information based on a list of search terms that consisted of the names of surveillance programs identified in the Snowden documents. If an individual's profile contained terms or names from one of these programs, the scraper collected the entire profile, including job history (job title, company, start/end date, and description for each job), skills, educational history, and location/area. These historical job data were used to create organizational networks, where individuals' employment histories were used to identify organizations involved in surveillance programs from the 1970s to the 2000s.

3 Results

As noted above, the data are analyzed in four distinct sections representing each decade from the 1970s through the 2000s. This study begins its analysis with a mapping of each decade. These visualizations are helpful for providing a glimpse into the scope of these networks and how the relations between public and private surveillance organizations have evolved over time. These visualizations alone are insufficient to draw robust empirical conclusions, however. To compensate for this, network statistics and measures are also provided to allow for a more clear, empirical means of comparison between each network. I thus rely upon statistical network analyses, including measures of network structure as well as measures of homophily for each decade, to support and expand upon the insights gained from the graphical network representations.

The most striking feature of the network graph in Figure 1, below, is the overwhelming presence of private organizations as the dominant nodes in terms of degree within the network. There is strong clustering of public-private ties, most noticeable on the private side of the graph, with two major groups of public-private ties clustered together. There also seem to be, proportionately speaking, fewer public-public and private-private ties than there were in previous decades. This would seem to indicate that connections between public and private organizations were more frequent during this period. In the following section, I rely upon statistical network analytic methods to support these graphical illustrations and to draw more direct empirical conclusions about the structure of the US surveillant assemblage from the 1970s to the 2000s.

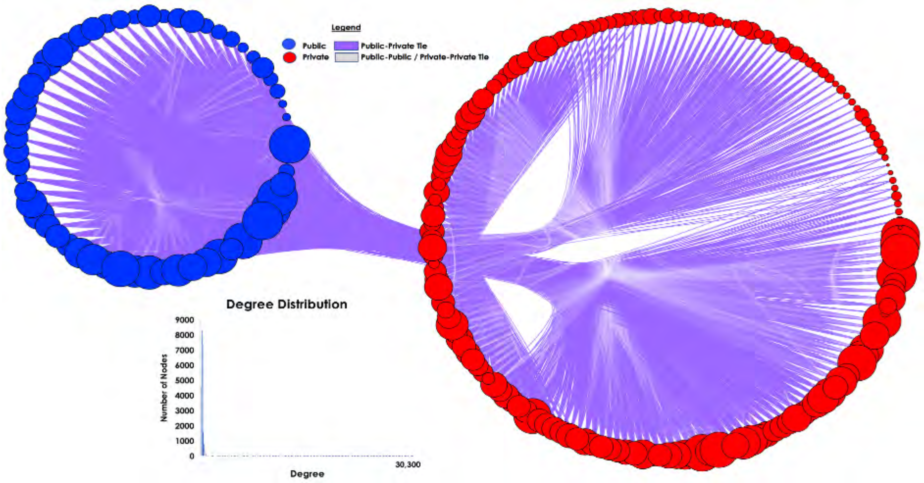


Fig. 1. Network Graph, 2000-2009

Table 1, below, displays the results of the quantitative network analyses with regard to network connectivity and homophily for the 1970s, 1980s, 1990s, and 2000s networks. The results suggests that the US surveillant assemblage became more centralized and highly connected over time. Measure such as the number of connected

Table 1. Measures of Network Connectivity and Homophily.

Measure	1970-1979	1980-1989	1990-1999	2000-2009
Nodes	743	2443	6961	31,015
Edges	10,189	95,939	708,517	1,048,575
Connected Components	13	9	14	1
Clustering Coefficient	0.838	0.856	0.834	0.966
Network Centralization	0.783	0.911	0.919	0.976
Attribute Assortativity	-0.001	0.081	0.090	-0.064
Degree Assortativity	-0.171	-0.144	-0.112	-0.590
Public-Public Ties (%)	22%	21%	22%	17%
Private-Private Ties (%)	47%	58%	57%	47%
Public-Private Ties (%)	46%	38%	38%	48%
Odds-Ratio: Public-Private Tie	0.72	0.38	0.36	0.85

components, clustering coefficient, and network centralization suggest that there is a considerable shift in the structure of the US surveillant assemblage over time. Also, the measures of homophily and heterophily displayed in Table 1 seem to indicate that there were a greater tendency and likelihood of public-private connections in the 1970s and 2000s networks compared to the 1980s and 1990s. This supports the graphical evidence that suggested the 2000s network is quantitatively different from prior decades. The 2000s network is highly centralized, well-connected, and heterophilic – that is, the 2000s network had a greater tendency for public-private partnerships than previous years, as well as a greater tendency for connections between nodes of differing degree values.

4 Discussion

The results of the network analysis indicate that the US surveillant assemblage is becoming increasingly privatized; indeed, the line between "public" and "private" is becoming blurred as private organizations are, at an increasing rate, partnering with the US government to engage in mass surveillance. More work, by activists and scholars alike, is needed to continue to unmask these actors and to allow the public to better understand their entanglement with this vast surveillant assemblage.

References

1. Crampton, J.W., Roberts, S.M., Poorthuis, A.: The New Political Economy of Geographical Intelligence. Annals of the Association of American Geographers 104 (1), 196–214. (2014)
2. Hayes, B.: The Surveillance-Industrial Complex. In: Kirstie Ball, Kevin D. Haggerty, and David Lyon (eds.) Routledge Handbook of Surveillance Studies. pp. 167–175. Routledge (2012).
3. Mann, S., Nolan, J., Wellman, B.: Sousveillance: Inventing and Using Wearable Computing Devices for Data Collection in Surveillance Environments. Surveillance and Society. 1, 331–355 (2003).

Quarantined world through SoundCloud hashtags network

Vitalba Macaluso[1], Clara D'Apoli[1], and Giulio Rossetti[2]

[1] University of Pisa, Italy,
[2] ISTI-CNR, Pisa, Italy,
v.macaluso@studenti.unipi.it
c.dapoli3@studenti.unipi.it
giulio.rossetti@isti.cnr.it

1 Introduction

In March and April 2020 all the world was involved in a pandemic lockdown due to the Covid-19 emergency. During that period, people used social networks not only as entertainment channels, but above all as a place for expressing moods, collecting news and sharing passions through 'hashtags'. The research is aimed to verify, by examining the hashtags network of the online music platform 'SoundCloud' [1], whether the 'hashtag' is not only an index of the track, but a real form of communication between users. To achieve this target, DEMON [1] was used as a Community Discovery (e.g., CD) algorithm; the choice is based on the sociological hypothesis that a node can belong to different communities, accurately reproducing the multiple scenario in which the quarantined world was divided and united at the same time and hashtags can have the explanatory power of covering a number of semantic areas. The network on which the analysis was carried out consists of 29862 nodes and 406651 edges, where the latter are given by the co-presence of two hashtags within a track. The CD was started on it and at first parameters ε, the merging threshold in [0,1], and κ, the minimum community size, have been set: the value $\kappa = 10$ was set in order to have a minimum size of the communities not too high ensuring a proper level of coverage, the value $\varepsilon = 0.4$ has been set for several reasons: first, to guarantee an accurated semantic study of them and then to bring out more than one aspect of the same phenomenon.Therefore, DEMON found 1422 communities and we selected and analyzed those containing the hashtag 'quarantine' to verify their ability to reflect the multiple realities, in terms of interests, passions, moods, that the quarantine was able to put together.

2 Results and Discussion

The algorithm found 152 communities containing the hashtag 'quarantine', among which thirteen categories have been discovered from a more in-depth study of community partition. All the categories found are shown in Fig. 1 but we are analyzing the most interesting ones.

[1] An online audio distribution platform and music sharing that enables its users to upload, promote, and share audio, https://soundcloud.com/

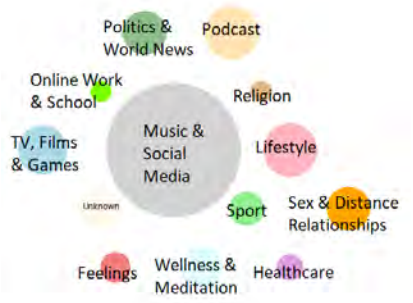


Fig. 1. Categories of communities

Music & Social Media. As expected, this is the most frequent category, with 72 communities. Music has always been said to break down many barriers and this is exactly what happened during the lockdown period. The CD highlighted the hashtags 'Travisscott' and 'Fortnite', the first a famous American rapper who during the quarantine shared with his fans a concert on the gaming platform 'Fortnite', recording a world record and becoming one of the possible future scenario for music. Among hashtags concerning musical genres and artists, the hashtags 'GetMorePlays' and 'SCxiamOTHER' stand out. SoundCloud launched these two promotional activity for users, linked to these hashtags, in order to give more visibility to emerging and unknown artists. The company, together with the well-known singer Pharrell Williams, decided to publish a compilation in aid of charity, through the hashtag 'SCxiamOTHER', the entire proceeds of which have been donated to some musician organizations.

Lifestyle. Another interesting aspect that emerged from CD concerns everyday life. The lockdown period has forced people to change their habits and lifestyles. Many of them took the opportunity to rediscover the pleasures of family life, and this has been shared on digital platforms. Hashtags in related communities include words such as 'daughter', 'mum', 'stay home', 'family', 'netflix' 'workingfromhome', 'coloring', 'cooking', 'drawing', 'home and family', 'sleep', 'keeping kids engaged', 'nature walks', 'student life', 'unemployment', and 'thingstodo' that effectively sums up their purpose. Obviously, mixed feelings also come out: 'boredathome', 'isolation', 'funny', 'positivity' etc. In an emergency like COVID-19 pandemic, the fear of the new and unexpected situation and its potential impact on health, combined with the need for social isolation, causes an inevitable feeling of desolation with time running slowly, but can help people regain control of their lives, increasing the capacity to respond in a positive way at distress caused and to 'fill the gaps' left by the hectic life to which they were accustomed.

Podcast. During the Lockdown, the world of podcasting recorded an increase in ratings of 53% across Europe [2] . The topics most listened to were mainly family, politics and health and hashtags like 'family', 'epidemic' and 'conversation' were the most shared in this category of communities. On the other hand, there has been an opposite trend in the USA, where podcasting is mainly linked to commuting, stopped due to the pandemic. However, one argument that affected many communties was the discussion about Microsoft's founder, Bill Gates, with hashtags like 'BillGatesDidIt', '5G', 'conspiracy' and 'QAnon'.

Sex & Distance Relationship. The category 'Sex & Distance Relationships' highlights new and safe ways that people found to reconnect intimately and sexually using digital technologies, due to forced isolation. In fact, hashtags like 'tinder', 'online dating', 'facetime', 'Facebook', 'zoom', 'bumble', reflect exactly this new virtual reality. Also 'online dating', 'virtualdating', 'sex uninterrupted', 'socialdistancing', 'love lost', 'safe sex', 'acceptance', 'physicalhealth', put in evidence that in times of technological advancement, the lockdown has not prevented from forming romantic and platonic connections with other people, albeit virtually.

Feelings. Several are the communities identified that enclose a set of emotions and moods, even very contrasting with each other, which faithfully reflect the tumult of feelings from which the world community was invaded suddenly. The hashtags themselves, such as 'emotions', 'anxiety', 'happiness', 'frustrated', 'courage', 'sad' and 'fun' emphasize the nature. Not surprisingly feelings such as 'hope', the externalization of well-being, 'stress' and 'vulnerability', but also more challenging issues such as 'suffering', 'support' and 'depression' were strongly felt and shared by users, through the numerous presence of the respective hashtags.

Online work & School. The closure of schools and offices due to the pandemic has revolutionized the way we work and teach all over the world. Some companies already knew smart working and distance learning, others were unprepared for this kind of work and had to adapt quickly: it has a strong presence of hashtags such as 'remoteworking', 'working remote', 'onlinelearning', 'homeschooling' and 'distancelearning' and slogan as 'DoBusinessBetter'.

The research is only the first step of a larger study of the role of hashtags within musical tracks. However, as noted at least in part, DEMON's communities partition, and the categories identified starting from this, exactly reflects the quarantined people's lives and feelings all over the world. Therefore, given the use that users make of hashtags on social platforms, as a further means of moods expression and ways of living, the analysis could be considered an alternative way to investigate society, increasing the target sample and deepening the identified thematic categories.

Acknowledgements

This work is supported by the scheme 'INFRAIA-01-2018-2019: Research and Innovation action', Grant Agreement n. 871042 'SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics'.

References

1. Coscia, M., Rossetti, G., Giannotti, F., & Pedreschi, D. (2012, August). Demon: a local-first discovery method for overlapping communities. In ACM SIGKDD.
2. Forbes. (2020) How Is The Podcasting Business Under Coronavirus Quarantine? Listen Up. Retrieved from forbes.com/sites/howardhomonoff/2020/06/15/how-is-the-podcasting-business-under-coronavirus-quarantine-listen-up/#222e399f3ff8.

Inferring political opinion and its relationship with use of language in a Twitter conversation around a territorial conflict

Julia Atienza-Barthelemy[1][2], Samuel Martin-Gutierrez[1], Juan C. Losada[1], and Rosa M. Benito[1]

[1] Grupo de Sistemas Complejos, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Universidad Politécnica de Madrid, Av. Puerta de Hierro, 2, 28040, Madrid, Spain
[2] julia.martinezatienza@upm.es

1 Introduction

In many political scenarios, public opinion is often divided into two extreme and opposite positions. In sociological terms, this process is called polarization, and has important social consequences. Political polarization generates strong effects on society, driving controversial debates and influencing the institutions. Territorial disputes are one of the most important polarized scenarios and have been consistently related to the use of different languages [4]. In this work [1], the polarized system studied is centered around the Catalan independence issue and has been analyzed through a Twitter dataset. The dataset is made up of tweets about the topic published in the period between 09/15/2017 and 03/11/2017. During this period important events occurred, the most relevant being the celebration of a referendum on independence not approved by the Spanish Government. Some of these events were clearly reflected as activity peaks in the time series of the conversation.

2 Results

In order to infer the users' opinion, we analyze their interactions. Among them, we choose the retweet interaction because it is a broadcasting mechanism that usually implies that the user endorses the original tweet. We build the retweet networks and adopt a model[2] based on the De Groot process to infer the opinion of a user in a network as the average of her neighbor's opinions. This model requires to define two sets of users: users with a fixed opinion to be used as opinion seeds, called elite, and users whose initial opinion is neutral and it is to be inferred by their connections in the social network they are embedded, called listeners. The first step is to find these elite users. To do that, we first search for highly infuential and engaged users, i.e., users with a high number of retweets that participate over 95% of the days, and we study the community structure of these users with a hierachical stochastic block model [3]. The elite users are made up by the two communities that include the largest number of users where all politicians contained are from political parties completely in favor of the Catalan independence, or completely against it, respectively. The elite users of the pro-independence community have an opinion index of 1 and those from the against indepedence community have an opinion index of -1. The rest of users, called listeners, iterativety update their opinions as the average of their neighbors' opinions. This process quickly converges to an opinion index between -1 and 1.

As we estimate the users opinions based on their retweets, in order to minimize noise in the opinion distribution, we only consider users that have posted at least 10 times so they provide a less noisy opinion distribution. This distribution is shown in Fig 1 and it presents a mainly bimodal behavior with an intermediate third pole that shows a less polarized society due to the presence of not only antagonist opinions. In this case, the third pole can be explained by the existence of political parties that defend a middle ground. Complementarily, we have studied users that tweet a high number of tweets (active users) users that participate a high proportion of days (engaged users) and users that received a large number of retweets (influential users). These three types of users are practically the same subset and hold more extreme positions.

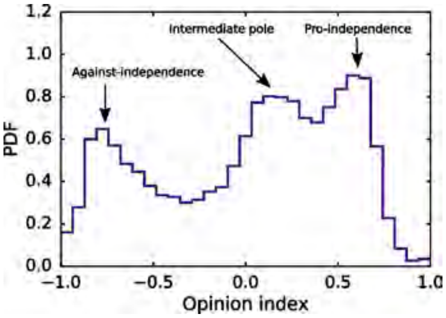


Fig. 1. Opinion distribution of the users with an activity corresponding to more than 10 tweets in the Catalan independence conversation.

By looking at the temporal behavior, we have seen that, whenever the number of users that participate in the conversation is low, the daily opinion distributions are mostly bimodal with most of the users concentrated on the two extremes. However, when the number of users increases, the opinion distribution becomes more diverse.

There are two co-official languages in Catalonia that are involved in the conversation: Catalan and Spanish. It has been proven that, as in many other social categories, the use of language in the context of a territorial conflict is related to certain political positions. For this reason, we have studied the interplay between the inferred political opinion and the language used. This use of language is quantified by the language index that measures the Spanish-Catalan usage ratio of each user. We have shown that there is a clear relationship between the political opinion and the use of language (see Fig 2): users with an opinion index ranging from completely against-independence to slightly pro-independence speak almost exclusively Spanish. On the other hand, for users with an opinion index closer to the pro-independence pole, the range of values of the language index is wider, i.e., they speak Catalan and Spanish almost indistinctly. Finally, we have applied three different activity thresholds to the analysis of the opinion-language relationship. Although the global opinion- relationship pattern remains similar for the three thresholds, there are three regions that stand out, corresponding to the three political poles. Users of the most against-independence pole and users of the pole with the most central opinion index speak almost exclusively Spanish. Conversely, users of the pole with an opinion index nearer to the pro-independence extreme present a wider range of language use between Catalan and Spanish.

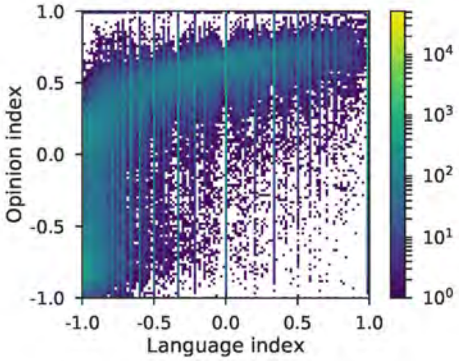


Fig. 2. Interplay between language and opinion indices of the users of the Catalan independence Twitter conversation. Language index = -1 when the user speaks only in Spanish and 1 when she speaks only in Catalan. Opinion index=1 when the user is completely pro-independence and -1 when she is completely against it. Color measures the number of users located in the 2D bin.

Summarizing, our results show that the proposed methodology is able to reveal the complex patterns of the ideological landscape in a polarized context. It is worth emphasizing the detection of a third pole that naturally emerged from the application of the opinion inference model, although the elite considered was only bipolar, as well as the richness of the information extracted from the study of the interplay between language and ideology.

References

1. Atienza-Barthelemy, J., Martin-Gutierrez, S., Losada, J.C., Benito, R.M.: Relationship between ideology and language in the catalan independence context. Scientific reports 9(1), 1–13 (2019)
2. Morales, A.J., Borondo, J., Losada, J.C., Benito, R.M.: Measuring political polarization: Twitter shows the two sides of venezuela. Chaos: An Interdisciplinary Journal of Nonlinear Science 25(3), 033114 (2015)
3. Peixoto, T.P.: Hierarchical block structures and high-resolution model selection in large networks. Physical Review X 4(1), 011047 (2014)
4. Shabad, G., Gunther, R.: Language, nationalism, and political conflict in spain. Comparative Politics 14(4), 443–477 (1982)

Mapping the Global Scientific Landscape of Virology Before the COVID-19 Pandemic: A Large-Scale Document Analysis with the Representation Learning and Network Visual Representation

Feifan Liu[1], Shuang Zhang[1] Shuangling Luo[2], and Haoxiang Xia[1]

[1] Institute of systems engineering, Dalian University of Technology, Dalian 116024, China
hxxia@dlut.edu.cn,
WWW home page: http://faculty.dlut.edu.cn/hxxia/en/index.htm
[2] School of Maritime Economics and Management, Dalian Maritime University, Dalian 116026, China

1 Introduction

The COVID-19 pandemic prompted all sectors of society to recognize the importance of comprehensively establishing medical science and technology innovation systems. However, the empirical analysis built on scientific literature data focusing on virology is still insufficient, given the severity of the COVID-19 pandemic. Understanding virology provides important basic information on science and technology to establish virological discipline and public health systems. On the other hand, limited by the traditional methods of topic identification, the current studies generally lack a worldwide comparison on topic identification, domain development patterns, and specialties built on a relatively complete corpus of papers in the virology field. In this study, we use a well-established academic dataset, MAG, and the document representation learning methods, Doc2Vec and UMAP, as literature analysis tools to detect Virology's global research topics from 1989 to 2019.

2 Results

We finally identified 27 core research subfields, and compared various countries' key research directions and development trends in virology and its subfield, ”Coronavirus”. Figure 1 shows the accumulated topical landscape of various virology subfields based on document representation learning. Figure 1a presents the transformation of literature information in virology into a semantic knowledge map of this field. Figure 1b is a large-scale citation graph in virology obtained by the edge bundling technique. Depending on the ability of manifold learning algorithm UMAP in preserving the global semantic information of the text and that of the edge bundling method in the optimization of the topological structure of the large-scale network, the citation network among 140,000 papers on different subfields explicitly manifests the research sub-branch within virology. Figure 2a shows the topical density of virological papers in six countries, including the top five most-cited countries (the United States, the United Kingdom, France,

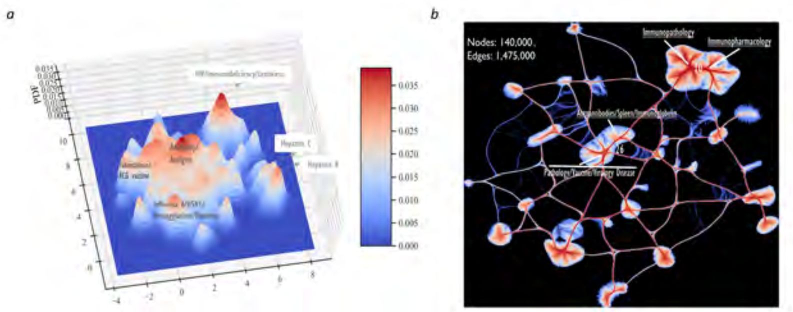


Fig. 1. Topical landscape in virology (a) The 3D topical landscape of virology based on the Doc2Vec,UMAP and kernel density estimation (KDE); (b) A citation network of virology grouped by the paper citation relationships among 27 identified subfields based on edge-bundling techniques (the width and color depth of edges reflect the citation strength between the subfields)

Germany, and Japan) and China, which shows the global virological research status. It can be seen from the overall landscape of virology in Figure 1 and Figure 2 that the United States leads this field with the widest range of research subfields and the deepest scientific inquiry in all 27 subfield areas. The second-tier countries, Britain, France, Germany, and Japan experience a wide range of research, but they each focus differently on specific research directions. Besides, China's current research focuses only on a few subfields such as ”Swine fever,” ”H5N1 influenza,” ”Coronavirus,” and ”Enterovirus/Poliovirus.” Figure 2b shows the distribution of 1,553 research institutions which publish virological papers, as well as the citation network among papers published by each institution. In this citation graph, the node represents each research institution, and the number of cited research papers of that institution determines the strength of the edge. The institutions ranked in the top 100 most citation counts are marked with circles. As Figure 2b shows, in virology, the world's influential scientific research institutions are mainly located in North America and Europe, especially along the east coast of the United States. In Asia, Japan occupies four of the top 100 research institutions.

In summary, this overview investigates the science community's global awareness and preparedness before this pandemic outbreak. Through examining the content of explosive growth, we identified 27 core research subfields, their mutual knowledge connections in this field, compared the breadth and depth of each country's scientific exploration in this field, and presented the distribution of the global influential scientific research institutions. The worldwide significantly unbalanced development in virology is revealed by comparing the global virological subfields, publications and impact. Comprehensive measures such as strengthening international cooperation and collaboration, adjusting the discipline layout, and increasing resource investment are still very urgent since the impact risk of new viruses in the future still exists under the current circumstances.

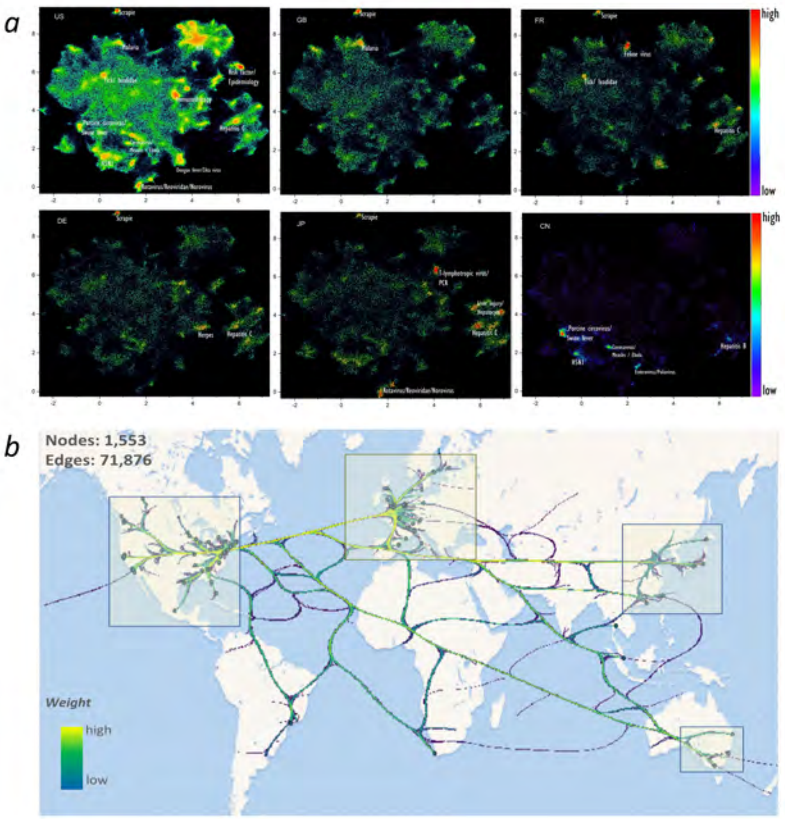


Fig. 2. The worldwide development status in the virology discipline (a) Topical landscape of the top five most cited countries (US, GB, FR, DE, JP) and China in virology based on the identified subfields in fig 1(a)(the main subfields are labeled for each country); (b) A citation graph of papers constructed from citation relationships of 1,553 institutions in virology using the edge bundling technique; circled 100 research institutions with the most cited papers and squared the worldwide main research regions

Immigration in the Italian Public Debate: Dynamics of Interactions in a Segregated Network

Salvatore Vilella[1], Mirko Lai[1] Daniela Paolotti[2], and Giancarlo Ruffo[1]

[1] Computer Science Department, University of Turin, Turin, Italy,
salvatore.vilella@unito.it,
[2] ISI Foundation, Turin, Italy

1 Introduction

Migration has been one of the most debated topics in Italy during the last decade [1], hogging the attention of politics and citizens,thanks to the ever extensive media coverage, to the social-economic challenges of migrants' integration in a complex society, and to the rising of right-wing nationalist political parties. In general, immigration is a very polarising topic: the Italian public debate is sharply divided between those in favour and those against, often transcending the actual topic and escalating into ideological quarrels. We decided to study the interactions between social media users posting about immigration during the period between August 2018 and July 2019, when it was one of the main focus of the Government - and, therefore, of news media. In particular, we want to investigate the structural aspects of the interactions, focusing on some key research questions:

- is this really a strongly divisive topic? Who are the actors driving the debate, and are we able to identify factions with different stances? Is the size of such clusters somehow proportional to the popularity of their leaders?
- *What is the role of clusters in the diffusion of opinions and news*? Do locally popular news and contents have the chance to reach also distant clusters? Do clusters act as barriers or accelerators for content diffusion?

2 Methods and Results

To answer those questions we rely upon Twitter data, the renowned micro-blogging platform, widely used by politicians and news media as well as by common citizens. We collected almost 6 millions tweets posted between 2018 and 2019, and we build and study the *retweet* network, since retweets constitute the vast majority of the collected data. This allows us to identify communities whose stance should be internally coherent, since retweets (and quotes) are often an expression of endorsement [2–4]. Community detection is carried out using the Louvain method: as expected, we find a small number of macro-communities (Tab. 1), whose stance can be easily guessed by inspecting the top 10 nodes for **in-degree**, all famous politicians and journalists. We enrich this qualitative information by combining it with an analysis of the most used bi-grams in each community (Tab.1) to corroborate our results: we find communities (RT2-3-4) **aligned**

ID	Size	Internal link density	No. of #migranti	Highest in-degree nodes (usernames)	Top Bi-grams (EN)
RT1	116,831	$1.5 \cdot 10^{-3}$	51,639	Gad Lerner, Roberto Saviano, La Repubblica, Linkiesta, Udo Gümpel, Fabio Niccolò Zancan, jacopo iacoboni, Nello Scavo, laura boldrini.	Safe Harbor, Closed Harbors, Human Beings, Human Rights
RT2	34,174	$1.93 \cdot 10^{-2}$	62,980	Giorgia Meloni, Cesare Sacchetti, Francesca Totolo, Diego Fusaro, La Verità, ImolaOggi, Giank-deR, Claudio Perconte, Il Sofista, Antonio M. Rinaldi .	Abetment Immigration, Uncontrolled Immigration, Economic Migrants, Illegal Migrants, Human Beings
RT3	27,845	$2.4 \cdot 10^{-3}$	4,575	Matteo Salvini, Lega - M. Salvini Premier, Noi con Salvini, TG2, Attilio Fontana, Generazione Identitaria, Marco Morini, Cittadina Italiana, Don Alphonso, Matteo SALVINI	Abetment Immigration, Uncontrolled Immigration, Economic Migrants, Illegal Migrants
RT4	9,553	$3.5 \cdot 10^{-3}$	8,887	Il Fatto Quotidiano, Danilo Toninelli, Peter Gomez, Carlo Sibilia, Movimento 5 Stelle, Franco Bechis, Andrea Franchini, Le Frasi di Osho, Elio Lannutti.	Abetment Immigration, Uncontrolled Immigration, Economic Migrants, Illegal Migrants, *Fatto Quotidiano* (Italian newspaper)
RT5	9,225	$2.4 \cdot 10^{-2}$	6,193	SkyTg24, ANSA, Tgcom24, RaiNews, Agorà Estate, Agi Agenzia Italia, Adkronos, Dagospia, Ultime Notizie, Il Messaggero.	*Merde Alors* (French), Thousands Italians, Colorful *Merde*, Italians Came, Ends Expression, Patience Snaps, Luxembourg Dear

Table 1: Table describing the communities of the RT network and the top bi-grams in each.

on a **negative stance** towards migrations, while RT1 is the only cluster with a positive stance. RT5 can be traced back to news media.

This alignment in the stance is evident also by inspecting the weighted adjacency matrix of the community-induced directed RT network (Fig. 1). RT2-3-4 display a much higher number of connections between themselves rather than towards other communities with different stance: there is a clear segregation in the network, due to the relatively small communication between factions with different stance.

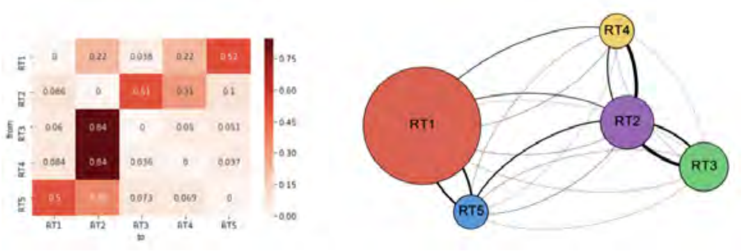


Fig. 1: The adjacency matrix of the RT community-induced network (normalised by row) and its graphical representation. The direction in the graph is given clockwise.

The diffusion of content is not unaffected by segregated behaviour. To measure it, we select all the URLs with at least 100 shares and we look at the distribution of their shares among the communities. To each URL we can associate a sequence (a vector) of communities sharing it over time. Inspired by [5], in order to quantify the diversity in communities sharing an URL, i.e. how heterogeneous is the reach of an URL, we can compute the entropy $\mathrm{H}(\mathrm{URL}_i)$ of the associated vector simply as

$$\mathrm{H}(\mathrm{URL}_i) = -\sum_{c \in C} \mathrm{u}_c \ln(\mathrm{u}_c)$$

where u_c is the fraction of shares by community c over the total. Now that every URL is assigned its own entropy, we can finally characterise the communities by the entropy distribution of all the URLs they shared (Fig.2). We observe how two areas that are at the odds with respect to their stance towards the topic of immigration (RT1 and RT2) share a similar behaviour: their stance and their cultural background are different, yet they both seem to be tightly folded in on themselves, sharing URLs that can hardly ever be found in other communities and giving rise to some sort of **echo-chamber** effect that slows down and limits the diffusion of content.

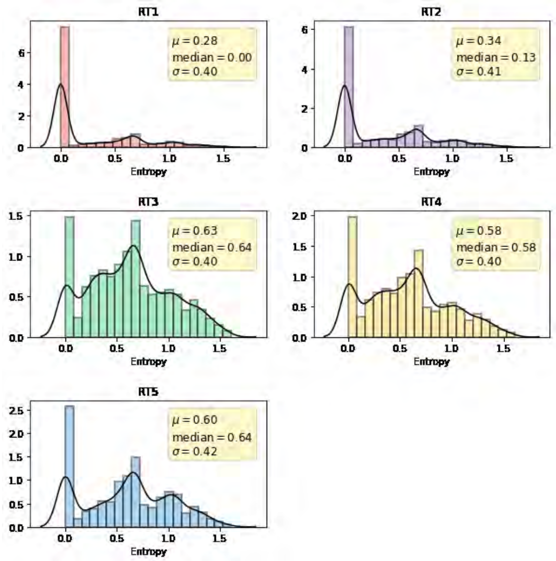


Fig. 2: Distribution of the URL entropy for each community.

References

1. Fasano, L. M., and N. Pasini: Tra frammentazione e polarizzazione del sistema politico italiano: interpretazioni e casi empirici. (2014): 109-142.
2. Conover, Michael D., et al. ”Political polarization on twitter.” Icwsm 133.26 (2011): 89-96.
3. Feller, Albert, et al. ”Divided they tweet: The network structure of political microbloggers and discussion topics.” Fifth International AAAI Conference on Weblogs and Social Media. 2011.
4. Lai, Mirko, et al. ”Stance polarity in political debates: A diachronic perspective of network homophily and conversations on Twitter.” Data & Knowledge Engineering 124 (2019): 101738.
5. Weng, Lilian, Filippo Menczer, and Yong-Yeol Ahn. ”Virality prediction and community structure in social networks.” Scientific reports 3 (2013): 2522.

Avres: visualising multilayer networks for analysing human-trafficking networks

Norbert Féron[1*], Jason Vallet[1], Guy Melançon[1], Bénédicte Lavaud-Legendre[1],
Cécile Plessard[1], Benjamin Renoust[2], Alexander Freeland[3]

[1]University of Bordeaux, [2]Osaka University, [3]FSP
*corresponding author: *norbert6feron@gmail.com*

1 Introduction

While complex networks have shown great applications for the analysis of human phenomena, recent trends now bring a new level of complexity to handle even more complex systems through the use of multilayer networks [8]. Visualizing and interacting with networks, especially when complex, tremendously help to understand their structure, and to identify particular entities or links [11].

Social network modelling is a powerful tool to investigate criminal networks [6][3][4], in particular human trafficking, such as in [2] for which affiliation networks are used to model individual–event participation, and derive co-participation one-mode projections. In [10], the roles of madam in a telephone-based sex trafficking social network is investigated. In [12], the complex balance between control and cooperation is studied from the position of prostitutes in the construction of their relationships with pimps and other prostitutes. More recently [7], the combination of wiretap ego-networks has been supporting the investigation a large human smuggling network.

While the roles of certain actors in these networks is clearly identified (the sex worker, the client, and the procurer), extended social networks also show us third parties at the edge of the networks. These persons are rarely directly involved with the sex workers and remain in a grey area as they profit from these activities in a manner that is not punishable by the law. Identifying those actors is key to dismantle a human trafficking network.

The data accumulated to find these individuals is often sparse, incremental, and extremely varied. While Borgatti [1] 4 types of ties (social relations, interactions, flows, and homophily) notes captured to document these networks often belong to the first three categories. The challenge then becomes to propose a holistic approach to, from such varied and sparse information, accompany experts to build and interrogate the network such that they can identify specific individuals from their role and similarities. Multilayer networks are natural candidates to capture such varied interactions.

2 Model and visualization

This work is made in close collaboration with jurist and law enforcement experts. Their first step is to obtain from the Prosecutor Office access to material from closed legal

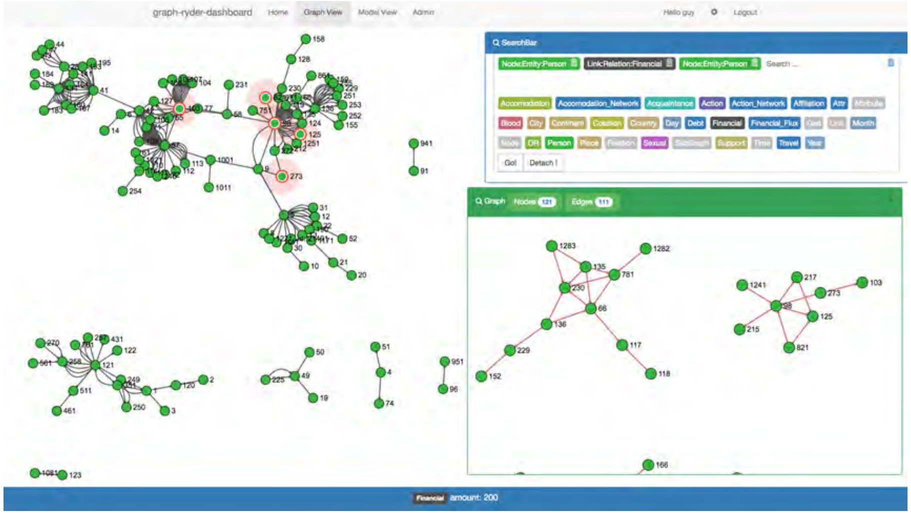


Fig. 1. User interface of Avres

cases (*e.g.* transcripts of the court proceedings, witness interviews, *etc.*). These are gathered by hand, and modelling the problem as a network requires a shift of perception. As a consequence, several rounds of model refinements have been necessary to converge toward a stable model. The nature of nodes, and of the edges, families of nodes and edges, forming the multilayer network are in constant evolution, requiring a flexible way to model the issue.

We have identified 3 different node type: *persons* representing physical actor of the network, *geolocations* which can have different levels of precision (i.e. a street name, a region or a country), and *time*, which also can be either a specific day or a period. We have also identified 6 most important types of interactions between people: *accommodation*, *acquaintance*, *action*, *blood*, *financial*, *sexual*, and *support* forming our 6 different layers.

We denote the **meta-model** as the graph $\mathbb{G} = (V, E)$ where V is the set of meta nodes that represent different node types from our network (i.e. Person, Cotation, Geolocation), and a meta edge $e = (v_1, v_2) \in V^2$, $e \in E$ where e exists if at least one edge between a node of type v_1 and a node of type v_2 exist in the network. This meta-model has many roles: it helps the experts to keep uniformity during the input process, each new element or relation being required to fit to the current meta-model. It also provides basic typing to help customise the interface (i.e. building inline queries and store user visual preferences including element colour and size). This meta-model can be changed interactively by the user to add any new type of entities or relations.

To explore such networks, our tool sits on a network node-link visual representation. The expert uses this layer to visualise the network but also to input new elements while staying in context, that helps him remaining focus during the reading process and still observe peripheral information. In addition, we rely on the use of a non-relational database to provide the necessary flexibility to refine the model during the query process. Such database – i.e. Neo4J, https://neo4j.com – can be explored with declarative graph query language – such as Cypher, https://www.opencypher.org/.

Once the information is stored in the database, users can explore the network by querying any particular layer – i.e. family of relationships – group of layers, or all the network by using an inline query based on this meta-model. For example the query: *Person - Financial - Person* will display the financial layer, i.e. any interaction belonging to the *Financial* family that connects two entities of the type *Person*.

In order to support their exploration, users dispose of multiple graph views. Each query creating a new window that can be operated independently. Basic operations can be applied independently: zoom, pan, and details on-demand by clicking on individual elements (displaying its characteristics).

The tool is built on top of javascript and Sigma http://sigmajs.org for the front-end, with a back-end written in Python using Tulip https://tulip.labri.fr/TulipDrupal/, and Flask https://palletsprojects.com/p/flask/. It is freely available on github at https://github-.com/norbertFeron/graph-ryder-dashboard-v2. An example of Avres interface is available in Fig. 1. This interface has helped the analysis of a Nigerian human trafficking networks in France [5] [9].

References

1. Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G.: Network analysis in the social sciences. science 323(5916), 892–895 (2009)
2. Campana, P.: The structure of human trafficking: Lifting the bonnet on a nigerian transnational network. British Journal of Criminology 56(1), 68–86 (2016)
3. Campana, P., Varese, F.: Studying organized crime networks: Data sources, boundaries and the limits of structural measures. Social Networks (2020)
4. Cavallaro, L., Ficara, A., De Meo, P., Fiumara, G., Catanese, S., Bagdasar, O., Song, W., Liotta, A.: Disrupting resilient criminal networks through data analysis: The case of sicilian mafia. PloS one 15(8), e0236476 (2020)
5. Détruy, M.: La technologie au secours de la lutte contre l'exploitation sexuelle. LeFigaro (2019), 13/14 April - p15
6. Ficara, A., Cavallaro, L., De Meo, P., Fiumara, G., Catanese, S., Bagdasar, O., Liotta, A.: Social network analysis of sicilian mafia interconnections. In: International Conference on Complex Networks and Their Applications. pp. 440–450. Springer (2019)
7. Gollini, I., Caimo, A., Campana, P.: Modelling interactions among offenders: A latent space approach for interdependent ego-networks. Social Networks 63, 134–149 (2020)
8. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. Journal of complex networks 2(3), 203–271 (2014)
9. Lavaud-Legendre, B., Melançon, G., Pinaud, B., Plessard, C., Feron, N.: Analyse et visualisation des réseaux criminels. Tech. rep., COMPTRASEC, LaBRI, LISST (2019)
10. Mancuso, M.: Not all madams have a central role: analysis of a nigerian sex trafficking network. Trends in Organized Crime 17(1-2), 66–88 (2014)
11. McGee, F., Ghoniem, M., Melançon, G., Otjacques, B., Pinaud, B.: The state of the art in multilayer network visualization. In: Computer Graphics Forum. vol. 38, pp. 125–149. Wiley Online Library (2019)
12. Morselli, C., Savoie-Gargiso, I.: Coercion, control, and cooperation in a prostitution ring. The ANNALS of the American Academy of Political and Social Science 653(1), 247–265 (2014), https://doi.org/10.1177/0002716214521995

The opinions of a few: A cross-platform study quantifying usefulness of reviews

Osnat Mokryn[†]

University of Haifa

1 Introduction

Online reviews for products and services have surged in popularity, becoming a trusted and influential factor in consumers' decision process [1]. The fast adoption of reviews in the online world by both sites and consumers introduced scalability problems for sites and an overload of information to the searching consumers. As a means to ease the load, sites placed a useful voting button per review, enabling readers to indicate which reviews they find useful. The number of votes is counted and presented, and reviews indicated as useful are commonly presented at the top of the relevant page. This voting system had an enormous effect on both sites and consumers. Consumers prefer useful reviews, indicating they contribute significantly to their decision process; useful reviews also have a positive impact on sales [2]. Given their clear influence, both researchers and the industry are intrigued with the following question: What characteristics, if any, make a review useful?

Previous research has identified predictive characteristics of the review metadata and text, reviewer, specific emotions expressed, style of writing and information quality, and peripheral cues. Recent works, however, identified that the literature contains *conflicting findings* [3, 4].

Here, we investigate the attributes of over 1.2M reviews written by more than 327K users over three different review platforms. We explore three hypotheses.

Our first hypothesis, H_1*: Emotional reviews are useful*, was previously investigated for specific emotions, Anger and Anxious [5]. Emotional experiences trigger people to share them with others [6]. Social media has become an additional venue in which people are willing to share their personal experiences and emotions with strangers [7]. To find whether emotional reviews are more useful we employ a lexicon-based approach, word-emotion association lexicon (NRC) [8] for detecting eight basic emotions (Joy, Sadness, Anger, Fear, Disgust, Surprise, Trust, and Anticipation) and the sentiment of the review.

Our second hypothesis builds on previous research and correlations we found in the data[1]. H_2*: Longer reviews containing more nouns than other parts of speech are more useful*.

Our third hypothesis concerns the impact of the reviewers themselves. Review platforms award active reviewers and display this awarded reviewer reputation in a prominent way. In Yelp, 'Elite members' are elected each year based on 'good Yelp citizen-

[†] Corr. author: ossimo@gmail.com

[1] See full paper for relevant references [9]

ship'; In Amazon the 'Top-1000' reviewers have a special tag displayed next to their name; In IMDb reviewers have a special badges, for example 'IMDb Champ', and 'Top Reviewer'. To capture the *perceived impact* of a reviewer, we borrow from a known metric used to assess the impact of scientists and scholars, the *h-index* [10].We define two variables to capture a reviewer's history: *h-index*, i_5*-index*.

Reviewer *h-index* : Reviewer has index h if h of her N_p reviews have at least h useful votes each, and the other $(N_p - h)$ reviews have no more than h useful votes each. Hence, an *h-index* of 10 means the reviewer has at least 10 reviews, which were voted useful by at least 10 other people, and the rest of her reviews have less than 10 useful votes each.

Reviewer i_5*-index* : The number of reviews that received at least 5 votes.

Our third hypothesis is then H_3*: High impact reviewers' reviews are more useful.*

2 Results

The data in our research is comprised of four different datasets, as depicted in Table 1. Our study is comprised of two different experiments. The first is a binary classification of the usefulness of reviews (over SMOTE balanced datasets). The second is an exact score prediction algorithm.

Table 1. Summary of statistics for each of the datasets - original, and after SMOTE

	Total	Amazon	Yelp Bay Area	Yelp challenge	IMDb
Users	327638	21087	95296	43873	167382
Items	84274	2103	66803	11537	3831
Reviews	1242580	23868	488805	229907	500000
Original	Useful	12877	63074	14422	63789
	%Useful	53.9%	12.9%	6.2%	12.7%
SMOTE	Useful	13105	422595	209119	433744
	Not Useful	10763	425731	215479	436215
	%Useful	54.9%	49.8%	49.2%	49.8%

To classify whether a review is useful or not, we use a supervised learning based on a training set containing labeled data. Each review in the training set is represented by multi dimensional feature vectors, and is labeled as 'useful' or 'not useful' (normalized across sites). Reviewer information is calculated in the train set. The test set contains reviews represented as multi-dimensional feature vectors. The best results were obtained for XGBoost, reported here, on a stratified 6-fold cross validation. We performed different series of binary classification experiments for evaluating our different hypotheses separately and combined, as detailed here. Each of the hypotheses is evaluated separately, while the first experiment, referred to as *All*, uses all the features from all hypotheses.

Summary. We have shown over several datasets that a good predictor for predicting whether a review is voted useful is the reviewer's impact and that adding as features the emotions expressed in the reviews and textual features improves the prediction. Additional analyses showed that (1) Linear regression to predict the useful count further supports the above findings; (2) We find that reviewers that frequently write and for a

Dataset	Experiment	Accuracy	AUC
Yelp Bay Area	All	**0.944**	**0.886**
	Emotions	0.864	0.527
	Text	0.859	0.537
	Emotions-Text	0.869	0.521
	Reviewer	0.876	0.889
Amazon	All	**0.980**	**0.978**
	Emotions	0.582	0.578
	Text	0.584	0.571
	Emotions-Text	0.608	0.604
	Reviewer	0.978	0.977
Yelp Challenge	All	**0.925**	**0.878**
	Emotions	0.795	0.632
	Text	0.862	0.597
	Emotions-Text	0.917	0.555
	Reviewer	0.846	0.879
IMDb	All	**0.892**	0.740
	Emotions	0.842	0.507
	Text	0.866	0.500
	Emotions-Text	0.866	0.500
	Reviewer	0.757	**0.801**

Table 2. Accuracy and AUC for the Usefulness Classifiers

long period of time tend to garner useful votes for their reviews, and that they for a small, yet influential, subset of the reviewers. Further details can be found in the full paper [9].

References

1. Mudambi, S.M., Schuff, D.: What makes a helpful review? a study of customer reviews on amazon. com. MIS quarterly **34**(1) (2010) 185–200
2. Zhu, F., Zhang, X.: Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. Journal of marketing **74**(2) (2010) 133–148
3. Yin, D., Mitra, S., Zhang, H.: Research note—when do consumers value positive vs. negative reviews? an empirical investigation of confirmation bias in online word of mouth. Information Systems Research **27**(1) (2016) 131–144
4. Hong, H., Xu, D., Wang, G.A., Fan, W.: Understanding the determinants of online review helpfulness: A meta-analytic investigation. Decision Support Systems **102** (2017) 1–11
5. Yin, D., Bond, S.D., Zhang, H.: Anxious or angry? effects of discrete emotions on the perceived helpfulness of online reviews. Mis Quarterly **38**(2) (2014) 539–560
6. Christophe, V., Rimé, B.: Exposure to the social sharing of emotion: Emotional impact, listener responses and secondary social sharing. European Journal of Social Psychology **27**(1) (1997) 37–54
7. Tettegah, S.Y., Espelage, D.L.: Emotions, technology, and behaviors. Academic Press (2015)
8. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word–emotion association lexicon. Computational Intelligence (2012)
9. Mokryn, O.: The opinions of a few: A cross-platform study quantifying usefulness of reviews. Online Social Networks and Media **18** (2020) 100080
10. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences of the United States of America **102**(46) (2005) 16569

A Framework for Interaction-based Propagation Analysis in Online Social Networks

Daniel Thilo Schroeder[1,2], Pedro G. Lind[3], Konstantin Pogorelov[2], and Johannes Langguth[2]

[1] Technical University of Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany,
[2] Simula Research Laboratory, Martin Linges vei 25, 1364 Fornebu, Norway
daniels|konstantin|langguth@simula.no,
[3] Oslo Metropolitan University, Pilestredet 46, 0167 Oslo, Norway
pedro.lind@oslomet.no,

1 Introduction

Online social networks create a digital footprint of human interaction naturally by the way they function. Thus, they allow a large scale analysis of human behavior which was previously infeasible for social scientists. Consequently, social networks have been studied intensely in the last decade. The core of most social networks is the relationship between users which can be described as a graph. The graph can be either undirected, as is the case for the friendship relation of Facebook, or directed, which is the case of the follower relation on Twitter. The relationship is readily visible, e.g. on the user interface the social networks themselves. However, these edges are unweighted expressions of interest and reflect how individuals have chosen to relate to each other rather than how they actually interact with each other. For studying information propagation, comparing interaction properties is crucial and, therefore, using models based on connections that reflect different dimensions and strengths of acquaintance seems appropriate. Thus, there is a need for obtaining weighted edges from the communication that occurs on the social network. In this paper, we present a novel method to calculate an acquaintance score between pairs of Twitter users and use the resulting networks to enable the analysis of interaction based information propagation. By understanding the frequency and velocity with which individuals share content as a measure of acquaintance, it becomes possible to predict, compare communication patterns, and detect unusual communication. In contrast to previous work which assigns edge weights based on tie strength [3], our score considers the response time as a crucial factor and, therefore, enables time-based spreading comparisons.

2 Building an empirical social network with weighted acquaintances: beyond the follower-network of Twitter

To build an empirical social network with weighted acquaintances from Twitter, we introduce two functions, which estimate a "strength" to measure the level of acquaintance between pairs of Twitter users. The first function evaluates the effect of the acquaintance between two users, i and j, in the waiting time T_{ij} between retweets among them. We assume that as more acquainted two users are, as faster they will retweet each other. Thus, we define the function

$$\alpha_{ij} = \frac{1}{n_{ij}} \cdot \sum_{k=0}^{n_{ij}} \frac{1}{T_{ijk}} \,. \tag{1}$$

where n_{ij} is the total number of messages that j shared from i and $T_{ijk} = t_{ik} - t_{jk}$ where t_{ik} and t_{jk} are the timestamps of user i or j for a tweet k. In Figure 1b we show the intervals between successive retweets in a sample of 60k retweets. We observe that the distribution of waiting times T_{ij} follows approximately a power-law with an exponent of $-3/5$.

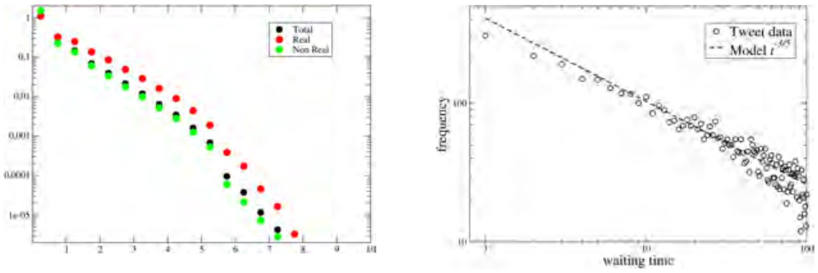


Fig. 1. The figure on the left shows to what extent the existence of an edge in the Twitter follower network correlates with the existence of a $c_i j$-score > 0. The figure on the right shows to what extent the distribution of time differences between tweets and retweets corresponds to the model function $t^{-3/5}$.

The second function evaluates the effect of the acquaintance between two users on the frequency of retweets. Here, we assume that the more acquainted two user are, the more frequently they will retweet each other. Thus, we define the function

$$\beta_{ij}(t) = \frac{2n_{ij}}{n_i + n_j} \,. \tag{2}$$

where n_i and n_j represent the amount of messages authored by i and j, respectively. Having these two functions, we define the “strength of acquaintance” between two users as

$$C_{ij}(t) = \alpha_{ij}\beta_{ij} \,. \tag{3}$$

Even though the C_{ij}-score is semantically different from the follower-network, there is a correlation between C_{ij}-score > 0 and Twitter’s follower edges. Figure 2a supports this hypothesis by plotting the histogram of all C_{ij}-scores and compare it to the corresponding histogram containing just the scores which are backed-up by an actual follower relationship.

3 Discussion and conclusions

We have applied this method to approx. 830 million statuses (532,254,060 tweets, 299,053,952 retweets), building a network with approx. 120 million estimated weighted acquaintances among 73,061,739 users. Such network can now be taken as a framework for investigating the structure of real social networks, in what concerns its degree

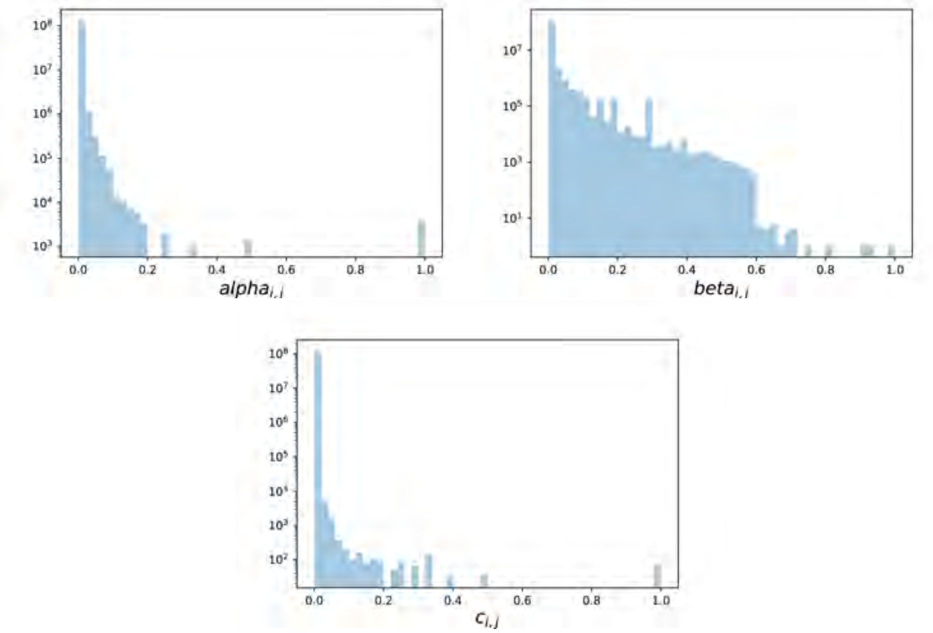


Fig. 2. This figure shows in the upper left corner the distribution of the average time between tweets and retweets of a user pair (see equation 1). In the upper right corner the distribution of the frequencies (see equation 2) is shown and in the lower right corner. Finally, the distribution of all $C_i j$-scores (equation 3).

distribution, degree-degree correlations, average shortest paths and betweenness. The resulting scores constitute a dense representation of acquaintances for all pairs of users. Due to the size of typical social networks, such a representation is only practical for communities of interest in the study of social phenomena, such as the spread of misinformation or hate speech. However, within such communities, the likely spread of information can easily be derived from the C_{ij}-scores, which is not possible from the link structure of the social network alone. For larger communities, it will be necessary to sparsify the representation by omitting negative scores from the representation.

References

1. Schroeder, Daniel Thilo, Konstantin Pogorelov, and Johannes Langguth. ”FACT: a Framework for Analysis and Capture of Twitter Graphs.” 2019 International Conference on Social Networks Analysis, Management and Security
2. Sadri, A.M., Hasan, S., Ukkusuri, S.V. and Lopez, J.E.S., 2018. Analysis of social interaction network properties and growth on Twitter. Social Network Analysis and Mining, 8(1), p.56.
3. Gilbert, E. and Karahalios, K., 2009. Predicting tie strength with social media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’09). Association for Computing Machinery, New York, NY, USA, 211–220.

A popularity model for information spreading: Twitter as a case study

Lília Perfeito[1] and Joana Gonçalves-Sá[1,2]

[1] LIP - Laboratório de Instrumentação e Física Experimental de Particulas, Lisboa, Portugal
[2] Departamento de Física, Instituto superior Técnico, Universidade de Lisboa

1 Introduction

Extreme inequality is present in multiple human activities. This inequality gives rise to so-called heavy tail distributions whereby few elements are responsible for the vast majority of the observations. Examples include the size of cities, the distribution of friends in a social network, wealth and even the frequency with which words are used [1, 2]. There has been intense discussion on the exact shape of the tail of the size distributions: whether they are closer to lognormal, power law or others [3, 4]. The choice of statistical distribution may be solved by using more data and more robust tests but it is not clear whether that is the case. Moreover, even if we know the shape of the distribution, estimating its parameters or performing statistical tests on heavy tails is extremely challenging. An alternative approach to heavy tail distributed data is to build mechanistic models from basic principles. A number of them have been proposed for specific examples [5, 1, 2] but quantitative fits are rare and a general formulation applicable to different systems is lacking. Here we focus on the spread of information in social media to build a model that explains the observations well and whose parameters are readily interpretable.

With the advent of the internet, and particularly of social media, the creation of information (whether factual or not) and its massive spread are at the fingertips of many. Whether in social media [7] or academia [8], the attention information receives is distributed according to a heavy tail: most posts and scientific articles receive the attention of few people, while a hand-full are extremely popular. One of the most used social media platforms is Twitter where the easiest way to measure information spreading is to quantify re-tweet cascades. These are groups of tweets which originate in one user and contain the exact same text, along with the indication that it is a “re-tweet” and the original author’s identifier. Most tweets are not retweeted at all, while others give rise to very long cascades. When looking at the overall distribution, again we find a heavy tail which is well approximated by a power law [6, 7]. It has been shown that the structure of the network is sufficient [7, 9] but not that it is necessary to generate the heavy tailed distribution.

2 Results

Our dataset comprises tweets obtained using Twitter’s REST API [10] between the 26th of March and the 25th of May, 2020 with the keywords “Covid” or “Corona” in

Portuguese. The choice of keywords was related to finding a popular topic; the choice of language was to make sure we could get full cascades without reaching Twitter's rate limits. Tweets were extracted daily with the free version of the API without reaching the maximum daily limit. Therefore we expect to have obtained nearly all tweets meeting the criteria [11]. In total we obtained 4 970 489 unique tweets, where unique means it contained a unique text. Of those, 85% were single copy and hence did not generate any cascades. We analyzed the 1516 (or 0.3%) cascades which contained at least 1000 re-tweets. As can be seen in figure 1 A, the total size of the cascades follows a power law with exponent -1.96. We then grouped data hourly and counted how often each tweet was re-tweeted per hour.

We model tweet cascades as growing exponentially (the more copies of a tweet there are, the bigger the probability it will get re-tweeted) with an exponential popularity decay: as time passes the rate growth decreases. As such the size N of cascade i follows:

$$\frac{dN_i}{dt} = N_i(t)a_i e^{-g_i t} \tag{1}$$

The parameters a_i and g_i measure the initial popularity and its subsequent decay and k_i is the size of the cascade at infinity. We fit all three of these parameters to each tweet. Much like any other human activity, twitter usage is not constant through time and shows a strong circadian rhythm (figure 1B). In order to account for that, the time scale we use is not that measured by our clocks but that measured by how much twitter activity there is. As such, during the night barely any time passes, while during the day it goes faster. Time in "generations" of tweets is then given by:

$$t = \sum_{n=0}^{t_c} \frac{N_{total}(n) - N_{total}(n-1)}{\widetilde{\Delta N}} \tag{2}$$

where N_{total} is the sum of all tweets that were produced in that hour, $\widetilde{\Delta N}$ is the mean number of total tweets produced per hour and serves as a normalizing constant; t_c is chronological time measured in hours. With this transformation, the total number of tweets increases by $\widetilde{\Delta N}$ per time unit.

Equation 1 was fitted to each of the 1516 tweet cascades independently and in 98.7% of them the R^2 of the fit was larger than 0.90. The 18 cascades where it was not showed two waves, a small first one followed by a much higher second one. The fact that these cascades do not fit the model is interesting in itself but is beyond the scope of the current work. We focused on the 1498 where the fit was good. Figure 1C shows four examples of cascades and their fits. The parameters g and a measure different aspects of popularity but are highly correlated (Spearman correlation 0.52 $p < 0.001$), figure 1D, in large part due to the structure of equation 1. On the other hand, a is not correlated with the total size of the cascade (Spearman correlation $p > 0.1$ - figure 1E) and g is weakly correlated (Spearman correlation -0.18 $p < 0.001$)), indicating they are representing something else than just the the number of re-tweets. Importantly, it means that looking at the size of twitter cascades is different from looking at popularity. Indeed the distribution of a is not well approximated by a simple power law (figure 1F). Instead, the distribution is much closer to a lognormal which has finite variance.

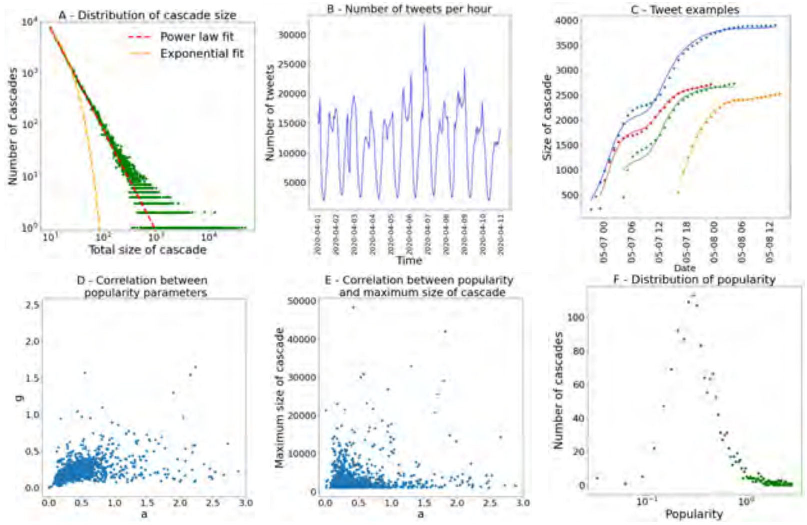


Fig. 1. **A** Distribution of sizes of twitter cascades larger than 10 re-tweets. The distribution is well approximated by a power law with exponent -1.96. **B** Number of new tweets in Portuguese with either the word Covid or Corona. Counts are grouped by hour. **C** Example of three tweets (dots) and their respective fits to equation 1 (lines). **D** Relationship between the fitted parameters a and g. Each dot represents the value of the two variables in a cascade. **E** Relationship between the maximum size of each cascade and its fitted parameter a. **F** Distribution of parameter a.

References

1. Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. science, 286(5439), 509-512.
2. Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary physics, 46(5), 323-351.
3. Perline, R. (2005). Strong, weak and false inverse power laws. Statistical Science, 68-88..
4. Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. SIAM review, 51(4), 661-703.
5. Yule, G. U. (1925). II.—A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FR S. Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character, 213(402-410), 21-87.
6. Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011, February). Everyone's an influencer: quantifying influence on twitter. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 65-74).
7. Gleeson, J. P., Ward, J. A., O'sullivan, K. P., & Lee, W. T. (2014). Competition-induced criticality in a model of meme popularity. Physical review letters, 112(4), 048701.
8. Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. Science, 342(6154), 127-132.
9. Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. Scientific reports, 2, 335.
10. https://developer.twitter.com/en/docs
11. Kim, Y., Nordgren, R., & Emery, S. (2020). The Story of Goldilocks and Three Twitter's APIs: A Pilot Study on Twitter Data Sources and Disclosure. International Journal of Environmental Research and Public Health, 17(3), 864.

Fair Comparisons Among Network Sampling Strategies

Yitzchak Novick[1,2] and Amotz BarNoy[3]

[1] CUNY Graduate Center, New York NY 10016, USA
ynovick@gradcenter.cuny.edu
[2] Touro College and University System, New York NY 10018, USA
[3] Brooklyn College, Brooklyn NY 11210, USA
amotz@sci.brooklyn.cuny.edu

1 Introduction

In 2003, in a landmark paper, Cohen et al [2] introduced a new random vertex sampling method that gives a higher expected degree than naïve random sampling. Instead of using a randomly selected vertex, the random vertex's neighbors are collected and one of these is selected at random. Based on Feld's 'friendship paradox' [3], the neighbor's degree is presumed to be higher than the original vertex, and it has been demonstrated that this is in fact the case [2, 4].

It seems to be ignored, however, that the method is computationally expensive compared to naïve random sampling. One is required to sample twice as many vertices in order to accumulate a collection of the same size. The method was originally conceived in a scenario where there are a limited number of immunizations that need to be administered and the goal is to give them to the highest degree vertices possible. Under these circumstances, it is of course natural to ignore a constant factor increase in computational cost. But if the method is generalized to apply to scenarios where computational efficiency cannot be ignored, a fairer comparison should be sought.

The concept is best explained with an example. Consider a scenario where some budget, b, allows us to sample vertices from the network, with the goal of maximizing the highest degree found when the budget has been exhausted. The typical comparison between random neighbor sampling and naïve vertex sampling would compare b random vertices to b neighbors of vertices. However, the correct comparison should be between b random vertices verses a collection comprised of $b/2$ random vertices and $b/2$ neighbors. In Fig. 1 we plot the average max-degree found over multiple experiments for a given budget in a Barabási Albert graph [1] with $n = 10,000$ and $m = 20$. The green dots are max-degrees found with naïve random sampling. The blue dots are max-degrees found for a full b neighbors, implying no cost for sampling the first vertex. The red dots represent the true results of a budget b, a collection comprised of one-half random vertices and one-half random neighbors. It is telling that the red dots lie closer to the blue dots than the green dots. One should not attribute the entire gain in neighbor sampling to the fact that more vertices have been sampled. As we noted, it has been proven that the half of the collection containing neighbors will have a higher expected degree. However, the red dots tell a more complete story then the blue dots because they reflect the effectiveness of the method in light of the extra cost.

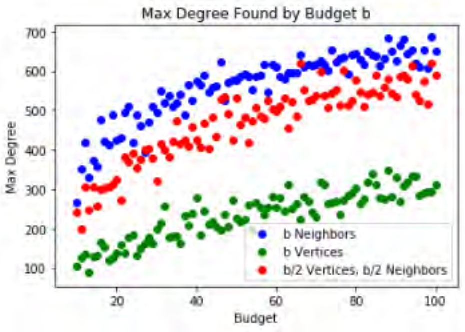


Fig. 1. The average max-degree returned for multiple experiments where vertices are collected up to a given budget.

2 Results

Our study considers multiple possible costs of sampling methods, such as the cost of collecting neighbors or the cost of compromising privacy by learning about neighbors. Using these concepts, we build cost models that allow for a robust comparison between methods.

We also consider additional sampling methods that address these new costs. For example, we find an advantage in sampling methods that take more than one neighbor of a selected vertex, which allows us to further capitalize on the 'price' already paid for the first vertex. Alternatively, in situations where the degrees of the individual vertices can be determined, we consider taking the highest-degree vertex from some subset of the first vertex's neighbors. We find a significant improvement with the inclusion of a single additional neighbor, but the gains quickly become insignificant with the inclusion of additional neighbors.

Through these analyses, we are able to better understand the strengths and weaknesses of different methods and determine the scenarios to which they are best suited.

References

1. Barabási, A. L., Albert, R.: Emergence of Scaling in Random Networks, Science 286, 509–512 (1999))
2. Cohen, R., Havlin, S., Ben-Avraham, D.: Efficient Immunization Strategies for Computer Networks and Populations, Phys. Rev. Lett. 91(24), (2003)
3. Feld, S.: Why Your Friends Have More Friends Than You Do, Am. Jour. of Soc. 96(6), 1464–1477 (1991)
4. Kumar, V., Krackhardt, D., Feld, S.: Network Interventions Based on Inversity: Leveraging the Friendship Paradox in Unknown Network Structures, https://vineetkumars.github.io/Papers/NetworkInversity.pdf, (2018)

Relationship graph for organisations using communication tools

Abdel-Rahmen Korichi[1,2], Hamamache Kheddouci[1], and Daniel James West[2]

[1] Université Claude Bernard, Lyon 1, France,
WWW home page: `https://www.univ-lyon1.fr/`
[2] Panalyt Pte. Ltd.
WWW home page: `https://www.panalyt.com/`
`ak@panalyt.com`

1 Introduction

The rise of digitalisation and remote working, accelerated by the COVID-19 crisis, has created an environment where the volume and the adoption of online communication and productivity tools have increased rapidly. As an example, Microsoft has revealed that its Teams daily active users have jumped by 70 per cent to 75 million [1]. Effectively understanding and managing employee interactions and having a real-time view on how an organisation is functioning across its underlying networks was already trendy, but it has now become critical.

Organisational Network Analysis, or ONA, is arguably one of the best solutions to tackle this issue. ONA is the study of how communications, information and decisions flow through an organisation. As a consequence of the growing remote and digital culture, organisations are now sitting on a huge and untouched amount of data, including emails exchanges, calendar invites, instant messaging applications, and other productivity applications. By collecting the data, and examining the strength, frequency and nature of interactions between people in those networks, practitioners can gain a better understanding of the relationships that affect the effectiveness of individuals and groups, helping them to make better decisions for their business.

By presenting organisational networks using graph structures, where nodes are actors and edges are relationships between them, it is possible to apply graph theory methods and to deduce meaningful information about an organisational network.

There are two challenges in this problem. Firstly, we need to evaluate the relationship between people to know how to connect them in the graph. Although many authors have proposed different methods for building a communication graph, the measures proposed are usually only deduced from the volume of communications between people - as described in [2] - and a lot of information that can be found in the logs is ignored. Secondly, to query partial subgaphs with specific characteristics - department, grade, location, etc. - it is necessary to add information based on other datasets.

In our approach, we define a relationship as a function based on multiple measures deduced from the communication logs such as the response time and the tenure of the relationship. We only focus on metadata based on the logs of the communication without looking at the content. Additionally, we explain the value of crossing the data from different sources to answer very fast specific business questions.

2 Construction of the relationship graph

Over the last few years, more and more online communication and management tools have started to provide API access for their users, making it possible for any organisation to access and work on the data very easily. ([3], [4], [5], [6]). Organisations can extract metadata from emails, instant messaging applications and video conference applications and they can know who communicated with whom, when, for how long, etc. The data can be pulled and stored in a database and organisation can have a real-time view of their organisational health.

Unfortunately, the methods used to draw connections between people based on communication are usually only deduced from the volume of messages (count of messages sent by a node i to a node j, number of e-mails sent by a node i to a node j divided by the total number of e-mails sent by member i [7], geometric mean of sent-received counts [8], etc.). The problem with this approach is that it doesn't tell you much about the relationships' quality. Are your people responding to each other? How long do they take to reply? Is the employees' communication balanced?

Organisations should be able to encompass whatever they consider as factors for their employees' relationship in a unique score (the relationship score) and not just the volume of message. This is why in our approach, we propose to create a custom formula taking into account any measure deduced from the logs.

$$R(t_1,t_2)((\alpha_1,M_1),(\alpha_2,M_2),...,(\alpha_n,M_n)) \tag{1}$$

Where $M_1,M_2,...M_n$ are measures of the communication network (deduced from employees' logs, e.g: response time, reciprocity, tenure of the relationship, etc.), $\alpha_1,\alpha_2,...\alpha_n$ are weights of importance of these measures, and t_1,t_2 are two timestamps where $t_1 < t_2$ that define the time range of the communications.

Once the function R (1) is defined and the values have been computed for all pairs of employees, the next step is to choose different thresholds for different categories of relationships based on the distribution of the relationship scores and limits sets by the experts. For example:

$$R(t_1,t_2)((\alpha_1,M_1),(\alpha_2,M_2)) = \begin{cases} Negligible\ relationship & R < x_1 \\ Weak\ relationship & x_1 \leq R < x_2 \\ Medium\ relationship & x_2 \leq R < x_3 \\ Strong\ relationship & x_3 \leq R \end{cases}$$

Where $x_1 < x_2 < x_3$.

3 Integration of other sources of data

Being able to measure the relationships in an organisation is great, but giving more background to the employees' data brings even more value. Indeed, organisations usually have access to a diverse type of data points about their employees coming from human resource information systems (HRIS), applicant tracking systems (ATS), payroll systems, task management systems, performance systems, customer relationship management (CRM), and so on and so forth. For instance, by bringing together the relationship score and data from an HRIS - data points such as the starting date, gender, department, location, job title, managers etc. - organisations can do very specific queries to retrieve a partial subgraph with specific characteristics. Once we have merged the data together, we see that it is very easy to query a partial subgraph that verifies a list of properties. The variety of questions that can then be answered is huge:

- How is the relationship between managers and their direct reports by department?
- How well engaged are employees with up to 3 months tenure by manager?
- How are different departments collaborating together?

4 Conclusion

By building a relationship score, organisations have access to new ways of measuring engagement and the collaboration between employees over time. Additionally, by merging the relationship with other employee data, organisations can perform powerful queries to generate a partial subgraph.

There are several benefits:

- The relationship formula (1) has been defined by the organisation itself and therefore is not a black box, managers can understand the exact reason why a relationship score is low or high, allowing them to take appropriate and actionable decisions.
- Analysis can done very fast, as we don't work on the whole graph.
- Analysis are very precise, as we work only on a partial subgraph that verify specific proprieties.

Eventually, managers have a real-time view of their organisational health and can see the consequences that can have an internal or an external decisions or events like the COVID-19 pandemic.

References

1. Windows Central, `https://www.windowscentral.com/microsoft-teams-hits-75-million-daily-active-users`
2. M. Duczynski, B. Yin, Hierarchy detection through email network analysis
3. Microsoft Graph API, `https://docs.microsoft.com/en-us/graph/use-the-api`
4. Gmail API, `https://developers.google.com/gmail/api`
5. Slack API, `https://api.slack.com/`
6. Zoom API, `https://marketplace.zoom.us/docs/api-reference/zoom-api`
7. R. Michalski1, S. Palus1, P. Kazienko1, Matching Organizational Structure and Social Network Extracted from Email Communication. Lecture Notes in Business Information Processing 87:197-206 (Jun 2011)
8. Munmun De Choudhury, Winter A. Mason, Jake M. Hofman, Duncan J. Watts, Inferring Relevant Social Networks from Interpersonal Communication (Jan 2010)

Part XIII

Urban Systems and Networks

Urban gentrification as an avalanche process

Diego Ortega[1], Javier Rodríguez-Laguna[1], and Elka Korutcheva[1,2]

[1] UNED, Departamento de Física Fundamental, Paseo Senda del Rey 9, E-28040 Madrid, Spain
dortega144@alumno.uned.es
[2] Bulgarian Academy of Sciences, G. Nadjakov Inst. Solid State Physics, 1784 Sofia, Bulgaria

1 Introduction

Gentrification occurs in urban zones where a financial gap between people exists. In this situation real state investors tend to expand its area looking for economical profit. This creates pressure over the less economically favoured class who is socially and monetarily coerced to move out. As can be inferred, gentrification implies a segregationist trend. A comprehensive model of segregation was introduced by Schelling in [7]. The model considered two different social groups (*red* and *blues*) distributed over a square lattice with some vacancies. Agents were able to be relocated in an empty place if this movement increased their *happiness*. Several variations of this pioneer work have been developed: in [3] the key parameter is the tolerance, T, defined as $T = N_d/(N_d + N_s)$ where N_d and N_s are the number of different and similar neighbors, respectively. Another interesting variant is the so called *open city* [4], where agents could leave or enter the lattice, thus internal and external moves are possible. Whereas these models consider fixed values for wealth or tolerance, here we characterize how changes in the economic environment for both kind of agents affect the urban structure.

In this communication, which essentially is a brief overview of our recently published work [6], we extend the model of open city [4] by adding an external magnetic field H. This field can be interpreted as a financial gap between the agents. The considered network is a 100 × 100 square lattice with free boundary conditions, where each agent is connected with its actual closest neighbors (*Moore neighborhood*). The transference rule that allows the displacement of agents is defined as a function of a dissatisfaction index I_{dis} that can be written as:

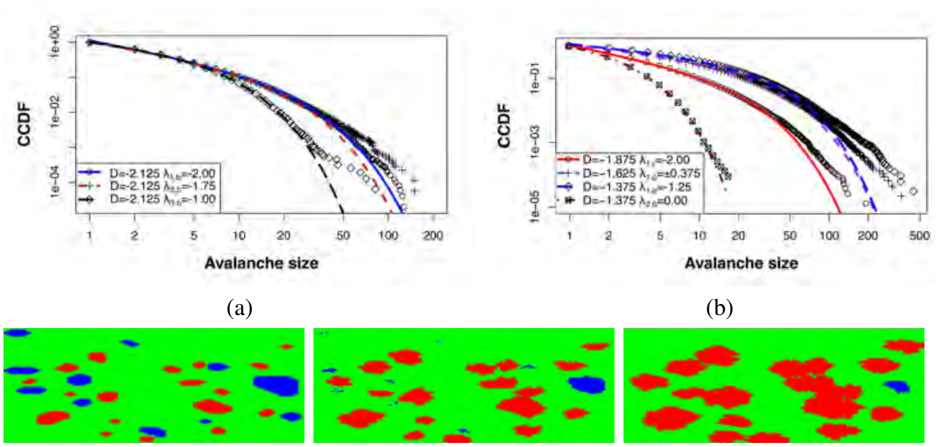
$$I_{dis} = N_d - T(N_s + N_d) + D \pm H \leq 0, \tag{1}$$

Lower values of I_{dis} correspond to a high value of *happiness* and contrarily. The first two terms of Eq. (1) accounts for social preferences, considering the type of agents in his/her neighborhood. D is associated with the average economic status of the system. If $D < 0$,the system is attractive for agents, and some unhappiness arising from neighbors can be balanced. Finally, we assume that wealth levels of the two kinds of agents considered are not equal, so H is substracted from red agents and added to blue ones. In this way, a situation where blue agents leave the lattice and red agents are coming into can be described easily, using $H > 0$.

As we increase H, some blue agents may feel dissatisfied, being $I_{dis} > 0$, and are forced to leave the city. The empty places generated are occupied by red agents coming from outside. After that, some blue agents close to them may feel frustrated making the process goes on in a self-sustained way, giving rise to an avalanche.This is what we called a *blue avalanche*, because it originates with blue agents leaving the lattice. Yet, there is another way to generate an avalanche. We depart from an equilibrium situation with some vacancies in the lattice. Before H is strong enough to force blue agents out, red agents may be able to fill these vacancies up, provoking an avalanche. As this process starts with the incoming red agents is defined as a *red avalanche*. Finally, a *purple avalanche* occurs when both kind of avalanches, red and blue ones, act simultaneously over different neighborhoods.

2 Results

To characterize avalanches we have depicted the histogram (*CCDF*) and fitted a power-law cutoff function, $Cx^{-\alpha}e^{-\frac{x}{x_0}}$. The power-law exponents from Fig. 1a and Fig. 1b are in the range $[-1.781, -1.381]$. Values in this numeric interval were previously reported in references [1, 2, 5], concerning self-organized criticality in different systems.



(a)

(b)

(c)

Fig. 1. In (a) and (b) the CCDF of the avalanche distribution sizes for $T = 1/4$ are depicted. For each curve the D value is specified. Treshold values are given as $\lambda_{n,k}$ where n is the avalanche set index and k its kind: r for reds, b for blues and p for purple ones. For purple avalanches upper and lower tresholds are expressed as $D \pm \lambda_{1,p}$. The fitted power-law cutoff functions are depicted with lines. In Fig (c) system evolution for $T = 1/4$ and $D = 0.875$. From left to right: equilibrium (left), 6 MC steps (center) and 13 MC steps (right).

For different values of T and D a wide range of phenomena with social interpretations appear. For $T = 1/4$ and $D = -2.125$ (see Fig. 1a), up to three blue avalanches are needed to deplete the system from blue agents. The social meaning is clear: the less favoured agents will try to stand on an economic advantageous environment despite their income gap with the other group. To achieve this goal diverse neighborhood structures will be created (ghettos). Blue, red and purple avalanches arise for $T = 1/4$ and greater values of D (see Fig. 1b). These red solid and blue dashed curves are dissimilar, pointing out different neighborhoods involved in their respective avalanches. The last analyzed values in this communication are $T = 1/4$ and $D = 0.875$. Now the city can be considered an economically deprived area which produces a *predominant vacancy state* [6]. In this regime only small clusters remain on the system (see left panel on Fig. 1c). Socially, the situation might be compared to the one in the Chicago suburbs where the population could increase their personal ties via a community network [3] to overcome monetary issues. However, in our model, the evolution of the system departs in two opposite directions: blue agents, which does not cooperate, are removed from the system, in contrast to red agents, which increase their population.

Although several connections with social situations are established in [6] future extensions of this research might study its application over real-world scenarios. Another way to complete this study is to characterize the system behaviour with different economical zones in the lattice or consider another type of neighborhood.

Summary. In this communication we have extended the Schelling segregation model, including in it economic terms. The studied network is a 100 × 100 lattice with free boundary conditions. Each agent is connected with its eight closest neighbors. When an agent is relocated only the actual neighborhood is considered. For different parameters values a wide range of phenomena with social interpretations appear, i.e. gentrification. This social reality occurs in urban zones where a economical gap between people exists and the financially handicapped ones are forced to leave. Another interesting scenario happens when the environment is economically deprived and only small clusters remain on it. This case higlights the importance of cooperation to overcome monetary issues. Both situations evolves trough avalanches which are fitted to power-law curves. The exponents of these curves are in the range of other works related to self-organized criticality.

References

1. Batac, R., Paguirigan, Jr. A., Tarun, A., Longjas, A.: Sandpile-based model for capturing magnitude distributions and spatiotemporal clustering and separation in regional earthquakes. Nonlinear Processes in Geophysics 24, 179 (2017).
2. Beggs, J. M., Plenz, D.: Neuronal Avalanches in Neocortical Circuits. The Journal of Neuroscience 23, 11167 (2003).
3. Gauvin, L., Vannimenus, J., Nadal, J.P.: Phase diagram of a Schelling segregation model. Eur. Phys. J. B. 70, 293–304 (2009).
4. Gauvin, L., Nadal, J.P., Vannimenus, J.: Schelling segregation in an open city: A kinetically constrained Blume-Emery-Griffiths spin-1 system. Phys. Rev. E 81, 066120 (2010).
5. Munoz, M. A., Dickman, R., Vespignani, A., Zapperi, S.: Avalanche and spreading exponents in systems with absorbing states. Phys. Rev. E 59, 6175 (1999).
6. Ortega, D., Rodríguez-Laguna J., Korutcheva, E.: Roughness and avalanches in an extended Schelling model: an explanation of urban gentrification. arXiv:2007.10767v1, (2020).
7. Schelling, T.: Dynamic models of segregation. J. Math. Sociol. 1, 143–186 (1971).

On the breakup patterns of urban networks under load

Marco Cogoni and Giovanni Busonera

CRS4 - 09010 Pula (CA), Italy
marco.cogoni@crs4.it

1 Introduction

Traffic has been extensively studied in recent years with a focus on the free-flow to congestion transition. Several models have been employed, from microscopic to coarse-grained approaches [1]. Very recently, a new perspective, based on percolation [2], has been proposed to study traffic flows in large cities with real GPS data. This approach disregards vehicle dynamics, by focusing on the ability of each road to guarantee transportation efficiency above some minimum threshold. What emerged from these studies was that, beyond the well known passage from free-flow to the congested state, a percolation transition exists: when observing the network as a whole, it progressively decomposes from a single giant (strongly connected) component to a set of separated clusters, each able to sustain traffic within its boundaries above some threshold speed, but functionally disconnected from the others [2]. It is known that a urban network graph undergoes a critical percolation transition when a fraction of its edges is removed. The resulting strongly connected components form a structure of clusters whose size distribution follows a power law with critical exponent τ [2]. This exponent changes under different traffic regimes [3]. A complete explanation of which factors affect τ is still lacking. We argue that spatial correlations may play a major role influencing the τ behavior since their range is known to grow when going from free-flow to congested real traffic [5]. We show that values of τ for random noise with increasing spatial correlations and those from GPS data for increasing congestion follow a similar behavior.

2 Methods

We obtained the transportation networks (only roads open to cars) from OpenStreetMap in the form of directed weighted graphs centered on New York City and London (1600 km^2 each). To perform random percolation, edges are uniformly removed whereas, for real data, they are deleted when the local average speed is under a critical threshold (congested). In both cases, we localize the critical probability threshold p_c, which determines the average number of remaining edges, by performing a set of percolation instances for $p \in (0,1)$ and choosing p that leads to the maximum size of the second largest cluster. In the synthetic, uncorrelated case we use a uniform random generator in $(0,1)$ to assign a ”speed” value to each edge, while to induce spatial correlation, we compute the graph Laplacian and its spectrum via a fast approximated method [7], then we perform a Graph Fourier Transform of the uncorrelated noise and filter it [10] by multiplying the eigenvalues with a power-law $f(q) \sim q^{-\lambda}$: a rapid decay ($\lambda \to 2$) leads to long spatial correlations, while $\lambda \to 0$ is equivalent to uncorrelated noise. For the real data, we first compute the average speed value on each edge for the relevant time span (UBER data has a one hour granularity) and then divide it by the maximum speed observed on that edge over six months. We compared the existing results from Ref. [2] (based on undisclosed data) with our percolation analysis conducted both on real traffic (from the UBER Movement datasets) and for synthetic data with increasing spatial correlation. To compute the critical exponent τ we perform a linear fit on a log plot of the binned distribution of the cluster sizes for each percolation instance [12]. Finally, to produce a spatial map of the city areas associated to the main clusters, we plot the number of times that, over several percolation replicas, each edge belongs to the 1st, 2nd and 3rd largest cluster, respectively.

3 Results

We first use real data to perform percolation runs on a monthly basis for London and New York City networks: during low congestion periods (0AM-3AM), τ is close to the one associated to percolation in mean-field networks ($\tau \sim 2.5$), while during rush hours (6AM-9AM) the exponent approaches the one observed for a regular square lattice ($\tau \sim 2.05$), as shown in Fig. 1. We subsequently analyze random percolation with increasing spatial correlations. For each city, τ slowly decreases with longer correlations. For both cities τ ranges from ~ 2.2 for $\lambda = 0$ to ~ 2.0 for $\lambda \sim 2.0$. These results show that a simple random percolation is able to capture the basic properties of real uncongested traffic (higher τ similar to mean field networks), and that increasing spatial correlations leads to a different cluster size distribution (lower τ similar to lattice percolation), typical of rush hours. We believe this to be a useful result for a better understanding of how traffic clusters behave over large urban areas under different congestion levels.

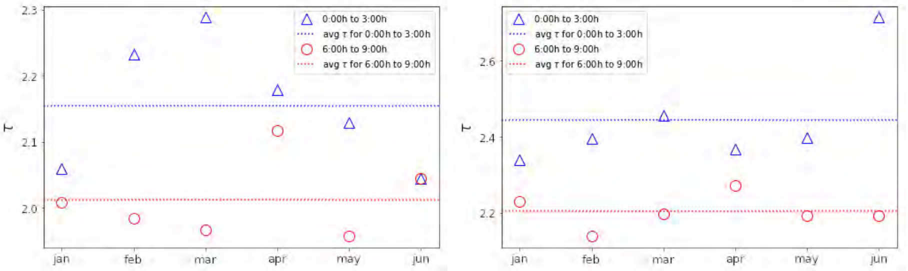


Fig. 1. Monthly average τ for London(left) and NYC(right) (rush hours(red) and off-peak(blue))

We finally present some preliminary evidence that, at criticality, the largest clusters display spatial predictability and that cluster configurations from 2000 replicas lead to a small number of breakup patterns, specific for each city. This holds both for random percolation and real traffic. Results obtained from synthetic percolation (uncorrelated) are shown in Fig.2 for London and NYC, where the three largest clusters are well localized and spatially distinct. Results from correlated percolation and from real traffic data are qualitatively similar, but the statistics for the latter is limited to one sample per week. This consistent cluster organization appears to be strongly influenced by local topographical structures such as rivers and bridges. The maps help visualizing how frequently city areas belong to the main functional traffic clusters and could be useful for city planners to improve city connectivity by easily comparing different road topologies. Moreover, citizens could better choose where to buy a house by selecting an area belonging (on average) to the same efficiently connected cluster as their workplace.

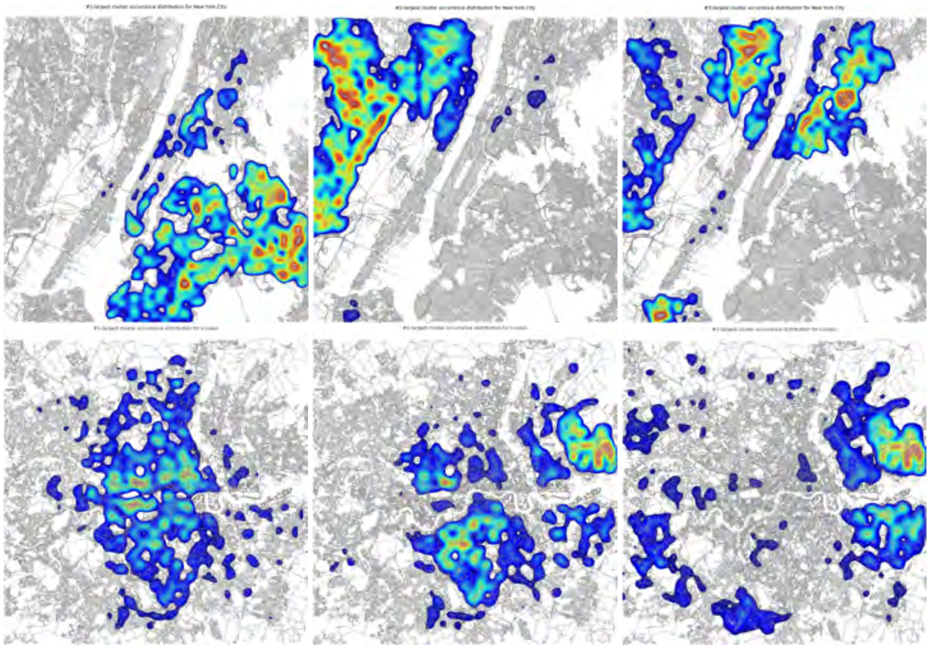


Fig. 2. From left to right: 1st, 2nd and 3rd largest cluster spatial distributions for NYC (top) and London (bottom). Blue (red) means that an edge was associated to the cluster in 80% (100%) of the replicas.

References

1. Helbing, Dirk. Traffic and related self-driven many-particle systems. Reviews of modern physics 73.4 (2001): 1067.
2. Li, D., Fu, B., Wang, Y., Lu, G., Berezin, Y., Stanley, H.E. and Havlin, S., 2015. Percolation transition in dynamical traffic network with evolving critical bottlenecks. Proceedings of the National Academy of Sciences, 112(3), pp.669-672.

3. Zeng, G., Li, D., Guo, S., Gao, L., Gao, Z., Stanley, H.E. and Havlin, S., 2019. Switch between critical percolation modes in city traffic dynamics. Proceedings of the National Academy of Sciences, 116(1), pp.23-28.
4. Haklay, M. and Weber, P., 2008. Openstreetmap: User-generated street maps. IEEE Pervasive Computing, 7(4), pp.12-18.
5. Rempe, F., Huber, G. and Bogenberger, K., 2016. Spatio-temporal congestion patterns in urban traffic networks. Transportation Research Procedia, 15, pp.513-524.
6. Shuman, D.I., Narang, S.K., Frossard, P., Ortega, A. and Vanderghenyst, P., 2013. The emerging filed of signal processing on graphs. IEEE Signal Processing Magazine.
7. Defferrard, M., Martin, L., Pena, R. and Perraudin, N., 2017. Pygsp: Graph signal processing in python. URL https://github. com/epfl-lts2/pygsp.
8. De Martino, A., Marsili, M. and Mulet, R., 2004. Adaptive drivers in a model of urban traffic. EPL (Europhysics Letters), 65(2), p.283.
9. Prakash, S., Havlin, S., Schwartz, M. and Stanley, H.E., 1992. Structural and dynamical properties of long-range correlated percolation. Physical Review A, 46(4), p.R1724.
10. Prakash, S., Havlin, S., Schwartz, M. and Stanley, H.E., 1992. Structural and dynamical properties of long-range correlated percolation. Physical Review A, 46(4), p.R1724.
11. Makse, H.A., Havlin, S., Schwartz, M. and Stanley, H.E., 1996. Method for generating long-range correlations for large systems. Physical Review E, 53(5), p.5445.
12. Campi, X. and Krivine, H., 2005. Zipf's law in multifragmentation. Physical Review C, 72(5), p.057602.

Passenger Delay and Topological Indicators in Public Transport Networks

Oded Cats[1][0000-0002-4506-0459] and Anne Mijntje Hijner [1]

[1] Department of Transport & Planning, Delft University of Technology, the Netherlands

o.cats@tudelft.nl

1 Introduction

To effectively improve service reliability, it is essential to understand how delays spill over the network. However, little is known about the properties of delay propagation in public transport networks as experienced by passengers. Past work on train delay propagation has approached the problem from the perspective of the operator of the network, rather than the passenger [1,2]. Past work on robustness analysis were performed based on an analytical or simulation-based public transport assignment model rather than using empirical passenger flow data, e.g. [3,4]. Few past studies investigated delay propagation in train networks from the passenger perspective, even though a shift towards interest in measuring passenger delays rather than vehicle delays has been observed [5]. However, such efforts were until recently limited but are now increasingly enabled by passively collected smart card data. There is lack of empirical knowledge on passenger delay properties and in particular the relations between delays occurring across the network and their relation to the underlying complex network structure. To this end, this study develops a passenger delay propagation model by estimating a Bayesian Network using extensive empirical data. In the context of train operations,

We adopt a data-driven method without making assumptions about the underlying phenomenon of passenger delay propagation in public transport networks. This allows us to potentially unravel unexpected relationships and consider all possible dependencies. Data on passenger delays at different stations can be used by the Bayesian Network method to determine the relationships with respect to delay between the different stations in the network. This would establish the extent to which a given station state in terms of the amount of passenger delay observed there can provide information on the state of other stations. These relations are a way of representing the propagation of passenger delay, as such relations will only exist if stations experience delays from the same cause and at approximately the same time, meaning those delays have propagated. Knowledge on the relations between stations can then be condensed into a set of indicators, hereafter called *informativity indicators*. These indicators quantify stations' capability of providing information on the delay state of the rest of the network.

2 Method

Our method consists of two key steps. First, based on the available data, a Bayesian Network (BN) is constructed. It is assumed that passenger delay attributed to each station is available as input to the process of constructing the BN. In this study, the mean passenger waiting time delay per station was available as input and was obtained from an estimation algorithm applied to individual passenger trajectories as detailed in [6]. Using the conditional probability tables, the arcs of the BN can be labelled, according to the strength of the dependence.

Second, the structure of the BN and the arc labels can then be used to calculate the informativity indicators for each node, which can be considered the second step in the process. We introduce a set of original indicators. Three different indicators are proposed as they could have different uses, and might lead to different observations: (i) *outgoing node degree* (d_n^+) - indicates how many nodes information can be provided on; (ii) *expected direct informativity* (e_n) - describes how informative this node is expected to be on any node it is connected to, and; (iii) *total informativity* (t_n^{min}, t_n^{max})- the upper and lower bounds of the total information a node can provide on the delay state of the rest of the network by integrating information from further descendants through higher order relations. This can be accounted for by multiplying the link weights along the path to a descendant. When multiple possible paths exist, we consider two options: the most informative path can be taken into account, or all paths can be taken into account as different paths can provide novel information.

3 Application and Results

For this study, a year's worth of train movement and passenger-train assignment output from the Washington DC metro system managed by the Washington DC Area Metro Transit Authority (WMATA) is available. This data was pre-processed to obtain estimated passenger delay per network element as detailed in [6]. The input available to this study consists of the estimated average initial and transfer passenger delay per station for 30 minutes time window throughout the analysis period.

The results of the Bayesian Network construction, displayed using a geographical map of the metro network, can be seen in Figure 1. Several observations can be made, namely that connections exist predominantly between nearby stations, and between nodes that are going into the same direction. This corroborates the underlying assumption in constructing BN for sectors. Notwithstanding, there are also a few connections over large distances. Furthermore, the results for different folds of the data set (the data was split into several training and test sets for a k-fold approach to calculate the errors of the model) showed that the mean percentage errors for all nodes were below 10% when compared for different partitioning of the dataset. We also examined the resulted attained for different folds and in all results similar observations were made regarding the types of dependencies found. namely that most dependencies occur between nearby stations and nodes going into the same direction.

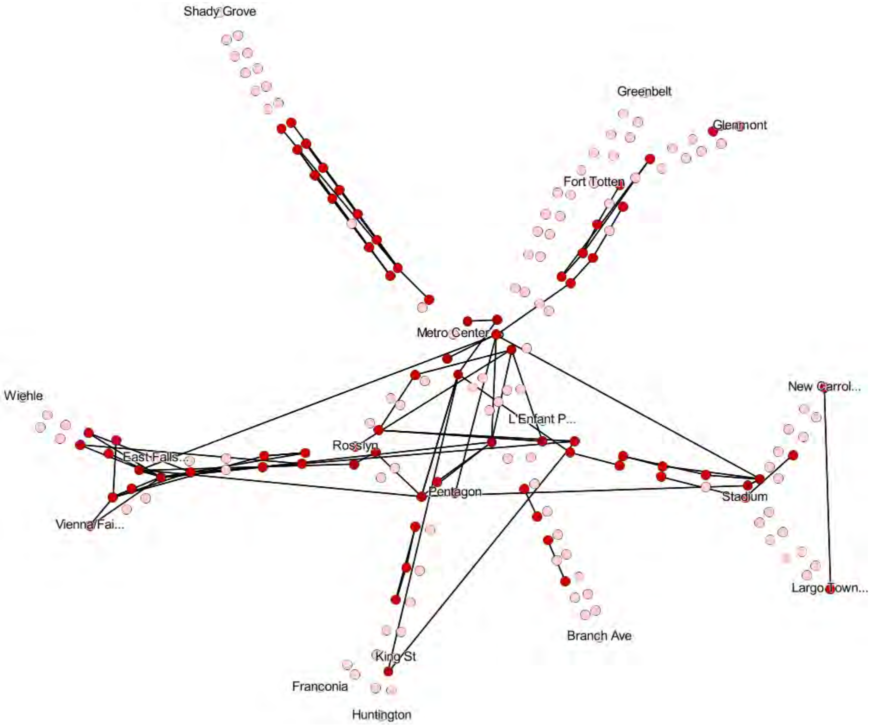


Fig. 1. The recombined BN mapped onto the geographical map of the stations. No distinction is made between transfer and initial stations in the presentation. Nodes that are not connected are shown in light pink, while nodes that are connected are dark red.

References

1. Berger, A., Gebhardt, A., Müller-Hannemann, M., and Ostrowski, M.Stochastic delay prediction in large train networks. In OASIcs-OpenAccess Series in Informatics, volume 20. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2011).
2. Kirchhoff, F. and Kolonko, M.: Modelling delay propagation in railway networks using closed family of distributions. Technical Report (2015).
3. Cats O., Koppenol G-J. and Warnier M.: Robustness assessment of link capacity reduction for complex networks: Application for public transport systems. Reliability Engineering & System Safety, 167, 544-553 (2017).
4. Malandri, C., Fonzone, A., and Cats, O.: Recovery time and propagation effects of passenger transport disruptions. Physica A, 505, 7–17 (2018).
5. Nielsen, O. A., Landex, O., and Frederiksen, R. D. Passenger delay models for rail networks. In Schedule-Based Modeling of Transportation Networks, pages 1–23. Springer (2019).
6. Krishnakumari P., Cats O. and van Lint H. Estimation of network passenger delay form individual trajectories. Transportation Research Part C, 117, 102704 (2019).

The role of geography in the complex diffusion of innovations

Balazs Lengyel[1], Eszter Bokanyi[2] Riccardo Di Clemente[3], Janos Kertesz[4], and Marta Gonzalez[5]

[1] ANET Lab, Centre for Economic and Regional Studies, Budapest 1097, Hungary,
lengyel.balazs@krtk.mta.hu,
home page: anet.krtk.mta.hu
[2] Corvinus University of Budapest, Institute of Advanced Studies, Budapest 1093, Hungary
[3] University of Exeter, Computer Science Department, Exeter, EX4 4QF, United Kingdom
[4] Central European University, Department of Network and Data Science, Wien, Austria
[5] University of California at Berkeley, Department of City and Regional Planning, Berkeley
CA, 94720, USA

1 Introduction

Collective behavior, such as massive adoption of new technologies is a complex social contagion phenomenon [1]. Individuals are influenced both by media and by their social ties in their decision-making. This feature was first modelled in the 1960s with the Bass model of innovation diffusion [2]. The model distinguishes between exogenous and peers' influence and reproduces the observation that few early adopters are followed by a much larger number of early and late majority adopters, and finally, by few laggards [3].

Only in the past two decades, the importance of the social network structure has become increasingly clear in the mechanism of peers' influence. In spreading phenomena, individuals perform a certain action only when a sufficiently large fraction of their network contacts have performed it before [4]. Complex contagion models, in which adoption depends on the ratio of the adopting neighbors, often referred to as adoption threshold [1], have been efficiently applied to characterize the diffusion of online innovations [5]. In order to incorporate the role of social networks in technology adoption, the Bass model has been implemented through an agent-based model (ABM) version [6]. This approach is similar to other network diffusion approaches regarding the increasing pressure on the individual to adopt as network neighbors adopt; however, spontaneous adoption is also possible in the Bass ABM. Nevertheless, understanding how physical geography affects social contagion dynamics is still lacking

2 Results

In this paper, we analyze the adoption dynamics of iWiW, a social media platform that used to be popular in Hungary, over its full life cycle (2002-2012). This unique dataset allows us to investigate two major geographical features that characterize spatial contagion dynamics: town size described by the urban scaling law [7] and distance decay described by the gravity law [8].

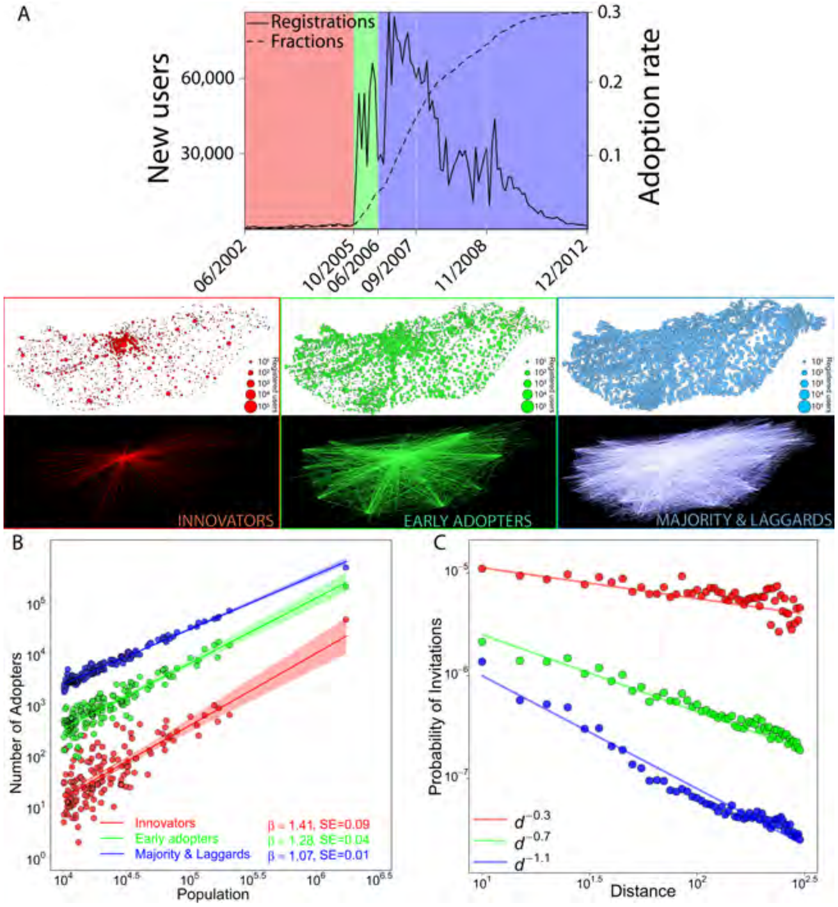


Fig. 1. Spatial diffusion over the OSN life-cycle. A. *Top:* Number of new users and the cumulative fraction of registered individuals among total population over the OSN life-cycle. *White background maps:* Coloured dots depict towns; their size represent the number of adopters over the corresponding period. *Black background maps:* Links depict the number of invitations sent between towns over the corresponding periods. **B.** Adoption scales super-linearly with town population. **C.** The Probability of Invitations to distant locations is relatively high in the Inventors stage but decreased over the product life-cycle while diffusion became more local.

We find empirical evidence (Figure 1) that early adoption is concentrated in large towns and scales super-linearly with town population but late adoption is less concentrated. Diffusion starts across distant big cities such that distance decay of spread is slight and becomes more local over time as adoption reaches small towns in later stages when distance decay becomes strong.

To better understand the spatial characteristics of complex contagion in social networks, we develop a Bass ABM of new technology's adoption on a sample of the empirical network preserving the community structure and geographical features of connections within and across towns. The data allows us to measure individual adoption thresholds that we can use to parameterize the likelihood of adoption at given fractions of infected social connections. We compare how the ABM and the Bass differential equation (DE) model fit to the empirical urban scaling and distance decay characteristics. Finally, we evaluate model accuracy in predicting the time of local adoption peaks and assess the bias induced by local network structures, or geographical features of towns. These analyses enable us to evaluate the role of geography in complex contagion models at local scales.

We find that the scaling of the number of earliest adopters with town population is best reflected by the ABM when threshold parameters are incorporated. None of our models can reproduce the high probability of diffusion across distant peers in the early stages of the life-cycle. Certain features of the network within towns - eg. high network density and transitivity - accelerate the ABM diffusion and make predictions of adoption peaks early, which can be overcome when controlling for threshold distributions. Meanwhile, other features of the network - eg. modularity and average path length - delay the prediction of adoption peaks, and cannot be eliminated with the threshold control. Nonetheless, we assess that contagion models cannot cure the bias of physical geography, such as distance from the innovation origin and town size, on the predictions of adoption peaks.

The threshold mechanisms introduced to the Bass ABM allow us to reproduce aggregated effects in relation to the number of adopters per population size. However, as expected, it is hard to predict the location of the social ties when an adoption occurs. This is in turn, affects the prediction of when the different towns reach their tipping point. Unfolding these aforementioned empirical features, we were able to capture the limitations of the standard model of complex contagion in predicting adoption at local scales and to describe key elements of diffusion in geographical space through the contact of local and distant peers.

References

1. Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.
2. Frank M Bass. A new product growth for model consumer durables. *Management science*, 15(5):215–227, 1969.
3. Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.
4. Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.
5. Márton Karsai, Gerardo Iñiguez, Riivo Kikas, Kimmo Kaski, and János Kertész. Local cascades induced global contagion: How heterogeneous thresholds, exogenous effects, and unconcerned behaviour govern online adoption spreading. *Scientific reports*, 6, 2016.
6. William Rand and Roland T Rust. Agent-based modeling in marketing: Guidelines for rigor. *International Journal of Research in Marketing*, 28(3):181–193, 2011.
7. Luís MA Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences*, 104(17):7301–7306, 2007.
8. David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

Characterising road networks through subgraph graphlet analysis

Andrew Elliott[1]*, Stephen Law[24]*, and Luis Ospina-Forero[3]*

[1] School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ
[2] UCL Department of Geography, University College London, London, WC1E 6BT
[3] Alliance Manchester Business School, The University of Manchester, Manchester, M15 6PB
[4] The Alan Turing Institute
*All authors contributed equally to this work.

1 Introduction

Cities are artefacts of many human interventions over long periods of time, that can either come from top-down planning where the placement of every road, junction and park are dictated, or grown organically from the bottom-up where places evolve and change by adding a street, a square or a bridge where required. In this work we explore, to what extent we can capture these differences leveraging works from transport geography (walkability), network science (graphlet counts) and complexity (entropy).

Until fairly recently, research on urban street networks, a class of planar networks embedded in two dimensional Euclidean space, had largely been studied within the confined domains of architecture and transport geography [1, 5, 6]. Due to the growing availability of network databases and computation resources, there is a growing interest in network science in addressing the further characterization of urban street networks quantitatively. For example, [4] uses the number of dead-ends and unfinished crossings as an indicator to discriminate whether a city is planned or less planned. In another example, [3] characterised urban street patterns using both local and global network properties. Locally, the authors of [3] proposed a meshedness coefficient, which takes the ratio between the number of faces relative to its maximal number in a planar graph where grid-like and organic cities both exhibited high levels of meshedness in comparison to the more tree-liked counterparts. Globally, the authors found that the relative cost and efficiency of a street network have certain capacity to characterise cities from different periods. Following these earlier works, [2] was able to characterised a hundred cities around the world based on a measure of entropy using the orientation of the street network. He found that grid-like North American cities exhibited lower angular orientation entropy.

A limitation of the approach is that the orientation of a regular grid need not have a constant angular orientation. A case in point is Pittsburgh in the US, which has a regular grid system with different orientations. Some additional limitations remain. For example, recent research did not characterise cities using network statistics at a meso level, for example, based on a walking region in capturing how humans experience the city as inspired from the urban planning literature [**?**]. Furthermore, recent research also did not consider the appearance of more complex sub-structure configurations such as graphlet counts. This research could be important because the structural features of street networks, such as grids or cul-de-sacs, are more common and repeated in modern planned cities as opposed to cities that are less planned and have evolved over longer periods of time.

Our Approach To alleviate these limitations, we propose a new method to characterise the plannedness of a city. We note that planned cities may have repeated motifs or structures, and we further hypothese that these structures should occur at the scale that humans interact with a city, i.e. walking distances. Thus, we propose a set of novel meso level sub-structure street network statistics to capture the plannedness of cities based on graphlet counts (*4-nodes*) within a walkable region of each road (400m, and 800m equivalent to 5 and 10 minutes walking time) which we term a *walkinghood*. More specifically, for each *walkinghood* we count the occurances of each graphlet, and then we compute the entropy over the set of *walkinghood*s to give a measure of regularity. We compute the entropy independently for each type of graphlet and sum to obtain an overall measure of regularity between the *walkinghood*s thus helping us to characterise the plannedness, or cities or indeed parts of cities. We display our pipeline schematically in Fig. 1.

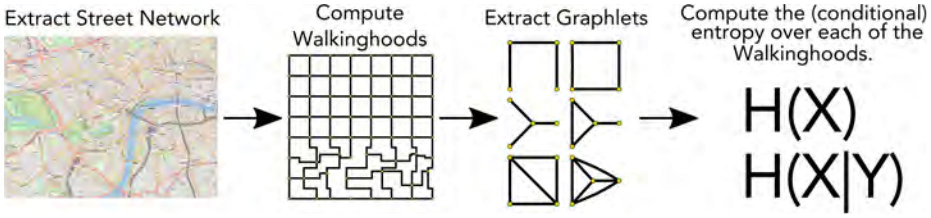


Fig. 1. Street network analysis pipeline. (image taken from OpenStreetMap under CC BY-SA see [7]).

2 Results

We first explore the performance of our method using a novel one parameter synthetic model which varies from highly planned to highly organic, and test how well our methods can uncover the underlying nature.

Second, we test the proposed measure with real data from OpenStreetMap [7]. We test how well our method can distingulish between the highly planned new towns in the United Kingdom and older less planned towns, and we compare to other several other approaches from the literature.

Finally, we explore the ability of our method to uncover similarities and differences between different areas of the same city, in this case the Boroughs of London, the Arrondissements of Paris and the Special Wards of Tokyo. Further research on comparison with other network statistics and on street networks across time is necessary to validate this exploratory research.

Summary. We construct a novel method to characterise the plannedness of a city based on walking regions, graphlet counts and entropy. To explore our novel measure we construct a novel one parameter synthetic model, and we further explore our measure on a real world dataset of very planned new towns and older less planned towns and cities.

References

1. Barthelemy, M.: Morphogenesis of Spatial Networks. Springer International Publishing (2018)
2. Boeing, G.: Urban Spatial Order: Street Network Orientation, Configuration, and Entropy. arXiv e-prints (Aug 2018)
3. Cardillo, A., Scellato, S., Latora, V., Porta, S.: Structural properties of planar graphs of urban street patterns. Phys. Rev. E 73, 066107 (Jun 2006)
4. Courtat, T., Gloaguen, C., Douady, S.: Mathematics and morphogenesis of cities: A geometrical approach. Phys. Rev. E 83, 036106 (Mar 2011)
5. Haggett, P., Chorley, R.J.: Network analysis in geography (1969)
6. Hillier, B., Hanson, J.: The social logic of space (1984)
7. Open Street Maps: https://www.OpenStreetMap.com/ (2018)

The effect of commuting on the structure and assortativity of online social ties

Eszter Bokányi*,[1,2], Sándor Juhász[1,2], Márton Karsai[3], and Balázs Lengyel[1,2]

[1] NETI Lab, Corvinus University of Budapest
1093 Budapest, Fővám tér 8., Hungary
[2] ANET Lab, Center for Economic and Regional Sciences
1097 Budapest, Tóth Kálmán u. 4., Hungary
[3] Department of Network and Data Science, Central European University
1100 Vienna, Quellenstraße 51-55., Austria
*eszter.bokanyi@uni-corvinus.hu

1 Introduction

There is an extensive literature studying and modeling human mobility at large scales. Models such as the gravity law [1] or the radiation law [2] have been successful at explaining mobility patterns across cities, administrative regions or countries. However, investigating human mobility inside urban areas is more challenging, as granular enough data is often proprietary or lacking. Moreover, the micro-geography of social networks inside cities is also less understood. Previous works suggest that distance or other geographical obstacles inside and across cities matter for network tie formation [3, 4], and that urban mobility influences contacts [5, 6].

Apart from the geographical constraints, assortativity is also a key property of both urban mobility and social networks. Driven by homophily, people having similar socio-economic status are more prone to maintaining a mutual social relationship [7], which in the end intensifies already existing segregation or opinion polarization patterns [8]. Thus, it is important to understand the extent of homophily in these relationships, and to characterize how physical mobility reinforces or reduces the homophilic effect.

In this work, we investigate geolocated urban social networks combined with mobility information for almost 1 million users. Our dataset is unique in its spatial resolution, in containing a rich social network structure, and in its large spatial coverage. Our findings show that commuting to distant locations increases the number of connections people can develop and acts against closed, highly clustered social ties. However, commuting to a distant workplace does not change the pattern that people most likely develop connections towards others from similar socio-econonomic background.

2 Results

We construct an aggregated daily timeline for the three most frequently visited locations of users posting geolocated messages on the online social networking platform Twitter. By identifying clusters of tweets where the time of tweeting is predominantly during or out of working hours, we were able to assign a home and work cluster to almost

1 million users in the top 50 US metro areas. Moreover, we also constructed their ego networks based on mutual followership on the platform. We connected each user's home and work location to census tracts and included annual household income information from the American Community Survey.

Additionally, we created income deciles for every metropolitan area, and sorted census tracts into 10 different income classes following these deciles. Home location of each user was connected to the income classes of their metro areas. Furthermore, we calculated the probability p_{ij} that a user from income decile j commutes to a tract, or forms a mutual follower relationship to a user in a tract with a particular income decile i. Thus, we get 10x10 commuting (C_{ij}) and friendship (F_{ij}) assortativity matrices for the top 50 metropolitan areas of the US.

First, we find that the commuting assortativity matrix C_{ij} shows less homophily than the friendship assortativity matrix F_{ij} for almost every metropolitan area. It means that even though physical mobility might connect people with different backgrounds within a city, this does not necessarily enforce the creation of ties across different income classes.

Second, we see that commuting distance affects the ability to create social ties in the income space. We divide our users into groups that commute more of less than 5 km inside cities. Figure 1 shows these restricted friendship assortativity matrices for 3 selected cities. As rows 3 and 4 show, the matrices mostly differ in the diagonal area, namely that people commuting less tend to form more assortative relationships. Because in most cities, the annual income values in neighboring census tracts are correlated, less commuting leads to more enclosed social relationships and more social segregation in our dataset.

To translate our findings to an individual level, we show that the ego networks of these two groups of people are also different in almost every investigated city. We observe that people commuting more also have more ties (have a higher degree), a less closed ego network structure (lower local clustering coefficient), and the difference to their friends' home income is higher. This later result suggests that they have a topologically and socio-economically more diverse network. The generality of the above trends across metro areas are also supported by regression models.

Our results suggest that even though commuting has a heterogeneizing effect on human relationships within a metropolitan area, a strong socio-economic homophily still exists in tie formation. Moreover, the social networks of users commuting more differs at the micro-level from those being spatially more confined. These observations also provide new insights for urban inequality or opinion divergence research.

References

1. D Liben-Nowell et al. Geographic routing in social networks. *PNAS*, 102(33):11623–11628, Aug 2005.
2. F Simini et al. A universal model for mobility and migration patterns. *Nature*, 484(7392):8–12, 2012.
3. N Eagle, AS Pentland, and D Lazer. Inferring friendship network structure by using mobile phone data. *PNAS*, 106(36):15274–15278, 2009.

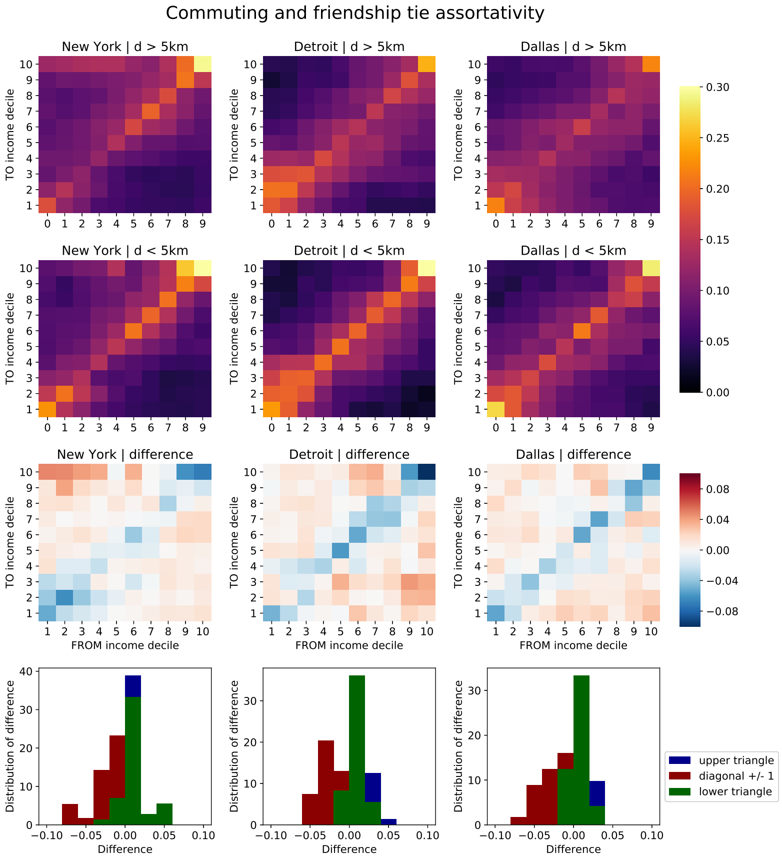


Fig. 1. Income assortativity matrices of users commuting more and less than 5 km. In rows 1 and 2, color represents the probability of a mutual followership, given that the ego is in the income decile given by the horizontal axis. Row 3 shows the difference in the probabilities of the two groups, row 4 is the distribution of the differences in the diagonal, upper and lower triangle areas of row 3.

4. F Calabrese, J Reades, and C Ratti. Eigenplaces : Segmenting Space through Digital Signatures. *Pervasive Computing, IEEE*, 9(1), 2010.
5. M Bailey et al. Social connectedness in urban areas. *Journal of Urban Economics*, 118, 2020.
6. AJ Morales et al. Segregation and polarization in urban areas. *RSOS*, 6(10), 2019.
7. Y Leo et al. Socioeconomic correlations and stratification in social-communication networks. *Journal of the Royal Society Interface*, 13(125), 2016.
8. M Bailey et al. Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives* 32(3):259–280, 2018.

Modeling the urban network

Romain Pousse, Claire Lagesse and Stéphane Douady

Laboratory MSC, Matter and Complex Systems, University Paris Diderot, Paris, France
romain.pousse@gmail.com

1 Introduction

Understanding the development of the urban fabric originally belongs to the human and social sciences (architecture, geography, history, town planning). Our aim is to study city growth through its roads networks with complex system science and graph theory. To do so, the first step is to reduce the road network to a geometrical graph (intersections as nodes and street segments, or arcs, as edges connecting them). We seek to find the existing, intuitive [4], roads through a geometric approach[2, 6] by creating the *way*, a continuous set of arcs and nodes producing a multi-scale object going from the dead end to the main avenues crossing the city. By applying classical graph theory indicators (closeness, betweeness...) and others (orthogonality, accessibility...) on this object, we obtain information on the structure and organization of this network (figure 1), without being constrained by the limits of the analyzed graph as opposed to a calculation performed only on arcs [5]. The meaningfulness of the results supports this approach and hints at the fact that the turns, defining a topological distance, are more meaningful than the metric distance along the network.

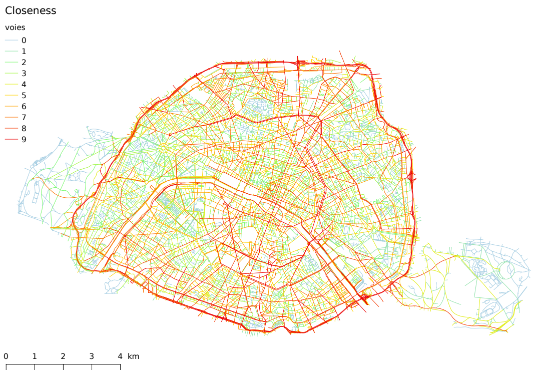
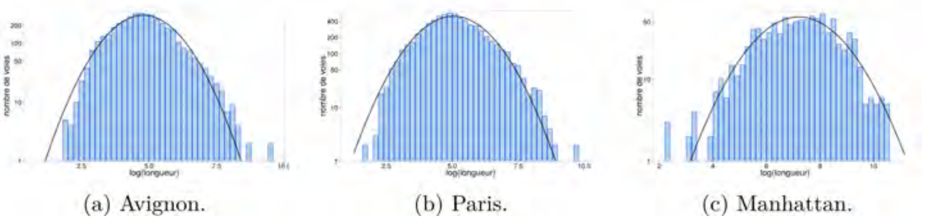


Fig. 1. Closeness indicator applied to the network of reconstructed ways of the city of Paris. This indicator reveals in parts the history of this network: The oldest ways and main piercings have a high score (in red), they are the most accessible in a small number of turn for the whole. network.

The distributions of indicators such as length, closeness or degree of ways each reveal patterns similar across the studied cities studied (log-normal distribution for length, see fig. 2, normal distribution for closeness, power distribution for degree). The question is to understand the appearance of such particular shapes of distribution for one of these indicators in order to better understand the dynamics of construction of such networks. We have chosen to look at the *length* of the *ways*. Indeed, contrary to closeness, it is a more accessible data, its calculation does not need to be integrated into the whole graph, it is also an indicator less studied in graph theory compared to the degree, which power laws can already be reproduced by several type of models (small world or citations[3]). It also presents a more original distribution (fig. 2). We seek to build a model that runs through all the processes of creating ways [1] and that can generate length distributions close to what is observed.



(a) Avignon. (b) Paris. (c) Manhattan.

Fig. 2. Ways' length distribution (log-log scale) for cities of Avignon, Paris and Manhattan island. We find a parabolic form (log-normal distribution in linear scale) for the three cities.

2 Results

Our models are based on the hypothesis that creating a ways by perpendicular cutting of parcels could explain this type of distribution. Each new way cuts a parcel into two smaller areas. This process is justified by the observations made in cities (the grid in the oldest urban areas in particular, like Saint-Michel or Chatelet district in Paris), and physical network like clay cracks - we measured for the cracks also a length distribution close to a log-normal. We rely on models with similar dynamics in order to find this statistic, just varying the criteria for the selection of next parcel to cut. Several processes have been carried out, dividing the plots according to different characteristics (fig. 3): the longest side, the density of intersections in the vicinity of each plot or the number of turns to be made to access them (the integrated topological distance over the entire network).

The results showed that our first hypothesis, a cut process to depend of the longest side is not enough to explain the log-normal distribution. With the density of crossroads in the proximity of each parcel, we obtain forms of networks similar to the cities (polar form) but the length distribution is still different. By substituting this density by the topological distance, we find a distribution closer to what is observed, validating the importance of this parameter in the development of roads networks.

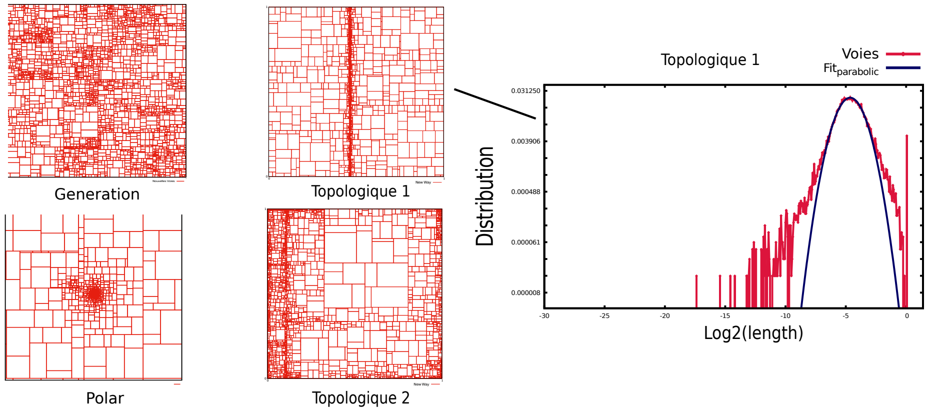


Fig. 3. Four models of ways network creations are represented. The lane creation processes depend on the density of intersections in the vicinity of each parcel as well as on its longer side (polar model), the topological distance (topological 2) which can also be coupled to the longer side (topological 1). Finally, we have also developed a model without selection process cutting all the plots at each iteration (Generation). On these 4 models, only the polar model giving more realistic network shapes is far from a log-normal distribution. For the three others, we obtain distributions as illustrated on the right for the topological model 1, distribution (in log-log scale) and parabolic fit are superposed on a large part around their maximum.

References

1. Courtat, T., Gloaguen, C., Douady, S.: Mathematics and morphogenesis of cities : A geometrical approach. Physical Review E 83, 036106 (2010)
2. Courtat, T., Gloaguen, C., Douady, S.: Hypergraphs and City Street Networks. Arxiv preprint, 1106.0297
3. de Solla Price, D. J.: Networks of Scientific Papers. Science. 149 (3683): 510–515 (1965)
4. Hillier, B. and Hanson, J.: The Social Logic of Space, Cambridge University Press, Cambridge (1984)
5. Lagesse, C., Bordin, P., Douady, S.: A spatial multi-scale object to analyze road networks. Network Science, 3(01), 156-181 (2015)
6. Perna A., Kuntz P., Douady S.: Characterization of spatial network like patterns from junction geometry. Physical Review E, 83(6), 066106 (2011)

9 782955 705049