
FRCCS 2024
Proceedings
French Regional
Conference on
Complex Systems
Montpellier, France
29-31 May

Contents

Foreword

Welcome to the French Regional Conference on Complex Systems (FRCCS 2024). It is our great honor to present these conference proceedings, which showcase the pioneering work of researchers dedicated to the study of complex systems.

This year's conference features an exceptional lineup of invited speakers, each a leader in their respective fields. Petter Holme from Aalto University, Finland, contributes his extensive expertise in network science and the dynamics of complex networks. Sonia Kéfi from Université de Montpellier, France, offers her profound insights into ecological networks and ecosystem resilience. Natasa Przulj, representing both the Catalan Institution for Research and Advanced Studies (ICREA) in Barcelona and the Supercomputing Center at University College London, brings her renowned research in computational biology and integrative data analysis. Boleslaw K. Szymanski from the Network Science and Technology Center at Rensselaer Polytechnic Institute shares his influential work in social networks and algorithmic solutions. Lastly, Ingmar Weber from Saarland University, Germany, presents his cutting-edge research on computational social science and digital behavior analytics.

The city of Montpellier, with its rich heritage and dynamic cultural scene, serves as an ideal location for FRCCS 2024. Known for its picturesque medieval streets, vibrant university atmosphere, and proximity to the Mediterranean, Montpellier offers a unique blend of history and modernity. The surrounding region, famous for its beautiful wetlands, is home to the charming flamingos—though we assure you, no flamingos have been enlisted for conference duties!

We express our sincere gratitude to the University of Montpellier for hosting this event, providing us with outstanding facilities and support. Our heartfelt thanks go to our esteemed publishers, Springer Nature and Frontiers, for their invaluable support in disseminating high-quality scientific knowledge. We also thank the EGC association, CIRAD and INRAE Research Organisations, Paul Valéry University, and the LIRMM Laboratory for their generous sponsorship and contributions.

The abstracts within this book reflect the exceptional quality and diversity of research presented at FRCCS 2024. Each contribution, rigorously reviewed, promises to foster insightful discussions, inspire new ideas, and promote collaborations among researchers worldwide.

Thank you for joining us in Montpellier. We wish you an enriching and enjoyable conference experience.

Sincerely,

Bruno Pinaud and Roberto Interdonato
The General Chairs
French Regional Conference on Complex Systems (FRCCS 2024)

Committes

Advisory Board CSS France

- Chantal Cherifi, *DISP, Lyon*
- Hocine Cherifi, *ICB, Dijon*
- Christophe Cruz, *ICB, Dijon*
- Roberto Interdonato, *CIRAD, Montpellier*
- Hamamache Kheddouci, *LIRIS, Lyon*
- Benjamin Renoust, *Median Technologies, Sophia Antipolis*

General Chairs

- Bruno Pinaud, *Université de Bordeaux, Bordeaux*
- Roberto Interdonato, *CIRAD, Montpellier*

Program Chairs

- Pascal Poncelet, *LIRMM, Université de Montpellier*
- Hocine Cherifi, *Université de Bourgogne, Dijon*

Poster Chairs

- Nicolas Dugué, *Université Le Mans*
- Maxime Lenormand, *INRAE , UMR TETIS*

Workshop Chairs

- Maria Malek, *LARIS, Cergy-Pontoise*
- Sophie Lèbre, *Université Paul Valéry Montpellier 3*

Publication Chairs

- Yoann Pigné, *Université Le Havre Normandie*

Finance Chair

- Benjamin Renoust, *Median Technologies, Sophia Antipolis*

Web Chair

- Christophe Cruz, *ICB, Dijon*

Sponsor Chair

- Lylia Abrouk, *Université de Bourgogne, Dijon*

Publicity Chairr

- Amira Mouaker, *Université de Perpignan, Perpignan*

Local Committee

- Elena Demchenko, *LIRMM, Université de Montpellier*
- Virginie Feche, *LIRMM, Université de Montpellier*
- Dino Ienco, *INRAE, UMR TETIS*
- Mathieu Roche, *CIRAD, UMR TETIS*
- Maguelonne Teisseire, *CIRAD, UMR TETIS*

Invited Speakers

Petter Holme
Aalto University, Finland

Petter Holme is a Professor of Network Science at the Department of Computer Science, Aalto University, Finland, and a Research Fellow at The Center of Computational Social Science, Kobe University, Japan. Formerly, Holme held faculty positions at the Tokyo Institute of Technology, Japan, Sungkyunkwan University, Korea, and Umeå University and the Royal Institute of Technology, Sweden. His research covers a broad scientific ground in the borderland between the social and formal sciences. Among many other things, Holme pioneered the study of temporal networks.

Keynote: Understanding the world from structures in time

Just like the graph structure of social networks can tell us much about social organizing and dynamic social processes, so can structures in the times when things happen. This also generalizes to other types of networks, so please replace « social » with your favorite topic. I will review the last two decades' research on temporal networks of human interaction and relate it to earlier thoughts about temporal structure in the social sciences and beyond. I will discuss why it is so hard to generalize concepts from static network analysis to temporal networks and the grand challenges of the future in this area. This topic also happens to be a point where many philosophical topics collide: the nature of time, the form of scientific explanations of complex systems, structuralism vs. universality, etc., so time permitting, I will also discuss those topics.

Sonia Kéfi
Université de Montpellier, France

I am a researcher at the CNRS based in the BioDICée team at the Institut des Sciences de l'Evolution de Montpellier (ISEM), France.

In an era of global change, my research aims at understanding how ecosystems persist and change under pressures from changing climate and land use. What makes ecosystems resilient to changes and what makes them fragile?

I combine mathematical modeling and data analysis to investigate the role of ecological interactions (in particular facilitation) in stabilizing and destabilizing ecosystems, but also to develop indicators of resilience that could warn us of approaching ecosystem shifts.

Keynote: The stability and resilience of ecological systems

Understanding the stability of ecological communities is a matter of increasing importance in the context of global environmental change. Yet it has proved to be a challenging task. Different metrics are used to assess the stability of ecological systems, and the choice of one metric over another may result in conflicting conclusions. While the need to consider this multitude of stability metrics has been clearly stated in the ecological literature for decades, little is known about how different stability metrics relate to each other. I'll present results of dynamical simulations of ecological communities investigating the correlations between frequently used stability metrics, and I will discuss how these results may contribute to make progress in the quantification of stability in theory and in practice.

Natasa Przulj
Catalan Institution for Research and Advanced Studies (ICREA), Barcelona
Supercomputing Center, University College London

Academician Professor Doctor Natasa Przulj holds a prestigious Catalan Institution for Research and Advanced Studies (ICREA) Research Professorship at Barcelona Supercomputing Center and is also a Full Professor of Computer Science at University College London.

She is a leader in network science and Artificial Intelligence (AI) algorithms for biomedical data analysis and fusion applied to precision medicine. She published 86 peer-reviewed journal and 18 peer-reviewed conference papers in the most prestigious venues, including four in Science, also 13 peer-reviewed book chapters and 2 books. Her research has been cited over 13,000 times, h-index=50, i10-index=78 (Google Scholar), and supported by over €25 million in competitive research funding. Notably, she received three prestigious, single PI, European Research Council (ERC) grants: Consolidator (2018-2025), Proof of Concept (2020-2023) and Starting (2012-2017). She has been elected into several academies: The European Laboratory for Learning and Intelligent Systems – ELLIS, in 2022; The Serbian Royal Academy of Scientists and Artists (SKANU), in 2019; Academia Europaea, The Academy of Europe, in 2017; and Fellow of the British Computer Society (BCS) Academy of Computing, in 2013. In 2014, she received the BCS Roger Needham Award, sponsored by Microsoft Research, in recognition of the potential her research has to revolutionize health and pharmaceutics. She obtained a PhD in Computer Science from the University of Toronto in 2005.

Acad. Prof. Dr. Przulj initiated utilization of non-negative matrix tri-factorization based AI / machine learning (ML) methodologies for fusion of heterogeneous, systems-level, molecular (multi-omics) networked data (the subject of her ongoing ERC Consolidator Grant) to aid to the development of personalized, or precision medicine. In addition, she initiated extraction of biomedical knowledge from the wiring patterns (topology) of omics network data to complement the genetic sequences as a source of new biomedical information (subject of her ERC Starting Grant). She is best known for introducing graphlets in 2004, a methodology now widely utilized to produce feature vectors capturing network topology, that are used as input into many AI/ML algorithms for network data analytics in various domains; graphlets are subject of around 21,400 research papers and 300 patents according to Google Scholar.

Keynote: Multi-omics network data fusion for enabling precision medicine

Increasing quantities of heterogeneous, interconnected, systems-level, molecular (multi-omic) network data are becoming available. They provide complementary information about cells, tissues and diseases. We need to utilize them to better stratify patients into risk groups, discover new biomarkers, re-purpose known and discover new drugs to personalize medical treatment. This is nontrivial, because of computational intractability of many underlying problems on large graphs (networks), necessitating the development of algorithms for finding approximate solutions (heuristics). We develop a versatile data fusion (integration) machine learning (ML) framework that utilizes the state-of-the-art network science methods to address key challenges

in precision medicine from these multimodal network data: better stratification of patients, prediction of biomarkers, and re-purposing of approved drugs to particular patient groups, applied to cancers, Covid-19, Parkinson's and other diseases. Our new methods stem from graph-regularized non-negative matrix tri-factorization (NMTF), a machine learning technique for dimensionality reduction, inference and co-clustering of heterogeneous datasets, coupled with novel network science algorithms. We utilize our new framework to develop methodologies for improving the understanding the molecular organization and diseases from the omics network embedding spaces.

Boleslaw K. Szymanski
Network Science and Technology Center, Rensselaer Polytechnic Institute

Dr. Boleslaw K. Szymanski is the Claire and Roland Schmitt Distinguished Professor, was the Director of the ARL Social and Cognitive Networks Academic Research Center and is the Director of the Rensselaer Network Science and Technology (NeST) Center. He received his Ph.D. in Computer Science from Institute of Informatics of National Academy of Science in Warsaw, Poland, in 1976. He published over 600 scientific articles, is a foreign member of the National Academy of Science in Poland and an IEEE Fellow and was a National Lecturer for the ACM. In 2009, he received the Wilkes Medal of British Computer Society, in 2003, William H. Wiley 1866 Distinguished Faculty Award from RPI and Service Award from Network Science Society in 2021. His current research interests focus on computer networks and technology-based social networks.

Keynote: Political polarization at the the age of social media

By now, it is common knowledge that social media has changed the way information spreads around the internet, but there is paucity of research on how exactly this new spread works. In this talk, we will start by discussing the new patterns of data flow and new roles for users in spreading information that coexist with remnants of the classic two-level propagation. Then, using statistical physics models, we discuss how the presence of social media increases polarization. The model reveals asymmetric hysteresis trajectories with tipping points that are hard to predict and that make polarization extremely difficult to reverse once the level exceeds a critical value. Political scientists have documented increasing partisan division, finding extremist positions to be more pronounced among political elites than among voters, raising the question of how polarization might be attenuated. In this talk, we introduce a general model of opinion change to see if the self-reinforcing dynamics of influence and homophily may enable tipping points that make reversibility problematic. The model applies to a legislative body or other small, densely connected organization, but does not assume country-specific institutional arrangements that would obscure the identification of fundamental regularities in the phase transitions. We also introduced exogenous shocks corresponding to events that create a shared interest against a common threat (e.g., a global pandemic). Phase diagrams of political polarization reveal difficult-to-predict transitions that can be irreversible due to asymmetric hysteresis trajectories. We focus on social media, which has been transforming political communication dynamics for over a decade. Using a billion tweets, we analyzed the change in Twitter's news media landscape between the 2016 and 2020 U.S. Presidential elections. We then identify influencers, users with the top ability to spread news in the Twitter network. The more influential 2016 users were, the higher was their rate of remaining active and keeping their level of influence in 2020. We also analyze changes in influencers' real-world affiliations, political biases, and in Twitter users' choices as to which influencers to retweet and which ideology to subsequently support. Despite the noted decrease in extremely biased content and fake news on Twitter, these results show increasing echo chamber behaviors and latent ideological polarization across the two elections at the user and influencer levels.

Ingmar Weber
Saarland University, Germany

Ingmar Weber is an Alexander von Humboldt Professor for AI and holds Chair for Societal Computing at Saarland University. His interdisciplinary research combines (i) computing of society, i.e., using non-traditional data sources and computational approaches to measure and understand societal phenomena, and (ii) computing for society, i.e., working with non-profit stakeholders to use technology to strengthen social development.

Keynote: From Screen to Sky: Monitoring Migration and Mobility Using Innovative Data Sources

What can advertising data tell us about cross-border mobility? And how can satellite imagery be used to monitor displacement during periods of war? In this talk, I'll present work from the last 10+ years on using innovative data sources to help monitor migration and mobility, in particular during humanitarian crises. First, I'll show how so-called "audience estimates" from Facebook's advertising platform can be used to nowcast cross-border migration, before extending this work to monitor country-internal displacement. In the second part, I'll show how different types of satellite imagery can be used to pick up on a particular signature of mobility: shifts in the geographic distribution of cars. The presented research is joint work with colleagues at the Qatar Computing Research, the Max-Planck Institute for Demographic Research, the University of Oxford, UNICEF Innovation, and others.

Epidemics, Rumors

FRCCS

Exploring Epidemiological Dynamics in a Social Dilemma

Francesco Bertolotti[1✓], Niccolò Kaddera[1], Luca Pasquino[1] and Luca Mari[1]

[1] *School of Industrial Engineering, LIUC Università Cattaneo, Corso G. Matteotti 22, Castellanza (VA), Italy ; fbertolotti@liuc.it, ni26.kadera@stud.liuc.it, lu28.pasquino@stud.liuc.it, lmari@liuc.it.*

[✓] *Presenting author*

Abstract. This paper presents a novel epidemiological extension of the El Farol Bar problem, utilizing an agent-based modeling simulation technique and exploring the interplay between social decision-making and epidemiological dynamics. The model simulates individual agents making binary decisions—to visit a bar or stay home—amidst an epidemic. Our study shows that even basic models can reveal complex dynamics in disease-spreading scenarios when the social dimension is also introduced.

Keywords. *Agent-based modelling; El-farol bar problem; Social dilemma; Epidemiological modelling*

1 Introduction

The advent of Covid-19 sheds new light on the spread of epidemics in social systems, which has ascended as a research imperative [45]. The pandemic has underscored the intricate interplay between disease dynamics and socio-behavioral patterns [34]. Consequently, understanding and strategizing against the spread of epidemics in interconnected social systems have become paramount to safeguarding global health and socio-economic stability.

Mathematical models [31], and subsequently, simulation models [5], have long been pivotal tools in the realm of epidemic management, offering the capacity to predict [18], analyze [17], and strategize [29] against the spread of infectious diseases. The computational implementation of an epidemiological model enables the analysis of disease transmission dynamics [39] through the systematic examination of epidemiological variables, even when they are not analytically tractable [12]. In this perspective, simulations can serve two main interrelated goals, although a more precise taxonomy can be defined [22, 21]. First, by incorporating real-world data and multifaceted parameters, simulations provide a computational platform to assess possible outcomes and interventions in real-world systems [10, 9]. Second, simulations can be employed to assess the reliability of hypotheses and to refine the objectives of empirical studies and treatments [24, 7, 8].

The El Farol Bar problem [2], a seminal example in complexity science [15], exemplifies the use of toy models to study the unpredictability of the dynamics of seemingly simple social systems [6]. In the original form of the problem, multiple agents all face the same binary decision, that each of them has to make without the possibility to agree or to share information with the

others: either to visit a bar with limited capacity or to stay home, where a threshold is set and known to all agents above which they no longer find it enjoyable to visit the bar. The binary outcome – either the visit was enjoyable if the bar was not too crowded or viceversa – is known to each attending agent after the event, and the time series of the outcomes of repeated events is the basis for predicting the next outcome and a decision of each accordingly. This dilemma can be coupled with the challenges posed by epidemiological scenarios, where individuals must decide whether or not to engage in social activities amidst a contagious disease outbreak, as the Covid-19 pandemic showed [33]. The interactions between the underlying mechanisms of social decision-making and the epidemiological dynamics in such scenarios are largely unexplored [37].

This paper presents an epidemiological extension of the El Farol Bar problem and aims to contribute to the understanding of the intertwined nature of the social and epidemiological facets of some systems. The model is implemented using the agent-based modeling (ABM) simulation technique, a computational approach that simulates individual agents and their interactions within a defined environment [13]. This methodology embodies a bottom-up approach, allowing for the representation of heterogeneous behaviors and leading to the emergence of complex system-wide phenomena [43]. This methodology is widely employed across multiple fields, including ecology [27], economics [3], social sciences [35], and epidemiology [45].

The model behavior suggests that even a seemingly simplistic model can exhibit profoundly intricate dynamics. Specifically, our analysis demonstrates that a simple setting, where each agent has only two states, is sufficient for a limit cycle, and therefore a dynamic attractor, to emerge within the state space of infection rate and event attendance. This observation underscores the potential for considerable complexity in real-world scenarios, emphasizing the need for more extensive investigations to improve how social systems should be managed during the spread of a disease.

This paper is structured as follows: The agent-based model is first introduced and a detailed description of its components provided. The model exploration process is then outlined, emphasizing the methodology employed to generate the results. Finally, we present and discuss the outcomes and draw conclusions from our research.

2 Related works

Social dilemma exists with the purposes of improve the understanding regarding how people interact in a resource-bounded environment [49], especially were there is a conflict between bounded rational entities which are metabolically dependent from a shared environment [48]. At the best of our knowledge, [20] was the first to introduce the concept of social dilemma, describing it as a scenario where individual decision-makers possess a dominant strategy that leads to non-cooperation and, if everyone adopt this dominant strategy, the outcome would be universally poorer, leading to a suboptimal equilibrium. The final rate of cooperation usually depends on the payoff structure [40]; even then, it is an approximation and usually requires a certain number of iterations to be reached [14].

Social dilemma are present in many fields. The literature on the use of social dilemmas in economics encompasses a range of perspectives and findings, such as investigating the difference between rational behaviour and social norm [50, 11]. Also, social norms have been employed for addressing environmental policies [16], conflict management strategies [44], social learning [36], and knowledge sharing [41]. Ecology has a long tradition of using social dilemmas [23, 52], especially in light of the pervasive presence of cooperation in natural species [26]. Social

dilemmas help in understanding the origin of sociality [38] or group foraging strategies [4]. Also, social dilemmas have been employed as the border between ecology and social sciences, to study how success in species conservation depends as much on individuals can collaborate to a common purpose [19] and to use classic economic concepts such as signaling and contract theory to interpret evolutionary biology [1].

Although the fields of epidemiology and social dilemmas have not traditionally intersected extensively, recent years have seen a burgeoning interest in the interplay between these disciplines, particularly highlighted by global health crises such as the COVID-19 pandemic [47]. The application of social dilemma frameworks has proven insightful for examining the relationship between individual behaviors and collective outcomes, notably in the context of vaccination rates within populations [46]. These analyses utilize various models to illuminate the impact of factors such as replicator dynamics [30], social efficiency [30, 32], and diffusion structures [51] on vaccination uptake. Furthermore, empirical studies highlight how pro-social behaviors may be amplified by the accelerated transmission of disease [42]. Additionally, the exploration of oscillatory behaviors within social dilemmas reveals how perceived infection risks can drive a collective shift towards more cautious approaches to social interaction, such as increased adherence to social distancing measures [25].

3 Methodology

In this section an agent-based model of an epidemiological version of the El Farol Bar problem is described and the method employed to explore the model is presented.

3.1 Agent-based model

In this section an agent-based model of an epidemiological version of the El Farol Bar problem is described and the method employed to explore the model is presented.

In addressing how epidemics affects the social dynamics in the El Farol Bar problem, agent-based modeling serves for two compelling reasons: as an approach traditionally employed to address social dilemmas, it is an effective means of communication within the scientific community; and it is particularly well-suited for capturing individual behaviors and their effects on an overall epidemic spread. This enables to get insights from the global co-effect of individual (i.e., agent-related) epidemic and social variables.

At each time step t, the model orchestrates a sequence of actions, as depicted in the flowchart (see Figure 1). These actions are divided into two sets. The first is about the decision-making process on bar attendance: evaluating agents' memory of past attendance, estimating the expected crowd level, and making a decision accordingly. The second is about the dynamics of infection as induced by the interactions among the agents given their health states, where an infectious pathogen could be transmitted to those who decide to attend the bar, influenced by the density of the crowd and the duration of exposure.

Together, these two sets of actions capture a dual aspect of agent behavior: social decision making influenced by past experiences and the epidemiological implications of these social choices. The model thus provides a framework for examining the interplay between individual decision making based on memory of previous states and the collective outcomes in terms of disease transmission, offering insights into how individual behaviors aggregate to impact public health.

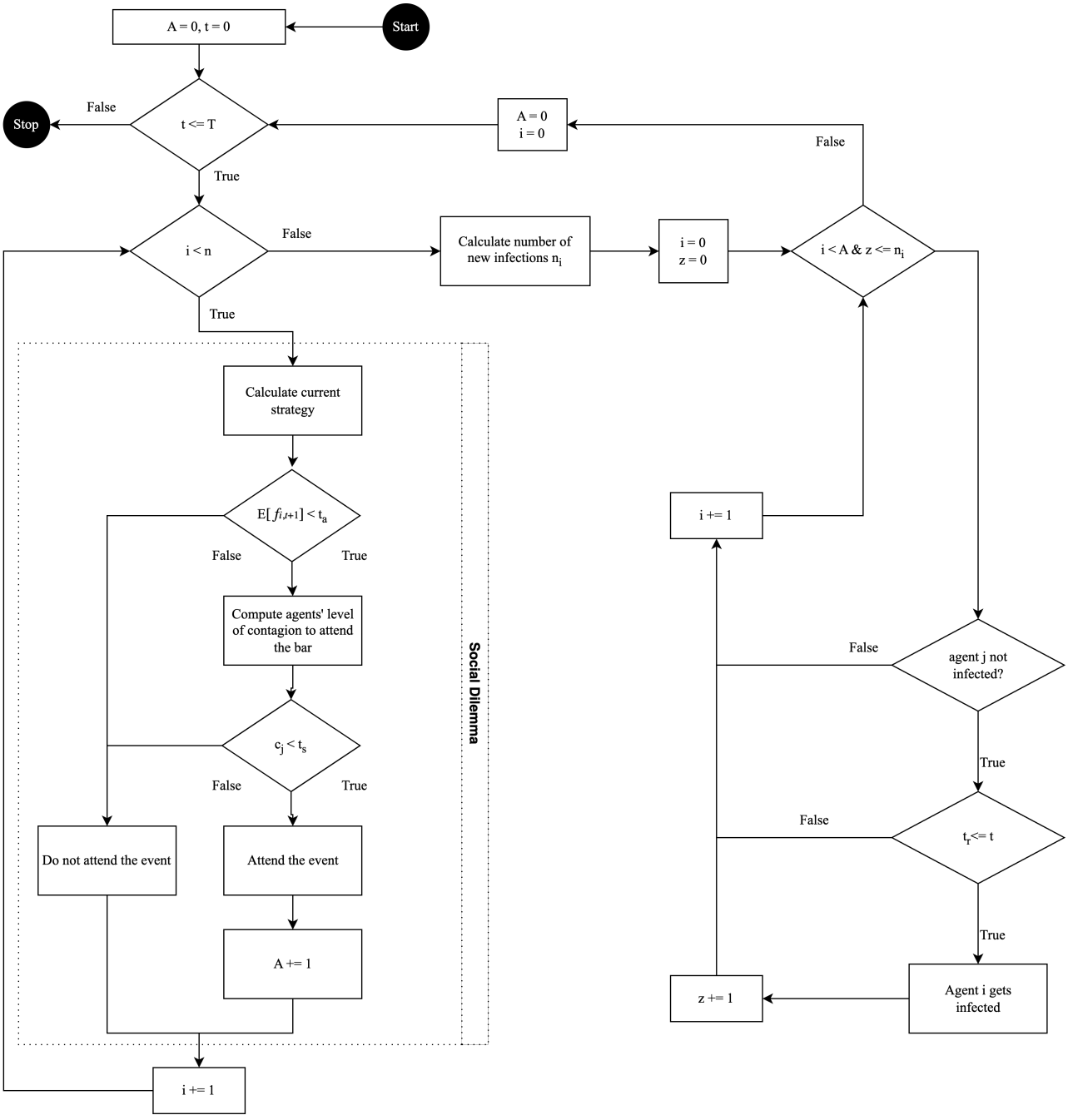


Figure 1: Scheduling process of the model.

3.1.1 Social dilemma

This model includes a single kind of agents, representing the individuals that could decide to attend the bar in any given week (the time step of the model) and thus possibly be infectious. The agents' behavior is modeled according to the hypotheses of the original El Farol Bar problem. First, the only decision each agent can make each week is whether to attend the bar, and the decision is always executed. Second, agents like to attend the bar, if it is not too crowded, and do it as much as they can: hence, each agent decides each week whether to attend the bar depending upon its expectation of the total number of agents who will attend. Third, agents interact with each other only at the bar, and therefore when their decision to attend has been already made.

The proposed model incorporates several hypotheses concerning agents' behavior. First, agents take a binary decision, as they can either choose to attend a bar or not; no other actions are included in the model, to focus on a specific aspect of social behavior. Second, agents inherently enjoy attending the bar and will do so as much as they can, but their preference is tempered by the bar's occupancy; agents are averse to overcrowding. Therefore, the decision to attend the event is influenced by their expectations regarding how crowded the bar will be. In time, this introduces a feedback loop where the average attendance of the bar inversely affects its attractiveness while it is directly influenced by it; a dynamic seen in many real-world social scenarios. Third, agents' interaction is solely defined by the shared presence in the crowded space of the bar, and there are no interpersonal communications or relationships affecting their attending decision.

The information about past attendance plays a crucial role in shaping agents' expectations, as it is used to estimate the number of agents likely to attend the bar in the subsequent week, as follows. For agents attending the event, the new value is the actual number of agents at the bar, while for agents that did not attend the new element of the memory is a random value, which stands for an educated guess made by agents which can not communicate with each other.

The agent's decision whenever to attend or not the bar is taken comparing an attendance threshold and the expectation regarding the future attendance. The attendance threshold t_a is a parameter of the model that depicts venue saturation level above which agents would consider unpleasant to be in the bar, consequently not attending the event.

Each agent i (where i goes from 1 to n, the total number of agents) generates an expectation regarding how many agents will attend the bar at the next time step memorizing the number of agents present at the bar the last m times it attended the bar, with m being the memory length, and weighing it to generate a prediction. In cases where the agent does not attend the bar, the value saved in memory is the one hypothesized by the agent, namely, the one generated with the 'expected attendance'. Let $s_{i,k}$ be the k-th element of the memory of agent i (with $k \in \{0, 1, ..., t\}$) and w the list of weight w_k used to the define importance of each memory element, which increases with k. So, the attendance – which is the number the agent i expects to be at the event at time $t + 1$ – is therefore given by:

$$E_i[\sum_{j=1}^{m} a_{j,t+1}] = \sum_{k=1}^{m} s_{i,k} w_k \tag{1}$$

where a_j is the participation of the agent j to the event at time $t + 1$, $\sum_{j=1}^{m} a_{j,t+1}$ is the total attendance at time $t + 1$ and $E_i[\sum_{j=1}^{m} a_{j,t+1}]$ is the expectation of the total attendance from agent i at time $t+1$. Consequently, from the expected attendance is possible to determine also the expected filling f of the venue at the time $t + 1$

$$E_i[f_{t+1}] = \frac{E_i[\sum_{j=1}^{m} a_{j,t+1}]}{C_{max}} \tag{2}$$

where C_{max} is the maximum capacity of the place. Given the expected filling, at each time step an agent i attends the event whenever $E_i[f_{i,t+1}] < t_a$.

3.1.2 Epidemiological transmission

In epidemiological models, each agent is typically into one of three states: susceptible, infectious, or recovered, a classification central to the SIRS (Susceptible, Infectious, Recovered, Susceptible) model of disease transmission dynamics. These class of models accounts for the possibility of waning immunity after an infection, and eventually become susceptible again, modelling diseases where immunity, either natural or vaccine-induced, can be acquired and diminishes over time.

The epidemiological dimension of this model is based on several key modeling hypotheses. Firstly, the contagion process is assumed to be uniform across all agents, characterized by a consistent duration and a uniform initial level of infectiousness. This simplification negates individual variations in disease progression and response to infection. Secondly, the model posits that the disease in question is non-lethal; agents cannot die as a result of contracting the illness. This assumption is critical as it focuses the model on the dynamics of disease spread rather than mortality rates, and the overall number of individual in the system remains the same. Furthermore, the model assumes the absence of long-term physical or psychological effects post-infection. Recovered agents are not hindered in their ability to participate in normal activities, such as attending a bar, indicating that the disease does not cause lasting health impacts. Psychologically, the model assumes that agents do not experience fear or behavioral changes as a result of the infection. They continue to frequent the events without any alteration in their behavior due to the experience of being infected. Finally, a crucial aspect of this model is the agents' ignorance of the epidemic. Agents lack information about the total number of infected individuals and do not consider the risk of infection in their decision-making process. This implies a lack of adaptive behavior in response to the epidemic, which significantly influences the model's predictions about disease spread. By ignoring potential changes in social behavior and risk assessment, the model strictly focuses on the mechanical spread of the disease under constant behavioral patterns. This approach simplifies the modeling process but may overlook important dynamics present in real-world scenarios where awareness and behavioral adaptations play a crucial role in disease transmission.

Agents can get infected only by participating to an event. So, the epidemic transmission happens solely at the bar, and only if at least an infectious is attending. The number of new infected agents i_t at time t is

$$i_t \propto \lfloor \frac{\sum_{j=1}^{n_i} c_j S_t}{C_{max}} \rfloor$$

where c_i is the level of contagion of each agent attending the bar (which is 0 when agents are not infectious) and S_t the number of agents in susceptible state attending the bar at time t. Notable, the contagious level is taken into account only for the n_i agents which are contagious enough, which is $c_j > t_c$.

In the proposed model, social relationships among agents are not considered, leading to a uniform infection probability for all individuals attending the bar at time t. Consequently, the selection of new i_t infected agents at each time step is randomized from those present, not considering individual interactions or relationships.

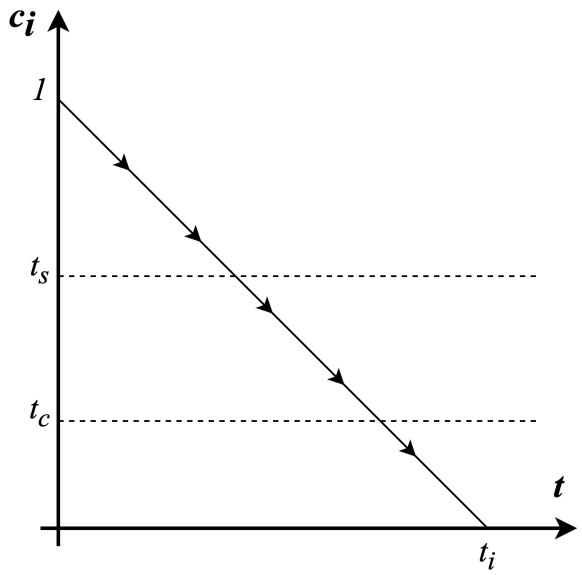


Figure 2: Decourse of contagious level for each agent

Whenever an agent become infected, the infection follows this dynamics. Initially, the contagion level of the newly infected agent i is set to $c_i = 1$. Given an infection duration d_i, the contagiousity of agents decreases linearly by $1/d_i$ at each time step.

In the progression of the disease modeled, two critical thresholds, t_s and t_c, play pivotal roles in influencing agent behavior and the spread of the infection. The first threshold, t_s, represents the infection level at which an agent exhibits sufficient symptoms to deter them from attending the bar. The second, t_c, indicates the infection level beyond which agents can spread the infection The spread of the infection is most influenced by the agents with $t_c < c_i < t_s$, so with a contagious level between these two thresholds. This is because it encapsulates the period when agents are infectious but may not have anymore the level of symptom severity or self-awareness to avoid social gatherings, thereby contributing to the disease transmission dynamics.

In the modeled scenario, infected agents undergo a recovery process after a duration of t_i time steps. Upon recovery, these agents are conferred a temporary immunity lasting t_r time steps. However, this immunity is not permanent; after the elapse of t_r time steps, the agents once again become susceptible to infection. This cyclical pattern of recovery and renewed susceptibility underscores the transient nature of immunity in the context of the model.

3.2 Model exploration

The model exploration consists of a grid sampling exploration of the parameter space, to collect the model outputs from different parameters' combination. Grid sampling from a parameter space involves systematically selecting a finite subset of parameter values that aims at comprehensively represent the entire parameter space. The idea underlying the use of this technique is to facilitate the exploration of system behavior across distinct parameter combinations, especially in cases where not a specific behaviour is expected or researched. Table 1 presents the parameters tested in the simulation and their explored ranges. The data are collected by simulation 10'000 times the agent-based model.

From each simulation, two time-series were collected: A, which is the number of attending agent at each time-step t, and I, the number of infected agents at each step of the simulation.

Table 1: Description of model parameters

Parameter	Description
t_a	Threshold of share of expected agents above which an agent does not attend the event
t_s	Threshold of infection above which the infected agents have symptoms and do not attend the bar
t_c	Threshold of infection below which an agent can not transmit the disease anymore
w_{t-1}	Weight of the last memory in the decision-making process of agents
t_i	Duration of the infection
t_r	Duration of the immunity

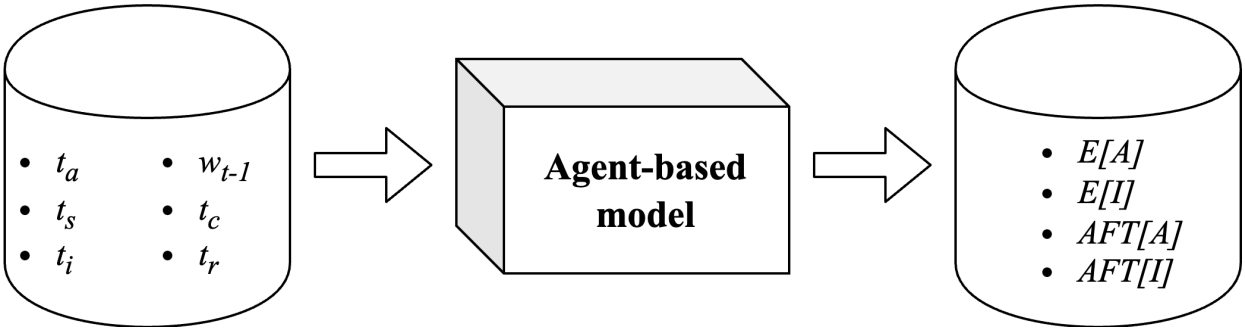


Figure 3: Black box diagram of the experimental setting

Each time-series was computed before to be stored, to extract two output of interest: the mean value of the series $E[A]$ and $E[I]$, which are used to assess the overall status of the system in time, and the autocorrelation $AFC[A]$ and $AFC[I]$. The autocorrelation is a statistical tool that quantifies and visualizes the degree of correlation of a given time series with its own past and future values as a function of time lag, and it is used to perceive seasonality in time-series. Specifically, for each simulation the higher value of the correlation between lagged sub-time-series is collected, and the lag windows used to computed it. Figure 3 depicts a black-box representation of the experimental setting. Even if the model is stochastic, each simulation was initialized with a specific random seed, that was stored as well. Consequently, the results were replicable later, and the time-series of each configuration of interest was observable.

The model, the exploration code and the data analysis are all implemented in Python 3.11. The code and the results are available upon reasonable request.

4 Results and discussion

In Figure 4, the output of the simulation is visualized, demonstrating the infection probability for each individual at the end of simulations. This figure is of interest for two reasons. Firstly, it is necessary to clarify whenever parameters affects the behaviour of the model. Secondly, the figure reveals how even in a simplified model, nonlinear effects become evident, particularly in the context of parameters such as t_a and w_{t-1}.

The parameter t_a, which in here is taken as example, provides a compelling example of this nonlinearity. A low value for t_a results in limited attendance, restricting the spread of infection to a smaller subset of agents. This, in turn, curtails the progression of the epidemic, as per the

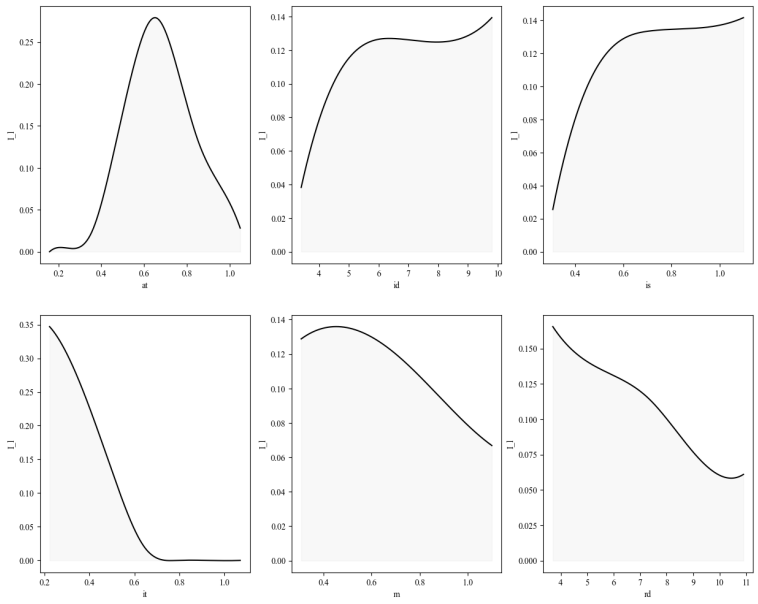


Figure 4: Probability for the infection to last until the final time step $t = 200$ for different parameter values, measured in shared of simulations in which the infection survives.

modelling hypothesis that infection transmission occurs exclusively through event participation, and, if insufficient numbers attend, the contagion cannot disseminate effectively. Conversely, a high t_a value implies near-universal attendance at events, leading to simultaneous infection of a larger agent population. This synchronous infection increases the likelihood of a concurrent rise in immunity, thus diminishing the likelihood that the infection lasted until the simulation end. Such dynamics underscore the critical role of parameter settings in shaping the outcomes of the model and highlight the complex interplay between individual behavioral patterns and the broader epidemiological trends in this simulated environment.

Upon establishing that social parameters significantly influence infection dynamics, it became pertinent to investigate whether these interrelated behaviors lead to the emergence of non-punctual equilibrium states, commonly referred to as limit cycles in two-dimensional scenarios. This exploration is crucial for understanding the temporal evolution of the system under varying social conditions. The presence of limit cycles in such a system suggests a cyclical pattern of infection spread and containment, influenced by social parameters, even in absence of any central control. Identifying and understanding these limit cycles can provide deeper insights into the long-term behavior of the infection, offering valuable perspectives for understanding and predicting the impact of social behavior on disease dynamics.

The investigation into the existence of periodic fluctuations focused on the relationship between the main epidemiological output, denoted as I, and the principal social output A. This analysis was predicated on the hypothesis that an increase in the number of attendees (A) at social gatherings might correlate with a rise in infection rates (I). While this relationship aligns with the conceptual underpinnings of the model, it can be considered an emergent phenomenon, arising from the complex interactions and decision-making processes of the agents within the model, without being explicitly encoded either in the micro nor the macro behaviour of the model. This emergent behavior holds profound implications for the management of infection spreading in scenarios influenced by social behaviors. Specifically, the possibility that social dynamics, driven by individual decision-making processes, could inherently lead to cyclical patterns of infection rates is a significant insight.

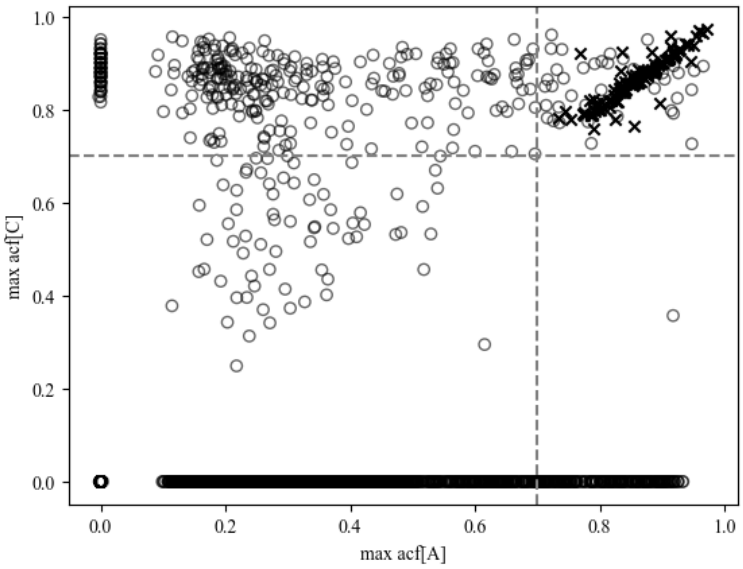


Figure 5: Scatter plot of the maximum values of auto-correlation for the time-series A and I for each simulation

The existence of a limit cycle is assessed by computing the maximum autocorrelation values for both time-series of the epidemiological output I and the social output A. The rationale behind this approach was to detect potential cyclic patterns within the data. A high maximum autocorrelation value in a time-series is indicative of cyclic behavior, signifying points in the series where periodicity or seasonality is pronounced. For the purposes of this study, a threshold value of 0.7 was established, above which autocorrelation is considered significantly high. The detection of high maximum autocorrelation values in both I and A time-series would be indicative of a non-punctual equilibrium within the system. Such a finding would imply that the dynamics of the system do not converge to a fixed point but rather exhibit ongoing cyclical fluctuations.

From the observation of Figure 4, which depicts the results of this analysis, three groups of simulation outcomes can be identified:

1. simulations S_a, that include all the results;
2. simulations S_i, that include all the simulations in which I does not go to 0 at the end of the 200 simulated time-steps;
3. simulations S_c, that include the results in which a limit cycle between A and I appears, so that the $max(AFC[C]) > 0.7$ and $max(AFC[I]) > 0.7$.

Analyzing Figure 4, which presents the results of this analysis from our simulation study, allows for the categorization of the simulation outcomes into three distinct groups based on their characteristics and behaviors. These groups are as follows:

1. Simulations S_a: This group encompasses all the simulation results, serving as a comprehensive dataset to which to make confrontations. It includes the entire range of outcomes observed during the study, providing a holistic view of the simulation's potential behaviors under various conditions.
2. Simulations S_i: This subset includes those simulations where the epidemiological output I remains non-zero at the conclusion of the 200 simulated time-steps. The persistence of

Table 2: Mean parameters values per each scenario

par	$E[par\|S_a]$	$E[par\|S_i]$	$E[par\|S_c]$	$\frac{E[par\|S_i]-E[par\|S_a]}{E[par\|S_a]}$	$\frac{E[par\|S_c]-E[par\|S_a]}{E[par\|S_a]}$
t_a	0.505900	0.582962	0.627943	0.152327	0.241241
t_s	0.506058	0.178483	0.140284	-0.647308	-0.722791
t_c	0.499642	0.594004	0.323688	0.188858	-0.352160
w_{t-1}	0.504864	0.445740	0.511489	-0.117108	0.013123
t_i	5.047981	5.895795	6.687943	0.167951	0.324875
t_r	5.552979	4.515539	4.652482	-0.186826	-0.162165

I beyond this duration indicates scenarios where the infection continues to be present in the system, suggesting incomplete containment or ongoing transmission dynamics. This category is crucial for understanding the conditions under which the infection sustains itself over extended periods.

3. Simulations S_c: The final group comprises simulations where a limit cycle between social output A and epidemiological output I is evident. This is characterized by both $max(AFC[C]) > 0.7$ and $max(AFC[I]) > 0.7$, indicating significant autocorrelation and, thus, the presence of cyclical patterns in both social behavior and infection rates. This group is particularly significant as it highlights the dynamic interplay between social behaviors and epidemiological outcomes, manifesting as cyclic fluctuations over time.

These categorizations provide a structured approach to analyzing the simulation data, enabling a clearer understanding of the different dynamics at play.

Table 2 depicts the mean parameters values for each of the scenario. The analysis of table shows that the most influential factor in the emergence of limit cycles in model's outputs is the contagiousness threshold t_c, the value above which individuals become infectious, since it is consistently lower in S_c than in S_i. This suggests that the observed seasonality in the model is at least partially driven by maintaining a low threshold for contagiousness. Additionally, an high the duration of contagion t_i is observed when cyclicity appears. This implies that within a socio-epidemiological context, a prolonged period of contagiousness might be a prerequisite for establishing stable oscillatory behavior in the whole population. It stresses the complex balance between the duration of infectiousness and the propensity to spread the disease, both significantly contributing to the emergence of limit cycles in this agent-based model.

Another influential parameter is t_s, which stands for the degree of health discomfort that prompts individuals to decide against attending the bar. Our findings suggest a difference in system behavior based on this parameter. Specifically, infections manifest without any noticeable cyclicity with an higher than average t_s.

Finally, a trend observed is the presence of cycles in scenarios where individuals better consider information from multiple past periods before making a decision, which in this model is given by an higher value of w_{t-1}, especially compared to the case in which infection is present. Essentially, when individuals incorporate a broader spectrum of historical data in their decision-making process regarding attendance, the system more frequently exhibits cyclical patterns. This suggests that the depth of historical context plays a significant role in shaping the system dynamics when there is an interplay between a social and an epidemiological dimension. By relying on a more extensive set of past data, individuals inherently introduce a delayed response mechanism. This delay can lead to periodic oscillations as individuals react to older information, causing a ripple effect in their collective behavior. The presence of these cycles underscores

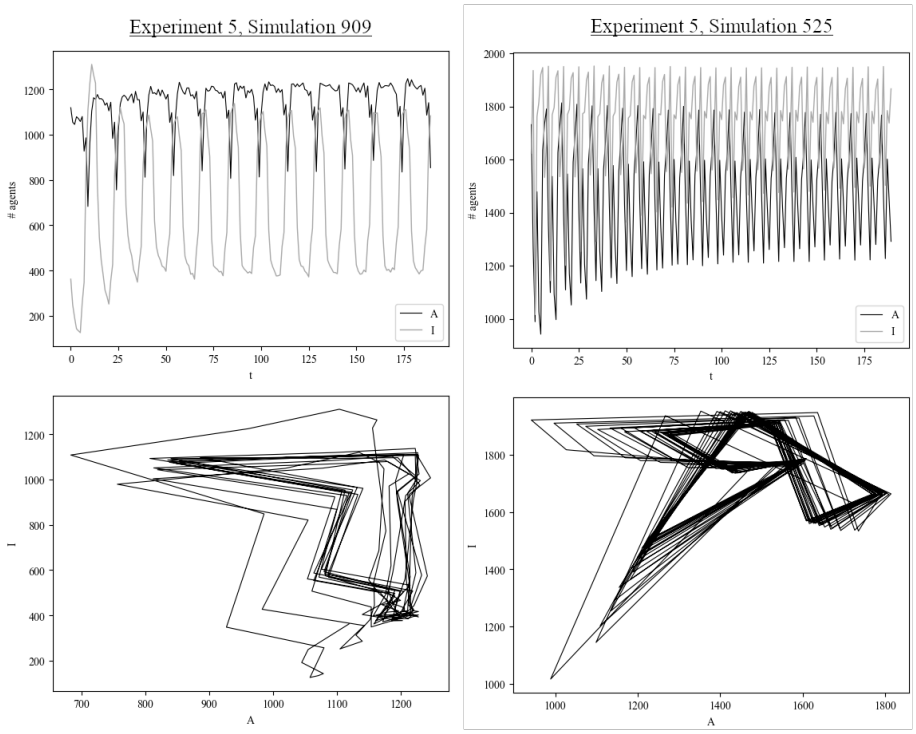


Figure 6: Two examples of simulations presenting a periodical behaviour (time-series and state-state)

the importance of understanding the temporal depth of decision-making processes in socio-epidemiological models. It indicates that not just the immediate past, but a more extended historical context, can have profound implications on the emergent dynamics of such systems.

Finally, Figure 6 presents two examples of simulations in which a cyclical equilibrium appeared. In both cases, it was considered more appropriated to present both the time-series representation and the behaviour on the state-space. The jagged nature of the observed cycles can be attributed to the high temporal granularity chosen for the study. As a consequence of this granularity, many discontinuities are apparent, which are not observed in classic limit cycles derived in continuous functions or in continuous time simulations. However, the very fact that we observe such sharp-edged cycles indicates the underlying dynamics generating these non-static equilibria are notably robust. In essence, despite the coarse temporal resolution introducing apparent irregularities, the inherent stability of the system's dynamics is evident. This robustness provides assurance in the reliability of the observed patterns, and suggest that an analogue real-world system could have a given resilience to external perturbations, given for example by policy-maker interventions or epidemiological setting variation. Nevertheless, further mathematical analysis and simulations are required to quantify the precise nature and stability of this limit cycle.

5 Conclusions

The results of this paper demonstrate how an intertwined socio-epidemiological toy model can be utilized to enhance the understanding of how individual behavior and its thresholds impact the spread of infections in socio-epidemiological models. More precisely, it shows that a non-stable equilibrium can exist in this type of system, and that these cycles are significantly influenced not only by the epidemic aspect of the system but also by the social aspect, even in

conditions where there is no central authority to implement controls and make decisions, and where the agents exhibit greediness without considering potential consequences.

The limitations of this work include a strong reliance on specific modeling assumptions, such as agent homogeneity and the specific rules of behavior, including when to attend the bar in case of infection levels above or below a certain threshold, or agents' inability to estimate the number of infected individuals who will attend the bar the following week. Furthermore, the results should be validated in more realistic scenarios.

Future developments entail the introduction of a social network to assess how the presence of specific relationships that determine when an agent attends the bar affects the intertwined relationship between the social and epidemiological components of the system. Additionally, the model could be employed to study potential healthcare policies, such as mandatory reductions in capacity at public places or awareness campaigns for citizens. Finally, an analytical treatment of the model could be performed to gain a better understanding of the oscillatory behavior observed in the model's output.

References

[1] Marco Archetti, István Scheuring, Moshe Hoffman, Megan E Frederickson, Naomi E Pierce, and Douglas W Yu. Economic game theory for mutualism and cooperation. *Ecology letters*, 14(12):1300–1312, 2011.

[2] W Brian Arthur. Inductive reasoning and bounded rationality. *The American economic review*, 84(2):406–411, 1994.

[3] W Brian Arthur. Out-of-equilibrium economics and agent-based modeling. *Handbook of computational economics*, 2:1551–1564, 2006.

[4] Lars A Bach, Torbjørn Helvik, and Freddy B Christiansen. The evolution of n-player cooperation—threshold games and ess bifurcations. *Journal of Theoretical Biology*, 238(2):426–434, 2006.

[5] Raul Bagni, Roberto Berchi, and Pasquale Cariello. A comparison of simulation models applied to epidemics. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

[6] Francesco Bertolotti, Angela Locoro, and Luca Mari. Sensitivity to initial conditions in agent-based models. In *Multi-Agent Systems and Agreement Technologies: 17th European Conference, EUMAS 2020, and 7th International Conference, AT 2020, Thessaloniki, Greece, September 14-15, 2020, Revised Selected Papers 17*, pages 501–508. Springer, 2020.

[7] Francesco Bertolotti and Riccardo Occa. "roads? where we're going we don't need roads." using agent-based modeling to analyze the economic impact of hyperloop introduction on a supply chain. In *Multi-Agent Systems and Agreement Technologies: 17th European Conference, EUMAS 2020, and 7th International Conference, AT 2020, Thessaloniki, Greece, September 14-15, 2020, Revised Selected Papers 17*, pages 493–500. Springer, 2020.

[8] Francesco Bertolotti and Sabin Roman. The evolution of risk sensitivity in a sustainability game: an agent-based model. 2022.

[9] Francesco Bertolotti and Sabin Roman. Risk sensitive scheduling strategies of production studios on the us movie market: An agent-based simulation. *Intelligenza Artificiale*, 16(1):81–92, 2022.

[10] Sukaina Bharwani, Mike Bithell, Thomas E Downing, Mark New, Richard Washington, and Gina Ziervogel. Multi-agent modelling of climate outlooks and food security on a community garden scheme in limpopo, south africa. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1463):2183–2194, 2005.

[11] Anders Biel and John Thøgersen. Activation of social norms in social dilemmas: A review of the evidence and reflections on the implications for environmental behaviour. *Journal of economic psychology*, 28(1):93–112, 2007.

[12] Georgiy V Bobashev, D Michael Goedecke, Feng Yu, and Joshua M Epstein. A hybrid epidemic model: combining the advantages of agent-based and equation-based approaches. In *2007 winter simulation conference*, pages 1532–1537. IEEE, 2007.

[13] Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America*, 2002.

[14] Valerio Capraro. A model of human cooperation in social dilemmas. *PloS one*, 8(8):e72427, 2013.

[15] John L Casti. Seeing the light at el farol: a look at the most important problem in complex systems theory. *Complexity*, 1(5):7–10, 1996.

[16] Nicola Cerutti. Social dilemmas in environmental economics and policy considerations: A review. *Ethics in Progress*, 8(1):156–173, 2017.

[17] Vittoria Colizza, Alain Barrat, Marc Barthelemy, Alain-Jacques Valleron, and Alessandro Vespignani. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS medicine*, 4(1):e13, 2007.

[18] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences*, 103(7):2015–2020, 2006.

[19] Graeme S Cumming. A review of social dilemmas and social-ecological traps in conservation and natural resource management. *Conservation Letters*, 11(1):e12376, 2018.

[20] Robyn M Dawes. Social dilemmas. *Annual review of psychology*, 31(1):169–193, 1980.

[21] Bruce Edmonds, Christophe Le Page, Mike Bithell, Edmund Chattoe-Brown, Volker Grimm, Ruth Meyer, Cristina Montanola-Sales, Paul Ormerod, Hilton Root, and Flaminio Squazzoni. Different modelling purposes. *JASSS*, 22(3), 2019.

[22] Joshua M Epstein. Why model? *Journal of artificial societies and social simulation*, 11(4):12, 2008.

[23] Ilan Eshel and Uzi Motro. The three brothers' problem: kin selection with more than one potential helper. 1. the case of immediate help. *The American Naturalist*, 132(4):550–566, 1988.

[24] Iulia Georgescu. Toy model. *Nature Physics*, 8(6):444–444, 2012.

[25] Alina Glaubitz and Feng Fu. Oscillatory dynamics in the dilemma of social distancing. *Proceedings of the Royal Society A*, 476(2243):20200686, 2020.

[26] Chaitanya S Gokhale and Christoph Hauert. Eco-evolutionary dynamics of social dilemmas. *Theoretical Population Biology*, 111:28–42, 2016.

[27] Jonathan R Goodman, Andrew Caines, and Robert A Foley. Shibboleth: An agent-based model of signalling mimicry. *PloS one*, 18(7):e0289333, 2023.

[28] Volker Grimm and Steven F Railsback. *Individual-based modeling and ecology*. Princeton university press, 2005.

[29] Danny Ibarra-Vega. Lockdown, one, two, none, or smart. modeling containing covid-19 infection. a conceptual model. *Science of the Total Environment*, 730:138917, 2020.

[30] KM Ariful Kabir. How evolutionary game could solve the human vaccine dilemma. *Chaos, Solitons & Fractals*, 152:111459, 2021.

[31] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[32] Md Mamun-Ur-Rashid Khan and Jun Tanimoto. Investigating the social dilemma of an epi-

demic model with provaccination and antivaccination groups: An evolutionary approach. *Alexandria Engineering Journal*, 75:341–349, 2023.

[33] Bruno Kluwe-Schiavon, Thiago Wendt Viola, Lucas Poitevin Bandinelli, Sayra Catalina Coral Castro, Christian Haag Kristensen, Jaderson Costa da Costa, and Rodrigo Grassi-Oliveira. A behavioral economic risk aversion experiment in the context of the covid-19 pandemic. *PLoS One*, 16(1):e0245261, 2021.

[34] Kurt Kreulen, Bart de Bruin, Amineh Ghorbani, René Mellema, Christian Kammler, Lois Vanhée, Virginia Dignum, and Frank Dignum. How culture influences the management of a pandemic: A simulation of the covid-19 crisis. *Journal of Artificial Societies and Social Simulation*, 25(3), 2022.

[35] Aviral Marwal and Elisabete A Silva. City affordability and residential location choice: A demonstration using agent based model. *Habitat International*, 136:102816, 2023.

[36] Markus Mobius and Tanya Rosenblat. Social learning in economics. *Annu. Rev. Econ.*, 6(1):827–847, 2014.

[37] Giulia Pullano, Eugenio Valdano, Nicola Scarpa, Stefania Rubrichi, and Vittoria Colizza. Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the covid-19 epidemic in france under lockdown: a population-based study. *The Lancet Digital Health*, 2(12):e638–e649, 2020.

[38] Jessica Purcell, Alan Brelsford, and Leticia Avilés. Co-evolution between sociality and dispersal: the role of synergistic cooperative benefits. *Journal of Theoretical Biology*, 312:44–54, 2012.

[39] Hazhir Rahmandad and John Sterman. Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models. *Management science*, 54(5):998–1014, 2008.

[40] David G Rand, Joshua D Greene, and Martin A Nowak. Spontaneous giving and calculated greed. *Nature*, 489(7416):427–430, 2012.

[41] Liana Razmerita, Kathrin Kirchner, and Pia Nielsen. What factors influence knowledge sharing in organizations? a social dilemma perspective of social media communication. *Journal of knowledge Management*, 20(6):1225–1246, 2016.

[42] Magdalena Rychlowska, Job van der Schalk, and Antony SR Manstead. An epidemic context elicits more prosocial decision-making in an intergroup social dilemma. *Scientific Reports*, 12(1):18974, 2022.

[43] Alexander F Siegenfeld and Yaneer Bar-Yam. An introduction to complex systems science and its applications. *Complexity*, 2020:1–16, 2020.

[44] Sim B Sitkin and Robert J Bies. Social accounts in conflict situations: Using explanations to manage conflict. *Human relations*, 46(3):349–370, 1993.

[45] Flaminio Squazzoni, J Gareth Polhill, Bruce Edmonds, Petra Ahrweiler, Patrycja Antosz, Geeske Scholz, Emile Chappin, Melania Borit, Harko Verhagen, Francesca Giardini, et al. Computational models that matter during a global pandemic outbreak: A call to action. *JASSS-The Journal of Artificial Societies and Social Simulation*, 23(2):10, 2020.

[46] Jun Tanimoto. *Sociophysics approach to epidemics*, volume 23. Springer, 2021.

[47] Jun Tanimoto and Jun Tanimoto. Social dilemma analysis of the spread of infectious disease. *Evolutionary Games with Sociophysics: Analysis of Traffic Flow and Epidemics*, pages 155–216, 2018.

[48] Vladislav Valentinov and Lioudmila Chatalova. Institutional economics and social dilemmas: a systems theory perspective. *Systems Research and Behavioral Science*, 33(1):138–149, 2016.

[49] Paul AM Van Lange, Jeff Joireman, Craig D Parks, and Eric Van Dijk. The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*,

120(2):125–141, 2013.

[50] J Mark Weber, Shirli Kopelman, and David M Messick. A conceptual review of decision making in social dilemmas: Applying a logic of appropriateness. *Personality and social psychology review*, 8(3):281–307, 2004.

[51] Yuting Wei, Yaosen Lin, and Bin Wu. Vaccination dilemma on an evolving social network. *Journal of Theoretical Biology*, 483:109978, 2019.

[52] Chris Woodall, Allison Handler, and Len Broberg. Social dilemmas in grassland ecosystem restoration: integrating ecology and community on a montana mountainside. *Ecological Restoration*, 18(1):39–44, 2000.

FRCCS

Identifying Sentinel Nodes and Communities in Nigeria: The Role of missing information

Asma Mesdour[1−2✓], Sandra Ijioma[3], , Muhammad-Bashir Bolajoko[3], Elena Arsevska[1−2], Mamadou Ciss[4],Mathieu Andraud[5], Stephen Eubank[6] and Andrea Apolloni[1−2]

[1] *CIRAD, UMR ASTRE, INRAE, 34398 Montpellier ; asma.mesdour@cirad.fr, andrea.apolloni@cirad.fr.*
[2] *UMR ASTRE, CIRAD, Université de Montpellier, Montpellier, France* [3] *National Veterinary Research Institute, Vom, Nigeria; ijioma.sandra@gmail.com* [4] *ISRA, LNERV BP 2057 Dakar-Hann, Senegal* [5] *ANSES, Ploufragan-Plouzané-Niort Laboratory, EPISABE Unit, Ploufragan, France* [6] *Biocomplexity Institute and Initiative, Virginia, USA*
✓ *Presenting author*

Abstract. The study investigates the spread of Peste des Petits Ruminants (PPR), a highly contagious disease in small ruminants in Nigeria, where livestock movements are crucial for trade and production. The main objective of this study is to identify sentinel nodes and their characteristics to improve PPR surveillance. As the data collected represents only a fraction of all existing links, we have reconstructed a network using the hierarchical random graph (HRG) method, which predicts the missing links. Simulations of PPR spread were conducted using a stochastic SI-weighted model, considering various transmission probabilities. Sentinel nodes, crucial for early epidemic detection, were identified, and contagion clusters were extracted using a novel community identification algorithm. Then, the optimal set of sentinel nodes' structural, socioeconomic, and environmental characteristics was assessed. In the predicted network, additional links led to higher prevalence and shorter outbreak duration. The number of sentinel nodes varied with transmission probabilities, peaking at Pinf = 0.01. Interestingly, socio-economic attributes played a more significant role than structural ones in sentinel node characteristics. Sentinel nodes and their characteristics remained consistent, but more emerged in the predicted network. Community analysis revealed geographically dispersed communities in both observed and predicted networks. The study underscores the imperative need for an integrative approach that merges field data, network analysis, and epidemiological modelling. This approach is essential for implementing targeted surveillance and effective control strategies, particularly in regions susceptible to PPR like Nigeria.

Keywords. *Mobility network, livestock diseases, Sentinel node.*

1 Introduction

Peste des Petits Ruminants (PPR) is a highly contagious disease affecting small ruminants, spreading primarily due to livestock mobility. In sub-Saharan Africa, livestock movements are essential for trade and production. Consequently, movements encompass several hundred kilometres, generating complex mobility patterns. Understanding these mobility patterns is vital

to developing efficient PPR surveillance systems. Useful epidemic detection relies on identifying a subset of nodes that could deliver accurate and timely insights into the spread of the disease (sentinel nodes)[2]. For PPR, this corresponds to identifying markets where infected animals could reach the epidemic's beginning. However, pinpointing these sentinel nodes is a complex task, contingent on various factors, including the availability and incompleteness of detailed data on animal mobility. In PPR-endemic countries like Nigeria, the disease seroprevalence varies geographically, ranging from 11Western Area[3]. Although the impact of PPR on livestock is widely recognized, no livestock identification system is currently available in Nigeria. Therefore, movement data, on which the present study relies, were collected using market surveys. In this work, we simulate the diffusion of PPR through the market network of three Nigerian states (Plateau, Bauchi and Kano) and their surroundings to describe diffusion patterns, detect communities (contagion clusters), and identify possible sentinel nodes and their characteristics.

2 Materials and methods

We used market survey data collected in the framework of the Lidiski project. A total of ten markets across three Nigerian States (Plateau, Bauchi, and Kano) were chosen for sampling: six markets were surveyed in Plateau, and two markets were sampled in both Bauchi and Kano States. The collected data included information on the origin and destination of animal movements and the number of animals involved. Using this dataset, we reconstructed the mobility network, referred to as the reference network, where nodes represent Wards (third-order administrative unit), and links indicate animals exchanged between two nodes weighted by the number of exchanged animals across the period. Market data collection is limited to specific regions and periods of the year. We have bolstered our conclusions by analyzing uncertainty and predicting missing links to gauge how incomplete data affects epidemic spread. To improve the reconstruction of the animal mobility network, we tested various methods, like neighbourhood-based predictors and the Structural Perturbation Method (SPM). Among these, the Hierarchical Random Graph (HRG)[1] method stood out with an AUC value of 0.9, indicating its suitability for prediction. HRG method reconstructs the network by exploring potential dendrograms, aiming to represent its hierarchical structure accurately. We simulated the spread of PPR disease through animal movements between the Wards of the three Nigerian States using a stochastic SI-weighted type model, where the cumulative probability for a Ward i of getting infected is:

(1)

$$\text{Pi} = 1 - (1 - \text{Pinf})^{Wi,j}$$

Pinf represents the probability that a single infectious animal from an infected Ward could transmit the disease to other animals in susceptible Ward i. Wi, j represents the number of animals moved from node j to node i. We tested several values for the Pinf, from 0.0001 to 0.3. We identify the sentinel nodes from simulation results, i.e., those likely to get infected before the epidemic peak. Following Nath et al.[4], we extracted contagion clusters, groups of nodes with potential mutual infection during an epidemic, using a new community identification algorithm based on the dynamical characteristics of the network instead of using Modularity maximization. Based on Shannon and Moore's reliability, the algorithm involves an edge ranking algorithm that scores edges based on their contribution to overall reliability[4]. Edges are sequentially removed based on their importance until the largest strongly connected component exceeds a specified size. The algorithm terminates when the maximum Strongly Connected Component (SCC) size remains within the desired limit. Then, we examined whether or not

the sentinel nodes identified previously are located within the same communities. Finally, we sought to understand our identified optimal set's structural, socioeconomic, and environmental characteristics using a Random Forest Classification. To assess the role of network structure and the interplay with transmission probability in identifying sentinel nodes, contagion clusters, and the extension of the epidemics, we ran simulations on both the reference and reconstructed network.

3 Results

The reference network comprises 233 nodes and 335 links. The HRG approach predicted 4670 additional links. The differences in network structures are also reflected in the dynamics of epidemics. In the predicted network, the prevalence of infected nodes is higher across all transmission probabilities compared to the reference network (figure 1A), and the duration of outbreaks—measured as the time taken to reach the maximum number of new infections—is shorter in the predicted network. This rapid spread can be attributed to supplementary links, introducing new pathways between nodes and facilitating the fast propagation of the epidemic. According to different probabilities, 1 and 11 sentinel nodes were identified in the reference network. Only one sentinel node was identified when Pinf was equal to 0.0001. The number of sentinel nodes increases to reach the maximum at Pinf = 0.01 before decreasing to 5 at Pinf = 0.1 and Pinf = 0.3 (figure 1B). This behaviour could be related to the fact that by increasing the transmission probability, the epidemic peak occurs very early, affecting only the most vulnerable nodes. During the disease simulation on the predicted network, we consistently identify the same set of nodes, and their behaviour remains unchanged (figure 1B). However, there is an increase in the number of additional sentinel nodes observed during the simulation.

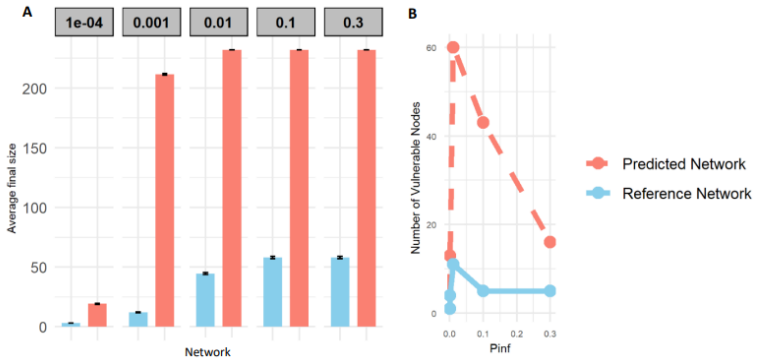


Figure 1: A represents the final size of the simulated epidemic, where the average was computed across all iterations. B illustrates the count of vulnerable nodes within the reference (sky blue colour) and predicted network (salmon colour).

Eight communities were identified in the reference network and 11 in the predicted network. The community sizes in the reference network range between 5 and 15, whereas those in the predicted network vary from 2 to 15 nodes (figure 2). Nodes forming the communities are geographically dispersed in the two networks. Although some communities overlap between the two networks, it is noteworthy that none of the identified sentinel nodes were part of these communities either in the predicted or in the reference network and often proved isolated. Analysis of sentinel node characteristics highlighted the role of socio-economic attributes over structural ones. The common vulnerable nodes between the two networks shared the population of animals as the most important characteristic, followed by the human population. Dry Matter Product (DMP) and eigenvector centrality ranked third in importance.

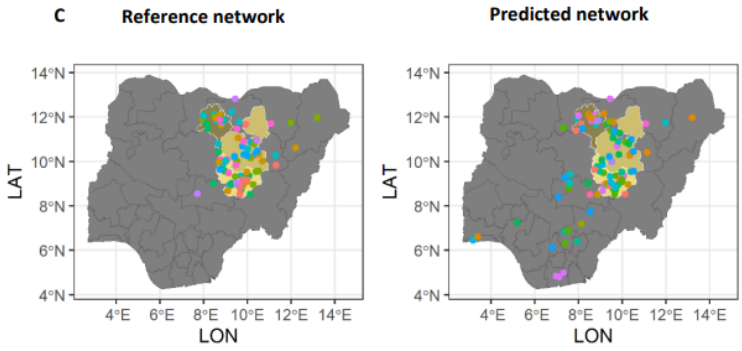


Figure 2: Location of communities of each network on a map (the yellow zone represents the Lidiski area, and each colored dot represents a community)).

4 Discussion

This work allowed for identifying vulnerable nodes from reference and simulated networks owing to sentinel node characteristics and revealed the existence of groups of nodes capable of infecting each other. It would be interesting to explore whether removing bridges connecting these communities could be a complementary prevention measure to enhance surveillance. Furthermore, this study considers a static network and an SI model, indicating that once infected, nodes could continue infecting for the duration of the study. Future work should include network evolution in time and using SIR models. Nevertheless, two dynamics could affect sentinel node characteristics. The network dynamics could change the propagation pattern and reduce the probability of reaching a particular node. At the same time, the SIR-like model could better recreate the long-term behaviour of the dynamics when nodes could lose their ability to infect (recover). Exploring the temporal behaviours becomes imperative to comprehensively understand the impact on the evolution of cluster structures and sentinel nodes over time. This step opens the way to an integrative approach, which could be fed with more exhaustive data collection. Combining field data, network analysis, and epidemiological modelling offers the opportunity for evaluation of the position of these nodes in the transmission chain and control strategies through targeted surveillance.

References

[1] David Allen, Lu Ching, Dave Huber, and Hankyu Moon. *'Hierarchical Random Graphs for Networks with Weighted Edges and Multiple Edge Attributes.* inproceedings, 2011.

[2] Paolo Bajardi, Alain Barrat, Lara Savini, and Vittoria Colizza. 'optimizing surveillance for livestock disease spreading through animal movements. *The Royal Society*, 2022.

[3] Samuel Mantip, Anthony Sigismeau, David Shamaki, Timothy Yusuf Woma, Olivier Kwiatek, Genevieve Libeau, Souabou Farougou1, and Arnaud Bataille. 'molecular epidemiology of peste des petits ruminants virus in nigeria: An update. *Transboundary and Emerging Diseases*, 2021.

[4] Madhurima Nath Manu Amundsen Abhijin Adiga Ritwick Mishra, Stephen Eubank. *'Community Detection Using Moore-Shannon Network Reliability: Application to Food Networks.* Springer, 2023.

FRCCS

Social Media Cross-Network Association and Prediction

Allison Gunby-Mann[1] and Peter Chin[1✓]

[1] *Thayer School of Engineering, Dartmouth College ; allison.mann.th@dartmouth.edu, peter.chin@dartmouth.edu*

[✓] *Presenting author*

Abstract. Separate social networks may be aligned via their shared users, and further, many users share friends between networks. This information presents an opportunity for modeling information flow between networks and other inter-network analyses. In this paper, we propose an unsupervised topological approach aimed at identifying cross-network associations among users. Our methodology leverages within-network link prediction techniques alongside cross-network alias detection. Empirical findings based on real-world social network data demonstrate that our proposed method outperforms baseline algorithms, highlighting its effectiveness in capturing inter-network relationships.

Keywords. *Cross-Network Alignment; Social Networks; Unsupervised Machine Learning; Generative Adversarial Network*

1 Introduction

In the modern era, online social networks have become an integral part of society, with their prevalence extending across diverse demographic groups and geographic regions. From personal relationships to professional networking, social networks serve as spaces for social interaction, information sharing, and community engagement. Further, many users seek a diverse array of platforms to satisfy various needs and often maintain multiple profiles across multiple social media networks simultaneously. This multi-platform presence creates a rich environment of interconnected digital footprints, offering a wealth of data for analysis, modeling, and integration.

By leveraging techniques such as data fusion, cross-network association, and user profiling, researchers can harness the collective information dispersed across multiple platforms to gain a more comprehensive understanding of user behavior, preferences, and interactions. Additionally, the integration of data from multiple social networks enables the construction of models of social dynamics such as information diffusion, facilitating the development of applications like personalized recommendations. The abundance of users across multiple social networks not only underscores the complexity of contemporary digital ecosystems but also offers ample ground for interdisciplinary research and innovation in network analysis.

However, these social networks are inherently uneven and heterogeneous which poses difficult challenges when attempting to compare and merge network data. Unlike controlled experi-

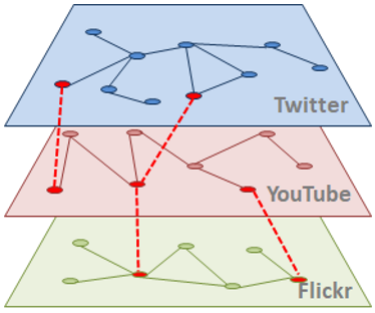


Figure 1: A visualization of the cross-network alignment problem, where red edges between networks representing the same entity are what we are hoping to uncover.

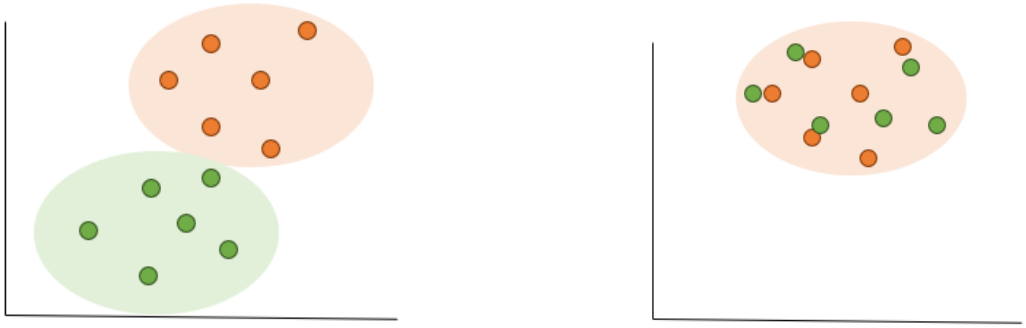
mental settings where sample test networks may exhibit perfect isomorphism, not all users are present in every social network, leading to incomplete and sparse data. Users will also engage in content and form relationships in connected but not identical ways depending on the function of the specific social network. To make matters more complex, researcher's tools for gathering social network data also often result in incomplete data captures of the networks due to space limitations and potential privacy settings.

In this paper, we consider the problem of cross-network association. Informally, the classic formulation of the cross-network association problem is that there are multiple graphs that represent similar or related data, and the goal is to associate nodes between the graphs. A simple example of this is a graph of Facebook users and a graph of Twitter users where our goal is to discover a partial mapping between accounts on these social networks. This mapping would indicate that for a pair of users in the map, the Facebook account is an alias for the same user on the mapped Twitter account, as shown in Fig 1. Successful cross-network alignment can provide a wider picture of a user and their network, and it has important applications to ad recommendations, modeling the spread of information across-networks, and surveillance for security threats.

Despite the difficulty of the problem, it has been a topic of interest in graph theory over the last two decades and significant advances have been made. There are three common categories of approaches to the problem: 1) A spectral approach where the output of the algorithm is a score for each pair of nodes. Some examples include REGAL [7], FINAL [15], IsoRank [14], MAD [9] and BigAlign [8]. 2) Algorithmic or combinatorial approaches which are often greedy and rely on neighborhood similarity and semantic data such as username scores such as UIA [3], the work of Buccafurri et. al. [2], and FRUI [18]. 3) Graph embedding approaches which attempt to align the vector space of the graphs, typically with machine learning techniques. Examples of this technique include PALE [11], IONE [10], Deeplink [17], and COSNET [16].

A common limitation of some of these algorithms is that they rely heavily on semantic data, which may be unreliable if the user is attempting to conceal their identity or if semantic data is unavailable. Additionally, many of these aforementioned methods are supervised meaning they rely on ground truth data for analysis, which is not often available in practice.

In this paper, we consider the problem of cross-network alignment from an unsupervised lens, with no semantic data and no ground truth during training. In this view, we are truly looking at a network alignment solution based solely on the connections present in the underlying graph structure. We employ a node embeddings approach enriched by a link prediction algorithm.



(a) (b)

Figure 2: a) This graphic represents the embedding spaces of the two networks before mapping, and we can see that they are incomparable based on standard distance metrics and thus difficult to align. b) The result post-mapping the green network's embeddings to the orange network's embedding space. Now we can directly compare them for cross-network entity resolution.

Finally, we utilize a Generative Adversarial Network (GAN) [4] to create an initial mapping between the two network's embedding spaces. A GAN is a type of deep learning framework comprising two neural networks, a generator and a discriminator, engaged in a competitive process. The generator creates synthetic data samples, in this case mapped embeddings to the other network's vector space, while the discriminator evaluates whether these embeddings are genuinely sampled from that network or generated by the model. Through iterative training, the generator learns to produce an increasingly realistic mapping, while the discriminator becomes more adept at distinguishing between real and mapped embeddings. This adversarial process will theoretically drive the model to generate a high-quality mapping between the two spaces in an unsupervised setting, as shown in Figure 2.

The benefits of this approach are that it is less reliant on access to high quality labeled data to train the model, and is able to find anchor pairs of aliases within the network on its own, without needing seeds to act as anchors to explore the network and expand the alias set.

2 Dataset

Table 1: Overview statistics of the two networks used in this analysis

Network	Nodes	Edges
Flickr	215,495	9,114,557
Last.fm	136,420	1,685,524

Throughout this project we used the COSNET dataset [16], which is a large dataset that contains a variety of social network data among many platforms. It contains within network friendship edge information between five social networks: flickr, lastfm, likedin, livejournal, and myspace, as well as truth data for edges which link the same user between social networks. It contains a large number of nodes, and the nodes are well connected, leading to a rich topological landscape. This fact along with the rare inclusion of cross-network truth data cements it as a popular choice for this problem. In this paper, we primarily focused on the flickr and lastfm datasets, details of which can be observed in Table 1. In many of the experiments, we used a

sampled version of the dataset generated from random walk sampling of size 5000 nodes and approximately 100,000 edges. The dataset contains 510 pairs of known aliases which we use for evaluation purposes.

3 Method

3.1 Problem Definition

We consider the problem of aligning exactly two social networks by predicting alias anchor pairs with an unsupervised model. Let a social network graph be denoted $G = (V, E)$ where V is the set of N nodes, or users, and $E \subseteq VxV$ is the set of within-network edges, or friendship links. For this paper, we will assume that the graph is undirected meaning that $(u, v) \in E \rightarrow (v, u) \in E$. Each vertex $v_i \in V$ will have a representation in d-dimensions $R_i \in \mathbb{R}^d$, where R is a matrix containing all of the representations of V.

Here we consider two networks, a source and a target network which we will denote as $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ respectively. For each node in the source graph, we want to find its alias in the target graph assuming that it exists, which manifests as a mapping $M : u \in V_1 \rightarrow v \in V_2 \cup \{\emptyset\}$.

3.2 Overview

The approach proposed in this paper follows three main steps: 1) We learn node embeddings for each network with a single network link-prediction by performing random walks on both networks to learn node embeddings for G_1 and G_2 independently. 2) Then we topologically align those embedding spaces using a Generative Adversarial Network (GAN) to approximate a mapping M. 3) Finally, we select anchor pairs based the cosine similarity of the mapped embeddings from M to the source graph embeddings.

3.3 Link Prediction

Table 2: The result of our link prediction model on the COSNET dataset

Network	F-Measure	AUC
Flickr	0.78	0.90
Last.fm	0.86	0.94

The link prediction technique utilized in this study draws its foundation from Word2Vec [12], where embedding vectors are learned for each word in a corpus. By setting a window size within which words frequently appearing in similar contexts are positioned closely in embedding space, Word2Vec facilitates semantic similarity representation. Node2Vec [5] applies the concepts from Word2Vec on graph structured data, but instead of a window of words, a random walk in the graph is used as the context. This process allows relational similarities between nodes traversed in these walks to be captured and consequently, nodes that frequently occur in the same random walk are close in the embedding space. We use a Multi-Layer Perceptron (MLP) model with the embeddings as input to predict future edges for link prediction. The advantage of this approach is that it provides an edge-centric view of the graph via link prediction which will be valuable for cross-network association.

For this paper, we used the method developed by Google Research [1] which learns node embeddings and MLP weights simultaneously. This approach demonstrated superior performance over various conventional link prediction methods, particularly with smaller embedding dimensions.

To apply the Google Research [1] method to the COSNET [16] data, we create a test set by removing edges randomly while maintaining connectivity in the graph. After the edges in the test set are removed and stored, the remaining graph is the training set. To create negative edges for training, random edges are proposed and checked to ensure they were not already in the edge set. For training we made 50% of the edges positive and 50% negative examples. Random walks are then conducted on the training graph, with the resulting train/test sets and random walks serving as inputs to the MLP.

This embedding method achieves high link prediction accuracy for all COSNET networks, as seen in Table 2.

3.4 Cross-Network Topological Alignment

After computing embeddings enriched by the link prediction technique, we then attempt to align the networks using only the learning embeddings based on structural features of the graph as apposed to semantic features such as username and biographical data.

For this step, we utilize a Generative Adversarial Network (GAN) [4] as the model to align the network's embedding spaces. The GAN consists of models: a generator and a discriminator. The generator takes as input embeddings from the source graph G_1 and attempts to output corresponding embeddings in the embedding space of G_2. The generator uses the following loss function, that it learns to minimize during the course of training:

$$L = -\frac{1}{N}\sum_{i=1}^{N} log(1 - D(G(r_1^i))$$

Where D is the output of the discriminator and G is the output of the generator.

Concurrently, the discriminator evaluates the authenticity of these mapped embeddings compared to true embeddings from G_2. Through iterated training, the generator refines its output to produce a more realistic mapping and the discriminator becomes increasingly adept at discerning real and generated embeddings. It is trained with the following loss function that it wants to minimize:

$$L = -\frac{1}{N}\sum_{i=1}^{N} log(D(r_2^i) + log(1 - D(G(r_1^i))$$

The first term refers to the probability of incorrectly classifying real embeddings as fake and the second is the case where generated embeddings are classified as genuine.

Details of the architecture of the model can be seen in Figure 3. This adversarial learning format facilitates mutual improvement.

3.5 Association and Alias Prediction

Given a node $v_1 \in G_1$ and its mapped embedding $M(r_1)$ our goal is to find a corresponding $v_2 \in G_2$ that could represent its alias. To achieve this, we identify the embedding in R_2 that is

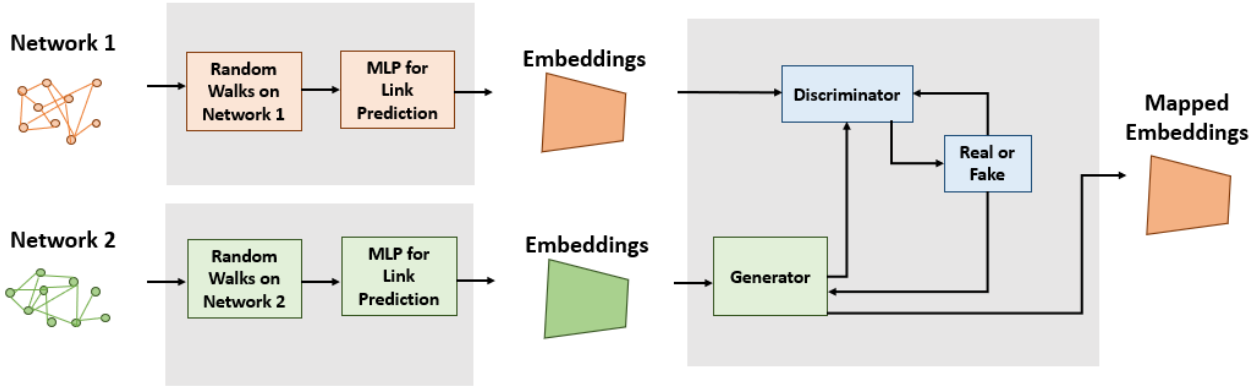


Figure 3: The architecture of the Link Prediction and GAN system for cross-network association.

closest to it. We tested with the Frobenius norm:

$$min_{r_2}||M(r_1) - r_2||, r_2 \in R_2$$

as well as the cosine similarity:

$$min_{r_2} cos(M(r_1), r_2), r_2 \in R_2$$

We then take the top k scoring pairs of nodes in the network based on an experimentally determined threshold to assign aliases. Alternatively, using these same metrics rather than simply taking the top score, we can generate a list of the top scoring aliases for each node to narrow exploration depending on the application.

4 Experiments

To evaluate our model for cross-network alignment we compared it to baseline methods on the COSNET dataset for Flickr and Last.fm. We used the metric of Precision@N based on the aliases discovered by the algorithm which is defined as follows.

$$P@N = \frac{\sum_{i=1}^{q} (\text{success@k(i)})}{q}$$

where success@k(i) is a binary variable representing whether or not the true alias of user i is within the top-k scoring candidates and q is the number of pairs for which we have ground truth evaluation data available.

We chose unsupervised methods of comparison including:

- **Degree-Based Alignment**: A trivial baseline where users in networks are aligned by relative degree rank.
- **Embedding Distance Alignment**: A baseline computed by applying closest distance matching to align the networks using the embeddings after link prediction without the GAN model that maps the embedding spaces. We used the cosine similarity for evaluation because it consistently performed better experimentally for the mapped embeddings.

- **MAD** [9]: A spectral approach to cross-network alignment that matches nodes via singular value decomposition, another comparable unsupervised model to predict aliases.

We also tested two different distance measures for the mapped embeddings of G_1 to G_2 embeddings as discussed in the previous section:

- **Frobenius Norm**
- **Cosine Similarity**

Finally, we tested the last.fm vs flickr alignment as well as sampled last.fm vs sampled last.fm with 80% common edges to test alignment in a less challenging setting.

4.1 Evaluation

In our first experiment, we aimed to assess the cross-network association by comparing random sampled versions of the Last.fm network. This setup ensures ample overlap in the graphs, in this case 80%, resulting in a test set that maintains equal core structure and similar characteristics while introducing variability and disrupting the isomorphism. By conducting the analysis on these sampled versions, we sought to evaluate the performance of our method in a more structured environment. For this experiment, we used Precision@1, Precision@5, and Precision@10. We chose such low values for k due to the heavy overlap resulting in the task being easier for the model.

The results of this experiment can be found in Table 3. It is evident that our method, Link Prediction + GAN outperforms both baseline methods- embedding and degree rank- for all tested cases. The substantial gap in performance in particular between the embedding and our full method demonstrates the power of the GAN model in aligning the two embedding spaces. We observe that the cosine similarity metric and the frobenius norm have very similar comparable results. This is expected, due to the fact that they operate on the same mapped embeddings. Consequently, the metrics are making the same decisions in most cases, leading to closely aligned scores. We do see that cosine similarity consistently slightly edges out the frobenius norm, which is an interesting result, indicating that the cosine similarity better captures the similarity of the embeddings. The MAD approach does beat our method for Precision@10 but their results are much worse for Precision@1 and comparable for Precision@5.

Table 3: Performance comparison between baseline methods for predicting aliases between sampled **Last.fm** and **Last.fm** networks with 80% overlap.

Model	Metric	P@1	P@5	P@10
Embedding	Cosine Similarity	0.0092	0.0379	0.1242
Degree		0.0081	0.0341	0.1039
MAD		0.1508	0.4040	**0.6905**
LP + GAN	Frobenius Norm	0.3009	0.4733	0.5704
LP + GAN	Cosine Similarity	**0.3056**	**0.4884**	0.6005

The results of our next experiment are summarized in Table 4. These results were generated from the alignment of the Last.fm and Flickr data with Flickr being mapped to the Last.fm

embedding vector space, meaning Flickr is the source network and Last.fm is the target network. In this experiment, we considered Precision@10, 20, and 30 to accommodate the more difficult task. Across all metrics, our method surpassed all other tested approaches in the cross-network alignment task, outperforming degree rank by the largest margin. We do observe for the first time, the frobenius norm performing better than the cosine similarity for one particular case.

Table 4: Performance comparison between baseline methods for predicting aliases between **Last.fm** and **Flickr** social networks.

Model	Metric	P@10	P@20	P@30
Embedding	Cosine Similarity	0.0506	0.1021	0.1262
Degree		0.0339	0.0975	0.1128
MAD		0.1341	0.1962	0.2870
LP + GAN	Frobenius Norm	0.1734	**0.2742**	0.3162
LP + GAN	Cosine Similarity	**0.1767**	0.2499	**0.3214**

Overall, these strong performances compared to baseline demonstrate the value in the LP + GAN approach proposed in this paper. Given that the method is wholly unsupervised, the results are especially strong. We suspect that supervised methods would likely result in higher performance, but labeled training data is seldom available in real world networks. Additionally, the best pairs found by this unsupervised method could be used as a seed set or training set for other supervised methods.

5 Future Work

There are many potential avenues for further growth of the method presented in this paper. First, the aliases found by this method are not necessarily internally consistent, meaning that multiple nodes in G_1 can be mapped to the same node in G_2 and vice versa if we only consider top scoring aliases. We could implement a measure for global consistency derived from considering the set of distances in a systematic, non-greedy fashion.

In order to improve accuracy we could also experiment with additional embedding techniques including DeepWalk [13] and GraphSage [6]. We could alternatively attempt to link the embedding model to the alignment model and implement an iterative approach where we refine the original embeddings based on the mapping found by the GAN. The embeddings could be trained in tandem with the GAN to represent the graphs for the specific task of cross-network alignment.

Additionally, we are interested in investigating extensions and applications of this research. For example, the application to community detection within cross-network datasets. In each graph, only a subset of users in the entire multi-network are represented. More broadly, we can imagine identifying specific communities. Some may be friends on slack, some may connect on LinkedIn, but not everyone in the community is present on both networks. In this case, the community is most completely defined as the union of the sets within multiple social networks.

We could also extend this research for multiple networks, not just pairs. Currently to accommodate additional networks, a mapping would have to be trained pair-wise for every combination of two networks in the set. We would then have to assert global consistency in the results, making this scheme much more complex but very intriguing and potentially powerful.

6 Conclusion

This result has broad impacts and important consequences due to its practical application to real-world data sets such as social networks and its adaptability. The graph-based data model of social networks is commonly used in modern computer science research. This model is useful because it allows people's profiles, activities, and information to be directly related with links, represented by edges. In these virtual societies, relationship links (such as friendship, following, commenting/reposting, or any other valid form of engagement) are the main form of expression individuals use to participate in the community. Studying social networks can allow us to track events, activities, information flow, and values within a community. However, social networks in the modern age are becoming increasingly complex. They are neither isolated nor independent; a user will generally have accounts on multiple different websites and engage in content and form relationships in connected but not identical ways. These various social networks are not the same, each may have different main functions or provide a space for niche content and we need to consider the entire story to perform in-depth analysis. An obvious example is if we are tracking a target of interest and we know their account on one social network, we could use this technology to discover their corresponding accounts on other social networks, allowing for increased surveillance of activity. Alternatively, if our goal is to track the spread of some information of interest, considering multiple networks is necessary because the information is going to be spread among different social networks potentially by the same users their followers. In general, we can use this association for increased accuracy and performance in our models. The implications of this research have a strong impact especially in the field of defense and security. In addition to being used directly in analyses for downstream tasks, the discovered aliases from the method proposed in this paper could be used as seed sets for alternative cross-network alignment algorithms that take advantage of additional features but require ground truth for training.

References

[1] Sami Abu-El-Haija, Bryan Perozzi, and Rami Al-Rfou. Learning edge representations via low-rank asymmetric projections. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17. ACM, November 2017.

[2] Francesco Buccafurri, Gianluca Lax, Antonino Nocera, and Domenico Ursino. Discovering links among social networks. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 467–482, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[3] Kaikai Deng, Ling Xing, Longshui Zheng, Honghai Wu, Ping Xie, and Feifei Gao. A user identification algorithm based on user behavior analysis in social networks. *IEEE Access*, 7:47114–47123, 2019.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[5] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks, 2016.

[6] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs, 2018.

[7] Mark Heimann, Haoming Shen, Tara Safavi, and Danai Koutra. Regal: Representation learning-based graph alignment. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18. ACM, October 2018.

[8] Danai Koutra, Hanghang Tong, and David Lubensky. Big-align: Fast bipartite graph

alignment. In *2013 IEEE 13th International Conference on Data Mining*, pages 389–398, 2013.

[9] Chung-Yi Li and Shou-De Lin. Matching users and items across domains to improve the recommendation quality. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 801–810, New York, NY, USA, 2014. Association for Computing Machinery.

[10] Li Liu, William K. Cheung, Xin Li, and Lejian Liao. Aligning users across social networks using network embedding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 1774–1780. AAAI Press, 2016.

[11] Tong Man, Huawei Shen, Shenghua Liu, Xiaolong Jin, and Xueqi Cheng. Predict anchor links across social networks via an embedding approach. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 1823–1829. AAAI Press, 2016.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[13] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '14. ACM, August 2014.

[14] Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.

[15] Si Zhang and Hanghang Tong. Final: Fast attributed network alignment. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1345–1354, New York, NY, USA, 2016. Association for Computing Machinery.

[16] Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip S. Yu. Cosnet: Connecting heterogeneous social networks with local and global consistency. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1485–1494, New York, NY, USA, 2015. Association for Computing Machinery.

[17] Fan Zhou, Lei Liu, Kunpeng Zhang, Goce Trajcevski, Jin Wu, and Ting Zhong. Deeplink: A deep learning approach for user identity linkage. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 1313–1321, 2018.

[18] Xiaoping Zhou, Xun Liang, Haiyan Zhang, and Yuefeng Ma. Cross-platform identification of anonymous identical users in multiple social media networks. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):411–424, 2016.

Foundations of complex systems

I'm polite(r) because I think you are. Elucidating the impact of contextual information on the expression of emotions by users of Wikipedia

Sasha Piccione[1✓] and Nicolas Jullien[1,2]

[1] *Ca' Foscari University of Venice, 873, Fondamenta San Giobbe, 30121, Venice, Italy; sasha.piccione@unive.it*
[2] *IMT Atlantique, LEGO-M@rsouin, Technopole de Brest Iroise 29238 Brest, France; nicolas.jullien@imt-atlantique.fr*
[✓] *Presenting author*

Abstract. The objective of this research is to examine the emotions conveyed in comments within online cooperative contexts that rely on asynchronous textual interactions, and to identify the factors that contribute to their emotional intensity. To achieve this, the study relies on Appraisal and Communication Accommodation theories to explore the interaction between the individual, the context, and the outcome of this interaction in the comments posted on Wikipedia's Talk Pages. It has been found that the context has an impact on the emotions expressed by users. Specifically, the thread level has a greater impact compared to wider or narrower levels of analysis. Additionally, we highlight that emotion intensity varies due to the context around a baseline which is the emitting subject's one.

Keywords. *Sentiment Analysis; Wikipedia; Commons*

1 Extended abstract

In recent years, organisations have implemented various approaches to improve their virtual team working efforts for knowledge production. Team members increasingly work independently and communicate through digital infrastructures such as Team Communication Platforms (TCPs) like Slack, Wiki systems, or version control systems such as Github. This shift entails several challenges. For example, the technical characteristics and the affordances of such infrastructures can enable certain collective actions while limiting others in the coordination of collaborative work [1].

If the platform and the text produced by interacting users can signal to others what needs to be done through stigmergic mechanisms, it has also been proven that more complex actions, such as decisions on how things should be done or negotiations on certain content, require explicit communication [3].

Communication is not only important for its content, but also for its form, including the emotions and social cues with which it is enriched. Research has highlighted the impact that the expression of certain emotions can have on the effectiveness of cooperative efforts. Affective ex-

pression is used to communicate information on the goals, beliefs and intentions of an individual towards social exchange [8].

The expression of emotions in a cooperative context is influenced by two major categories of factors. The first group pertains to the subject's demographics, while the second group examines the subject's context and their reaction to it. Research on computer-mediated communication (CMC) in online environments has demonstrated that individuals communicate primarily through asynchronous text. This text-only interaction omits typical social clues used by a group to interpret the messages, such as the demographics or social roles. The given textual corpus is the sole source of contextual information and has been found to constitute a set of tacit guidelines to which users tend to adhere [7]. Considering the significance of expressed emotions in a working group, it is crucial to comprehend the role that context plays in the expression of emotions in cases where context is reduced to the bare essentials, *i.e.* to the text of asynchronous interaction. In particular, considering the aggressiveness witnessed in online communities, it is important to understand the factors that influence positive or negative expressed emotions. This understanding can help improve comprehension of cooperative dynamics in CMC environments.

We will thus focus on the second group that encompasses theories such as appraisal, mimicry, and Communication Accommodation Theory (CAT). Appraisal theory suggests that an individual's expressed emotion is a result of their reaction and evaluation of an external stimulus, which is influenced by their personal beliefs and values. Mimicry theory argues that individuals tend to imitate the vocabulary, emotions, and expressions of those around them. According to CAT, the ultimate goal is to signal shared appraisals and values [2].

This research aims to investigate the factors that influence the emotions expressed by users through text in Wikipedia Talk Pages. Specifically, we aim to examine the relationship between individual emotional tendencies and their reactions to the context in which they interact, using the concepts of Appraisal and Communication Accommodation theory. Our decision to use Wikipedia as a research setting is motivated by three factors. Firstly, users are responsible for managing conflicts themselves as there is no hierarchical conflict resolution mechanism [6]. Secondly, users voluntarily choose to participate in the collaborative process as there are no standard incentives such as financial or career benefits [5]; furthermore, there are no hierarchical obligations, such as job requirements, that force users to moderate their tone or reach an agreement [6]. Finally, the interactions can be traced as the whole history of comments and modifications is publicly available, providing a significant amount of data. In contrast to previous studies on Wikpedia talk pages[4], and for the reasons given above, this research will focus on the contextual characteristics of the comments rather than the socio-demographic characteristics of the subjects. The aim is to understand how the context influences the emotions expressed in the comments. Specifically, this study will use the dataset developed by [9] to examine two emotional dimensions: valence and arousal. The first term denotes the degree of pleasantness of the word, while the second term denotes its emotional charge, whether positive or negative.

The aim of this research is to investigate the effect of the contextual valence (and arousal) on the emotional expression of a new comment posted in that context. In addition, we aim to explore the relationship between a subject's emotional tendencies and the context in which they discuss by examining the effect of perceived valence (arousal) on emotional expression.

The analysis was conducted on 1.2 million comments gathered from 312.000 Wikipedia articles. Our strategy for selecting articles aimed to create a sample that represented a variety

of subjects, team sizes and levels of conflict. Each observation - a comment posted on a talk page - is enriched with contextual characteristics at the level of the page, thread, and the possible previous relationships between sender and receiver. Additionally, characteristics of the overall activity of the comment author are taken into account. In line with [4], we employed two techniques, namely Word Frequency Averaging (WFA) and Word Embeddings (EMB), to assess the emotions conveyed in each comment and its surrounding context. The purpose of using these techniques was to address any potential data sparsity in the dataset created by [9]. Logistic regression models were employed to analyse comments with exceptionally high or low valence (arousal) and identify the contextual or behavioural factors that may have contributed to the expression of extreme emotions. During the analyses, we considered three contextual levels of analysis, *i.e.* the discussion page, the thread and the dialogue between the author and the receiver of the observed comment. The results support the hypothesis that the valence (arousal) of the context positively influences the likelihood of an extremely high valence (arousal) comment. The thread level appears to have a greater effect than other levels. It is suggested that this is because authors must take into account previous comments in order to participate in the discussion. The limited impact of the dyadic level may be due to its limited occurrence in the dataset. A similar approach was utilized for the study of the variation of the comments authors' valence (arousal) compared to their usual valence (arousal). The analysis yielded similar results. There is indeed an overall positive impact of the context on the variation. The thread level appears to have the stronger impact compared to the other contextual levels.

There are limitations to this research. On the one hand, only two emotions are considered, which limits the scope of the managerial implications of this study. On the other hand, the setting is limited to a single project, with the risk of context-specific effects. But it proposes a new method to investigate, among other things, the emotional discrepancy between the subject's idiosyncratic emotional tendency and the emotion conveyed in the context that can be transposed to other context. The data set can also be reused with other, more complex emotion analysis tools.

References

[1] Anders A. (2016) Team communication platforms and emergent social collaboration practices. International Journal of Business Communication 53(2):224–261.

[2] Bernhold Q. S., Giles H. (2022) Emotional mimicry: a communication accommodation approach. Cognition and Emotion 36(5):799–804.

[3] Dipple A., Raymond K., Docherty M. (2014) General theory of stigmergy: Modelling stigma semantics. Cognitive Systems Research 31:61–92.

[4] Gallus J., Bhatia S. (2020) Gender, power and emotions in the collaborative production of knowledge: A large-scale analysis of wikipedia editor conversations. Organizational Behavior and Human Decision Processes 160:115–130.

[5] Klapper H., Reitzig M. (2018) On the effects of authority on peer motivation: Learning from wikipedia. Strategic management journal 39(8):2178–2203

[6] Lerner J., Lomi A. (2020) The free encyclopedia that anyone can dispute: An analysis of the micro-structural dynamics of positive and negative relations in the production of contentious wikipedia articles. Social Networks 60:11–25.

[7] Rösner L., Krämer N. C. (2016) Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments. Social Media+ Society 2(3).

[8] Van Kleef G. A. (2009) How emotions regulate social life: The emotions as social information (easi) model. Current directions in psychological science 18(3):184–188.
[9] Warriner A. B., Kuperman V., Brysbaert M. (2013) Norms of valence, arousal, and dominance for 13,915 english lemmas. Behavior research methods 45:1191–1207.

The Atlas of Social Complexity

Brian Castellani1[1✓], Lasse Gerrits[2]

[1] *Durham University, UK; brian.c.castellani@durham.ac.uk*
[2] *University of Rotterdam, Netherlands; gerrits@ihs.nl*
[✓] *Presenting author*

Abstract. The purpose of this presentation is to introduce attendees to our forthcoming book, *The Atlas of Social Complexity* (June 2024, Edward Elgar, https://www.atlassocialcomplexity.org), which maps the latest advances in the study of social complexity – including five major transdisciplinary themes and 24 leading-edge areas of research – based on what we see as the new *social science turn* in the complexity sciences.

Keywords. *Social complexity history; Cognition, emotions and consciousness; Psychology of complexity; Living in social systems; Complexity methods*

1 EXTENDED ABSTRACT

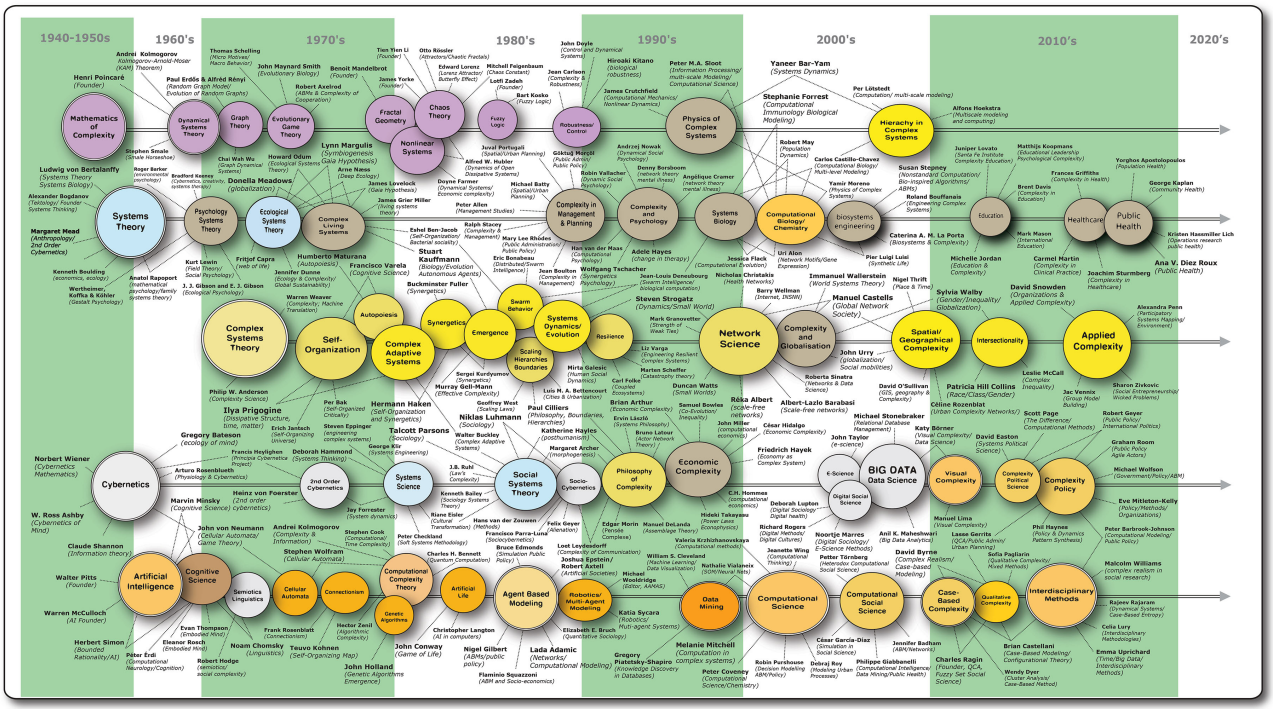


Figure 1: Map of the Complexity Sciences.

1.1 The Challenges of Studying Social Complexity

As Figure 1 shows, the complexity sciences have evolved over the past several decades into an expansive field of study that crosses over just about every major area of academic research. It also presently involves almost all of the latest advances in method and multi-methods, particularly computational modelling (e.g., Mitchell, 2009).

Within the social sciences, the complexity sciences have had an impact on social inquiry, going all the way back to the Macy Conferences and the emergence of systems theory and second-order cybernetics – or, in some ways, even further, with the work of Pareto and Spencer. Still, the most recent and longstanding impact started with the *complexity turn* in social science research in the late 1990s (Byrne and Callaghan 2022; Urry 2005).

Since then, while the complexity sciences have done much to advance social science, over the last decade the field has run into some considerable situations – thirteen to be exact. As shown in Figure 2, examples include complexity scientists ignoring social science; privileging computational modelling over qualitative inquiry; failing to address issues of power and inequality in social-ecological systems; and being tone deaf about the real world. These situations prevent the study of social complexity from becoming the disruptive, transdisciplinary field it originally sought to be in the 1950s and, more recently, the 1990s when the complexity turn in the social sciences took place.

Situation	Characteristics
1. No philosophy of complexity	Few attempts to define an epistemology and ontology for social complexity
2. A failure to engage the wider social sciences	Assumption that the social sciences can be ignored because the complexity sciences would offer superior insights
3. Reinventing the wheel	Reinventing existing insights from the social sciences that are then presented as new insights
4. Old words, new words	Rebranding existing insights using terms from the complexity sciences
5. Obscurantism and mystification	Scientific overreach and complicated jargon combine to suggest that life's biggest questions are uncovered
6. Forgetting multilevel thinking and modelling	Despite the transdisciplinary approach of social complexity, almost all research focuses on a single level of analysis.
7. Technique in the absence of theory	Focus on computational methods and big data pushes social theory out of sight
8. Learning tools vs. predictive machines	The ability to learn from simulations is replaced by a desire to predict and control social complexity
9. Minor role of qualitative research	Dominance of quantitative research and quantification of data established a blind spot for qualitative data and methods
10. Methodological closing of social scientific mind	Shying away from advances in computational methods sees many social scientists becoming illiterate with such methods
11. The dire sound of technicalities	Going into a spiral of ever-smaller technical refinement while losing the bigger picture out of sight.
12. Being tone-deaf about the real world	Advanced analyses are coupled to crude recommendations that fail to appreciate the complexity in the target domain
13. Practice does not make perfect	Pragmatic and rushed adoption of the complexity sciences by practitioners constitutes verbal detritus

Figure 2: The Thirteen Situations of Social Complexity Research.

1.2 The Social Science Turn in Complexity Studies

Fortunately, a small but growing global network of scholars are charting new territories for the study of social complexity. We call this the social science turn. This 'turn' fosters a transdisciplinary, social complexity imagination that, in one way or another, addresses the field's thirteen situations to create new areas of disruptive and highly innovative social inquiry. The Atlas of social complexity charts this new territory.

1.3 The Future History of Social Complexity

The Atlas of social complexity organises the future history of social complexity research into six major themes – (1) understanding the history of social complexity research and its current 13 situations; (2) Cognition, emotion and consciousness, (3) Dynamics of human psychology, (4) Living in social systems, (5) Advancing a new methods agenda, and (6) The unfinished space. Within and across these themes, the Atlas surveys over twenty-four leading-edge research areas (some still under construction) that readers can variously combine and develop to pursue their own work. As shown in Figure 3, which highlights the research covered in Theme 4, topics range from immune system cognition and network theories of psychopathology to configurational and intersectional social science to the complexities of place and governance to resilience and economics in an unstable world.

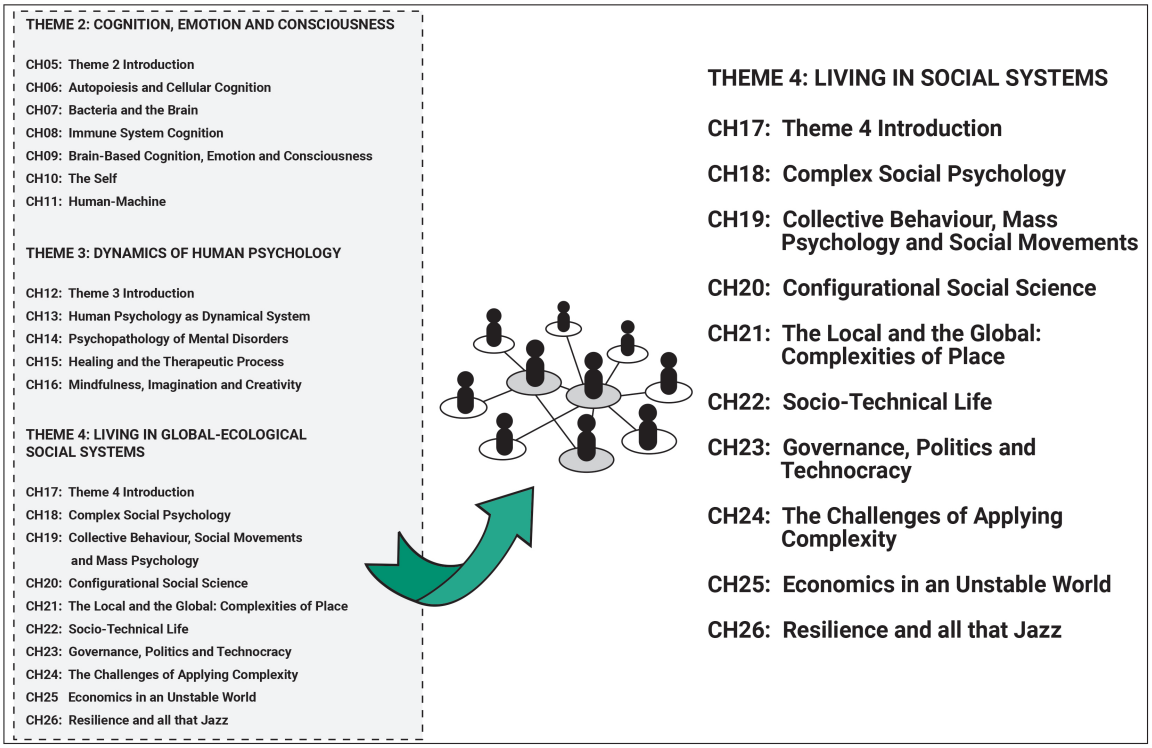


Figure 3: Theme 3: Living in Social Systems.

For those looking to get past the normalising conventions of the complexity sciences (particularly postgraduate students and early career researchers) in search of new ideas and new ways of working, the tour taken by the Atlas should prove of some value.

1.4 Current Presentation

The purpose of our presentation is to introduce attendees to *The Atlas of Social Complexity*. We cannot obviously survey the entire book, so instead will focus on a brief introduction to the 13 situations and then introduce the five major themes, focusing in on Theme 4 (Living in social systems), which is the most congruent with the focus of FRCCS2024.

1.5 Endorsements for the Atlas of Social Complexity

*The Atlas of Social Complexity*was peer reviewed. Here are endorsements by leading figures in social complexity research:

'Many have observed that in the social sciences, everything is connected to everything else but so far we have been singularly unsuccessful in attempting to explain the richness and diversity of this interconnected world. Glimpses of such explanations have come from the sciences of complexity but much of this reasoning has been contained within the traditional straitjacket of the physical sciences. What Castellani and Gerrits have done is to produce an Atlas of this world, through a series of maps that guide the reader to a great array of disciplines that can be informed by a multitude of ideas that they define as social complexity. This is a remarkable commentary on our progress in dealing with complex systems in all their guises and it is essential reading for everyone who seeks an understanding of our interconnected world.' – Michael Batty, University College London, UK

'This book is not just an invaluable Atlas to the extensive and fascinating literature on social complexity, but also an opinionated (in the best way) tour of the landscape, its heights and its depth and its quirks. The authors have read widely, thought carefully and explained clearly a broad sweep of research and practice on the idea of complexity and its application in the social, psychological and economic sciences. The book will be invaluable to academics, researchers, and policy analysts intrigued by how a social complexity approach might aid in the understanding of our complex world.' – Nigel Gilbert, University of Surrey, UK

'An inspiring read for believers and non-believers. Whether or not you agree that complexity is what it is all about, this book formulates a great set of challenges to spark renewal in the interdisciplinary social sciences. Covering a wide terrain from cognition to ecology and intersectionality, it charts a set of adventurous routes through recent research to show how a sociologically informed complexity science can meaningfully address the questions that matter.' – Noortje Marres, University of Warwick, UK

'This is a superb review of the development of social complexity in the social sciences and is a must read for anyone interested in cutting-edge social theory. Castellani and Gerrits convincingly show that this set of concepts is being transformative of social science thinking across multiple disciplines, even if it is developing too slowly.' – Sylvia Walby, Royal Holloway, University of London, UK

"This book stands as a formidable achievement, a true tour de force wherein the authors delve into our complex social world. They unravel the intricacies from the molecules comprising our cells to those shaping our bodies and ultimately forming us as conscious individuals. These individuals, in turn, have pioneered, discovered, and advanced technologies such as electronics, thinking robots, nuclear power, the contraceptive pill, and antibiotics. Collectively, they shape a complex society that, with an accelerating pace of change, achieved remarkable feats like landing a man on the Moon, eradicating smallpox, and establishing the World Wide Web. And, yes, also a society that kills its brothers and sisters and destroys its own natural environment much faster than it can reason about it. A profound sense of urgency emerges to comprehensively grasp these intricacies of our human society, considering its multifaceted interactions. Much like the Greek Titan, this Atlas bears the weight of the world and its remarkable inhabitants, and – guided by the science of complexity – offers new qualitative and quantitative avenues to make sense of this amazing world we live in." – Peter Sloot, University of Amsterdam, the Netherlands

1.6 REFERENCES

Byrne, D., and Callaghan, G. (2022). Complexity theory and the social sciences: The state of the art. Routledge.

Castellani, B. and L. Gerrits (2024). The Atlas of Social Complexity. Edward Elgard.

Mitchell, M. (2009). Complexity: A guided tour. Oxford university press.

Urry, J. (2005). The complexity turn. Theory, culture and society, 22(5), 1-14.

FRCCS

What Are Data-Driven Methods Missing? A Physics-Guided Learning Approach for Predicting Chaotic Systems Dynamics

Liu Feng[1,2✓], Yang Liu[2], Benyun Shi[1] and Jiming Liu[2*]

[1] *College of Computer and Information Engineering, Nanjing Tech University, Nanjing, China*
[2] *Department of Computer Science, Hong Kong Baptist University, Hong Kong, China*
{cslfeng,csygliu,jiming}@comp.hkbu.edu.hk, benyunshi@outlook.com
[*] *Corresponding Author*
[✓] *Presenting author*

Abstract. Predicting the dynamics of chaotic systems is crucial across various practical domains, including the control of infectious diseases and responses to extreme weather events. Such predictions provide quantitative insights into the future behaviors of these complex systems, thereby guiding the decision-making and planning within the respective fields. Recently, data-driven approaches, renowned for their capacity to learn from empirical data, have been widely used to predict chaotic system dynamics. However, these methods rely solely on historical observations while ignoring the underlying mechanisms that govern the systems' behaviors. Consequently, they may perform well in short-term predictions by fitting the data, but their long-term predictive reliability is compromised, particularly in chaotic systems, where slight initial variations may result in substantial differences in a finite number of time steps. To address this challenging issue, in this paper, we propose a novel Physics-Guided Learning (PGL) method. The proposed method aims to synergize observational data with the governing physical laws of chaotic systems to predict the systems' future dynamics. By fusing data-driven insights with physics-guided principles, this method utilizes a deep neural network architecture to enhance prediction accuracy. Empirical validation on four dynamical systems, each exhibiting unique chaotic behaviors, demonstrates that PGL achieves lower prediction errors than existing benchmark predictive models. The results highlight the efficacy of our design of data-physics integration in improving the precision of chaotic system dynamics forecasts.

Keywords. *Physics-Guided; Data-Driven; Chaotic Systems; Dynamics Prediction.*

1 Introduction

Chaotic systems are ubiquitous, from academic research in physics [1, 2] and chemistry [3, 4] to real-world domains such as epidemiology [5, 6] and climatology [7, 8]. By predicting the dynamics of these systems, we can gain valuable insights into their future behaviors, which can not only help us understand the underlying mechanisms of these systems but, more importantly, effectively inform and guide the decision-making process in real-world problems within the respective fields. For example, forecasting the dynamical behaviors in the spread of epidemics can help us uncover the disease transmission patterns and, accordingly, deploy effective inter-

vention strategies to control the infectious diseases [9]. Predicting the dynamics of variables in the climate system, such as temperature and precipitation, can help us be well prepared for extreme weather events [10].

In recent years, with the availability of large amounts of data and the advancement of computing power, many studies have utilized data-driven approaches to analyze and predict the dynamics of chaotic systems. These methods generally utilize the given data to learn the mapping function between historical observations and the future value of the target variable, and then use the learned mapping function to conduct the prediction. Typical data-driven methods that have been widely used in chaotic system dynamics prediction include long short-term memory networks(LSTM) [11, 12], reservoir computing [13, 14], etc. The above methods have been proven to be effective for the short-term prediction of chaotic systems, demonstrating an ability to capture the instantaneous dynamics [15]. However, their ability to make long-term predictions is limited, especially for those rapidly evolving chaotic dynamical systems, where even a slight initial variation can result in significant differences as the evolution over time [16]. The reason could be that such data-driven methods rely solely on historical observations during the learning process but ignore the underlying mechanisms of chaotic systems, which are, in fact, of great importance in characterizing the systems' dynamical behaviors.

To overcome the limitations of data-driven models in predicting chaotic system dynamics, we introduce a novel method in this paper, designated as Physics-Guided Learning (PGL). Inspired by a recently developed physics-informed neural network (PINN), which was originally designed for solving forward and reverse problems in nonlinear partial differential equations [17], our PGL method seeks to synergize observational data with the governing physical laws of chaotic systems. Specifically, the architecture of PGL is composed of three integral components: a data-driven component that learns the dynamical patterns and mapping functions from historical observations, a physics-guided component that exploits and represents systems' governing mechanisms, and a nonlinear learning component that integrates the output from the data-driven component and that from the physics-guided component in a proper way. The objective functions of these three components will be jointly optimized to achieve the desired goal of chaotic dynamics prediction.

The remainder of this paper is organized as follows. Section 2 outlines the proposed methodology, with a detailed explanation of its core principles, architecture design, and learning processes. In Section 3, we present the settings and results of our experiments on four typical chaotic systems, which are designed to validate the effectiveness of the proposed method in the task of chaotic dynamics prediction. Finally, we conclude our work in Section 4.

2 Methodology

In this section, we will outline the formalism and computational mechanism of the proposed PGL method. We begin by defining the learning problem and providing an overview of the method. Subsequently, we present the mathematical definition and formulation of the proposed method for chaotic system dynamics prediction, which integrates data and physical understanding. To enhance the clarity, we detail the method's structure, workflow, and objective function.

2.1 Problem Statement

First, we state the definition of chaotic system dynamics prediction. For a chaotic system with N state variables, we represent the system's state observations at time t as $\mathbf{X}_t = [x_t^1, x_t^2, \ldots, x_t^N]$.

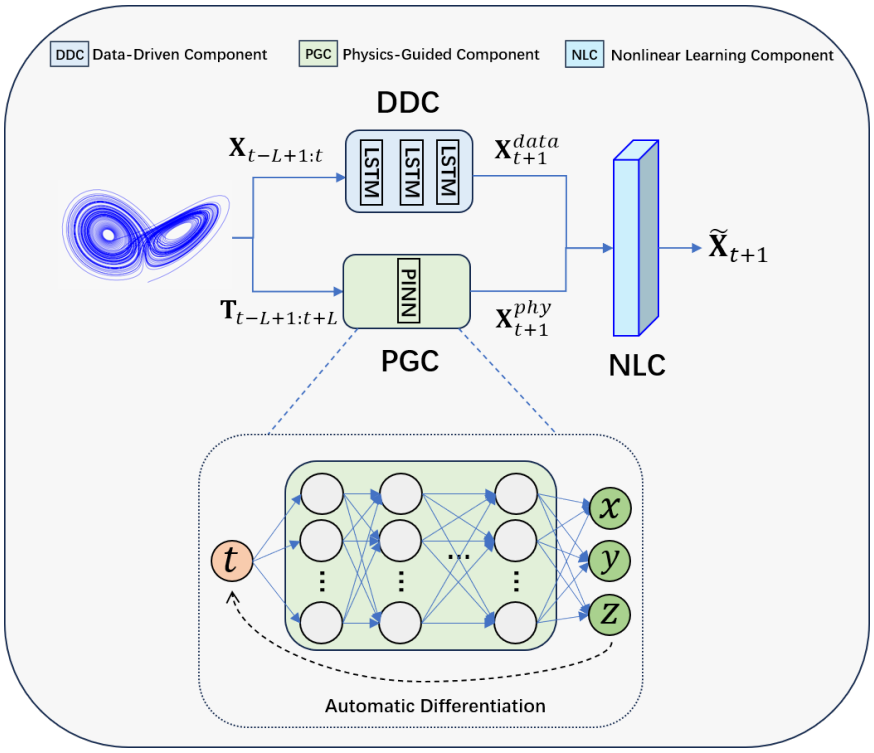


Figure 1: Illustration of the architecture of the proposed method PGL, which is composed of three core components: a data-driven component (DDC), a physics-guided component (PGC), and a nonlinear learning component(NLC).

$\mathbf{X}_{t-L+1:t} = [\mathbf{X}_{t-L+1}, \mathbf{X}_{t-L+2}, \ldots, \mathbf{X}_t]$ denotes the historical data containing L time steps. Meanwhile, the time point sequence $\mathbf{T}_{t-L+1:t} = [t-L+1, t-L+2, \cdots t]$ corresponding to the system's state value sequence $\mathbf{X}_{t-L+1:t}$ is also recorded. The target of chaotic system dynamics prediction is to learn the underlying state transition function and the potential dynamics of the system based on the historical data and governing physical laws, and then forecast the subsequent state of the chaotic system, denoted as $\tilde{\mathbf{X}}_{t+1}$. To achieve this goal, we devise a PGL method that makes use of both the observational data and the underlying dynamical mechanism of the chaotic system. Specifically, the proposed method comprises three core components: a data-driven component (DDC), a physics-guided component (PGC), and a nonlinear learning component (NLC). In the subsequent section, we will furnish a more detailed exposition of our design.

2.2 Physics-Guided Learning

Figure 1 illustrates the architecture of the proposed method PGL, consisting of DDC, PGC and NLC. For the DDC, we use a three-layer LSTM with 20 hidden units each, followed by a dense layer. For the PGC, we refer to the PINN configuration [17], using a 10-layer neural network with 32 neurons in each layer. The structure of the NLC is a multilayer perception (MLP) [18] with two layers: one input layer and one output layer. Next, we will elaborate in detail on how these three components work together to predict the dynamical behaviors of chaotic systems.

Data-Driven Component Firstly, we obtain the prediction of the data-driven branch for the next time step, denoted by $\mathbf{X}_{t+1}^{data} = DDC(\mathbf{X}_{t-L+1:t})$. We expect the long short-term memory (LSTM) structure in the DDC to capture both short-term and long-term temporal

dependencies in the historical state sequence through its unique gating mechanism and make predictions for the next time step.

Physics-Guided Component Afterward, we extend the $\mathbf{T}_{t-L+1:t}$, turning it into $\mathbf{T}_{t-L+1:t+L}$, which is further fed into the PGC. The PGC generates the system state predictions that are of equal length to the extended time sequence $\mathbf{T}_{t-L+1:t+L}$. This process is shown in the following equation:

$$\mathbf{X}^{phy}_{t-L+1:t+L} = PGC(t-L+1, t-L+2, ..., t+L), \tag{1}$$

where $\mathbf{X}^{phy}_i = \left[x^{phy}_i, y^{phy}_i, z^{phy}_i\right]$. We expect that, with the guidance of physical knowledge, the PGC can learn the dynamics of the system and assist the entire model in making predictions. Note that the design of PGC is general and can be used in various chaotic systems. Here, for a better explanation, we use the typical Lorenz system [16] as an example to show how the PGC works. The only information that we have is the form of the system's equations shown in the following Eq. (2), and we do not know the crucial initial values and system parameters.

$$\begin{aligned} \frac{dx}{dt} &= a(y-x), \\ \frac{dy}{dt} &= cx - y - xz, \\ \frac{dz}{dt} &= xy - bz. \end{aligned} \tag{2}$$

Following the work of physics-informed neural networks in [17], we utilize the automatic differentiation tools within the deep learning framework PyTorch [19] to compute the derivative of the PGC's output $\mathbf{X}^{phy}_{t-L+1:t+L}$ with respect to its input $\mathbf{T}_{t-L+1:t+L}$, yielding the following:

$$\frac{\partial \mathbf{X}^{phy}_{t-L+1:t+L}}{\partial t} = \left[\frac{\partial \mathbf{X}^{phy}_{t-L+1}}{\partial t}, \frac{\partial \mathbf{X}^{phy}_{t-L+2}}{\partial t}, \cdots, \frac{\partial \mathbf{X}^{phy}_{t+L}}{\partial t}\right], \tag{3}$$

where $\frac{\partial \mathbf{X}^{phy}_i}{\partial t} = \left[\frac{\partial x^{phy}_i}{\partial t}, \frac{\partial y^{phy}_i}{\partial t}, \frac{\partial z^{phy}_i}{\partial t}\right]$. We expect that the approximate derivatives conform to the definition of the Lorenz system, and therefore, we have calculated the residuals with respect to the physics-guided component, as shown below.

$$\begin{aligned} loss_{phy} &= \lambda_1 loss_x + \lambda_2 loss_y + \lambda_3 loss_z, \\ loss_x &= \sum_{i=t-L+1}^{t+L} \left| \frac{\partial x^{phy}_i}{\partial t} - \tilde{a}\left(y^{phy}_i - x^{phy}_i\right) \right|^2, \\ loss_y &= \sum_{i=t-L+1}^{t+L} \left| \frac{\partial y^{phy}_i}{\partial t} - \left(\tilde{c} x^{phy}_i - y^{phy}_i - x^{phy}_i z^{phy}_i\right) \right|^2, \\ loss_z &= \sum_{i=t-L+1}^{t+L} \left| \frac{\partial z^{phy}_i}{\partial t} - \left(x^{phy}_i y^{phy}_i - \tilde{b} z^{phy}_i\right) \right|^2, \end{aligned} \tag{4}$$

where λ_1, λ_2, and λ_3 are hyper parameters, $\tilde{a}$, $\tilde{b}$, and $\tilde{c}$ are trainable parameters of the model. Note that the true parameters of the chaotic systems remain unidentified for the PGC and for the proposed PGL model, a scenario that is typical in real-world applications. It is our expectation that the proposed model is capable of learning and characterizing the systems'

dynamics even in the presence of such uncertainties. Additionally, since we have the ground truth $\mathbf{X}_{t-L+1:t}$, we conduct supervised learning by minimizing the following $loss_{data}$:

$$loss_{data} = \frac{1}{L} \sum_{i=t-L+1}^{t} \left| \mathbf{X}_i^{phy} - \mathbf{X}_i \right|^2 . \tag{5}$$

By incorporating penalty terms based on physics and data, we hope that the PGC can rely on known physical knowledge and work in collaboration with the DDC to predict chaotic systems.

Nonlinear Learning Component Next, a nonlinear learning component will balance the predicted $\mathbf{X}_{t+1}^{data}$ and $\mathbf{X}_{t+1}^{phy}$ from DDC and PGC to provide the final prediction $\tilde{\mathbf{X}}_{t+1}$ for the system at the time step $t+1$. The loss is formulated as follows:

$$loss_{NLC} = \left| \tilde{\mathbf{X}}_{t+1} - \mathbf{X}_{t+1} \right|^2 , \tag{6}$$

where $\tilde{\mathbf{X}}_{t+1} = NLC\left(concatenate\left(\mathbf{X}_{t+1}^{data}, \mathbf{X}_{t+1}^{phy}\right)\right)$, and $\mathbf{X}_{t+1}$ denotes the ground truth value of the system's state variable at time step $t+1$, which serves as the label in our supervised learning. It is important to note that the data for $\mathbf{X}_{t+1}$ in Eq. (6) is exclusively accessible during the training phase. This information is not available during the testing phase, where the model must predict $\mathbf{X}_{t+1}$ without the aid of ground truth values.

In our implementation, the NLC utilizes the Rectified Linear Unit (ReLU) activation function to capture the nonlinear dependencies inherent in the data. The architecture is intentionally designed to be straightforward to affirm the feasibility of the proposed idea of integrating data-driven and physics-guided components. It should be noted that real-world data often exhibit more complex nonlinear relationships. Our model is designed with flexibility, allowing for the incorporation of more sophisticated neural network architectures to accommodate and adapt to these higher levels of complexity.

Objective Function The final optimization objective function, which takes account of both data and physics, is given as follows:

$$min(w_1 loss_{NLC} + w_2 loss_{data} + w_3 loss_{phy}), \tag{7}$$

where w_1, w_2, and w_3 are hyper parameters.

3 Experimental Results

In this section, we use four dynamical systems with different chaotic behaviors, i.e., the Rossler, Lorenz, Chua, and Chen systems, which are widely used in chaotic systems dynamics prediction [20–24], to validate the performance of the proposed PGL method in long-term forecasting of chaotic dynamics. We also perform an ablation study to analyze the contributions of different components of the proposed method to the chaotic dynamics prediction.

3.1 Descriptions of Chaotic Systems

Rossler System In 1976, Otto Rössler proposed the well-known Rossler system, which exhibits chaotic phenomena and nonlinear dynamical behavior. The system is defined by the

following differential equations [25]:

$$\begin{aligned} \frac{dx}{dt} &= -y - z, \\ \frac{dy}{dt} &= x - ay, \\ \frac{dz}{dt} &= b + xz - cz. \end{aligned} \tag{8}$$

Lorenz System In 1963, Edward Lorenz discovered the existence of a peculiar "butterfly effect" in meteorological systems when studying convective instability. The Lorenz system can be described by the following equations [16]:

$$\begin{aligned} \frac{dx}{dt} &= a(y - x), \\ \frac{dy}{dt} &= cx - y - xz, \\ \frac{dz}{dt} &= xy - bz. \end{aligned} \tag{9}$$

Chua System In 1986, Chua et al [26] introduced the Chua system, marking an advancement in the study of chaotic systems by linking chaos and nonlinear circuits. The equations of the Chua system are given as follows:

$$\begin{aligned} \frac{dx}{dt} &= a(y - x - G(x)), \\ \frac{dy}{dt} &= x - y + z, \\ \frac{dz}{dt} &= -by, \\ G(x) &= cx + (d + c)(|x + 1| - |x - 1|). \end{aligned} \tag{10}$$

Chen System In 1999, Chen et al [27] identified a chaotic attractor that bears similarities to the Lorenz system, but is topologically distinct in their research on chaotic control. The Chen system can be described by the following equations:

$$\begin{aligned} \frac{dx}{dt} &= a(y - x), \\ \frac{dy}{dt} &= (c - a)x - xz + cy, \\ \frac{dz}{dt} &= xy - bz. \end{aligned} \tag{11}$$

All the above four dynamical systems have nonlinear and chaotic behaviors, posing great challenges for long-term prediction. We use the fourth-order Runge–Kutta method with a step size of 0.01 to obtain the chaotic time series containing $10,000$ steps, which are divided into training, validation, and testing datasets in a ratio of $6 : 2 : 2$. Specifically, we utilize the data from the initial $6,000$ time steps for training purposes. This is followed by the subsequent $2,000$ time steps, which are designated for the validation process. Finally, we employ the data from the concluding $2,000$ time steps to test the performance of our model. Table 1 provides the details of system parameters and initial values.

Table 1: The system parameters and initial values of four chaotic systems used in our study.

System	Parameters	Initial values
Rossler	$a = 0.2, b = 0.2, c = 5.7$	$(x_0, y_0, z_0) = (1.0, 1.0, 1.0)$
Lorenz	$a = 10.0, b = 8/3, c = 28.0$	$(x_0, y_0, z_0) = (1.0, 1.0, 1.0)$
Chua	$a = 15.6, b = 25.28, c = -0.75, d = 0.47$	$(x_0, y_0, z_0) = (0.1, 0.1, 0.1)$
Chen	$a = 35.0, b = 3.0, c = 28.0$	$(x_0, y_0, z_0) = (0, 1.0, 0)$

3.2 Comparison Models and Evaluation Metrics

We select four representative methods as the baselines for performance comparison in our experiments. They are the long short-term memory (LSTM) [11], the echo state network (ESN) [28], the next generation reservoir computing method (NG-RC) [29], and DLinear [30]. Here, LSTM is a classic recurrent neural network model for time series prediction; ESN and NG-RC are representative methods specifically designed and widely used for chaotic system dynamics prediction, and DLinear is a state-of-the-art deep learning method developed for complex time series forecasting. For LSTM, we use a three-layer architecture with uniform hidden state size. To achieve its optimal performance, we experiment with a variety of hidden state sizes, specifically 8, 16, and 32, and report the best result. For ESN, we implement it with a spectral radius of 1.4 and a reservoir size of 300. For NG-RC and DLinear, we follow the default settings reported in their original papers.

When assessing the effectiveness of the methods in capturing and forecasting the dynamical behavior of chaotic systems over the long term, it is a common practice to employ the model's own prediction as the input for forecasting subsequent time steps during the test phase. This iterative process can result in an increase in errors as the forecast horizon extends, especially in chaotic systems, where small deviations at the beginning can lead to significant differences in later outcomes. The mean absolute error (MAE) and root mean square error (RMSE) are used as evaluation metrics to measure the prediction performance. The MAE and RMSE are defined as follows:

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|\hat{y}_t - y_t|, \tag{12}$$

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}, \tag{13}$$

where $\hat{y}_t$ denotes the predicted value of the model, y_t denotes the ground truth, and T is the corresponding forecast horizon.

3.3 Analysis of Results

Figure 2 demonstrates the comparison result of the ground truth of dynamics of the Rossler, Lorenz, Chua, and Chen systems in $2,000$ time steps, which is illustrated in blue in each subfigure, and the predictions generated by the proposed PGL method, which are shown in red. We can observe that the proposed PGL method is able to capture the dynamical patterns of these four chaotic systems. Although employing an iterative prediction process in the prediction phase brings great challenges to the task of long-term forecasting, the integration of data and physics enables our method to produce predictions that are consistent with actual dynamics.

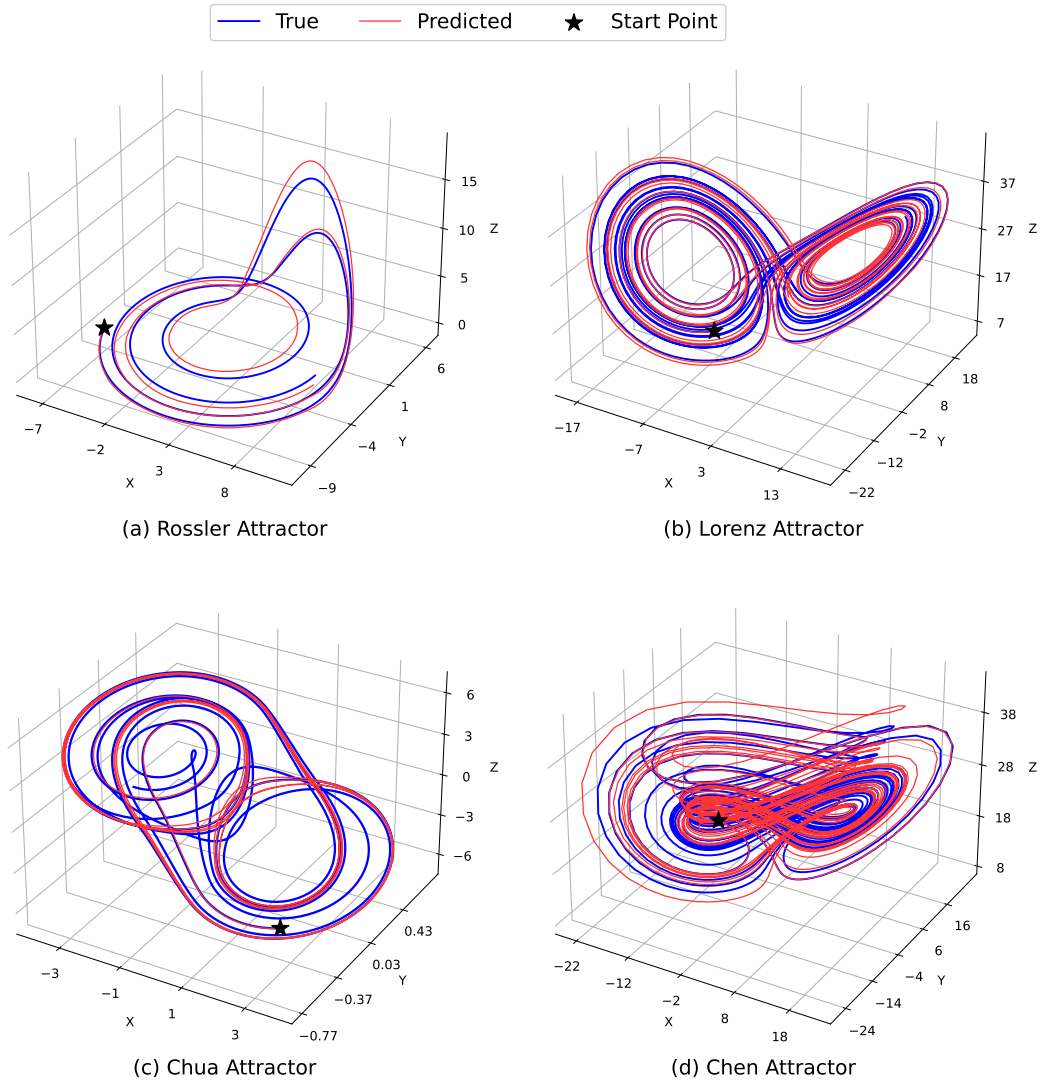


Figure 2: Comparison between the ground truth of dynamics of the Rossler, Lorenz, Chua, and Chen systems (blue) and the predictions generated by the proposed PGL method (red).

To further evaluate the performance of our predictions, we also conduct an analysis by visualizing the temporal evolution of the ground truth and predictions of the state variables in these chaotic systems in Fig. 3. For the Rossler system, the predicted curve closely resembles the ground truth, even for the irregular patterns in $Z(t)$, indicating that the proposed method successfully captures the dynamics of this chaotic system and thus is able to make accurate predictions in such a long-term period. For the Lorenz system, the predictions generated by our method are very close to the ground truth in the first $1,000$ time steps. Notable discrepancies between the predicted and actual values of the $X(t)$, $Y(t)$, and $Z(t)$ components are evident after around $1,000$, $1,000$, and $1,200$ steps, respectively. A possible reason is that the two nonlinear terms xz and xy in the Lorenz system make the dynamics more complex and harder to capture than the Rossler system, which has only one nonlinear term xz. We can observe similar patterns in the Chua and Chen systems: the long-term prediction results in these systems with multiple nonlinear terms, especially those after $1,200$ time steps, are not as accurate as the those in the Rossler system. We further observe that, compared to the Lorenz system, the

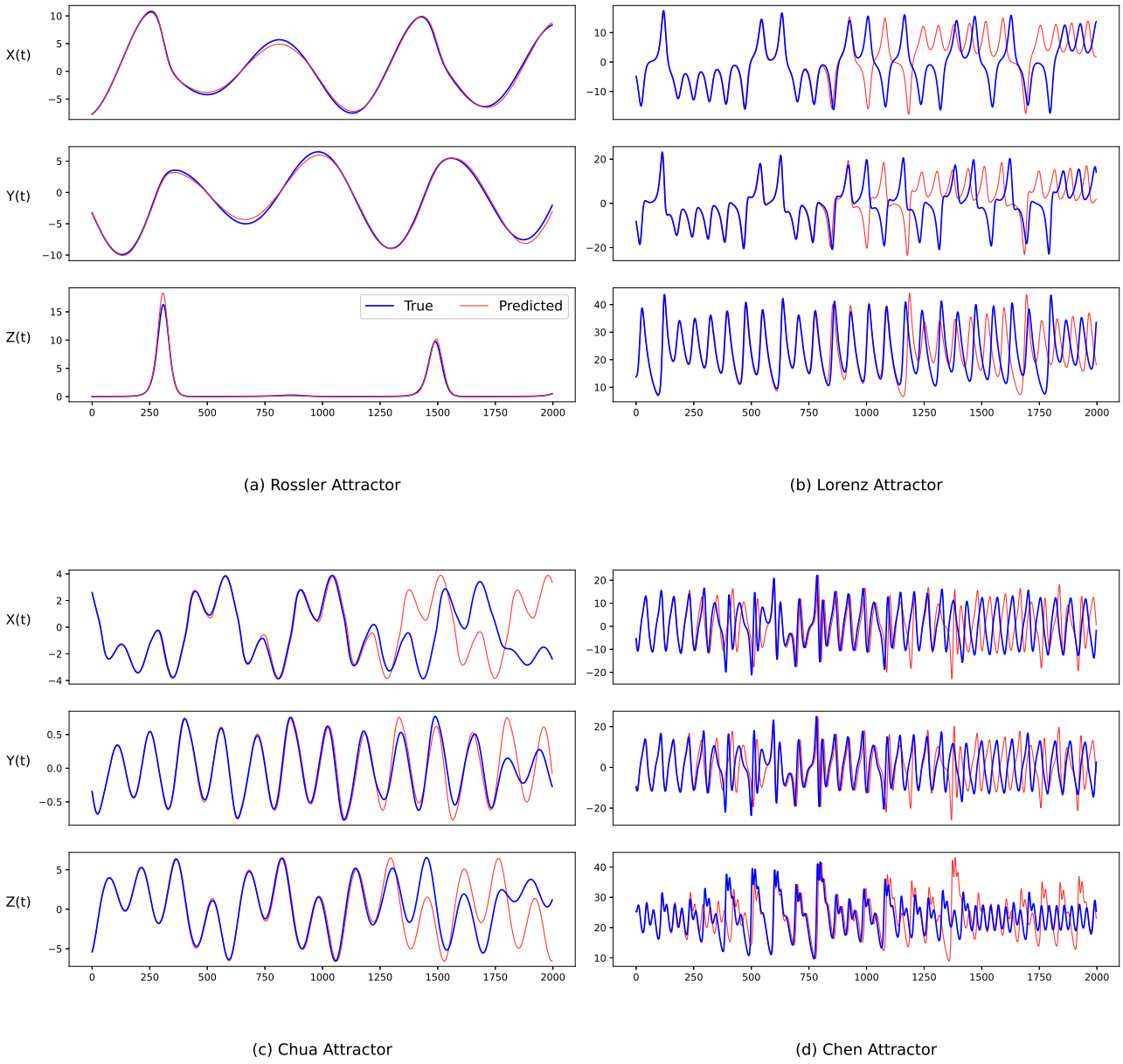


Figure 3: Comparison between the ground truth of the state variables of the Rossler, Lorenz, Chua, and Chen systems (blue) and the predictions generated by the proposed PGL method (red) over time.

model's predictions on the Chua system exhibit higher accuracy. Significant deviations in the components $X(t)$, $Y(t)$, and $Z(t)$ are observed only after $1,300$, $1,800$, and $1,300$ time steps, respectively. For the Chen system exhibiting more complex chaotic behavior, the proposed model generates reliable predictions within the first 250 time steps, followed by significant disturbances between 250 and 500 time steps. Fortunately, due to the proposed model's ability to balance between data and physical knowledge, it regains accuracy in predictions after the disturbances, maintaining precision up to about $1,300$ time steps.

To quantitatively compare the performance of our method with that of existing methods, we report the MAE and RMSE of all methods for different prediction horizons in Tables 2 and 3, respectively. The results demonstrate that the proposed method achieves the lowest prediction errors in most of the settings, demonstrating the effectiveness of our method in making long-

Table 2: MAE of LSTM, ESN, NG-RC, DLinear, and the proposed PGL in different prediction horizons on four chaotic systems. The best performance in each setting is highlighted in bold.

Systems	Horizon	LSTM	ESN	NG-RC	DLinear	PGL(Ours)
Rossler	200-horizon	0.295	0.228	0.046	2.374	**0.055**
	600-horizon	0.934	1.411	0.865	4.318	**0.182**
	1000-horizon	0.705	3.126	1.084	4.625	**0.236**
	1400-horizon	0.583	3.919	1.199	5.111	**0.208**
	2000-horizon	0.744	4.713	1.942	5.341	**0.225**
Lorenz	200-horizon	2.668	0.631	1.105	7.748	**0.338**
	600-horizon	4.828	5.462	5.088	6.602	**0.616**
	1000-horizon	5.996	7.103	4.287	7.316	**1.189**
	1400-horizon	6.579	8.110	5.285	7.541	**3.685**
	2000-horizon	7.103	8.174	6.774	7.984	**5.536**
Chua	200-horizon	0.163	0.104	1.057	1.007	**0.023**
	600-horizon	1.372	0.999	1.815	1.321	**0.079**
	1000-horizon	1.382	1.417	1.967	1.568	**0.117**
	1400-horizon	1.242	1.855	2.047	1.622	**0.285**
	2000-horizon	1.281	2.187	1.948	1.620	**0.912**
Chen	200-horizon	5.531	3.093	3.770	4.876	**0.300**
	600-horizon	8.085	6.480	7.906	6.182	**4.439**
	1000-horizon	7.349	7.859	8.888	6.576	**4.526**
	1400-horizon	8.026	8.323	8.731	6.604	**5.146**
	2000-horizon	7.996	8.012	9.166	**6.484**	6.535

term predictions of the chaotic system dynamics.

3.4 Ablation Study

In this subsection, we conduct an ablation study to understand the individual contributions of the different components within our proposed method to the prediction of chaotic dynamics. Specifically, we examine the performance of the Lorenz system dynamics prediction using four distinct configurations of our method: (1) employing only the DDC, which is an LSTM network; (2) using solely the PGC, represented by a PINN structure; (3) integrating both DDC and PGC through a simple linear combination, termed PGL-Linear; and (4) implementing the full proposed method as described in this manuscript, referred to as PGL. The PGL-Linear setting is essentially a simplified version of PGL, where it linearly aggregates the outputs of DDC and PGC using a fixed weight to generate the final prediction. In this ablation study, we simulate the Lorenz system for 500 steps, with the same system parameters and step size as in prior experiments. To better capture the breadth of the chaotic dynamics within a limited timeframe, we select a different initial condition of $(5.0, 5.0, 5.0)$. We allocate the initial 300 steps of data for training the model and the remaining 200 steps for testing its predictive capacity.

Table 4 presents the results of the ablation study with respect to MAE and RMSE across various forecast horizons. The results indicate that the data-driven component (denoted as DDC) alone yields satisfactory predictions in the short term, specifically for the initial 60 steps. However, beyond this range, the prediction error increases significantly. In contrast, the integration of a physics-guided component (denoted as PGC), as implemented in our proposed PGL framework,

Table 3: RMSE of LSTM, ESN, NG-RC, DLinear, and the proposed PGL in different prediction horizons on four chaotic systems. The best performance in each setting is highlighted in bold.

Systems	Horizon	LSTM	ESN	NG-RC	DLinear	PGL(Ours)
Rossler	200-horizon	0.354	0.277	0.053	3.103	**0.064**
	600-horizon	1.692	2.435	1.483	5.370	**0.315**
	1000-horizon	1.371	5.050	1.637	5.562	**0.373**
	1400-horizon	1.186	5.789	1.723	5.963	**0.331**
	2000-horizon	1.352	6.670	3.209	6.133	**0.341**
Lorenz	200-horizon	4.511	0.922	1.624	9.884	**0.561**
	600-horizon	7.834	8.372	8.693	8.550	**0.891**
	1000-horizon	8.900	9.667	7.773	9.410	**2.864**
	1400-horizon	9.401	10.560	8.689	9.638	**7.405**
	2000-horizon	9.967	10.595	10.067	10.050	**8.982**
Chua	200-horizon	0.209	0.123	1.398	1.274	**0.027**
	600-horizon	2.021	1.617	2.274	1.569	**0.111**
	1000-horizon	1.992	2.038	2.402	1.854	**0.162**
	1400-horizon	1.800	2.515	2.468	1.919	**0.535**
	2000-horizon	1.795	2.855	2.350	1.918	**1.495**
Chen	200-horizon	8.174	5.487	5.591	6.229	**0.483**
	600-horizon	10.476	8.944	9.890	7.649	**7.298**
	1000-horizon	9.738	10.251	10.824	8.165	**6.953**
	1400-horizon	10.448	10.593	10.778	8.138	**7.680**
	2000-horizon	10.151	10.322	11.147	**7.910**	9.124

Table 4: MAE and RMSE of DDC, PGC, PGL-Linear, and PGL in different prediction horizons on the Lorenz system. The best performance in each setting is highlighted in bold.

Metrics	Horizon	DDC	PGC	PGL-Linear	PGL
MAE	20-horizon	0.329	**0.100**	0.531	0.244
	60-horizon	0.742	1.689	0.945	**0.639**
	100-horizon	1.421	7.511	1.422	**0.679**
	140-horizon	1.733	9.243	1.953	**0.828**
	200-horizon	2.856	12.306	3.274	**1.465**
RMSE	20-horizon	0.363	**0.110**	0.568	0.300
	60-horizon	0.937	3.312	1.091	**0.838**
	100-horizon	1.825	10.854	1.831	**0.842**
	140-horizon	2.180	12.214	2.514	**1.005**
	200-horizon	3.800	15.126	4.510	**2.072**

consistently delivers substantially lower prediction errors across all considered horizons, thereby affirming the effectiveness of the proposed design. The pure PGC demonstrates remarkable accuracy in the very short-term predictions (in 20-40 steps), as detailed in Table 4 and illustrated in Fig. 4(b). This is attributed to its powerful capability in simulating and learning the governing differential equations that determine the dynamics of the system. However, the PGC itself is highly sensitive to initial perturbations; in the absence of data-driven regularization, even minor discrepancies at the onset of the testing phase can lead to substantial deviations

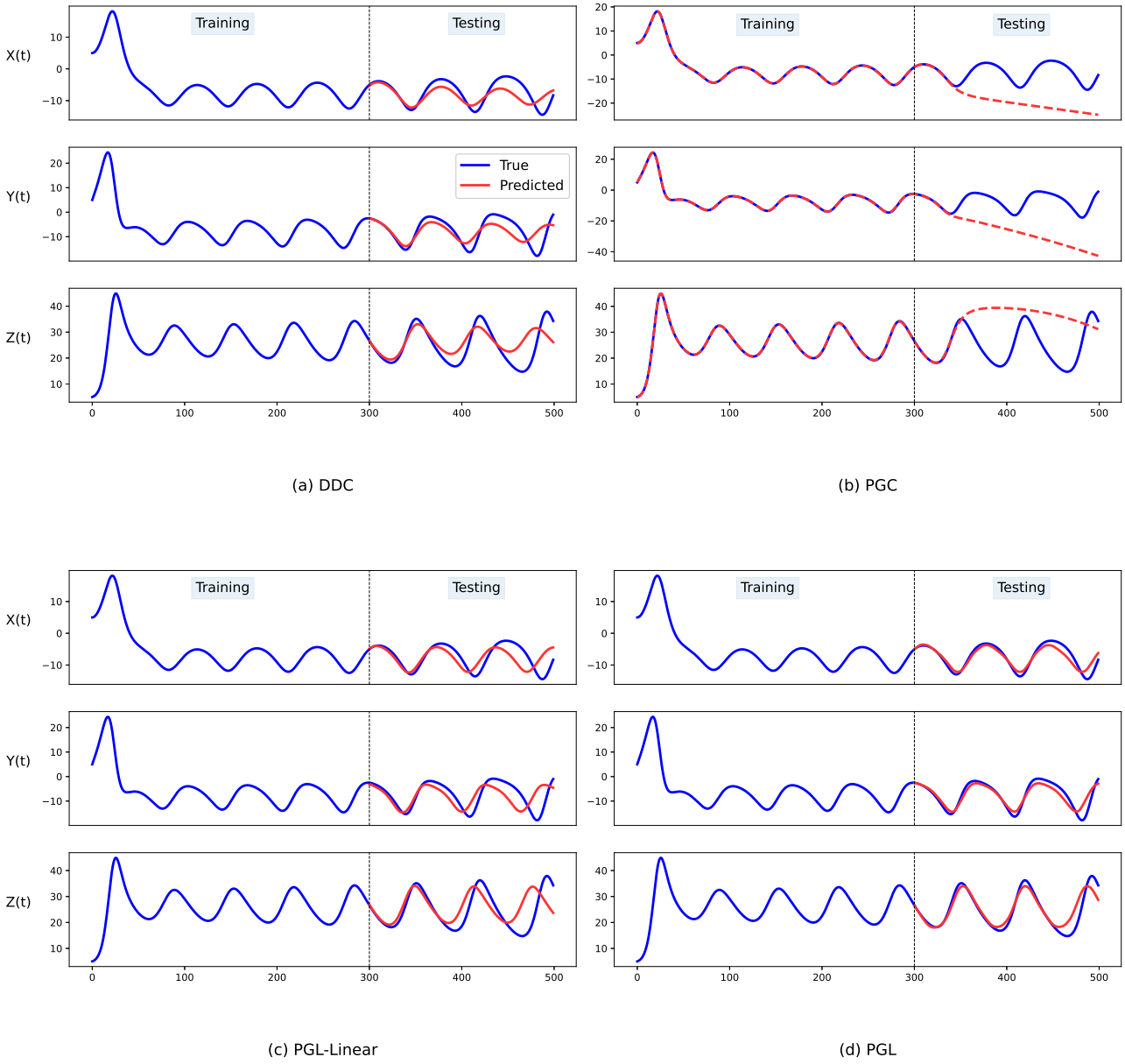


Figure 4: Comparison between the ground truth of the state variables of the Lorenz system (blue) and the predictions (red) generated by DDC, PGC, PGL-Linear, and PGL.

in subsequent predictions. Notably, the hybrid approach, PGL-Linear, which employs a linear combination of outputs of DDC and PGC, does not perform as well as anticipated. This underperformance suggests that the relationship between observational data and the physical principles governing the system's dynamics is likely to be inherently nonlinear. Consequently, a simplistic linear combination may not be adequately equipped to capture the complexity of such interactions, showing the necessity of the design of the proposed NLC in integrating the DDC and PGC to improve prediction accuracy.

4 Conclusion and Discussion

In this paper, we proposed a physics-guided learning approach to predict the dynamics of chaotic systems. We experimentally evaluated the performance of our method on the Rossler,

Lorenz, Chua, and Chen dynamical systems. The experimental results demonstrated that our method outperforms other baselines in terms of prediction accuracy.

Some existing studies have already made efforts in combining data and physical mechanisms for chaotic dynamics prediction. For example, PIESN [31] and its variant [32] encode the systems' governing equations into models' loss functions to penalize the predictions that do not obey the physics. Furthermore, several methods use physical knowledge to help reconstruct and predict the dynamics of a chaotic system with unmeasured variables [33, 34]. These methods, however, typically require complete and precise knowledge of the governing differential equations of the systems, including the equation parameters, to guide the predictive models effectively. In contrast, our research aims to forecast the dynamics of chaotic systems under the more challenging condition of incomplete knowledge of the true system parameters. By relaxing the requirement for full physical understanding, our proposed method offers practical advantages for a wide range of applications where only partial knowledge of the system's dynamics is available.

To our knowledge, PINN is among several representative techniques that employ neural networks to solve ordinary and partial differential equations. Other noteworthy methods include those based on the Deep Galerkin Method (DGM) [35, 36] and Neurodifferential approaches [37, 38], each offering unique contributions to the field. In our work, we utilize PINN as a typical example to demonstrate the efficacy of integrating data-driven structures with physical knowledge to accurately predict the dynamics of chaotic systems. This exemplification paves the way for further exploration into the integration of other physics-guided modules with data-driven components, potentially leading to enhanced predictive capabilities.

In our future work, we aim to attempt alternative ways to incorporate data and physical knowledge and extend our work to scenarios where observations are noisy and the underlying governing differential equations are not fully known in advance. Further, we intend to apply the proposed method to various real-world applications, such as infectious disease risk prediction, climate forecast, and traffic flow prediction.

Acknowledgements

This work was supported in part by the Ministry of Science and Technology of China (2021ZD0112502), in part by the National Natural Science Foundation of China and the Research Grants Council (RGC) of Hong Kong Joint Research Scheme (No. 62261160387, N_HKBU222/22), in part by the Hong Kong Research Grants Council General Research Fund (RGC/HKBU12201619, RGC/HKBU12202220, RGC/HKBU12203122), and in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant no. SJCX23_0435). We acknowledge the use of ChatGPT (GPT-4, OpenAI's language model: http://openai.com) in polishing some of the wordings in the manuscript. The final manuscript was edited and approved by all authors.

References

[1] Pecora, L. M., & Carroll, T. L. (1990). Synchronization in chaotic systems. Physical Review Letters, 64(8), 821.

[2] Grassberger, P., & Procaccia, I. (1983). Characterization of strange attractors. Physical Review Letters, 50(5), 346.

[3] Hess, B. (1990). Order and chaos in chemistry and biology. Fresenius' Journal of Analytical

Chemistry, 337, 459-468.
[4] Field, R. J. (1993). Chaos in chemistry and biochemistry. World Scientific.
[5] Aguiar, M., Kooi, B., & Stollenwerk, N. (2008). Epidemiology of dengue fever: A model with temporary cross-immunity and possible secondary infection shows bifurcations and chaotic behaviour in wide parameter regions. Mathematical Modelling of Natural Phenomena, 3(4), 48-70.
[6] Mishra, A. M., Purohit, S. D., Owolabi, K. M., & Sharma, Y. D. (2020). A nonlinear epidemiological model considering asymptotic and quarantine classes for SARS CoV-2 virus. Chaos, Solitons & Fractals, 138, 109953.
[7] Palmer, T. N. (1993). Extended-range atmospheric prediction and the Lorenz model. Bulletin of the American Meteorological Society, 74(1), 49-66.
[8] Olsen, P. E., Laskar, J., Kent, D. V., Kinney, S. T., Reynolds, D. J., Sha, J., & Whiteside, J. H. (2019). Mapping solar system chaos with the Geological Orrery. Proceedings of the National Academy of Sciences, 116(22), 10664-10673.
[9] Mangiarotti, S., Peyre, M., & Huc, M. (2016). A chaotic model for the epidemic of Ebola virus disease in West Africa (2013–2016). Chaos: An Interdisciplinary Journal of Nonlinear Science, 26(11).
[10] Toreti, A., Naveau, P., Zampieri, M., Schindler, A., Scoccimarro, E., Xoplaki, E., ... & Luterbacher, J. (2013). Projections of global changes in precipitation extremes from Coupled Model Intercomparison Project Phase 5 models. Geophysical Research Letters, 40(18), 4887-4892.
[11] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.
[12] Chattopadhyay, A., Hassanzadeh, P., & Subramanian, D. (2020). Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods: Reservoir computing, artificial neural network, and long short-term memory network. Nonlinear Processes in Geophysics, 27(3), 373-389.
[13] Jaeger, H. (2001). The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, 148(34), 13.
[14] Pathak, J., Hunt, B., Girvan, M., Lu, Z., & Ott, E. (2018). Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. Physical Review Letters, 120(2), 024102.
[15] Chantry, M., Christensen, H., Dueben, P., & Palmer, T. (2021). Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft AI. Philosophical Transactions of the Royal Society A, 379(2194), 20200083.
[16] Lorenz, E. N. (1963). Deterministic nonperiodic flow. Journal of Atmospheric Sciences, 20(2), 130-141.
[17] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686-707.
[18] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.
[19] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., ... & Lerer, A. (2017). Automatic differentiation in pytorch.
[20] Nasiri, H., & Ebadzadeh, M. M. (2022). MFRFNN: Multi-functional recurrent fuzzy neural network for chaotic time series prediction. Neurocomputing, 507, 292-310.
[21] Cheng, W., Wang, Y., Peng, Z., Ren, X., Shuai, Y., Zang, S., ... & Wu, J. (2021). High-efficiency chaotic time series prediction based on time convolution neural network. Chaos,

Solitons & Fractals, 152, 111304.
[22] Na, X., Ren, W., & Xu, X. (2021). Hierarchical delay-memory echo state network: A model designed for multi-step chaotic time series prediction. Engineering Applications of Artificial Intelligence, 102, 104229.
[23] Wu, G., Tang, L., & Liang, J. (2024). Synchronization of non-smooth chaotic systems via an improved reservoir computing. Scientific Reports, 14(1), 229.
[24] Kennedy, C., Crowdis, T., Hu, H., Vaidyanathan, S., & Zhang, H. K. (2024). Data-driven learning chaotic dynamical system using Discrete-Temporal Sobolev Networks. Neural Networks, 106152.
[25] Rössler, O. E. (1976). An equation for continuous chaos. Physics Letters A, 57(5), 397-398.
[26] Chua, L. E. O. N. O., Komuro, M., & Matsumoto, T. (1986). The double scroll family. IEEE Transactions on Circuits and Systems, 33(11), 1072-1118.
[27] Chen, G., & Ueta, T. (1999). Yet another chaotic attractor. International Journal of Bifurcation and Chaos, 9(07), 1465-1466.
[28] Pathak, J., Lu, Z., Hunt, B. R., Girvan, M., & Ott, E. (2017). Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data. Chaos: An Interdisciplinary Journal of Nonlinear Science, 27(12).
[29] Gauthier, D. J., Bollt, E., Griffith, A., & Barbosa, W. A. (2021). Next generation reservoir computing. Nature Communications, 12(1), 5564.
[30] Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023, June). Are transformers effective for time series forecasting?. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 9, pp. 11121-11128).
[31] Doan, N. A. K., Polifke, W., & Magri, L. (2020). Physics-informed echo state networks. Journal of Computational Science, 47, 101237.
[32] Na, X., Li, Y., Ren, W., & Han, M. (2023). Physics-informed hierarchical echo state network for predicting the dynamics of chaotic systems. Expert Systems with Applications, 228, 120155.
[33] Özalp, E., Margazoglou, G., & Magri, L. (2023, June). Physics-informed long short-term memory for forecasting and reconstruction of chaos. In International Conference on Computational Science (pp. 382-389). Cham: Springer Nature Switzerland.
[34] Racca, A., & Magri, L. (2021, June). Automatic-differentiated physics-informed echo state network (API-ESN). In International Conference on Computational Science (pp. 323-329). Cham: Springer International Publishing.
[35] Sirignano, J., & Spiliopoulos, K. (2018). DGM: A deep learning algorithm for solving partial differential equations. Journal of Computational Physics, 375, 1339-1364.
[36] Aristotelous, A. C., Mitchell, E. C., & Maroulas, V. (2023). ADLGM: An efficient adaptive sampling deep learning Galerkin method. Journal of Computational Physics, 477, 111944.
[37] Lagaris, I. E., Likas, A., & Fotiadis, D. I. (1998). Artificial neural networks for solving ordinary and partial differential equations. IEEE Transactions on Neural Networks, 9(5), 987-1000.
[38] Ramuhalli, P., Udpa, L., & Udpa, S. S. (2005). Finite-element neural networks for solving differential equations. IEEE Transactions on Neural Networks, 16(6), 1381-1392.

A methodological approach to map complex research systems to the Sustainable Development Goals: Analysis of CIRAD publications

Audilio Gonzalez Aguilar[1][0000-0001-8693-2076], Francisco Paletta[2][0000-0002-4112-5198]

Juan Camilo Vallejo-Echavarría[3][0000-0002-9159-8292]

[1] Université Paul-Valéry Montpellier III, Montpellier, França
[2] Francisco Carlos Paletta, Universidade de São Paulo, São Paulo, Brasil
[3]Juan Camilo, Universidad de Antióquia: Medellín, Antióquia, Colômbia

Abstract. This study proposes a methodology for mapping complex research systems to the Sustainable Development Goals (SDGs) using publications as an indicator of research activity. The approach is applied to analyse the research focus of CIRAD, a French agricultural research organisation, by mapping its publications to the SDGs. The article highlights the challenges associated with mapping publications to the SDGs, including the complexity of research systems, ambiguity in classifying the SDGs and data availability. The methodology uses advanced search functionalities and data extraction techniques from Agritrop databases, followed by visualisation tools such as Gephi to explore publication trends, collaboration networks and thematic priorities. The analysis reveals that CIRAD's research is aligned with several SDGs, with a primary focus on food security and nutrition (SDG 2), sustainable agriculture and rural development (SDGs 1 and 8), biodiversity conservation and environmental sustainability (SDGs 13 and 15) and gender equality and women's empowerment (SDG 5). This research offers valuable insights into the potential of mapping publications to understand the contribution of research institutions to the SDGs and guide evidence-based decision-making aligned with sustainable development.

Keywords: Sustainable Development Goals, Complex Research Systems, CIRAD.

1 Introduction

The United Nations' 2030 Agenda for Sustainable Development established 17 Sustainable Development Goals (SDGs) as a shared framework for tackling the biggest global

challenges. Achieving these ambitious SDGs requires mobilising and aligning complex social systems such as agriculture, energy, transport, and health. Research organisations play a significant role in generating knowledge, technologies and evidence-based solutions that can accelerate sustainable development. Assessing the orientation of large, decentralised research systems towards the SDGs remains a methodological challenge.

One approach is to use academic publications as tangible outputs that provide indicators of the focus and priorities of research institutions. Mapping scientific publications around themes associated with the SDGs provides a consistent basis for assessing how research activity aligns with the SDG framework. It also makes it possible to understand interconnections between the goals based on co-occurrences in publication mappings. Comparative analysis can also reveal variations in the focus of the SDGs between departments within the same institution. In this way, mapping research publications for the SDGs contributes to presenting guidelines and networks embedded within complex research systems.

This study presents a methodology for mapping research publications to the SDGs using the institutional repository of CIRAD, a French agricultural research organisation. The network analysis visually represents the connections between the SDGs based on CIRAD's publication mappings. A comparative analysis highlights the differences in emphasis on various SDGs between CIRAD departments. The methodology consists of a universally applicable approach to systematically track, assess, and strengthen the orientation of complex systems of research organisations towards the SDGs. Mapping scientific publications around the SDGs provides valuable insights into the structure and focus of multifaceted research activities. The technique can thus instrumentalise institutions in assessing their alignment with the sustainability goals and guiding their research activities.

2 CIRAD's contribution to the SDGs

By providing scientific and practical solutions to improve agricultural production and the management of natural resources, agronomic research is essential to address the interlinked issues of food security, environmental sustainability, and rural development, thus contributing significantly to the achievement of the United Nations Sustainable Development Goals. World Bank (2020).

CIRAD's contribution to the Sustainable Development Goals (SDGs) stands out, as it allows us to understand and evaluate the impact of its research and development activities at the international level. The SDGs provide a global framework for overcoming the world's most pressing obstacles, from eradicating poverty to protecting the environment and promoting gender equality.

2.1 The importance of agronomic research for sustainable development

Understanding agricultural systems and their interaction with the environment contributes to ensuring long-term sustainable food production and the well-being of local com-

munities. Agronomic research plays a strategic role in promoting sustainable development by addressing the main barriers related to food production, natural resource management and climate change mitigation.

Agronomic research focuses not only on increasing crop yields, but also on improving resource efficiency, reducing the use of pesticides, and promoting environmentally friendly agricultural practices. According to Kell et all (2009), agronomic research provides innovative solutions to increase the resilience of agricultural systems to climate and environmental issues.

With an active role in the fight against hunger and malnutrition around the world, agronomic research, according to the FAO (2021), has worked to ensure the food and nutritional security of the world's population by helping to improve agricultural productivity and access to quality food. Agronomic research contributes to the development of more nutritious and disease-resistant crops, as well as promoting agricultural practices that increase food availability in vulnerable areas.

Strategic in tackling the interrelated barriers of food security, environmental sustainability and rural development, agronomic research, by providing scientific and practical solutions to improve agricultural production and natural resource management, contributes significantly to achieving the United Nations Sustainable Development Goals.

The use of techniques to visualise CIRAD's contribution to the SDGs makes it possible to identify areas of competence and opportunities to improve research investments. According to Gunning et al (2020), mapping the contribution of research institutions to achieving the SDGs makes it possible to ensure effective strategic planning and to understand and evaluate the consequences of their research activities for the SDGs.

Using techniques to visualise CIRAD's contribution to the SDGs helps raise awareness among key stakeholders, such as policymakers, donors, and civil society, of the importance of investing in agricultural research and rural development. According to Rastoin and Chiffoleau (2016), data visualisation is a powerful tool for communicating the effectiveness of research activities in achieving the SDGs and mobilising public policies and financial support.

By visualising CIRAD's contribution to the SDGs, transparency, social responsibility and the effectiveness of its research and development activities can be improved. By providing a clear and accessible representation of the results of CIRAD's contribution to the implementation of the SDGs, greater collaboration, coordination, and funding can be promoted to tackle the problems associated with sustainable development worldwide.

2.2 Mapping publications by SDGs

Mapping publications for the Sustainable Development Goals (SDGs) presents a multifaceted challenge due to the inherent complexity of research systems and the broad scope of the SDGs themselves (Sachs, et al, 2019). In the context of analysing CIRAD's publications, several difficulties arise and must be addressed to ensure an accurate and meaningful mapping, Table 1, where we can highlight:

Table 1. Challenges in mapping SDG publications

Complexity and Interconnection of Research Systems	One of the main problems in mapping publications to the SDGs lies in the complexity and interconnectedness of research systems. Research often addresses multiple aspects of sustainable development simultaneously, making it difficult to categorise publications into distinct SDGs. For example, a study on agricultural practices can contribute to several SDGs, including zero hunger, sustainable agriculture, and climate action. Thus, determining the most relevant SDGs for a specific publication requires careful consideration of their various dimensions and outcomes.
Ambiguity and Subjectivity in the Classification of the SDGs	Categorising publications according to the SDGs can be subjective and open to interpretation. While some publications may clearly align with specific objectives, others may address multiple objectives or fall into a grey area where their relevance to the SDGs is less obvious. This ambiguity complicates the mapping process and can lead to inconsistencies in the classification, especially when different analysts assess the same publication. Standardising criteria for classifying the SDGs and providing clear guidelines can help mitigate this impasse.
Granularity and Details of Level	The SDGs cover a wide range of topics, from broad thematic areas to specific targets and indicators. Mapping publications to the SDGs requires finding a balance between granularity and comprehensiveness. At an important level, publications can align with general objectives such as poverty eradication or gender equality. However, to provide meaningful insights, it is often necessary to delve deeper into the targets and indicators associated with each goal. This level of detail increases the complexity of the mapping process and may require substantial resources and expertise.
Evolutionary nature of the SDGs	The SDGs are dynamic and subject to revision as global impasses and priorities evolve. New targets and indicators may be added, and existing ones may be modified to reflect emerging issues and knowledge gaps. As a result, mapping publications to the SDGs requires keeping up to date with changes in the SDG framework and ensuring that mapping methodologies remain relevant and adaptable over time. Continuous monitoring and refinement of mapping approaches are essential to

	capture the evolving relationship between research outputs and sustainable development priorities.
Data Availability and Quality	Effective mapping of publications to the SDGs depends on access to comprehensive and reliable data. However, data availability can vary between different research domains and geographical regions, presenting difficulties to the mapping process. In addition, the quality and consistency of data sources can influence the accuracy of SDG classification. Addressing data gaps and improving data quality are critical steps in improving the robustness of mapping methodologies and ensuring the credibility of mapping results.

Mapping publications to the SDGs presents several challenges that need to be overcome in order to facilitate meaningful analyses and decision-making. Dealing with the complexity of research systems, navigating ambiguity in the classification of the SDGs, balancing granularity with comprehensiveness, keeping up to date with evolving SDG frameworks, and ensuring data availability and quality are key considerations in devising methodological approaches to map complex research systems to the SDGs effectively. Despite these adversities, mapping efforts have enormous potential to advance our understanding of the contributions of research to sustainable development and guide public policy and evidence-based practice. World Bank Group (2019).

3 Theoretical Framework

The 17 Sustainable Development Goals (SDGs) were established by the United Nations (UN) in 2015 as a universal call to action to end poverty, protect the planet and ensure prosperity for all by the year 2030. Building on the Millennium Development Goals (MDGs) that preceded them, the SDGs represent a comprehensive and interconnected framework for tackling the world's most pressing issues. United Nations (2015).

Each of the 17 SDGs encompasses specific targets and indicators designed to address key aspects of sustainable development, including economic growth, social inclusion, and environmental sustainability. The goals cover a wide range of issues, from eradicating poverty and hunger to gender equality, climate action, peace, and justice.

3.1 Introduction to the 17 Sustainable Development Goals (SDGs)

The SDGs recognise that sustainable development must be holistic, addressing the interconnections between the social, economic, and environmental dimensions. They emphasise the importance of leaving no one behind, ensuring that progress is inclusive and reaches the most vulnerable and marginalised populations.

Achieving the SDGs requires collaboration and partnership between governments, civil society, the private sector, and other stakeholders at local, national, and global

levels. It requires innovative approaches, transformative policies and integrated solutions that address the root causes of poverty, inequality, and environmental degradation. The 17 SDGs are:

Fig. 1. The 17 SDGs

These goals are interconnected and mutually reinforcing, recognising that progress in one area often depends on progress in others. By addressing these interconnected obstacles in a coordinated way, the SDGs aim to create a more sustainable and equitable world for present and future generations. Le Blanc (2015).

The 17 Sustainable Development Goals represent a global commitment to building a better world for all. They provide a roadmap to address the most pressing issues facing humanity and offer an opportunity to create a future in which prosperity is shared. Achieving the SDGs requires collective action, political will, and a renewed commitment to leaving no-one behind.

3.2 Links between CIRAD's research activities and the SDGs

The Centre for International Cooperation in Agricultural Research for Development (CIRAD) works to tackle global complications related to agriculture, food security and sustainable development. Founded in 1984, CIRAD is at the forefront of carrying out research and promoting innovation to improve livelihoods, promote environmental sustainability and contribute to achieving the Sustainable Development Goals (SDGs) established by the United Nations.

The SDGs provide a comprehensive framework for overcoming the world's most pressing social, economic and environmental issues by 2030. CIRAD's research activities are aligned with several of these goals, reflecting its commitment to promoting sustainable development worldwide, Table 2.

Table 2. CIRAD's research activities and the SDGs

Objective 1 Eradication of Poverty	CIRAD's research focuses on improving agricultural productivity and rural livelihoods, especially in low-income countries. By developing innovative agricultural techniques, promoting inclusive value chains, and supporting small-scale farmers, CIRAD contributes to poverty reduction and economic empowerment among vulnerable communities.
Objective 2 Zero Hunger	CIRAD works to improve food and nutrition security by conducting research into sustainable agriculture, crop diversification and food systems. Through partnerships and capacity-building efforts, CIRAD helps to increase agricultural productivity, increase resilience to climate change and ensure access to food for all.
Objective 3 Health and Well-being	CIRAD's research into agroecology, sustainable livestock management and disease control contributes to improving human health and well-being. By promoting sustainable agricultural practices and reducing the use of pesticides and harmful chemicals, CIRAD helps to mitigate health risks and promote a healthier environment for farming communities.
Goal 5 Gender Equality	CIRAD is committed to promoting gender equality and empowering women in agriculture. Through research projects and training initiatives, CIRAD seeks to address gender disparities in access to resources, decision-making and agricultural productivity, thereby promoting more inclusive and equitable development outcomes.
Goal 13 Climate Action	CIRAD's research into climate-smart agriculture, agroforestry and sustainable land management contributes to climate change mitigation and adaptation. By developing resilient agricultural practices and promoting biodiversity conservation, CIRAD helps build adaptive capacity and reduce the vulnerability of farming communities to climate-related risks.
Objective 15 Terrestrial Life	Through efforts to promote agro-ecological practices and sustainable forest management, CIRAD helps protect and restore ecosystems, thus safeguarding biodiversity and ecosystem services vital to human well-being.

CIRAD's research activities are closely aligned with the Sustainable Development Goals, reflecting its commitment to promoting sustainable agriculture, food security and rural development worldwide. Through its interdisciplinary approach, collaborative partnerships, and innovative solutions, CIRAD contributes to advancing sustainable development and building a more resilient and equitable future for all.

4 Complex systems and the Sustainable Development Goals (SDGs)

Implementing the Sustainable Development Goals (SDGs) requires understanding and addressing the complexities inherent in socio-economic and environmental systems. The conceptualisation of complex systems provides a framework for understanding the interconnectedness, feedback loops and non-linear dynamics that characterise the impediments and opportunities associated with sustainable development. (Pradhan, 2017).

4.1 Conceptualising complex systems in the context of the SDGs

Complex systems theory emphasises the emergence of properties and behaviours at the system level that cannot be fully understood by analysing individual components in isolation. This holistic approach recognises the interdependence and interactions between various elements within a system, such as ecosystems, economies, and societies, and acknowledges their dynamic nature over time.

To formulate effective strategies and public policies to achieve the SDGs, it is necessary to understand the key concepts that underpin the conceptualisation of complex systems in the context of the SDGs. Table 3 summarises the main concepts:

Table 3. Key concepts of complex systems in the context of the SDGs

Systems Thinking: Systems thinking involves examining the relationships and interdependencies between various components of a system to understand how they influence the system's behaviour and outcomes. By adopting a systems perspective, stakeholders can identify leverage points for intervention and design holistic solutions that address interconnected pathways.
Interdisciplinary Approaches: Addressing sustainable development challenges requires the integration of insights from various disciplines, including ecology, economics, sociology, and engineering. Interdisciplinary approaches allow for a more comprehensive understanding of complex systems by considering multiple perspectives and generating innovative solutions that transcend disciplinary boundaries.
Network Theory: Network theory provides a framework for analysing the structure and dynamics of interconnected systems, such as social networks, commercial networks, and ecological networks. By mapping relationships and flows of information, resources, and energy within and between systems, network theory helps to identify key actors, paths of influence and potential points of intervention to promote sustainability.
Resilience Theory: Resilience theory explores the capacity of systems to absorb disturbances, adapt to change and maintain functionality in the face of shocks and stresses. Understanding the resilience of socio-ecological systems is important for

building adaptive capacity, improving sustainability, and promoting long-term resilience in the context of global issues such as climate change, biodiversity loss and social inequality.
Participatory Approaches: Involving stakeholders and local communities in decision-making processes is essential for understanding the complexities of socio-ecological systems, incorporating diverse perspectives, and promoting ownership and legitimacy of sustainable development initiatives. Participatory approaches facilitate the co-creation of knowledge, collaborative resolution of difficulties and empowerment of marginalised groups, thus contributing to more effective and equitable results.

By understanding the conceptualisation of complex systems, stakeholders can navigate the uncertainties and complexities inherent in sustainable development, promote synergies between the SDGs and advance transformative change towards a more equitable, resilient, and sustainable future (We-ber et all, 2021).

4.2 Interrelationships and interactions between the different SDGs

The Sustainable Development Goals (SDGs) represent a holistic framework, i.e., they address various dimensions of sustainable development in a comprehensive and integrated manner, Table 4, taking into account their environmental, social, economic, and institutional aspects in an interconnected way, implying:

Table 4 - Holistic framework of the SDGs

Multidimensional Approach: The 17 SDGs cover a wide range of critical issues, from poverty eradication, food security, health, education, gender equality, to natural resource management, climate action, peace, and justice.
Interconnectedness and indivisibility: The SDGs recognise that the challenges of sustainable development are intrinsically interconnected and indivisible. Progress in one area is linked to progress in others.
Balancing the Three Dimensions: These seek to balance the environmental, social, and economic dimensions of sustainable development in an integrated and mutually reinforcing way.
Universal Application: The SDGs are applicable to all countries, rich and poor, requiring national efforts and global co-operation.
Systemic Approach: Addresses challenges holistically, considering their root causes, interrelationships and impacts on multiple sectors and actors.
Inclusive Participation: Involves a wide range of stakeholders, including governments, civil society, the private sector, and local communities

The holistic framework of the SDGs, designed to address interconnected global issues, recognises the complexity and multiple facets of development challenges, promoting an integrated and comprehensive approach to achieving truly sustainable development in all its dimensions. (Soleimani et al, 2020). Interrelationships and interactions between the different Sustainable Development Goals (SDGs) are inherently influenced by the concept of complex systems where it becomes necessary to understand their interrelationships, Table 5, for the effective implementation and achievement of the sustainable development goals.

Table 5 - Interrelations and Interactions between ODs

Synergies: Many SDGs have synergistic relationships, where progress on one goal positively influences progress on others. For example, investing in education (SDG 4) not only contributes to the eradication of poverty (SDG 1), but also promotes gender equality (SDG 5) and improves economic growth (SDG 8). Recognising and taking advantage of these synergies can broaden the scope of interventions and accelerate progress towards multiple goals simultaneously.
Trade-offs: The pursuit of certain SDGs can lead to trade-offs or unintended consequences for other goals. For example, promoting economic growth (SDG 8) through industrialisation and infrastructure development can lead to increased carbon emissions and environmental degradation, undermining efforts to combat climate change (SDG 13) and protect biodiversity (SDG 15). Understanding and managing these trade-offs is important to ensure sustainable development outcomes and avoid negative consequences in interconnected systems.
Complex feedback loops: Interactions between SDGs can give rise to complex feedback loops, where changes in one goal trigger cascading effects in several goals and systems. For example, investments in renewable energy (SDG 7) can reduce greenhouse gas emissions (SDG 13), mitigating the effects of climate change and promoting environmental sustainability. This, in turn, can improve agricultural productivity (SDG 2) and food security (SDG 3), contributing to poverty reduction (SDG 1) and better health outcomes (SDG 3). Recognising these feedback loops is essential to designing holistic and integrated strategies that address interconnected contingencies effectively.
Context Dependence: The interrelationships between the SDGs can vary depending on contextual factors such as geography, socio-economic conditions, and governance structures. What constitutes a synergistic or conflicting relationship between goals in one context may differ in another. Therefore, contextual analyses and localised approaches are key to adapting interventions to specific circumstances and maximising positive synergies while minimising trade-offs.
Policy Integration: Given the complex interdependencies between the SDGs, integrated policy approaches that consider multiple goals simultaneously are relevant.

Isolated approaches that focus on individual goals in isolation are insufficient to address the interconnected nature of sustainable development pathways. Instead, policymakers need to adopt intersectoral and multisectoral strategies that promote coherence, coordination, and alignment across different policy domains.

By recognising and embracing the interrelationships and interactions between the SDGs in the context of complex systems, stakeholders can promote synergies, manage trade-offs, and design integrated strategies that move effectively and holistically towards the sustainable development goals.

5 5 Methodological approaches to addressing complex systems in SDG-related research

The research methodology offers a systematic and interdisciplinary approach to understanding and addressing complex systems in SDG-related research, thus contributing to the advancement of sustainable development goals and the promotion of global prosperity, equity, and environmental preservation.

The use of advanced search functionalities and data extraction techniques enables the identification and collection of relevant information from Agritrop databases, facilitating the accurate analysis of research trends, collaboration networks and thematic priorities within the SDG research domain. (Valdano et al, 2019).

The use of advanced visualisation tools, such as Cosma, Gephi, Cytoscape and Tableau, allows researchers to transform complex data sets into visually informative representations. These visualisations offer a clear and intuitive way to explore relationships, patterns, and trends within SDG-related scientific networks, facilitating the identification of key players, research clusters and emerging topics.

Through network analysis algorithms and advanced mapping methods, hidden insights can be identified, and the dynamics of scientific collaboration and knowledge exchange within the SDG research area can be understood. Centre for Complex Systems Modelling. (2018).

The clear and informative visual representations generated by this methodology empower stakeholders, policymakers, and funding agencies to make decisions underpinned by research priorities, resource allocation and intervention strategies related to the SDGs. By providing actionable insights derived from data-driven analyses, this methodology supports evidence-based decision-making processes aimed at promoting sustainable development agendas. United Nations Sustainable Development Solutions Network. (2019).

The methodology uses advanced search functionalities and data extraction techniques from Agritrop databases, followed by visualisation tools such as Gephi to explore publication trends, collaboration networks and thematic priorities.

5.1 Data Collection: Agritrop

In Data Collection, we utilised the Agritrop website (https://agritrop.cirad.fr) which provides access to a rich source of data covering various aspects of agricultural research and development.

This comprehensive dataset from the Agrotrop database allows us to explore SDG-related issues in depth, providing a holistic understanding of the interconnectedness of the Sustainable Development Goals. Using the Agritrop website as a comprehensive source of data related to our research we have processed and prepared the data to perform our network analyses.

In extracting the information from the Agritrop databases, we included the publications, organized by each Sustainable Development Goal (SDG). Using search and export functionalities of the data in Zorero, Reference manager (RIS) format, we retrieved datasets for all 17 Sustainable Development Goals (SDGs).

By creating a dashboard of the publications, we separated specific data based on keywords, abstracts, publication dates, authors and thematic areas aligned with the SDGs from the publications and keywords into an excel sheet. We have made a treatment to the keywords of lowercase transformation of all words, deletion of all spaces to establish the relationship between the keywords and the 17 Sustainable Development Goals (SDGs).

From this excel file we have processed the 45779 keywords to eliminate the repeated words and to be able to elaborate a table of unique keywords (5638). From these two excel tables we created the nodes table and the links table to perform our network analysis with the Gephi software.

This comprehensive dataset allows for an in-depth exploration of topics relevant to the SDGs, providing a holistic understanding of the interconnection of agricultural systems with the sustainable development goals. Data collection was carried out according to the methodological procedures presented in Table 6:

Table 6 - Data Collection Methodology

Database: the Agritrop website as a source of data related to agricultural research and development https://agritrop.cirad.fr/recherches_odd.html
Relevant information was extracted from Agritrop's databases, including publications, projects, and partnerships, focusing on topics relevant to the Sustainable Development Goals (SDGs).
Advanced search functionalities were applied in order to filter and retrieve specific data sets based on keywords, auto-res, publication dates and thematic areas aligned with the SDGs.
The integrity and quality of the data was guaranteed by cross-referencing information from Agritrop with other reliable sources and conducting data validation procedures.
Excel (dynamic graphics) was used for statistical analyses.

5.2 Visualisation and mapping methods and tools

The methodological procedure used visualisation techniques to represent and analyse the complex scientific networks related to the SDGs, derived from the data collected from Agritrop.

To visualise the relationships between researchers, institutions, and publications within the SDG research domain, we used the Gephi platform. Network analysis algorithms were used to identify key players, influential research topics and collaboration patterns within the SDG scientific community.

Implemented the integration of Gephi and gexf.js to enhance visualisation capabilities and generate interactive and dynamic visualisations of scientific networks related to the SDGs. Advanced mapping methods were used to geographically visualise the distribution of research activities, funding sources and impact metrics related to the SDGs in different regions and countries.

Social network analysis (SNA) (Avila-Toscano, 2018) has been used as a methodology to visualise the data. SNA uses networks and graph theory (Andrienko et al., 2020; Otte; Rousseau, 2002). The software used to create these visualisations was Gephi (Bastian; Heymann; Jacomy, 2009): https://gephi.org.

Gephi is a program for visualising, exploring, and understanding all types of graphs and networks (Cherven, 2015). It is free and is based on ARS. The spatialisation algorithms used were Atlas Force 2 and Atlas 2-3D. It was combined with a visualiser that allows graphics made with Gephi to be exported to the web, called gexf.js (Velt, 2011), which is available on Github: https://github.com/raphv/gexf-js.

The final visualisation is displayed as an interactive map, which can be manipulated by the user to analyse the results by applying different integrated filtering strategies.

6 Results and Discussion

CIRAD's research activities are closely aligned with the Sustainable Development Goals, addressing various dimensions of sustainable agriculture, rural development, environmental conservation, gender equality and technological innovation,

Table 6. CIRAD Research Activities and the SDGs

Food and Nutrition Security Security (SDG 2)	One of CIRAD's main areas of research revolves around increasing food and nutrition security, particularly in regions facing complications such as poverty, climate change and resource scarcity. Through innovative agricultural practices, crop diversification and sustainable farming techniques, CIRAD aims to improve agricultural productivity and guarantee access to nutritious food for all, thus contributing to SDG 2 - Zero Hunger.

Sustainable Agriculture and Rural Development (SDG 1, 8)	CIRAD emphasises sustainable agricultural practices and rural development strategies to alleviate poverty (SDG 1) and promote inclusive economic growth (SDG 8) in rural communities. By empowering smallholder farmers, promoting value chains and fostering entrepreneurship, CIRAD strives to create resilient and prosperous rural economies.
Biodiversity Conservation and Environmental Sustainability (SDG 15, 13)	Biodiversity conservation and the preservation of natural ecosystems are integral components of CIRAD's research agenda. By studying agro-ecological systems, promoting agro-forestry, and advocating sustainable land management practices, CIRAD contributes to SDG 15 - Life on Land and SDG 13 - Climate Action, aiming to mitigate climate change and protect terrestrial ecosystems.
Gender Equality and Women's Empowerment (SDG 5)	CIRAD recognises the importance of gender equality and women's empowerment in agricultural development. Through gender-sensitive research initiatives, capacity-building programs, and inclusive policy interventions, CIRAD works to achieve SDG 5 - Gender Equality, ensuring equal opportunities for women in agriculture and rural livelihoods.
Innovation and Technology Transfer (SDG 9)	Promoting innovation and technology transfer is a key focus area for CIRAD to increase agricultural productivity, efficiency, and resilience. By facilitating the exchange of knowledge, fostering partnerships, and harnessing digital technologies, CIRAD contributes to SDG 9 - Industry, Innovation, and Infrastructure, boosting sustainable development through technological advancement.

By leveraging its expertise and partnerships, CIRAD continues to play a relevant role in advancing the global agenda for sustainable development and contributing to positive transformative change around the world.

6.1 Quantitative analysis of the distribution of publications by SDGs

Agritrop is a bibliographic database developed by the French agricultural research and education organization CIRAD (Centre de recherche agronomique pour le développement). It contains more than 1.5 million references covering a wide range of topics related to tropical and Mediterranean agriculture.

Table 7. Number of Publications per Sustainable Development Goal

SDGs	N° Publication	Unique Keywords	Keywords-Total
GOAL 1	87	399	1043

GOAL 2	367	865	3893
GOAL 3	307	1078	3332
GOAL 4	40	17	230
GOAL 5	125	47	1008
GOAL 6	54	26	434
GOAL 7	240	58	1833
GOAL 8	63	23	642
GOAL 9	57	16	553
GOAL 10	53	25	405
GOAL 11	386	70	2915
GOAL 12	293	66	2160
GOAL 13	2191	2796	17261
GOAL 14	42	22	399
GOAL 15	938	110	8941
GOAL 16	21	9	165
GOAL 17	43	11	565
Total	5307	5638	45779

Agritrop plays a crucial role in supporting research and knowledge dissemination aligned with the SDGs. The database's vast collection of references on various aspects of agriculture, food security, and rural development contributes to addressing critical challenges and promoting sustainable solutions.

The survey results provide an overview of CIRAD's main research areas related to the SDGs, highlighting the organisation's commitment to promoting agricultural sustainability, food security and socio-economic development in regions around the world, Table 7, Table 8, Figure 2, and Figure 3.

It can be seen that Goal 13 Climate Action and Goal 15 Life on Land, 2,191 and 938 respectively, represent 79% of CIRAD's publications in relation to the SDGs. Goal 16 Peace, Justice and Strong Institutions and Goal 17 Partnerships for the Goals have the lowest number of publications, with 21 and 43, respectively.

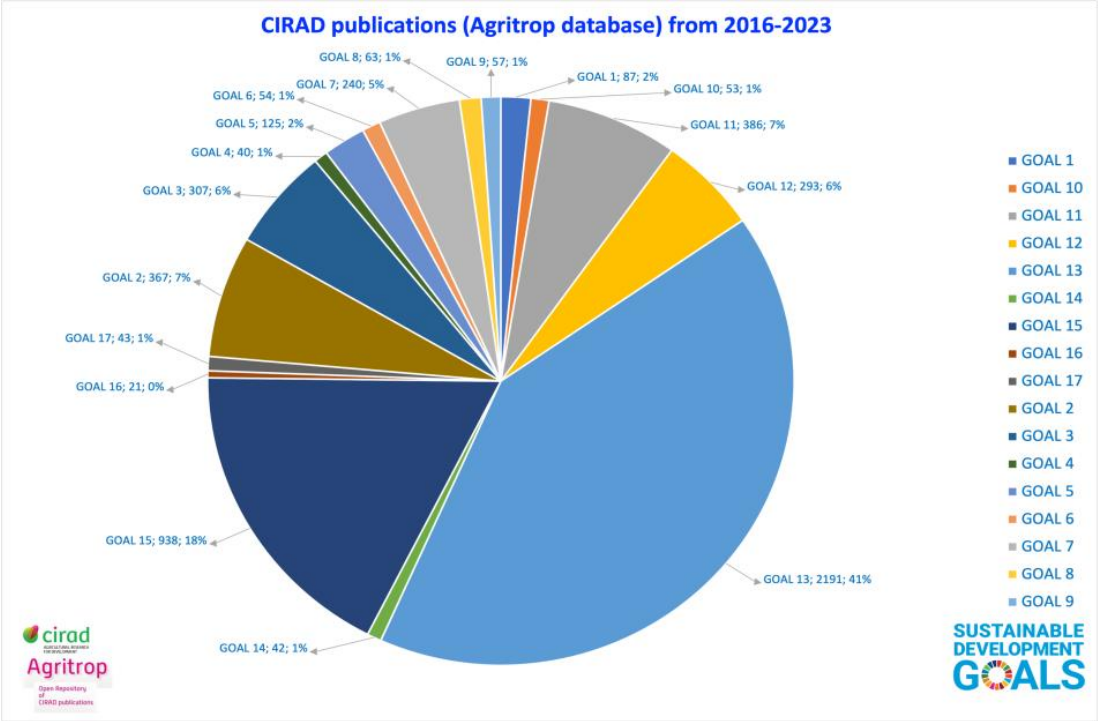


Fig. 2. CIRADS publications distribution for Goal

Table 8. CIRAD Publications Keyword Cloud

Word	TF-IDF	Weight
changement climatique	0.019	855
biodiversité	0.016	752
sécurité alimentaire	0.014	619
France	0.012	549
adaptation aux changements climatique	0.009	438
impact sur l'environnement	0.008	353
utilisation des terrer	0.007	327
gestion des ressources naturel	0.007	326
politique de développement	0.007	307
développement durable	0.007	299
forêt tropical	0.006	267
Brésil	0.006	255
services écosystémique	0.005	250
système de culture	0.005	242
agroécologie	0.005	241

étude de cas	0.004	203
agroforesterie	0.004	201
politique de l'environnement	0.004	197
évaluation de l'impact	0.004	197
agriculture durable	0.004	195
Afrique	0.004	195
séquestration du carbone	0.004	192
déboisement	0.004	190
développement agricole	0.004	189

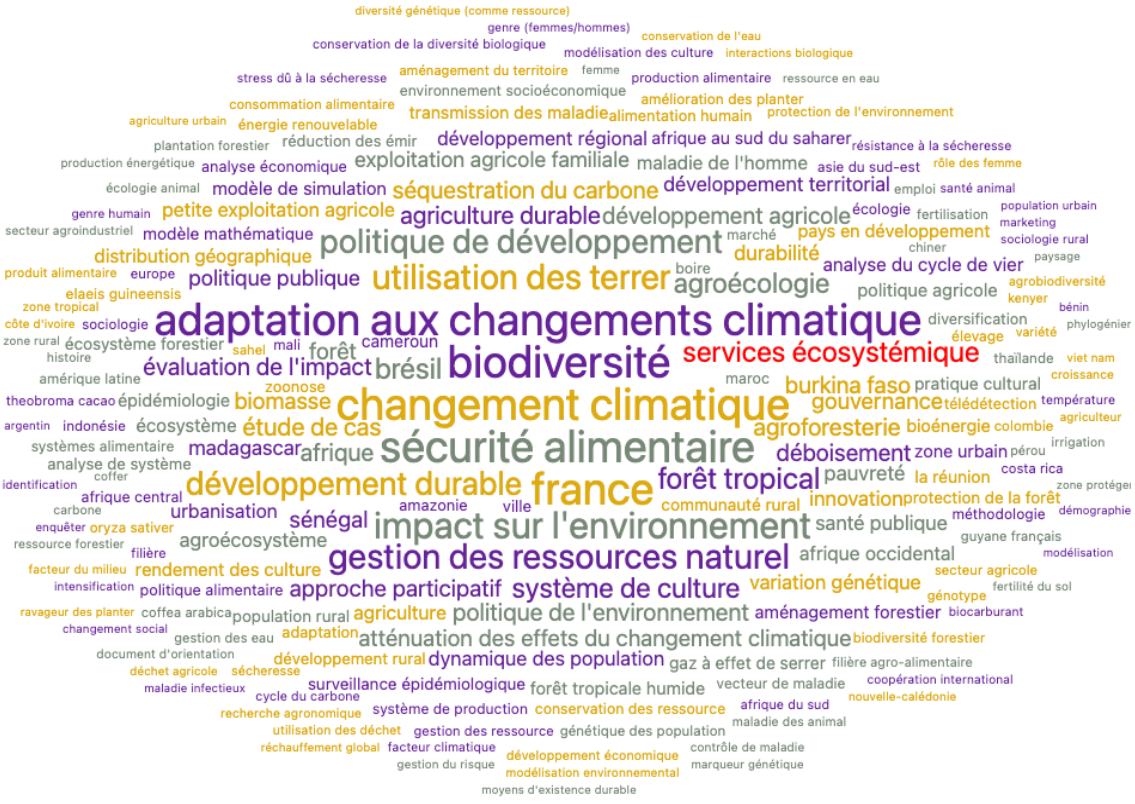


Fig. 3. CIRAD Publications Keyword Cloud

6.1 2016-2023 CIRAD (Agritrop database) SDG-related publications: interactive version

Below we present the results obtained using data visualisation techniques and the results of CIRAD publications using the Agri-top database from 2016 to 2023 related to the SDGs, Figure 4, 5, 6 and 7:

Table 7. Distribution of CIRAD Publications

Number of Publications	5307
Number of total keywords	45779
Number of unique keywords	4487
Number of nodes	9812
Number edges	45598
Densidad	0,001
Centralidad de Eigenvector	0,1062
Modularidad	0,459

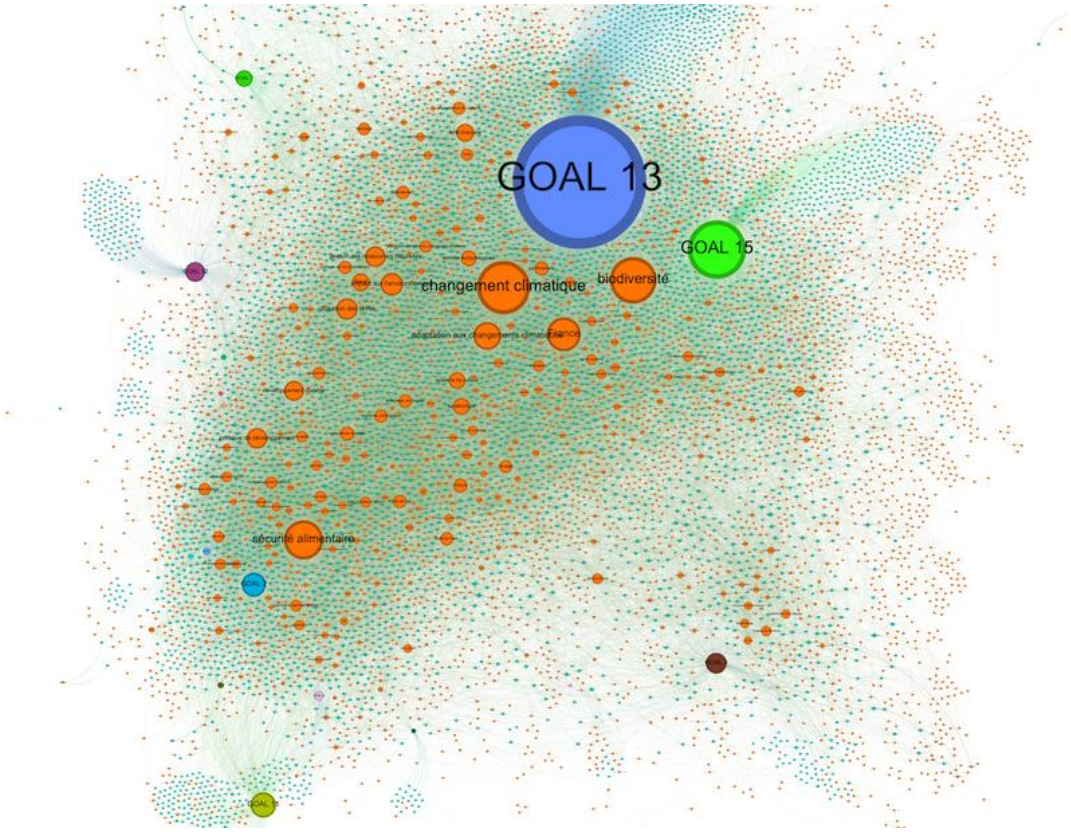


Fig. 4. Mapping of Cirad (Agritrop) publications 2016-2023. Interact Version : https://metroteach.com/SDG/index.html

Network Properties:

- Number of nodes: 9812
- Number of links: 48580
- Density: 0.001 (very sparsely connected)
- Eigenvector Centrality: 0.1062 (high value)
- Modularity: 0.459 (high modularity)

Interpretation of Properties:

- The extremely low density indicates that the network is extremely sparsely connected, meaning most nodes are not directly connected to each other.
- The high Eigenvector Centrality value suggests that influence is concentrated on a small number of key nodes.
- The high modularity indicates that the network is divided into distinct groups (modules) with stronger links within modules than between them. Fig. 6 and Fig. 7

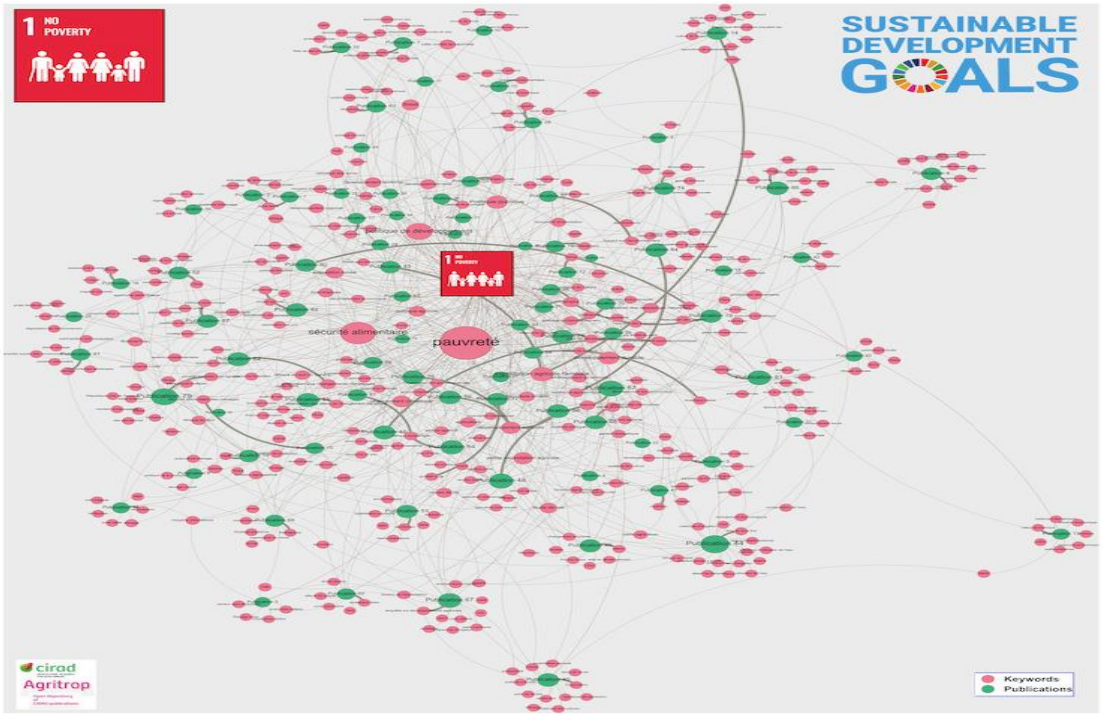


Fig. 5. Goal 1 - interactive tool allows you to explore CIRAD publications (http://metroteach.com/SDG/index.html)

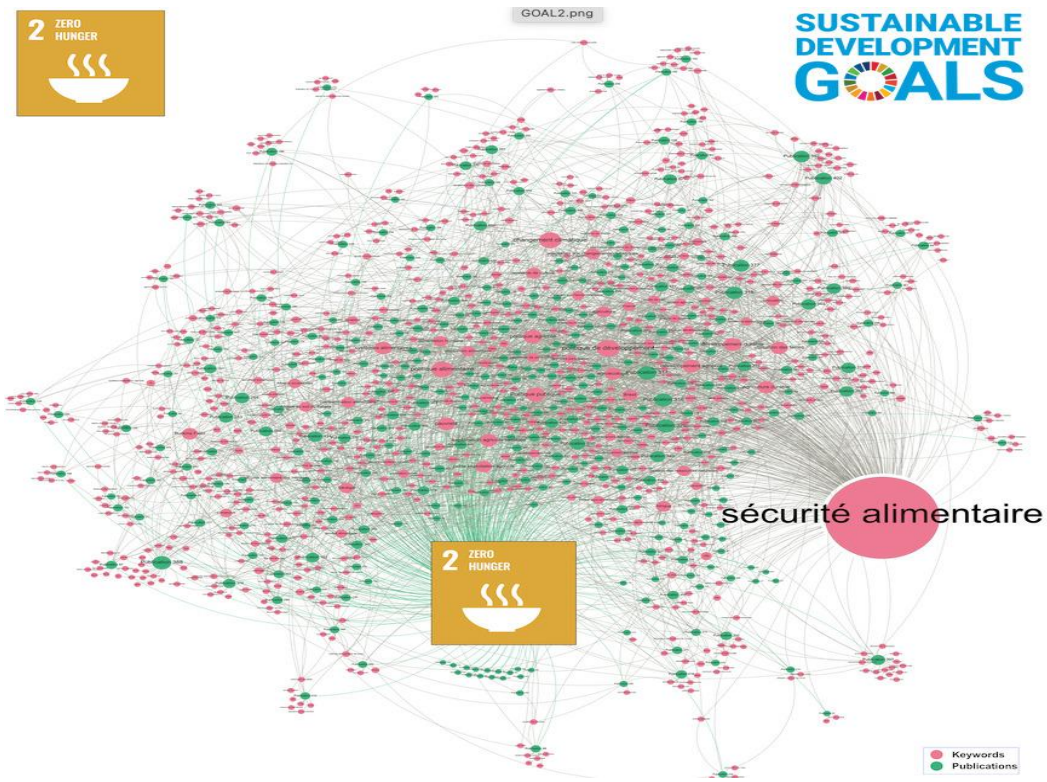


Fig. 6. Goal 2 - interactive tool allows you to explore CIRAD publications (http://metroteach.com/SDG/index.html)

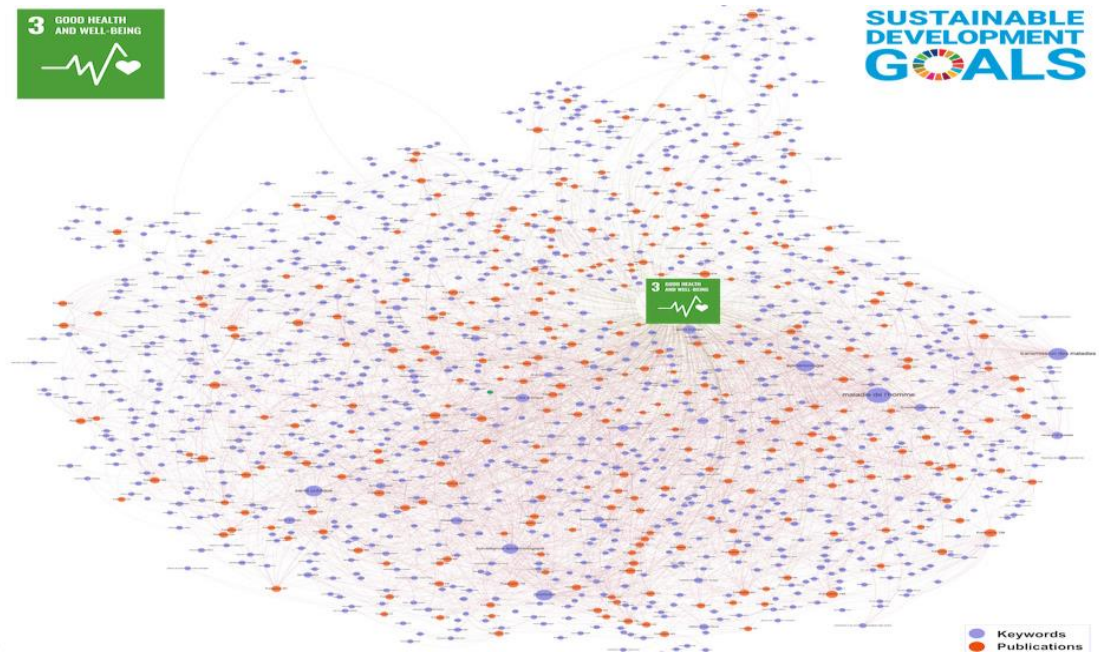


Fig. 7. Goal 3 - interactive tool allows you to explore CIRAD publications (http://metroteach.com/SDG/index.html)

7 Conclusion

The analysis of a Gephi network graph with 9812 nodes, 48580 links, a density of 0.001, an Eigenvector Centrality of 0.1062, and a modularity of 0.459 reveals a very sparsely connected network with a strong concentration of influence on a small number of key nodes. The network is also highly modular, suggesting the presence of distinct groups with stronger internal interactions.

The main conclusion of the research is to prove the importance of mapping publications for CIRAD and policymakers, considering complex systems within the SDGs. Publication mapping has significant value for both CIRAD and policymakers when considering complex systems within the SDGs, Table 7:

Table 7. Survey results.

BENEFITS FOR CIRAD
Identifying research gaps and opportunities: by mapping publications in different SDGs and research areas, CIRAD can identify areas where research is limited, highlighting potential areas for future focus and allocation of research resources. Overlaps or redundancies can occur, allowing for strategic collaboration and optimisation of resources. Emerging trends are developing, allowing CIRAD to anticipate future research directions and adapt its strategies.
Demonstrating the impact of research: mapping publications can help CIRAD visualise the breadth and depth of its research contributions to various SDGs. Track the reach and influence of its publications through citation analysis. Effectively communicate its research impact to stakeholders, including policymakers, donors, and the public.
Enhancing collaboration: mapping publications can reveal potential collaborators by identifying researchers working on similar topics within different disciplines or institutions. Facilitating connections and fostering interdisciplinary collaborations crucial to addressing the complex challenges of the SDGs.
BENEFITS FOR PUBLIC POLICY MAKERS
Inform evidence-based decision-making: mapping publications can help policymakers identify knowledge gaps and prioritise areas of research that require more investment to support policy development. Gain insights into emerging trends and anticipate future obstacles related to the SDGs. Evaluate the effectiveness of existing policies by analysing research on their implementation and results.
Monitor progress towards the SDGs: publication mapping can be used to track research efforts addressing different SDGs over time. Identify areas where considerable progress has been made through research breakthroughs. Highlight areas where additional efforts are needed to achieve the SDGs.

Facilitate international co-operation: mapping publications can identify research institutions and experts working on specific SDG targets globally. Promote knowledge sharing and collaboration between countries to address shared impediments

Considering the complex systems within the SDGs, mapping publications becomes even more relevant when considering the interconnected nature of the SDGs. By mapping publications on different SDGs, it is possible to identify synergies and potential conflicts between different SDGs, allowing for more holistic and integrated policy approaches as well as understanding the complex relationships between various SDG targets and the research efforts needed to achieve them. It is essential to promote transdisciplinary research that goes beyond disciplinary boundaries to address the interconnected dilemmas of the SDGs.

Acknowledge: Projeto FAPESP

References

Andrienko, Natalia; Andrienko, Gennady; Fuchs, Georg; Slingsby, Aidan; Turkay, Cagatay, Wrobel, Stefan (2020). "Visual analytics for understanding relationships between entities." In: Andrienko, Natalia; Andrienko, Gennady; Fuchs, Georg; Slingsby, Aidan; Turkay, Cagatay; Wrobel, Stefan. Visual analytics for data scientists. Cham: Springer, pp. 201-218. ISBN: 978 3 030 56146 8 https://doi.org/10.1007/978-3-030-56146-8_7

Ávila-Toscano, José-Hernando; Romero-Pérez, Ivón-Catherine; Marenco-Escuderos, Ailed; Saavedra-Guajardo, Eugenio (2018). "Identification of research thematic approaches based on keywords network analysis in Colombian social sciences." In: Thomas, Ciza (ed.). Data mining. London: InTechOpen. ISBN: 978 1 789235975 https://doi.org/10.5772/intechopen.76834

Bastian, Mathieu; Heymann, Sebastien; Jacomy, Mathieu (2009). "Gephi: an open-source software for exploring and manipulating networks." In: Proceedings of the International AAAI Conference on web and social media, v. 3, n. 1. Third international AAAI conference on weblogs and social media, pp. 361-362. https://ojs.aaai.org/index.php/ICWSM/article/view/13937

Centre for Complex Systems Modelling. (2018). Network Analysis for Sustainable Development**.** https://science.osti.gov/-/media/ascr/pdf/program-documents/docs/Complex_networked_systems_program_final.pdf

CIRAD Homepage. CIRAD. Retrieved from: https://www.cirad.fr/en

FAO. "The State of Food Security and Nutrition in the World 2021." Food and Agriculture Organization of the United Nations, 2021.

Gunning, Robert D., et al. "Mapping the contribution of agricultural research to the Sustainable Development Goals: A novel approach to understanding impacts and synergies." Global Food Security 26 (2020): 100404.

Kell, Douglas B., et al. "The sustainability of food production." Philosophical Transactions of the Royal Society B: Biological Sciences 364.1531 (2009): 907-918.

Le Blanc, D. (2015). Towards integration at last? The sustainable development goals as a network of targets. Sustainable Development, 23(3), 176-187. https://doi.org/10.1002/sd.1582

Pradhan, Prajal, et al. (2017). Network Analysis of Sustainable Development Goals (SDGs) Interdependencies: Implications for Policy Coherence and Integration. https://www.sciencedirect.com/science/article/abs/pii/S0195925523002615

Sachs, J. D., Schmidt-Traub, G., Kroll, C., Lafortune, G., & Fuller, G. (2019). Sustainable Development Report 2019. New York: Bertelsmann Stiftung and Sustainable Development Solutions Network. Retrieved from: https://www.sustainabledevelopment.report/reports/sustainable-development-report-2019/

Rastoin, Jean-Louis, and Yuna Chiffoleau. "Visualiser la contribution de la recherche agronomique au développement: une nécessité pour nourrir les débats et les politiques." Innovations Agronomiques 51 (2016): 89-99.

Soleimani, Nejat, et al. (2020). The Network of Global Sustainable Development Goals: A Network Science Approach". https://www.sciencedirect.com/topics/social-sciences/sdgs

United Nations. (2015). Transforming our world: the 2030 Agenda for Sustainable Development. Retrieved from: https://sdgs.un.org/2030agenda.

United Nations Sustainable Development Solutions Network. (2019). The Role of Network Analysis in Understanding and Addressing Sustainable Development Challenges**.** https://www.unsdsn.org/

Valdano, Daniele, et al. (2019). Using Network Science to Assess Progress Towards the Sustainable Development Goals. https://www.sciencedirect.com/science/article/abs/pii/S0195925523002615

World Bank Group. The World Bank Annual Report 2019: Ending Poverty, Investing in Opportunity. Washington, DC: World Bank. 2019. Retrieved from: https://documents1.worldbank.org/curated/en/156691570147766895/pdf/The-World-Bank-Annual-Report-2019-Ending-Poverty-Investing-in-Opportunity.pdf

World Bank. (2020). Using Network Science to Map and Measure Sustainable Development. https://datatopics.worldbank.org/sdgs

Weber, J.M., Lindenmeyer, C.P., Liò, P. and Lapkin, A.A. (2021), "Teaching sustainability as complex systems approach: a sustainable development goals workshop", International Journal of Sustainability in Higher Education, Vol. 22 No. 8, pp. 25-41. https://doi.org/10.1108/IJSHE-06-2020-0209

Cognitive Navigability: A philosophical invitation towards modelling cognitions

Hiott, Andrea

Universität Heidelberg and Brandenburg Institute of Technology
Andrea.hiott@uni-heidelberg.de

Abstract

Lack of a common understanding of 'cognition' within the cognitive sciences and a focus on finding its essence may be limiting our research. Agreeing upon a practical, theoretical orientation for cognition that holds across subjects of study (i.e., from insects to A.I.) may help us find answers to our challenges. Doing so requires general philosophy as well as practical models. Deeply interested in the latter but writing as a mere philosopher, I submit this paper as an invitation towards collaborating on complex models that trace cognitions across organisms and scales via navigability.

Keywords: cognition, navigability, landscape, affordance, regularities

Introduction

Many disciplines have come together in their common study of *cognitions* (intelligences, minds, mental states, etc.) to comprise the endeavor of cognitive science (CS). There are many exciting collaborations across its disciplines (McNamara, 2006; Vanney et al., 2023). In a strange twist, however, due to the excitement and passion generated by this interdisciplinary development, especially in the philosophical sections, it can seem at times that the focus has narrowed into a desire to find and define mind, intelligence, or consciousness rather than study its manifestations. That life is able to recognize itself as life is no small accomplishment, and it is no wonder we find enthusiasm in such revelries. Still, it is easy to become distracted by a search for essence. In continuing to debate 'hard problems' and 'mind-body problems' and determining what does or does not have intelligence or the 'mark of the mental' (Chalmers, 2007; Fodor, 1981; Neander, 2017; Hildt, 2019), we philosophers have gotten overly-entranced by our ability to recognize our own processes, locked into the wonders of turning cognition upon cognition. What we need is a way to visualize cognitions across scales and species, the way we have come to visualize forms of life. This modelling seems possible now via Bayesian inspired models, or new applications of graph theory and category theory, and the hope here is that we might agree on a general orientation for

cognitions, and then begin to model them from a non human-centric frame. One way to do this would be to orient cognition as navigabilities (connecting patterns) rather than parts dissected (Bateson, 1972; Northoff, 2018; van der Maas et al., 2017).

Our search to define cognitions has been long, and still there is no clear definition. Broadly speaking, these terms fit under the umbrella of *cognitions*, of that which is experienced (either by itself or by an observer) as *thought*, *mind, intelligence*, but each discipline and researcher uses these terms in diverse and overlapping ways, leaving what is meant by cognitions unexplained and its various terms used interchangeably, as if they were different but also as if they do not require clarification about such differences. There is a certain sloppiness accepted, a lack of light cast upon assumptions, as if we know what the cognitive is even as we search for it (Buzsaki, 2019). This ambiguity has led us philosophers into a familiar seductive circle—mind seems to be nowhere and everywhere the more one focuses on finding it, as if there is some 'vital essence' that we are at the brink of discovering. But what if there is not? What if, just as there is no one essence of life, there is no one essence of mind, no one kind of cognition, but only *cognitions* in the plural?

This potential circle is even more troublesome due to the sneaky ease of assuming definitions of cognitions from one's own experience of them, of defining other cognitions against one's own, which is partly at the heart of why the terms can be used so much even though there is no agreement about what they are or what they mean. Our assumed working definition of these terms can come from a personal experience of what we have experienced as *mind, mental*, or *thought* within ourselves, or it can be a particular line of education we have received through our field. In psychology, for example, we might have a very different reference base for 'cognitive' than we would have if we studied plants or Artificial Intelligence. If someone who has studied human psychology is now asked to accept that a plant or computer is cognitive, they will judge from their developmental position which is most likely a development that includes regularities (via books, papers, subjective experience, etc.) relative to a certain sort of cognition and so it may be impossible to accept a plant or A.I. system as cognitive within that definition. And in fact, there is no reason to believe they should accept this. What might be the case, however, is that an ant or A.I. system manifests cognition differently but in a similar pattern—that it is a matter of a different body and a different landscape, but that the patterns of the action itself are shared, and that we could model this in some way such that we can notice the patterns that are similar, the same way we might, in category theory, notice two very different paths leading to the same place. Still, without a common orientation of cognitive across disciplines, the researcher has

little choice but to default to their own understanding as developed within their trajectory and to reject or accept the plant or the A.I.'s cognition as valid or invalid based on that first-person clarity that is not brought into third-person awareness. In other words, as of now, there is no common orientation, what is expressed elsewhere as a System 3 representation (Hiott, 2022), of the term *cognitive* that allows them to step out of their own trajectory so to speak and look for patterns in common between that trajectory and the plant or A.I. trajectory without needing it to be the same in particulars. This is what a common orientation of the term *cognitions* can do for those studying its manifestations—give us an independent place from which to measure, a common orientation for assessment across fields, and parameters that must be at least loosely defined to manifest such patterns. This is, for example, what focusing on processes such as respiration or metabolism has done for science when it comes to the study of life: Rather than focus on finding the essence of life, one is able to focus on patterns that these activities have in common even when the particulars of those patterns are diverse. One is not left, as one once was, in trying to decide whether a plant or computer has 'vital essence'—the frame has simply moved away from this focus and instead we model patterns and processes in common among that which is studied as life, patterns such as respiration and reproduction. Today, the search for vital essence may have returned in the guise of a search for mind or intelligence, and it may have a similar answer.

Part of the wonder of cognitive science has been the sharing of diverse references; this sharing of knowledge between fields is itself partly how we have now come to a place where we can observe a common orientation for cognitions, one that can work for both plants and humans without needing their manifestations to look experientially similar, and one that can begin to help us understand whether or not we can discuss Artificial Intelligence as a cognition. This is an important step to take because solving problems relative to patterns of cognitions—be those depression, anxiety, climate change, technological advancement, or economic systems—is certainly in a state of urgency, as evidenced in various reports such as the latest from the United Nations (IPCC, 2023) and the mental health reports of both the European Union and the American Psychological Association for 2023. Finding the vital essence of intelligence is not necessary for addressing those urgencies, just as finding the vital essence of life was not necessary for progressing forward in better understanding lung disease, stomach ulcers, or insect migration. What is necessary is a common orientation that can allow for these processes to be modelled across fields.

Studying the cognitive without defining it

A rigorous study of life is now possible without defining the term "life", without agreement on any one definition. This was not always the case. There was a time when the main focus was on finding an objective definition or vital essence, but through technological developments such as microscopes, data collection, widespread printing and dissemination of literature, common patterns came into focus that allowed the search for essence to fade into the background (Northoff, 2018; Ginsburg, 2019). It became clear that rather than focus on finding some absolute definition or essence, one could accept an overall orientation of shared patterns with clear enough parameters such that the life sciences could observe and study its many divergent manifestations (Jablonka, 2022).

Today, 'life' is not defined, but there are models and external representations that allows us to better understand it. These models are models of processes that forms of life share in common, such as digestion, reproduction, aspiration, metabolization, growth, and excretion. These patterns look very different in different manifestations of life (in cells, in butterflies, in cats, in humans), and yet share basic processual patterns. When those patterns are modeled, studied, and shared, we gain greater insight in our research. It is now possible to orient the cognitive such that we also notice different manifestations of shared processual patterns. We can study very different manifestations of cognitions (mind, intelligence, thought) without needing those to fit some absolute definition or essence, and we can do this through accepting a common orientation. What we need then are basic guiding patterns present in all we study as cognitive that can orient our focus. Luckily, we have already started to observe them: There are already processes common to all we study as *cognitions*, and those processes can be understood as *the navigabilities* of the subject in question, as the representations (models) of its ongoing *way-making*, an orienting term explained below. Through navigability parameters of body, trajectory, and landscape, we can proceed with a common orientation of the term cognitive, one that applies regardless of the discipline and one that does not require a vital essence or definition of *cognition* (and all its associated terms).

Modeling navigabilities

As a body, there is no choice but to make way, so long as the body is a body: It arrives encountering and making its way through that encounter, and when it is no longer making its way, it is no longer a body. This is done in multiple landscapes at multiple scales according to body and goal. As we will see, this does not mean the body has to be literally moving through traditional geographical space; it only means the body is moving through what phenomenologists call its 'life space' or life world (Buttimer, 1976) as a body. At most basic, it is still processing the encounter: To breathe, for

example, is not to make way (though it may be necessary for it). Making way is continuing one's path through the spatiotemporal regularities of its encounter as a body (Barlow, 2001). We will get to what a body is soon enough, but for now we can say that we can assess (model, represent) a body's making-way as its navigabilities through its life space (DeClerck, 2018), and we can understand this as a common orientation for the term *cognitive*.

Navigability is purposively left general but this will need further unpacking, some of which will come below, but let me first give a general idea of what is meant: Much as we orient the study of life through studies of bodily processes (i.e. respiration, digestion), we can orient the study of cognition through its navigabilities (i.e. walking, swimming, crawling, remembering, conversing, reading, etc.). A body develops these navigabilities through historical and individual landscapes (statistically regularized life spaces), and these can be externally represented similar to the ways we represent life process via systems like digestion or respiration. When it comes to this orientation of cognitive, the frame then encompasses all that has traditionally been termed as mental or physical: Navigability can be modeled as embodied trajectories through any sensorily-regularized life space from the position of the body being studied, even when that space is the body's own awareness of its patterns, such as when we "notice our thoughts" or "become conscious of self." Processes of navigability relative to the cognitive can be understood as embodied movements through time and space, even when that time and space is not traditionally three-dimensionally geographical.[i] There are many statistically regular topographies of sensory regularities (Barlow, 2001) that a body makes its way through—emotional, historical, linguistic, conceptual, etc. It is important to emphasize that navigability is a matter of modeling and representing the dynamic process of a body's navigability in a modelled space—maps of a dynamic territory that must not be mistaken for the map. Way-making is the process, but this process could be modelled or mapped *as navigability*.

Once cognition is oriented (not defined as) making-way, it is the observer who clarifies which body (i.e., is one observing a cell body or the human body?) and which area of life space (landscape) will be the scale of study. Studying one landscape does not preclude or cancel the existence of others—it is only a model of one process. Just as modeling one part of the body as respiration and another as digestion does not mean those processes in reality can actually be separated (as they overlap), it only means that these parameterized parts of the process hang together and exhibit enough statistical regularity to be modeled as systems. We will come to a definition of a body soon enough, but for now, the point is that the observer designates the body being observed and the landscape it is observing the body navigate.

Within this encountering, there will be specific spaces relative to study—those might be geographical spaces like cities or forests, emotional spaces with regularities of feeling (Kraft, 2016), morphological spaces with regularities of form (Fields, 2022), conceptual spaces with regularities of linguist representations (Gardenfors, 2000), or the body's own patterns of movement as observed in neural patterns (Bellmund, 2019). There is no limit to where the scale can be set, so long as it is a body that is encountering a set of regularities. It is important to note that the body and its encountering can overlap in their regularities, just as the body can encounter itself in a mirror, it can encounter itself as a landscape. This orients *cognitions* to be the navigabilities of a body within a landscape, navigability that could be represented as trajectories within topographies. Just as there is no one kind of body that engages in digestion or respiration, there is no one kind of body that engages in navigabilities such as swimming, crawling, reading or conversing. Still, these are navigabilities that exhibit similar patterns within identifiable statistically regular landscapes (Barlow, 2001). How to mathematize this?

Navigabilities as laid out here means that walking and swimming can be understood as pre-reflective cognitive actions and that a body could also reflect on those actions as well: in both cases, a body is making way and those patterns can be modeled as its navigabilities within landscape. In the case of walking, the landscape may be a city or forest. In the case of thinking about walking, the landscape may be that body's own conceptual and linguistic landscapes—concept space. These thoughtful manifestations of the cognitive, though the first we as humans may recognize, are more like the growing tips of a very old and long process, the same one that is our walking, crawling, swimming, driving, and so on. We could study these as manifestations of navigability, however, ones that manifest very differently but still exhibit similar patterns according to body and landscape. Navigability is the pattern that connects (Bateson, 1972), and the connective pattern does not depend on the type of body or landscape.

In more general terms, cognitions can be understood through the orientation of what is navigating and can be observed (and perhaps dynamically formalized) as navigabilities—as regularities of body and landscapes, and as an explicated representation of that interactive trajectory, which is also a string into that body's experienced affordances. Let's turn now briefly to notice how navigability is an orientation for cognitions in various fields related to CS, then we will come back to the terms of the navigability framework—body, landscape, trajectory, affordance—for a closer look at what is really meant with them.

Precedents for the navigability framework

Hippocampal research provides us one way to orient the term cognitive at scales of navigability. This comes about due to the hippocampal formation's dual role: it is crucial to memory and knowledge acquisition (Eichenbaum, 2017), and it is also crucial to overall orientation, navigation, and wayfinding through physical environments (Buszaki & Moser, 2013; O'Keefe and Nadal, 1978). Research here illustrates that activities such as memory and knowledge acquisition are, for the body, matters of navigability. This area of the brain—the hippocampal formation and entorhinal cortex—is known to be crucial for memory and knowledge-acquisition through historic cases such as that of H.M. who lost the ability to remember and acquire durable new knowledge when his hippocampus was removed in the 1950s in a procedure done to cure him of his epilepsy. This same area of the brain was, a few decades later, discovered to be the location of the body's so-called GPS, with special cells (now known as place cells, grid cells, border cells, etc.) which fire specifically for the body as it makes its way through a landscape. In general, this allows us to understand navigability across traditional dividing lines of mental and physical. This was first thought to be only a geographical landscape, but in our most recent research it has become clear that these patterns hold for conceptual and virtual landscapes (Constantinescu, 2016; Bellmund, 2018) and computational models have been developed towards exploring the patterns that connect these different spaces. This offers a sketch of the potential scalability of manifestations of cognition that can be oriented as navigabilities.

An important note must be made, however: navigability does not have to be a matter of navigation for the subject—wandering, digression, being lost—all these are not navigation for the subject. Still, from an observer's point of view—and in any form of modeling, there is an observer, someone who is assessing this activity, it can be understood as navigability—as being able to find way, or not find way. Navigability thus includes specific and general uses of navigation and in neuroscience, and the different ways they are used in other fields. This is why the general term way-making is used for the process, such that all these terms are included: A body that is living, even when it is lost, even when it does not know that it is making way, even when it is moving to a goal, even when it is finding a previous path, even when it knows what it is making way towards, or when it cannot get there, is still making way and the movement still modellable as navigability.

I highlight the hippocampal example here because it is a clear example of how one can understand navigability across traditional bounds, but there are other examples in CS which use navigability as the pattern that connects for an orientation to study the

cognitive. Perhaps the most famous within developmental biology comes from biologist Michael Levin and physicist Chris Fields and their paper which looks at how cognition can be understood as embodied, observer-based competency of movement through a particular space. Though I have called these life spaces 'landscapes', because they have been shaped or molded in statistically regular ways, the idea is basically the same when they talk about metabolic space, transcriptional space, morphological space, and '3D motion spaces' (what most people think of as traditional behavioral space). Though the details would take too long to discuss, this is an engineering angle on understanding cognition across species and scales that uses navigation as its invariant (Fields, 2022; Levin, 2022).

In fact, there has long been a push in many of the evolutionary and biological sciences that suggests an orientation of cognition as navigability. In the work of Karola Stotz, for example, we find a progression towards an orientation of studying cognition as "how the mind grows" within the developmental system, whereby cognition is understood as developing along with the organism as it interacts with resources (Stotz, 2012). In other words, the mind is developing as the organism is developing *as the organism*. She and her co-authors write about regularities of cognitive development as something like trajectories of bodily interaction with the environment. To put this in my words, one could model this as a body developing cognition *as its own bodily navigability*.

There are also examples in complexity science and from those trying to understand how to think of Artificial Intelligence. In a recent book, for example, Max Bennet, the cofounder & CEO of Alby, an AI company, tries to understand intelligence, especially its relation to Artificial Intelligence, through looking at the evolution of the brain largely through an orientation of the body's need for steering and navigation, showing how this is continuous across a wide divergent of kinds of bodies and brains (Bennett, 2023).

We also find similar ideas across the fields of mathematics and applications of category theory (Baylor & Montero, 2023), and relative to complexity science. One recent example comes from a Sante Fe Institute partnership with Richard Solé, in which he and collaborators write about "the space of cognitions" and imagine different scales of cognitive systems that could exist beyond traditional bounds—a flock of birds for example as being something like a 'liquid brain' existing in a space where what would be neural units in a human brain are living units at a wider scale and move in more visible ways (Solé, 2019). We could approach cognition, they write, as numerous cognitive spaces and numerous kinds of bodies traversing them. Though I would argue that the

flock of birds is more like a 'liquid body' than a liquid brain, one can nevertheless see how the 'pattern that connects' here could still be understood through an orientation of navigability.

There are more such examples, and I am piecing these together as part of a larger work, but I present them generally here so that it is clear that this is already a pattern across CS fields and disciplines. To better understand this framework and how it applies across the many fields and disciplines of CS, it is helpful to briefly lay out the terms being used and clarify how they are used in the navigability framework. These terms are body, landscape, affordance, and trajectory.

Body

Through assessing navigabilities, we can understand *cognitions* as embodied trajectories through landscapes. It is also important we understand what is not navigability. An embodied act here means 'the action of a body' and the body is understood as 'that which has affordances specific and necessary to its unique biological trajectory—regardless how another body might need or encompass it.' A cell exists as embodied action within this definition whereas an internet or phone or rock does not: cells are actions specific and necessary to their biological motivation. Biological motivation here is meant in the most general sense but perhaps most closely associated with self-organization (Collier, 2004) and boundaries of self that are modeled as Markov blankets (Kirchhoff, 2018). In the case of the cell, for example, it is action that is specific to the cell's continuity, regardless of whether it is specific to the human body that encompasses it.

An internet or phone does not act as its own biological motivation: there is no action taken by it that is not referred to affordances outside it. If one were to try and say, for example, that the battery is the phone's affordance, it would mean the battery motivates the phone to do something that needs no reference to a trajectory other than the trajectory of the phone, but this cannot be done because the phone is itself an extension of some other (in this case human) trajectory. The same might be said of a brick. One cannot make the case of biological motivation when it comes to an object like a brick that has been created to build a house. The material might be pure granite, which some might say is an organic or natural material and thus confuse it as having biological motivation (because it is biological) but there is nothing about the brick that motivates that form to continue as a brick, there is no trajectory of affordances specific to that form—if it breaks or simply erodes away, there is nothing in the brick trying to counteract that entropy. The brick requires reference to some other body's trajectory for the motivation of its form and the continuance of its form. In this way, we avoid the notion that everything is cognition, and we have a

clear understanding of what is not cognition even without defining the term.

Landscape

A landscape is a set of encountered statistical regularities for a position. This can be in geographical spaces, in transcriptional spaces, in morphological spaces, in conceptual spaces, in mathematical spaces, etc. In all these various landscapes—in the context as specified by the field of study, cognitions can be understood as the ways the body in question (the subject being studied) makes through whatever it is encountering (wherever the scale has been set within the life space). A part of life space that is being observed for its regularities is a landscape because it has been molded or 'scaped' by the trajectories creating it (Strauss, 1956; Seamon, 2001). What is crucial is that we understand that a subject or position is defined by its affordances relative to those parameters. Navigability comes in different ways in different landscapes. Making way does not necessarily mean moving the entire subject through physical space but can also be a matter of moving through conceptual space, transcriptional space, morphological space, linguistic space, etc.—the landscape is the regularities of space and time that predominate for the subject in question as directed by the inquiry or interest of observation. And modeling the ways it makes is modeling its navigability.

Affordances

When making-way is modeled as navigability, the body has a pattern of unique affordances—it is the observation of what has been designated as the position co-developing with what has been designated as the landscape such that the affordances only apply to that position, to its orienting motivation. Something like a rake being blown by the wind, for example, would not be making-way or expressing navigability because it has no continuity of affordances relative to itself (as that position), the landscape, and its trajectory. In other words, the trajectory has no continuity with the position and its landscape—the wind is moving the rake, but the rake is not moving itself via the wind; the regularities of the interaction are not its affordances.

Trajectories

The trajectory is a dynamic modeling of navigabilities. It is the unbroken path of a body encountering in landscape. This can be in any landscape such that the path is simply the line of regularities as encountered by the body. We might model these or formulize them into standardized representations such that we can observe how different trajectories overlap or are closer or farther from others, be those between people, within the same person, within species or between them. The way or path is always a matter of the subject and

what is being encountered or 'moved through'. Cognitions can only be measured through observation of *ways made*.

Why a common orientation is needed

This orientation is meaningful and worth the effort for the following reasons: it offers a common baseline for *cognitions* that can be used across the growing disciplines of cognitive science, from technology and computer science to psychology and philosophy to developmental biology and physics; it unsticks *cognitive* from human-centric scaffolds and ruts, and helps humans avoid the subtle unnoticed assumption that all *cognitions* are something like their experience of them; it allows that the term *cognitive* to apply across species and scales in a specific way without collapsing cognitions together across those scales and species; and it allows a framework from which we can parse terms such as mind, intelligence, and thought that gives the many sciences of cognitions a common orientation.

This is a deep reorientation of current paradigms, requiring we reimagine our research, but it does not minimize that research or debunk it. Rather, it emboldens and strengthens it, so long as one is willing to accept the efforts of reorientation, efforts such as re-reading texts and rechanneling through this orientation of navigability. As part of this framework, if a way is found to mathematize and model navigability, researchers can compare patterns of cognitions across scales and species without needing to force particulars or definitions upon their areas of study. This provides a wide enough set of parameters such that all disciplines within cognitive science, from the biological to the neuroscientific to the computational, can work with the term in the background instead of focusing on finding its essence.

Summary

Cognitions *are* bodies making way. They can be studied via a body's navigabilities within landscapes, and hopefully modeled as categorial complex systems (or something of a better fit?). Navigabilties include actions such as crawling, swimming, walking, conversing, remembering, reading, and thinking—the activities we associate with any form of mind, intelligence or consciousness develop as processes of navigability within particular landscapes—geographical, linguistic, or social. These are different manifestations of a similar, dynamical pattern (navigability) which we can understand as cognitions. The pattern is similar even as the bodies and landscapes diverge in particulars. So long as we are clear about the body and landscape being observed, we can map the trajectory and (the hope is) find a way to mathematically model these dynamic patterns so as to compare and contrast cognitions across scales, body-types, and landscapes.

Taking this theoretical framework to heart means all cognitions, across scales, disciplines, and body-types, can be observed and one day modelled as navigabilities: actions are *cognitions* if they are embodied acts with modellable trajectories of autopoietic affordance, even when the regularities overlap with those of the observer, as is the case when a body observes its own patterns of making way through geographical, conceptual or social landscapes.

The focus here is on the patterns, not the types of form or environment. This allows us to orient through a complex systems approach towards studying "patterns that connect" rather than looking for parts that can be defined and dissected. Navigabilities are these patterns, which can be represented trajectories or paths—an expression of a particular body making way in a particular landscape, a trajectory of dynamic but statistically regular affordances. This orientation allows us to focus on patterns of bodies and landscapes rather than on what substantiates those patterns, to be specific without having to share particulars, to notice what shares observable regularities and processes, and to orient what cognitions in a way that will let us approach our study and research across body-types, scales, and disciplines and avoid the tempting habit of assuming other beings are *cognitive like us*. We can release ourselves from the enthusiastic Cartesian awakening (important as it is) that has led us to search for a definition or essence or 'mark of the mental' and agree on a common orientation: Cognitions are shared regularities that can be assessed across various subject matters, substrates, and scales of the cognitive sciences. To accept this is not to disregard or ignore the enthusiasm inherent in our awareness of our own cognition. Rather it is to accept its mystery.

References

Bateson, G. (1972). *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology* (pp. xxvi, 545).

Bennett, Max.(2023). *A Brief History of Intelligence: Evolution, AI, and the Five Breakthroughs That Made Our Brains***,** Mariner Books.

Barlow, H. (2001). The exploitation of regularities in the environment by the brain. *Behavioral and Brain Sciences*, *24*(4), 602–607.

Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*(6415). https://doi.org/10.1126/science.aat6766

Buttimer, A. (1976). Grasping the Dynamism of Lifeworld. *Annals of the Association of American Geographers*, *66*(2), 277–292. http://www.jstor.org/stable/2562470

Buzsáki, G. (2020). The Brain–Cognitive Behavior Problem: A Retrospective. *ENeuro*, *7*(4). https://doi.org/10.1523/ENEURO.0069-20.2020

Buzsáki, G., & Moser, E. I. (2013). Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature Neuroscience*, *16*(2), 130–138. https://doi.org/10.1038/nn.3304

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941

Craig, M. T., & McBain, C. J. (2015). Navigating the circuitry of the brain's GPS system: Future challenges for neurophysiologists. *Hippocampus*, *25*(6), 736–743. PubMed. https://doi.org/10.1002/hipo.22456

Chalmers, David (2007). The hard problem of consciousness. In Max Velmans & Susan Schneider (eds.), The Blackwell Companion to Consciousness. Chichester, UK: Blackwell. pp. 32–42.

Collier, J. (2004). Self-organization, Individuation and Identity. *Revue internationale de philosophie*, 228, 151-172.

Declerck, G., Lenay, C. (2018). Living in Space. A Phenomenological Account. In: Pissaloux, E., Velazquez, R. (eds) Mobility of Visually Impaired People. Springer, Cham. https://doi.org/10.1007/978-3-319-54446-5_1

Disessa, A. A. (2002). Why "Conceptual Ecology" is a Good Idea, (pp. 28–60). Springer Netherlands.

Eichenbaum, H. (2017). The role of the hippocampus in navigation is memory. *Journal of Neurophysiology*, *117*(4), 1785–1796.

Fields C, Levin M. Competency in Navigating Arbitrary Spaces as an Invariant for Analyzing Cognition in Diverse Embodiments. Entropy (Basel). 2022 Jun 12;24(6):819. doi: 10.3390/e24060819. PMID: 35741540; PMCID: PMC9222757.

Fodor, J. A. (1981). The mind–body problem. *Scientific American*, *244*(1), 114–123.

Fuchs, T. (2017). *Ecology of the Brain: The phenomenology and biology of the embodied mind* (1st ed.). Oxford University Press.

Griffiths, P., & Stotz, K. (2004). How the Mind Grows: A Developmental Perspective on the Biology of Cognition. Synthese, 122(2), 29-51.

Gärdenfors, Peter (2000). Conceptual Spaces: The Geometry of Thought. Tijdschrift Voor Filosofie 64 (1):180-181.

Gibson, J.J. (1966). *The senses considered as perceptual systems.* Houghton Mifflin.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, *105*(3), 416–434. https://doi.org/10.1016/j.neuron.2019.12.002

Hildt E. Artificial Intelligence: Does Consciousness Matter? Front Psychol. 2019 Jul

2;10:1535. doi: 10.3389/fpsyg.2019.01535. PMID: 31312167; PMCID: PMC6614488.

Hiott, A. (2022). *Ecological Memory: The spatiotemporal commons of physical and conceptual navigation*.

Ginsburg, S., & Jablonka, E. (2019). *The evolution of the sensitive soul: Learning and the origins of consciousness.* (A. Zeligowski, Illustrator). The MIT Press.

Jablonka, Eva & Ginsburg, Simona (2022). Learning and the Evolution of Conscious Agents. Biosemiotics 15 (3):401-437.

Kirchhoff M, Parr T, Palacios E, Friston K, Kiverstein J. The Markov blankets of life: autonomy, active inference and the free energy principle. J R Soc Interface. 2018 Jan;15(138):20170792. doi: 10.1098/rsif.2017.0792. PMID: 29343629; PMCID: PMC5805980.

Kraftl, P. (2016). Emotional Geographies and the Study of Education Spaces. In: Zembylas, M., Schutz, P. (eds) Methodological Advances in Research on Emotion and Education. Springer, Cham.

Levin, Michael: *Bioelectric networks: the cognitive glue enabling evolutionary scaling from physiology to mind:* Michael Levin

Levin M. Technological Approach to Mind Everywhere: An Experimentally-Grounded Framework for Understanding Diverse Bodies and Minds. Front Syst Neurosci. 2022 Mar 24;16:768201. doi: 10.3389/fnsys.2022.768201. PMID: 35401131; PMCID: PMC8988303.

McNamara, D.S. (2006), Bringing Cognitive Science into Education, and Back Again: The Value of Interdisciplinary Research. Cognitive Science, 30: 605-608.

Neander, Karen (2017). A Mark of the Mental: A Defence of Informational Teleosemantics. Cambridge, USA: MIT Press.

Northoff, G. (2018). The Spontaneous Brain: From the Mind-Body to the World-Brain Problem. MIT Press.

O'Keefe J. & Nadel L. (1978). *The hippocampus as a cognitive map*. Clarendon Press ; Oxford University Press.

Seamon, D. (2015). *A geography of the lifeworld: Movement, rest, and encounter*.

Seamon, D. (2018). *Life takes place: Phenomenology, lifeworlds and place making*.

Solé, Ricard; Moses, Melanie; Forrest, Stephanie, 2019 Liquid brains, solid brains*Phil. Trans. R. Soc. B***374**2019004020190040

Spiers, H. J., & Maguire, E. A. (2008). The dynamic nature of cognition during wayfinding. *Journal of Environmental Psychology*, *28*(3), 232–249. https://doi.org/10.1016/j.jenvp.2008.02.006

Stotz, K. (2010). Human nature and cognitive-developmental niche construction. *Phenomenology and the Cognitive Sciences*, *9*(4), 483-501. https://doi.org/10.1007/s11097-010-9178-7

Strauss, Erwin "Vom Sinn der Sinne" (1956), Engl. translation (1963) “The primary world of the Senses: A Vindication.” Glencoe/Ill.: Free Press.

Thomas, H. (2013). *The Body and Everyday Life* (1st ed.). Routledge.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind.* Belknap Press/Harvard University Press.

Van Der Maas HLJ, Kan K-J, Marsman M, Stevenson CE. Network Models for Cognitive Development and Intelligence. *Journal of Intelligence.* 2017; 5(2):16. https://doi.org/10.3390/jintelligence5020016

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, Mass: MIT Press.

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., Dolan, R. J., & Blakemore, S.-J. (2013b). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, *22*(1), 264–271. https://doi.org/10.1016/j.concog.2013.01.004

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2019). The Tolman-Eichenbaum Machine: Unifying space and relational memory through generalisation in the hippocampal formation. *BioRxiv*, 770495. https://doi.org/10.1101/770495

Zeithamova, D., & Bowman, C. R. (2020). Generalization and the hippocampus: More than one story? *Neurobiology of Learning and Memory*, *175*, 107317. https://doi.org/10.1016/j.nlm.2020.107317

Quammen D. (2018). *The tangled tree : a radical new history of life* (First Simon & Schuster hardcover). Simon & Schuster.

[i] Since I am using this analogy, one might ask how navigabilities are different from life processes. Something like digestion is not a navigability because it is not an embodied trajectory that can be separated from the trajectories of its peers. To be embodied means to have a trajectory of affordances that only apply to that body, and its navigabilities are those which can be shown as embodied trajectories relative to its peers and that unbroken trajectory. A digestive system is the path of a body, not a body. The digestive system is a path, not a body making paths.

Complex Networks: Structure & Dynamics I

FRCCS

Comparative Analysis of Structural Backbone Extraction Techniques

Ali Yassin[1✓], Hocine Cherifi[2], Hamida Seba[3] and Olivier Togni[1]

[1] *Laboratoire d'Informatique de Bourgogne, University of Burgundy, Dijon, France; ali.yassin@u-bourgogne.fr.*
[2] *ICB UMR 6303 CNRS - Univ. Bourgogne - Franche-Comté, Dijon, France;*
[3] *Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France;*

✓ *Presenting author*

Abstract. This study explores structural backbone extraction methods in diverse real-world networks, ranging from character to social networks. Eight techniques from the netbone package are analyzed using Jaccard and Overlap coefficients to measure similarities. The Kolmogorov–Smirnov statistic evaluates weight and degree distribution preservation. Results highlight correlations and hierarchical relationships among techniques, with Doubly Stochastic outperforming in mimicking original network distributions. The study's comprehensive insights contribute to refining and advancing structural filtering techniques, fostering a deeper understanding of complex systems.

Keywords. *Complex Networks; Backbone Extraction; Filtering Techniques; Sparsification*

1 Introduction

Networks are indispensable tools for representing and comprehending intricate systems, providing diverse applications such as pinpointing pivotal nodes [9, 10, 12], revealing communities [3, 8], and delving into the dynamics of networks [2, 1]. Nevertheless, the computational challenges become daunting when confronted with large networks.
Researchers have introduced several techniques to tackle this issue, focusing on diminishing network size while maintaining its fundamental properties. Two primary approaches have emerged to attain this objective: structural and statistical methods. Structural methods strive to retain a set of topological features inherent in the network while downsizing its overall scale. In contrast, statistical methods eliminate noise by selectively filtering out nodes or links that might obscure the network's underlying structure. Researchers have extensively compared statistical methods [14, 13], often against a few structural methods [7, 4]. However, there's been limited focus on comparing only the structural backbone extraction methods [5, 6, 11].
We compared eight structural filtering techniques in the World Air Transportation network in a previous work. This study extends it across 39 real-world networks of different sizes and covering various fields. Our initial experiment compares the similarities among eight structural filtering techniques from the netbone package [15]. We use the Jaccard Coefficient and Overlap Coefficient to assess the similarity between the edge sets of technique pairs. Sub-

sequently, we examine the weight and degree distributions of the resulting backbones. The Kolmogorov–Smirnov (KS) statistic evaluates deviations from the original distributions within each network. Then, we rank the methods based on their KS statistic values, reflecting their alignment with the original network's distributions.

2 Experimental Results

In Fig 1, the heatmaps of the Jaccard Score provide insights into significant correlations among specific pairs of techniques. Notably, there are strong correlations between Primary Linkage Analysis and Minimum Spanning Tree Filters (PLAM-MSP) and Ultrametric Backbone and Minimum Spanning Tree Filters (UMB-MSP). The H-Backbone and Doubly Stochastic Filters stand out due to their lack of correlation with other techniques, suggesting distinctive behavior.

Examining the heatmaps of the Overlap Coefficient, hierarchical relationships among backbones become apparent. For instance, the Minimum Spanning Tree is included in the Primary Linkage Analysis Backbone, the Minimum Spanning Tree Backbone is included in the Ultrametric Backbone, and the Ultrametric Backbone is included in the Metric Backbone. High percentages of overlap are observed in specific pairs, such as 90% and 97% of High Salience Skeleton edges overlapping with Planar Maximally Filtered Graph and Metric Backbones, respectively. Additionally, 93% of Primary Linkage Analysis Backbones overlap with the Planar Maximally Filtered Graph Backbone. The H-Backbone shows, on average, 84% and 86% overlap with Planar Maximally Filtered Graph and Metric Backbones, respectively. Ultrametric Backbone edges overlap, on average, by 85% with the Planar Maximally Filtered Graph Backbone. High Salience Skeleton edges overlap by, on average, 83% and 88% with Minimum Spanning Tree and Ultrametric Backbones, respectively. The Doubly Stochastic filter produces unique backbones with low overlap percentages with other methods.

The violin and box plots further illustrate the performance of backbone extraction methods. They highlight the superiority of Doubly Stochastic in extracting backbones that closely match the original network's weight and degree distribution. Conversely, H-Backbones exhibit the furthest weight distribution, and Primary Linkage Analysis backbones display the furthest degree distribution. These findings contribute valuable insights into the comparative effectiveness of structural filtering techniques across diverse real-world networks.

3 Acknowledgment

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

References

[1] Md Arquam, Anurag Singh, and Hocine Cherifi. Impact of seasonal conditions on vector-borne epidemiological dynamics. *IEEE Access*, 8:94510–94525, 2020.

[2] Debayan Chakraborty, Anurag Singh, and Hocine Cherifi. Immunization strategies based on the overlapping nodes in networks with community structure. In *Computational Social Networks: 5th International Conference, CSoNet 2016, Ho Chi Minh City, Vietnam, August 2-4, 2016, Proceedings 5*, pages 62–73. Springer International Publishing, 2016.

[3] Hocine Cherifi, Gergely Palla, Boleslaw K Szymanski, and Xiaoyan Lu. On community structure in complex networks: challenges and opportunities. *Applied Network Science*,

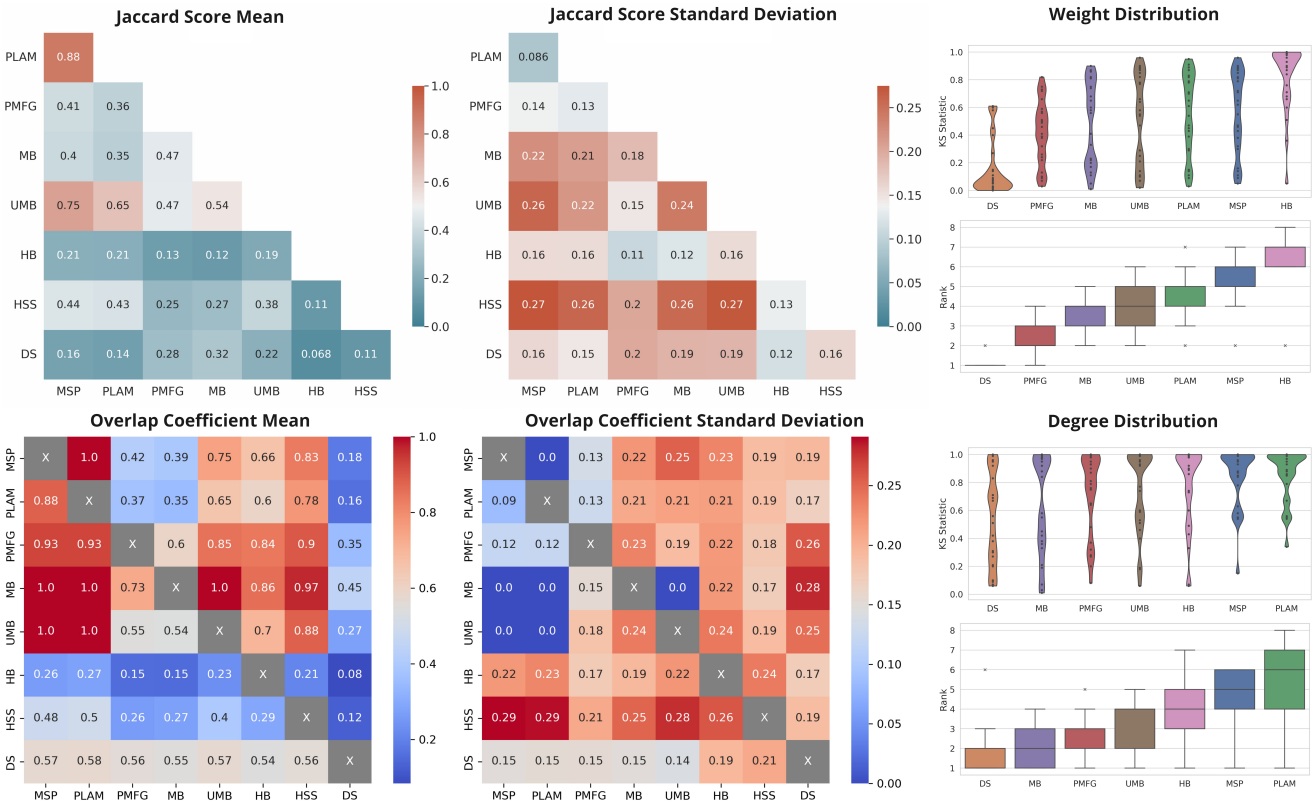


Figure 1: The top heatmaps display the mean and standard deviation of Jaccard Scores for each method pair, while the bottom heatmaps show the mean and standard deviation of Overlap Coefficient Scores. Additionally, the violin plot illustrates KS-static values between the backbone and original weight (degree) distribution across all networks. Boxplots depict method ranks based on the K-S statistic between each filtering technique's weight (degree) distribution and the original network. Evaluated techniques include MSP (Minimum Spanning Tree), PLAM (Primary Linkage Analysis Method), PMFG (Planar Maximally Filtered Graph), HSS (High Salience Skeleton), MB (Metric Backbone), UMB (Ultrametric Backbone), DS (Doubly Stochastic), and HB (H-Backbone).

4(1):1–35, 2019.

[4] Liang Dai, Ben Derudder, and Xingjian Liu. The evolving structure of the Southeast Asian air transport network through the lens of complex networks, 1979–2012. *Journal of Transport Geography*, 2018.

[5] Zakariya Ghalmane, Chantal Cherifi, Hocine Cherifi, and Mohammed El Hassouni. Extracting backbones in weighted modular complex networks. *Scientific Reports*, 10(1):15539, 2020.

[6] Zakariya Ghalmane, Chantal Cherifi, Hocine Cherifi, and Mohammed El Hassouni. Extracting modular-based backbones in weighted networks. *Information Sciences*, 576:454–474, 2021.

[7] Carlos Henrique Gomes Ferreira, Fabricio Murai, Ana P. C. Silva, Martino Trevisan, Luca Vassio, Idilio Drago, Marco Mellia, and Jussara M. Almeida. On network backbone extraction for modeling online collective behavior. *PLOS ONE*, 2022.

[8] Keziban Orman, Vincent Labatut, and Hocine Cherifi. An empirical study of the relation between community structure and transitivity. In *Complex Networks*, pages 99–110. Springer Berlin Heidelberg Berlin, Heidelberg, 2013.

[9] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Interplay between

hierarchy and centrality in complex networks. *IEEE Access*, 8:129717–129742, 2020.

[10] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Characterizing the interactions between classical and community-aware centrality measures in complex networks. *Scientific reports*, 11(1):10088, 2021.

[11] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Modularity-based backbone extraction in weighted complex networks. In *International Conference on Network Science*, pages 67–79. Springer International Publishing Cham, 2022.

[12] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Comparative evaluation of community-aware centrality measures. *Quality & Quantity*, 57(2):1273–1302, 2023.

[13] Ali Yassin, Hocine Cherifi, Hamida Seba, and Olivier Togni. Exploring Statistical Backbone Filtering Techniques in the Air Transportation Network. In *2022 IEEE Workshop on Complexity in Engineering (COMPENG)*. IEEE.

[14] Ali Yassin, Hocine Cherifi, Hamida Seba, and Olivier Togni. Air Transport Network: A Comparison of Statistical Backbone Filtering Techniques. In *Complex Networks and Their Applications XI*. Springer International Publishing, 2023.

[15] Ali Yassin, Abbas Haidar, Hocine Cherifi, Hamida Seba, and Olivier Togni. An evaluation tool for backbone extraction techniques in weighted complex networks. *Scientific Reports*, 2023.

FRCCS

Compression-based inference of network motif sets

Alexis Bénichou[1✓], Jean-Baptiste Masson[1] and Christian L. Vestergaard[1]

[1] *Decision and Bayesian Computation, Neuroscience department, Institut Pasteur, CNRS, Paris, France; alexis.benichou@pasteur.fr, jbmasson@pasteur.fr, christian.vestergaard@pasteur.fr.*

[✓] *Presenting author*

Abstract. Physical and functional constraints on empirical networks lead to topological patterns across multiple scales. A particular type of higher-order network feature that has received considerable interest is network motifs—statistically regular subgraphs. These may implement fundamental logical or functional circuits and have thus be referred to as "building blocks of complex networks". Their well defined structures and small sizes furthermore mean that their function is amenable to testing in synthetic and natural experiments. The statistical inference of network motifs is however fraught with challenges, from defining the appropriate null model and sampling it to accounting for the large number of candidate motifs and their potential correlations in statistical testing. We develop a framework for motif set inference based on lossless network compression using subgraph contractions. The minimum description length principle lets us select the most significant set of motifs as well as other prominent network features in terms of their combined compression of the network. Our approach overcomes the common problems in mainstream frequency-based motif analysis while avoiding repeated subgraph census on a sequence of randomly sampled graphs. The approach inherently accounts for multiple testing and correlations between subgraphs and does not rely on a priori specification of a null model. This provides a novel definition of motif significance which guarantees robust statistical inference. We apply our methodology to perform comparative connectomics, by evaluating the compressibility of diverse biological neural networks, including the connectomes of *C. elegans* and *Drosophila melanogaster* at different developmental stages, and by comparing topological properties of their motifs.

Keywords. *Network motifs, Network models, Statistical inference, Information theory, Network neuroscience, Statistical physics.*

1 Results

A network is classically described by dyadic models, which, by definition, only involve pairwise interactions [1]. The celebrated Erdős-Rényi model describes simple networks that share the same numbers of nodes, N, and edges, E. A second widespread model, the configuration model, characterises a network by its degree sequence. The randomness of such simplistic representations – measurable by their associated entropy – fail at capturing the mesoscopic nature of real networks [2], as their parameters are either fully global (the number of edges) or local (node degrees). A whole catalogue of topological measures constitutes attempts of assessing the non-trivial multiscale organization of real networks, e.g. the clustering coefficients, small-world properties, the betweenness centrality, motifs, etc. Statistical tools are necessary to judge if such properties are significant or randomly induced by some noise inherent to the data.

Hypothesis-based testing is currently the most popular approach for estimating the relevance of network statistics. However, a severe limitation of this approach is that it requires the definition of the appropriate null model *a priori* [3]. Statistical inference of a generative model, on the other hand, does not require choosing a null model [4]. If a network property is translatable into a model parameter, its statistical significance can be evaluated in a principled manner through model selection, thanks to the minimum description length (MDL) principle [5].

Here, focusing specifically on the task of network motif mining [6], we designed a statistical inference method that extracts *collectively* significant sets of motifs through a greedy algorithm that relies on subgraph contractions. Information theory favors the simplest representations, or, in other words, codes that are maximally compressive [7]. An encoding that coincidentally performs model selection is the Shannon-Fano code, which makes use of probability distributions. Networks, modeled as configurations sharing a set of constrained properties θ, can be encoded by uniform probability distributions P_θ [8]. The efficiency of such encoding, or the fitness quality of the related model, can be measured by the codelength of a network G. A so-called two part encoding defines the codelength as

$$L(G,\theta) = -\log P(G,\theta) = -\log P_\theta(G) - \log P(\theta). \tag{1}$$

where $-\log P(\theta)$ corresponds to the model complexity and prevents from overfitting. As the MDL suggests, the smaller the codelength, the more adequate is the representation. For incorporating candidate motifs into a network model, we take inspiration from and extend a previous approach [9] focused on fast independent network motif mining As opposed to the prevailing methodology for motif detection [10], which is purely based on graphlet frequencies, here motif occurrences are localized within the network as small modules, which we call *supernodes*, resulting from the contraction of subgraphs and leading to a reduced multigraph H, as depicted in Fig 1A. The parameter set of our model comprises a set of graphlets Γ, which gathers all candidate motifs found by a complete subgraph census, a set of supernodes $\mathcal{V}$, a multiset of graphlets $\mathcal{S}$, which identifies supernodes to their corresponding graphlets, and a dyadic model ϕ, that describes the structure of H.

To minimize the codelength $L(G, \{\Gamma, \mathcal{V}, \mathcal{S}, \phi\})$, we developed a greedy algorithm that selects, among induced subgraph samples, the subgraphs that maximize compression when contracted [1].

We applied our inference method on multiple biological neural networks, cf. Fig 1B, called *connectomes*, including all the currently largest completely known connectomes at the scale of the single neuron and synapse. We established the efficiency of our representation compared to four standard dyadic models of directed networks. The inferred motifs are consistent with neuroscience hypothesis and empirical observations: the most prevalent motifs of two symmetric brain regions of the Drosophila larva's connectome are topologically equivalent or at least very similar, cf. Fig 1C. We also have analyzed the structure of motifs through a series of network metrics that display the high density, recurrence and symmetry of the found motif circuits, cf. Fig 1D.

We recently uploaded an arxiv preprint of a paper, under review, detailing the overall methodology, which can be found at https://arxiv.org/abs/2311.16308.

[1] Other pattern mining algorithms rely on the MDL principle [16]. Note that their aim is quite far from our purpose. They usually consist of summarizing a graph with a limited prior set of pattern types, e.g. cliques and stars. We are here considering *any* candidate pattern, up to a certain size. Their inference is also sensitive to the choice of the graph connectivity encoding. For example, in [15], the latter is measured by the amount of connectivity overlap between the putative pattern dictionary and the input graph. In our approach, we consider only existing patterns, and we show how the graph encoding can drastically affect subsequent statistical interpretation. We consider multiple graph models, while insisting on their non-exhaustiveness and on the possibility of testing other graph models for encoding their connectivity, while their details will not affect the validity of the presented methdodology.

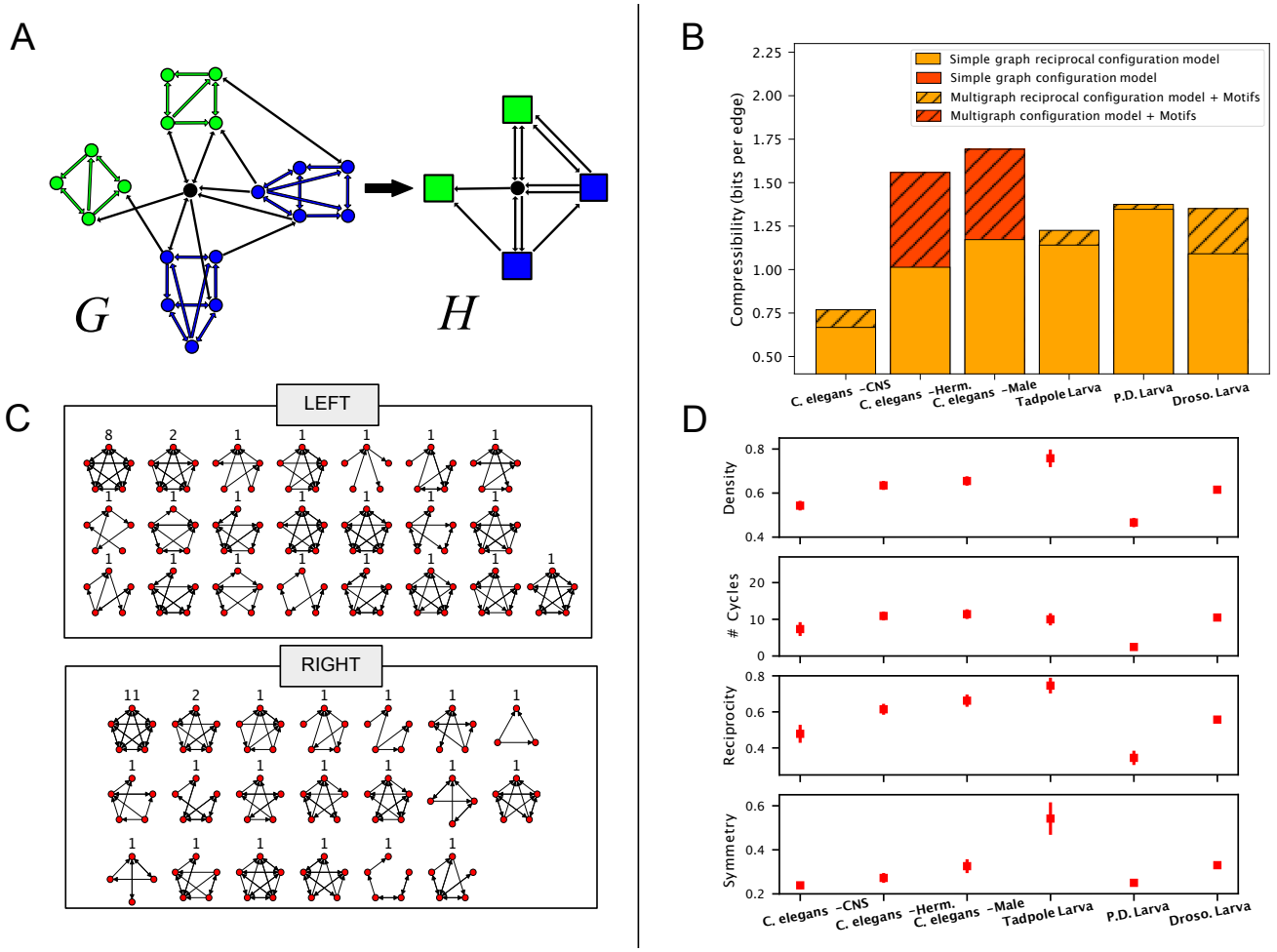


Figure 1: (A) Schematic of the graph reduction output by our greedy algorithm. The input graph G is simply directed, with labeled nodes of the same color. After a series of subgraph contractions, G is reduced into a multigraph with fewer nodes and edges H, but with multiple node colors. In the schematic, four subgraphs isomorphic to two distinct graphlets are selected. H contains three node colors, including one for the regular node. (B) We applied our motif-set-inference method on all known complete brain networks. The compressibility, $[L(G,\{N,E\}) - L(G,\{\Gamma,\mathcal{V},\mathcal{S},\phi\})]/E$, measures how well the motif-based model is more suitable than the Erdős-Rényi simple graph. The higher the compressibility is, the more significant is the found representation. In all cases at hand, a model that accounts for a motif set and degree sequences outperforms models deprived of higher-order features. (C) Motif sets inferred from two *a priori* symmetric subnetworks of the *Drosophila* larva brain, its *left* and *right* mushroom bodies. Numbers above the icons indicate the associated numbers of isomorphic contracted subgraphs. (D) Graphlet measures averaged over the contracted subgraphs constituting the motifs sets. The density, number of cycles and reciprocity are standard directed network metrics [13]. The "symmetry" measure is a continuous $(0,1)$-bounded graph polynomial root quantifying the topological symmetry of the graphlet connectivity [14]. A value of 0 corresponds to a least symmetric graphlet, while 1 corresponds to a clique, or a maximally symmetric graphlet.

References

[1] Bianconi, G. (2009). Entropy of network ensembles. Physical Review E, 79(3), 036114.

[2] Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., ... & Petri, G. (2020). Networks beyond pairwise interactions: Structure and dynamics. Physics Reports, 874, 1-92.

[3] McElreath, R. (2018). Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC.

[4] Peixoto, T. P. (2023). Descriptive vs. inferential community detection in networks: pitfalls, myths and half-truths. Cambridge University Press.

[5] Grünwald, P. D. (2007). The minimum description length principle. MIT press.

[6] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science, 298(5594), 824-827.

[7] Cover, T. M. (1999). Elements of information theory. John Wiley & Sons.

[8] Gauvin, L., Génois, M., Karsai, M., Kivela, M., Takaguchi, T., Valdano, E., & Vestergaard, C. L. (2022). Randomized reference models for temporal networks. SIAM Review, 64(4), 763-830.

[9] Bloem, P., & de Rooij, S. (2020). Large-scale network motif analysis using compression. Data Mining and Knowledge Discovery, 34, 1421-1453.

[10] Fodor, J., Brand, M., Stones, R. J.,& Buckle, A. M. (2020). Intrinsic limitations in mainstream methods of identifying network motifs in biology. BMC bioinformatics, 21(1), 1-11.

[11] Winding, M., Pedigo, B. D., Barnes, C. L., Patsolic, H. G., Park, Y., Kazimiers, T., ... & Zlatic, M. (2023). The connectome of an insect brain. Science, 379(6636), eadd9330.

[12] Eichler, K., Li, F., Litwin-Kumar, A., Park, Y., Andrade, I., Schneider-Mizell, C. M., ... & Cardona, A. (2017). The complete connectome of a learning and memory centre in an insect brain. Nature, 548(7666), 175-182.

[13] Hagberg, A., & Conway, D. (2020). Networkx: Network analysis with python. URL: https://networkx. github.io.

[14] Dehmer, M., Chen, Z., Emmert-Streib, F., Mowshowitz, A., Varmuza, K., Feng, L., ... & Tao, J. (2020). The orbit-polynomial: a novel measure of symmetry in networks. IEEE access, 8, 36100-36112.

[15] Koutra, D., Kang, U., Vreeken, J., & Faloutsos, C. (2014, April). Vog: Summarizing and understanding large graphs. In Proceedings of the 2014 SIAM international conference on data mining (pp. 91-99). Society for Industrial and Applied Mathematics.

[16] Liu, Y., Safavi, T., Dighe, A., & Koutra, D. (2018). Graph summarization methods and applications: A survey. ACM computing surveys (CSUR), 51(3).

FRCCS

Discovering temporal triadic closure patterns

Alessia Galdeman[1✓], Cheick Ba[2], Matteo Zignani[1] and Sabrina Gaito[1]

[1] *University of Milan - Via Celoria 18, 20133 Milan (Italy); alessia.galdeman@unimi.it, matteo.zignani@unimi.it, sabrina.gaito@unimi.it*
[2] *Queen Mary University of London - 285 Bancroft Rd, Bethnal Green, London (UK); c.ba@qmul.ac.uk*
[✓] *Presenting author*

Abstract. The dynamics within online social networks (OSNs) are influenced by numerous factors, encompassing user behavior, content generation, platform features, and technological advancements, with triadic closure standing out as a prominent and influential element. In this study, we focus on the temporal aspects of triadic closure and its role in the evolution of OSNs. We developed a comprehensive analytical pipeline to support the study of triadic closure patterns. This pipeline includes an efficient algorithm for the census of temporal triads, a vector-based model for representing temporal networks (temporal triadic profile), the identification of triadic closure rules (TERs), and the evaluation of the speed of the formation of closed triads. Our findings reveal significant variations in the impact of triadic closure across different OSNs, marked by diverse temporal triadic profiles and varying speeds of closed triad formation.

Keywords. *Triadic closure; Network evolution; Temporal networks; Web3; Blockchain*

Introduction. In recent years, there has been increasing interest in studying temporal patterns [6, 7], especially within online social networks (OSNs). These patterns are essential for predicting platform evolution and enhancing user experiences. Among these, *triadic closure*, deeply studied in social network theory, stands out as one of the most influential factors shaping network dynamics. Triadic closure reflects the tendency of individuals to form connections with friends of their friends, leading to the development of tightly interconnected clusters based on closed triads. This phenomenon, known as a universal process, plays a significant role in shaping the evolution of online social networks across various platforms and user communities [3]. However, much research overlooks the temporal dimension of interactions. This study aims to fill this gap by investigating the temporal aspects of triadic closure and closed triad formation in OSNs. We employ an efficient method for extracting and counting temporal triads. Then, we define an approach to create a network profile based on the characteristics of the triadic closure process, that unveil evolutionary footprints in social networks. Our focus extends to analyzing evolution patterns within the triadic closure process, encompassing topology/structure and closure speed. This is realized with triadic closure rules (TERs): inspired by graph evolution rules [2, 8] they provide human readable outcomes on the triadic closure process. Results show several interesting evolutionary insights, like the higher probability of a triad to close through the strongest (reciprocal) edge.

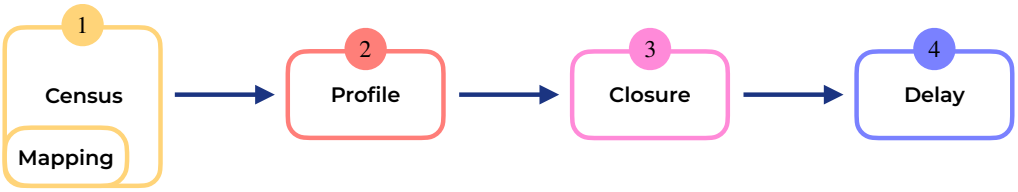


Figure 1: Overview of the methodology

Methodology. We first model social interaction's list $I = \{(x, y, t)\}$ as a directed graph $\mathcal{G} = (V, E)$ where E is the set of timestamped edges (x, y, t) that encodes the first connection from user x to user y (both belonging to the set of users V) at timestamp t [5]. The methodology pipeline of this work is shown in Figure 1. It starts with creating a *general mapping* structure to classify all potential 3-node temporal triads (assigning a unique ID), considering various timestamp orderings. This mapping is crucial for analyzing triadic closure across all temporal subgraphs. We initially generate all static triad structures and then incorporate temporal aspects by considering different edge orderings. Each potential graph is assigned a unique ID, with measures taken to handle isomorphic graphs. Since our focus is on triadic closure, we start with graphs of the minimum edge count $|E| = 2$, enabling us to generate all open and closed temporal triads, totaling 324 in number. Through the general mapping structure, the methodology proceeds with the proper census algorithm (that extends the original version of [1]). It maintains two counters: `Census` to track the frequency of encountered temporal triads and `Evolutions` to record temporal patterns leading to closed triads. The efficient enumeration method ensures that each triad is considered only once, and consequently classified with the general mapping structure to update the `Census` and `Evolutions` counter accordingly. These counters' information is then used to address the following phases of the methodology. Specifically, we first build the *profile* (step 2) of temporal closed triads as a vector representing the probability distribution of a specific triad class to occur. Then we obtain the triadic closure rules –TERs, computing the probability for an open triad to close in one of the two possible ways. Finally, on each TER, we compute the delay, i.e. the speed of triads to close. Technically, it is computed as the difference between the timestamp of the closing edge and the very previous one. In other words, the delay of a TER corresponds to the maximum timestamp of the closed triad minus the maximum timestamp of the corresponding open triad.

Datasets. This paper utilizes different datasets to explore social structures. These include UC-Social, capturing student message exchanges (19,573 nodes, 20,296 edges), Enron emails (87,273 nodes, 321,918 edges), Cryptokitties NFT exchanges (99,984 nodes, 481,540 edges), Sarafu complementary currency exchanges (40,343 nodes, 143,239 edges), and Steemit follow operations (23,493 nodes, 290,801 edges). Together, they offer insights into communication networks, market trends, alternative currencies, and online community engagement.

Results. The pipeline was applied to the 5 graphs previously described, and Table 1 shows the main quantitative results as far as concern the *census* step. The number of triads and isomorphic classes are shown (second and third column) for each dataset along with the computation time of both the census and delay steps. These are the results of execution on a server Dell Power Edge T620 equipped with 2 CPUs Intel Xeon 2.10GHz 16 Core (32 threads) and 376GB RAM, using 48 parallel jobs. Across datasets, closed triads vary in prevalence, with the cooperation-based network (Sarafu) exhibiting a higher tendency for closure. Moreover, the common basis made of 56 rules serves as initial proof that a common general process is acting on all the networks, i.e., the triadic closure, even if with varying impact in different temporal contexts.

Dataset	Closed triads (%)	Classes	Census time	Delay time
UC-social	14,319 (1.97%)	56	4min	3min
Enron	1,180,387 (2.5%)	56	3h9	17min
Cryptokitties	2,051,706 (1.02%)	56	12h	30min
Sarafu	273,001 (4.25%)	56	1h	2min
Steemit	2,902,199 (0.84%)	56	21h30	1h50

Table 1: Overview of the results of the algorithms introduced, with computation times. In parenthesis the percentage of closed triads over the overall volume of triads (closed or open).

By focusing solely on closed triads (extracted in the initial step of the pipeline of Figure 1), we gain insight into the dynamics of triadic closure within networks. For instance through our analysis using profiles (step 2 of Figure 1, results not shown) we find a reduced occurrence of closure patterns involving reciprocal edges, suggesting that triads often close without necessitating the creation of reciprocal edges. However, exploring the TER framework (steps 3 and 4) offers a fresh viewpoint on the evolutionary aspects of the triadic closure process, which holds particular importance in social networks. Figure 2 reports some examples worth noting. The rules in Figure 2(a) both start from the open triad with $ID = 0$, followed by the evolution of a triad through the establishment of a connection in one of two directions ($ID = 28, 30$). This connection could represent various interactions such as sending an email or following a user. Analysis of rule probabilities (step 3 of Figure 1) reveals that the identity of the triad closer is irrelevant, likely due to the symmetric nature of the pattern. Furthermore, examination of TER delays (step 4 of Figure 1) shows a balanced distribution, suggesting that the speed of triad closure is unaffected by the initial destination selected by the initiator (source node of the first edge in the triad). Another key insight concerns reciprocal/strong connections: frequently, closure occurs through the strongest link, namely, the reciprocated one. For example, in the rules presented in Figure 2(b), the likelihood of the rule $6 \rightarrow 42$ is consistently higher than the reverse case of $6 \rightarrow 54$.

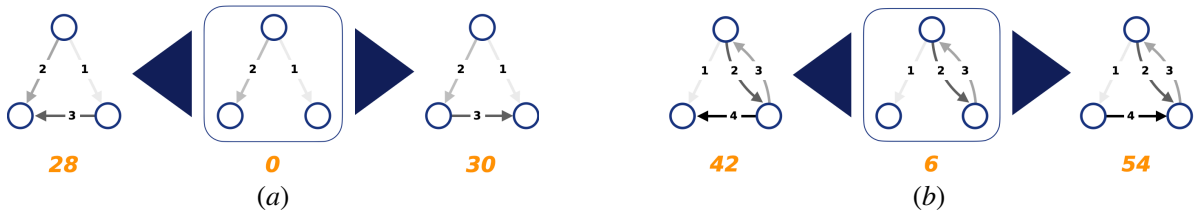


Figure 2: Examples of TER found on every dataset. Both (a) and (b) show a pair of rules sharing the same precondition.

This work is a short version of the paper [4] recently submitted to journal.

References

[1] Vladimir Batagelj and Andrej Mrvar. A subquadratic triad census algorithm for large sparse networks with small maximum degree. *Social networks*, 23(3):237–243, 2001.

[2] Michele Berlingerio, Francesco Bonchi, Björn Bringmann, and Aristides Gionis. Mining graph evolution rules. In *joint European conference on machine learning and knowledge discovery in databases*, pages 115–130. Springer, 2009.

[3] Ginestra Bianconi, Richard K Darst, Jacopo Iacovacci, and Santo Fortunato. Triadic closure as a basic generating mechanism of communities in complex networks. *Physical Review E*, 90(4):042806, 2014.

[4] Alessia Galdeman, Cheick Tidiane Ba, Matteo Zignani, and Sabrina Gaito. Triadic closure evolution rules, 2024. Submitted to journal.

[5] Petter Holme and Jari Saramäki. *Temporal network theory*, volume 2. Springer, New York City, NY, 2019.

[6] Lauri Kovanen, Márton Karsai, Kimmo Kaski, János Kertész, and Jari Saramäki. Temporal motifs in time-dependent networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(11):P11005, Nov 2011.

[7] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive data sets*. Cambridge university press, 2020.

[8] Erik Scharwächter, Emmanuel Müller, Jonathan Donges, Marwan Hassani, and Thomas Seidl. Detecting change processes in dynamic networks by frequent graph evolution rule mining. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1191–1196. IEEE, 2016.

FRCCS

Sampling based sequential dependencies discovery in Higher-Order Network Models

Julie Queiros[1✓], François Queyroi[1], Samuel Maistre[2]

[1] *Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004*
[2] *Université de Strasbourg, CNRS, IRMA, UMR 7501*
[✓] *Presenting author*

Abstract. Higher-order networks are a general form of network models that include memory nodes used to capture indirect dependencies between entities. When built from sequential or pathway data, random walks performed on these networks usually represent input sequences better than traditional first-order models. Unlike the latter, there are various ways to build higher-order networks that already exist in the literature. We introduce a new variable-order network where nodes can encode sequences of varying length. Nodes are selected by looking at sub-sequences that are unlikely extensions of already detected sub-sequences. Using experiments on real-world datasets, we demonstrate that our method produces smaller and as accurate models compared to the main variable-order model in the literature. We also study the differences achieved when ranking items using a higher-order reformulation of the PageRank metric.

Keywords. *Higher-order networks; Sequential data; Monte-Carlo; PageRank*

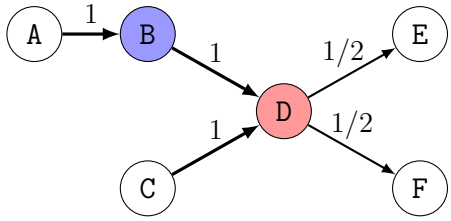
1 Introduction

Networks can be used to represent dependencies found in sequential data. One direct approach is to aggregate pairwise interactions between items in the input sequences. Examples include the number of clicks between two web pages or the number of times a ship navigates from one port to another. Most of the time, network mining algorithms use the indirect dependencies induced by the network topology, *e.g.* the PageRank metric is linked to the behaviour of a random walker on the graph.

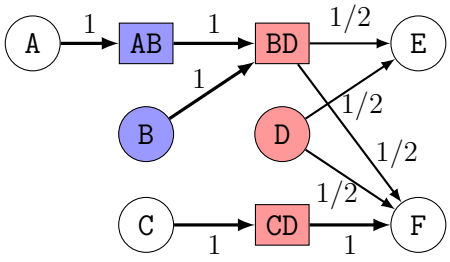
The use of this traditional representation raises an issue in the case of networks built from sequential data as the indirect relations induced by the network topology may not correspond to observed sub-sequences. For example, the graph in Fig. 1b suggests an indirect relation between item `C` and `E` going through `D`. But no such transitions exist in the input dataset the graph was built with (see Fig. 1a). Indeed, using this network representation presupposes that the modeled system respects the Markov property *i.e.* the information useful for predicting the future state is contained only in the current state (the process is also said to be "memory-less").

Sequences
D E
A B D E
B D F
C D F

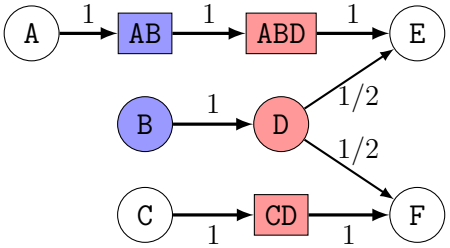
(a) Sequential Data



(b) FON$_1$



(c) FON$_2$



(d) VON

Figure 1: Examples of network representation of sequential data (1a). We assume here that each sequence is observed several times. There is a 50% chance to leave D to go to either E or F. First-order networks can encode these transitions probabilities (arc labels in 1b). However, when looking at the sequences, we can see that coming from C before D then its next destination is always F. Higher-order networks (1c and 1d) can encode these indirect dependencies by using multiple *representations* of each item. FON$_2$ network (1c) include all order-2 dependencies even those that do not help in predicting next item (*i.e.* BD). VON Network (1d) keeps only relevant dependencies (*i.e.* excluding BD but adding ABD) leading to a sparser and better representation of the sequential dependencies. The three red nodes are representations of the item D. Node labeled ABD is then a memory node of order 3.

In order to address this issue, *higher-order networks* can be used instead. In these networks, the memory on the previous steps is encoded into *memory nodes*. An item can then be represented not by one (as in first-order models) but by several nodes. On this new class of networks, a random walk will better simulate the input sequences. In the example, coming for C, the only possible destination for a random walker after D is F. We focus specifically on variable-order networks (see Fig. 1d) where there are no constraint on the length of memory nodes.

The main question is: how do we select the memory nodes to add to the network? In the Fig. 1d, the sequence ABD is included as a memory node while BD is not. Indeed, only knowing that B is observed before D does not add more information. We say that the sequence ABD is a *relevant context* as it adds information about the following items. The issue of *relevant context* definition is the central topic of this paper. The use of variable-order network was introduced by Xu *et al.* [23, 18] who provided a first answer to this question. However, as there are several ways to define "relevance", several variable-order network representations of the same sequential dataset are possible. In this paper, we introduce a new approach called MC-VON to construct variable-order networks where the relevance of a context is assessed using a statistical test. We show on real-world data that this new model can be as efficient as predicting sequences while being sparser that the model of Xu *et al.*. Furthermore, we evaluate the effect of model selection on higher-order PageRanks rankings.

2 Related works

The concept "higher-order networks" (HON) is used to describe graph models that are not limited to *dyadic* relations and design to capture different *system dependencies* (*e.g.* temporal, sequential, subset, spatial *dependencies*) [7, 21]. Like several authors [21], we use here the term "higher-order" to refer exclusively to networks that encode sequential dependencies (transition probabilities for *contexts* of length (or order) longer than 1). Applications of HON include the identification of overlapping modules or the evaluation of items centrality [23, 17, 13].

Most of the existing literature focus on fixed-order networks we call FON$_k$ [17]. These models generalize classic *memory-less* models as they are networks where the probability of a random walker to reach *item* σ depends on the previous k steps rather than only the last one. The parameter k can be set to a given value [17] but the estimation of the optimal order for a given dataset was also investigated [20, 19].

We focus in this paper on a less discussed family of models called variable-order networks (VON). Variable-order Markov models of discrete sequences have been studied in the past [3]. But, to the best of our knowledge, the only applications to network analysis is the seminal work of Xu *et al.* [23, 18]. Their main idea is that orders should be found locally rather than determining a global order for the system *i.e.* memory nodes should only be included if they indeed impact the behaviour of a random walker.

The number of memory nodes in a FON$_k$ grows exponentially with k. Therefore, VON models are useful when higher-order dependencies existing in the system are sparse. This is a safe hypothesis since real-world networks are considered sparse. Still, Xu *et al.* rightly suggest that one should find a balance between the resulting network size and the corresponding model's goodness-of-fit *w.r.t.* input sequences. In the present paper, we compare their network model (denoted D_{KL}-VON) to ours (denoted MC-VON). These models are defined in the following section.

3 Methods

Let $\mathcal{A}$ be the set of items. An input dataset corresponds to a set $\mathcal{S} = (s^1, s^2, \ldots)$ of sequences $s^i = \sigma_1^i \sigma_2^i \sigma_3^i \ldots$ where all $\sigma_j^i \in \mathcal{A}$. For a sequence s of symbols in $\mathcal{A}$, the *order* of s denoted $|s|$ is the length of s and the *support* of s denoted $c(s)$ is the number of occurrences of s in dataset $\mathcal{S}$. We will also use $C_s = (c(s\sigma), \sigma \in \mathcal{A})$ to denote the occurrences of items following the sequence s. Let $s = s_1 s_2$ be a sequence resulting in the concatenation of sequences s_1 and s_2. Here, s_1 is a *prefix* of s, s_2 is a *suffix* of s while we call s an *extension* of s_2. Based on a vector $K = \{K_\sigma, \sigma \in \mathcal{A}\} \in \mathbb{N}^{|\mathcal{A}|}$, where K_σ represents the number of occurences of element σ, we define the probability measure $Q_K(\sigma) = K_\sigma / \sum_{\sigma' \in \mathcal{A}} K_{\sigma'}$, *i.e.* the probability to choose σ when drawing a random element from K.

In discrete sequences prediction, we want to estimate $P(\sigma|\sigma_1 \ldots \sigma_k)$ *i.e.* the probability to encounter item σ after the sequence $\sigma_1 \ldots \sigma_k$. In a *higher-order* Markov model, we assume we have a set $\mathcal{R}$ (the relevant contexts) of sequences of items including $\mathcal{A}$. The probability $P(\sigma|\sigma_1 \ldots \sigma_k)$ will be estimated using $P^{\mathcal{R}}(\sigma \mid \sigma_1 \ldots \sigma_k) \coloneqq Q_{C_{s'}}(\sigma)$ where s' is the longest suffix

of $\sigma_1 \dots \sigma_k \in \mathcal{R}$. In a *memory-less* Markov model (or first-order model), we have $\mathcal{R} = \mathcal{A}$ and only the last visited item is taken into account *i.e.* $P^{\mathcal{A}}(\sigma \mid \sigma_1 \dots \sigma_k) = P^{\mathcal{A}}(\sigma \mid \sigma_k)$.

The HON models studied here are built from the same general procedure described in the Section 3.1. We defined known HON models using this procedure. The difference in the method we are proposing comes down to the identification of the set of relevant contexts $\mathcal{R}$ which is detailed in 3.2.

3.1 Generic Higher-Order Networks construction procedure

Higher-order networks aim at encoding transition probabilities of higher-order Markov models into a regular weighted directed graph (as the example in Fig. 1b). From a set of relevant contexts $\mathcal{R}$, the weighted directed graph $G(\mathcal{R}) = (\mathcal{R}, E, w)$ is built with each item $\sigma \in \mathcal{A}$ represented by multiple nodes corresponding to the contexts having σ as the last entry. We say that these *memory nodes* are the *representations* of item σ. The edge set E and the weights w are defined as follows. Let $s \in \mathcal{R}$ and $\sigma \in \mathcal{A}$ such that $P^{\mathcal{R}}(\sigma|s) > 0$, $G(\mathcal{R})$ contains a link $s \to s^*\sigma$ of weight $w(s \to s^*\sigma) = P^{\mathcal{R}}(\sigma|s)$ where s^* is the longest suffix of $s \in \mathcal{R}$. For example, let $s = abc$ and $s^*\sigma = bc\sigma$ be relevant extensions of c and σ respectively then there will be a link $s \to s^*\sigma$ if $abc\sigma \notin \mathcal{R}$ and $P^{\mathcal{R}}(\sigma|s) > 0$.

Algorithm 1: VON Generic Algorithm

Data: $\mathcal{S}$: set of sequences over itemset $\mathcal{A}$
Input: s_c, s_v: current and last relevant contexts
Result: R: set of relevant contexts
if *existRelevantExts*(s_c, s_v) **then**
  for $\sigma \in \mathcal{A}$ **do**
    if *isRelevant*$(\sigma s_c, s_v)$ **then**
      $R \leftarrow R \cup \text{VON}(\sigma s_c, \sigma s_c)$
    else
      $R \leftarrow R \cup \text{VON}(\sigma s_c, s_v)$
return $R \cup$ *prefixes*(s_v)

As previously said, the difference between HON models mainly comes from the way the set of contexts $\mathcal{R}$ is defined. Algorithm 1 is a general framework used in [18] to extract such a set. Relevant contexts are recursively found as *extensions* of contexts found at lower orders. For a dataset $\mathcal{S}$ and an itemset $\mathcal{A}$, the final set of contexts is defined as $\mathcal{R} \coloneqq \bigcup_{\sigma \in \mathcal{A}} \text{VON}(\sigma, \sigma)$.

The functions `isRelevant` and `existRelevantExts` depend on the model used. The test `isRelevant`(s_c, s_v) is passed when s_c is judged relevant when compared to the last relevant suffix identified s_v. The function `existRelevantExts` is used to identify situations where no relevant extensions of s_v are possible and, therefore, where the recursion must be stopped. As such, this function should not need to count sub-sequences $\sigma_1 s_c \sigma_2$. This operation is to be done only if a relevant extension is possible. If well designed, it should make possible the use of Algorithm 1 on large datasets [18].

Finally, function `prefixes` returns the set of prefixes of s_v (including itself). Indeed, a random

walker on a higher-order network can only reach memory node $s_1 s_2 \dots s_k$ if there is a path $s_1 \to s_1 s_2 \to \dots \to s_1 s_2 \dots s_k$. Therefore, every prefix of s_v are added in the network even if some are not relevant. In the example of Fig. 1d, knowing that we observed `A` before `B` does not provide more information than simply knowing it came from `B`. However, the representation `AB` is here as a prefix of `ABD` which is a relevant context.

Definition 1 *The fixed-order network* FON_k *is obtained by treating a sub-sequence as relevant if its order is lower or equal to* k*.*

With this definition, the network FON_1 is the traditional first-order network (Fig. 1b). Fixed-order networks are usually defined as subgraphs of the k order De Bruijn graph over $\mathcal{A}$ in the literature. But this formalism does not allow to keep track of transition probabilities for contexts of order lower than k. This model is also called *multi-order* model [19].

Definition 2 *The variable-order network* D_{KL}*-*VON(λ) *[18] is obtained by treating the sub-sequence* σs_c *as relevant w.r.t* s_v *iff*

$$D_{KL}\left(P_{\sigma s_c} || P_{s_v}\right) > \frac{\lambda |\sigma s_c|}{\log_2(1 + c(\sigma s_c))} \tag{1}$$

with $\lambda \in \mathbb{R}^+$ *the threshold multiplier,* $P_s \coloneqq Q_{C_s}$ *the distribution of items following* s *and* D_{KL} *the Kullback-Leibler divergence (in bits).*

One main advantage of D_{KL}-VON when compared to FON_k is that the length of contexts is locally defined in order to best fit the data. The right side of Eq. 1 makes longer and sparsely observed contexts harder to be recognized as relevant.
The parameter λ is not actually included in the original definition of [18] (we have an equivalent definition for $\lambda = 1$). The original definition of the authors is indeed parameter-free. However, the interpretation of the right-side threshold function in relation to the D_{KL} divergence is hard to grasp. We argue that the definition of the threshold function actually hides an arbitrary choice of "scale" made by the authors. Therefore, the "parameter-freeness" of D_{KL}-VON is limited in our opinion. We shall use different value of λ in order to compare D_{KL}-VON model to the one defined below.

3.2 MC-Von model definition

Our proposal to construct a variable-order network model is to use to the quantity $D_{KL}\left(P_{\sigma s_c} || P_{s_v}\right)$ as in [18], but as a test statistic in a hypothesis testing paradigm to avoid relying on an *ad hoc* threshold function. Indeed, if σs_c is not a relevant context, then $C_{\sigma s_c}$ should behave like a draw of $c(\sigma s_c)$ elements from C_{s_v} without replacement, i.e. from a multivariate hypergeometric distribution $\mathcal{MH}(C_{s_v}, c(\sigma s_c))$. Therefore, we will decide that σs_c is a relevant context when $C_{\sigma s_c}$ does not behave like a random draw, that is while we can reject the null hypothesis

$$H_0 : C_{\sigma s_c} \sim \mathcal{MH}(C_{s_v}, c(\sigma s_c)) \quad \text{vs.} \quad H_1 : C_{\sigma s_c} \nsim \mathcal{MH}(C_{s_v}, c(\sigma s_c)).$$

The nominal level $\alpha \in (0, 1)$ of the test allows us to choose how surprising we want the draw $C_{\sigma s_c}$ to be in order to consider s_c as a relevant context. It is also an upper bound for the probability of a context being considered relevant when it is not.

Definition 3 *The variable-order network MC-VON is obtained by treating the sub-sequence σs_c as relevant w.r.t s_v iff*

$$D_{KL}\left(P_{\sigma s_c}||P_{s_v}\right) > q_{1-\alpha}\left(c(\sigma s_c), s_v\right) \quad \textit{or equivalently} \quad p < \alpha \tag{2}$$

where $P_s \coloneqq Q_{C_s}$ the distribution of items following s, $q_{1-\alpha}\left(c(\sigma s_c), s_v\right)$ is the $(1-\alpha)$-th quantile of the distribution of $D_{KL}\left(Q_D||P_{s_v}\right)$ where D is a random draw from $\mathcal{MH}(C_{s_v}, c(\sigma s_c))$ and

$$p = \mathbb{P}\left(D_{KL}\left(Q_D||P_{s_v}\right) \geq D_{KL}\left(P_{\sigma s_c}||P_{s_v}\right)\right)$$

is the p-value of the test.

Example 1 *Assume we have an order 1 subsequence s_v with $C_{s_v} = (1, 2, 5, 0, \ldots)$ and we want to assess the relevancy of the extension σs_c with $C_{\sigma s_c} = (1, 0, 0, 0, \ldots)$. We have $D_{KL}\left(P_{\sigma s_c}||P_{s_v}\right) = -\log_2(1/8) = 3$. Since it is the hightest possible D_{KL}, the probability that a draw from $\mathcal{MH}(C_{s_v}, 1)$ has a divergence higher or equal is the probability to draw $C_{\sigma s_c}$ i.e. $p = \frac{1}{8}$. Taking a standard threshold of $\alpha = 10^{-3}$, we would accept H_0 and declare that this extension is not relevant. On the other hand, for D_{KL}-VON(1), the threshold function of Eq. 1 is equal to $\frac{2}{\log_2(2)} = 2$. The extension σs_c would here be considered as relevant.*

However, p (or $q_{1-\alpha}\left(c(\sigma s_c), s_v\right)$) can be difficult to compute, particularly if $c(\sigma s_c)$ is neither small nor close to $c(s_v)$. Therefore, we propose different approximations to estimate it and decide whether (2) holds or not. The first possibility is to use a Monte-Carlo algorithm that draws M independent replications $\{D_i, 1 \leq i \leq M\}$ from $\mathcal{MH}(C_{s_v}, c(\sigma s_c))$ and estimate p by $\widehat{p} = S_M/M$ where

$$S_M = \sum_{i=1}^{M} \mathbb{I}\left\{D_{KL}\left(Q_{D_i}||P_{s_v}\right) \geq D_{KL}\left(P_{\sigma s_c}||P_{s_v}\right)\right\}$$

follows a binomial distribution of size M and probability p, i.e.

$$\mathbb{P}(S_M = k) = \binom{M}{k} p^k (1-p)^{M-k} =: b(n, p, k).$$

The choice of M will affect the precision of our decision, particularly if p is close to α. On the contrary, if the conclusion is more obvious, i.e. $p \ll \alpha$ or $p \gg \alpha$, we might have chosen a smaller value for M to get a reasonable precision. Methods that adapt the number of replications to the distance between p and α were proposed, e.g. in [8] and [6] among others. These two papers both control the resampling risk defined by

$$RR_p(\widehat{p}) = \begin{cases} \mathbb{P}_p\left(\widehat{p} > \alpha\right) & \text{if } p \leq \alpha \\ \mathbb{P}_p\left(\widehat{p} \leq \alpha\right) & \text{if } p > \alpha. \end{cases}$$

This resampling risk measures the probability to take the wrong decision regarding (2). For a given $\epsilon > 0$, [8] and [6] propose procedures that ensures that $RR_p \leq \epsilon$. Nevertheless, there is no bound on the number of replications needed and the procedure might not end if $p = \alpha$, i.e. $D_{KL}\left(P_{\sigma s_c}||P_{s_v}\right) = q_{1-\alpha}\left(c(\sigma s_c), s_v\right)$. A maximum number of iterations must be chosen and the resampling risk is therefore not truly controlled. For this reason, we divide α into a lower value α^- and a higher value α^+ so that the number of iterations is always finite. The cost of this

finite number of iterations is made by accepting slightly less relevant sequences ($\alpha^- < p \leq \alpha^+$) rather than missing sequences that are relevant ($p \leq \alpha^-$) and define

$$\widetilde{RR}_p(\widehat{p}) = \begin{cases} \mathbb{P}_p\left(\widehat{p} > \alpha^*\right) & \text{if } p \leq \alpha^- \\ 0 & \text{if } p \in]\alpha^-, \alpha^+] \\ \mathbb{P}_p\left(\widehat{p} \leq \alpha^*\right) & \text{if } p > \alpha^+ \end{cases}$$

where $\alpha^* \in]\alpha^-, \alpha^+[$ is a critical value for $\widehat{p}$ such that we will reject H_0 iff $\widehat{p} < \alpha^*$. We construct a procedure that ends after a finite number of steps and such that

$$\sup_{p \in [0,1]} \widetilde{RR}_p(\widehat{p}) \leq \epsilon. \tag{3}$$

We define the algorithm 2 that ensures (3), where

$$\alpha^* = 1 - \frac{\log\left(\alpha^+/\alpha^-\right)}{\log\left(\dfrac{\alpha^+/(1-\alpha^+)}{\alpha^-/(1-\alpha^-)}\right)},$$

is such that $b(n, \alpha^-, n\alpha^*) = b(n, \alpha^+, n\alpha^*)$. α^* is also the value of p that will require the highest number of draws on average. The termination of the algorithm comes from the fact that the function $x \mapsto (n+1)b(n, x, n\hat{p})$ is the beta density with parameters $n\hat{p}+1$ and $n(1-\hat{p})$ and tends to a Dirac measure in p as $n \to \infty$. Therefore at least one of the two values $b(n, \alpha^-, S_n)$ and $b(n, \alpha^+, S_n)$ must tend to zero.

Algorithm 2: MC-Von Decision Algorithm

Data: $D_{KL,obs}$: observed KL divergence
Input: α^-, α^+, ϵ: test levels and bound for resampling risk
Result: $\widehat{p}$: estimated p-value
$S = 0$; $n = 0$
while $b(n, \alpha^-, S) > \epsilon/(n+1)$ *and* $b(n, \alpha^+, S) > \epsilon/(n+1)$ **do**
  $D \sim \mathcal{MH}(C_{s_v}, c(\sigma s_c))$
  if $D_{KL}\left(Q_D || P_{s_v}\right) \geq D_{KL,obs}$ **then**
    $S = S + 1$
  $n = n + 1$
return $\widehat{p} = S/n$

For the function `existRelevantExts`(s_c, s_v), we use a simple lower-bound on the p-value. Indeed, there are at most $z = \binom{c(s_v)}{\min\left(\frac{c(s_v)}{2}, c(s_c)\right)}$ draws from $\mathcal{MH}(C_{s_v}, c(\sigma s_c))$ for any $\sigma \in \mathcal{A}$. Therefore if $z^{-1} > \alpha^+$ there is no possible extensions of s_c that can be found relevant.

4 Experiments

We use four different datasets (see Table 1) for the experiments. They offer a variety not only in terms of origin of the data but also in terms of size of the itemsets and sequences. Two of them (AIR and PORTS) correspond to spatial pathway data. AIR contains the itineraries of US passengers for domestic flights extracted from the *RITA TransStat* database. PORTS contains the sequences of ports where shipping vessels stop over between April and October 2009.

Table 1: Summary of datasets used

Dataset	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{S}\rvert$	**min/max** $\lvert s\rvert$	**Refs.**
Shipping path. (PORTS)	909	4243	2/183	[23, 5]
US Airflight (AIR)	175	286,810	2/14	[19]
Wikipedia clicks (WIKI)	100	29,573	2/22	[20, 19]
MSNBC clicks (MSNBC)	17	388,434	2/1810	[20, 19]

This is an extract from the Lloyd's Maritime Intelligence Unit database. WIKI and MSNBC are clickstream datasets. WIKI is the result of Wikipedia navigation games. Following [19], we only retain the sequences going through the top 100 visited pages. MSNBC gathers click streams of visitors of the website of the channel. The pages are grouped into 17 categories (*e.g.* "business", "local", "sports",...). We also removed all consecutive repetitions from the input sequences.

To construct the networks MC-VON we use a standard value for the confidence threshold $\alpha^- = 10^{-3}$ with $\alpha^+ = \alpha^- + 2.10^{-3}$ to control a risk $\epsilon = 0.05$. This means we make the correct decision for p-values outside $)\alpha^-, \alpha^+($ with a probability at least $1-\epsilon$. We report results for the D_{KL}-VON [18] model as it is the the main other model of variable-order networks existing in the literature. In order to compare the contexts retained by each approach in terms of accuracy, we will also determine λ^* such that D_{KL}-VON(λ^*) contains a number of nodes equivalent to MC-VON. Similarly, we determine the α_*^- such that MC-VON(α_*^-) contains a number of nodes equivalent to D_{KL}-VON(1), the other parameters being equal. We also include the results obtained with the FON$_1$ network and the FON with the optimal order according to [19]. The `honyx` python package[1], developed by the authors, was used to generate the higher-order networks. The datasets and the scripts used for the experiments can be found at [15].

4.1 Networks size and models accuracy

We investigate here the difference between the constructed HON and whether a better or similar accuracy with a smaller model can be achieved using MC-VON. Table 2 reports the results for each constructed network on the four datasets using the whole set $\mathcal{S}$. Networks size is represented by the number of nodes $|V|$ in the networks. The order correspond to the maximum order among the vertices. The last columns reports each model accuracy score *Acc* (Eq. 4) when splitting $\mathcal{S}$ into a 90% training set and a 10% testing set $\mathcal{S}_T$. It corresponds to the average probability to identify the correct next item in $\mathcal{S}_T$:

$$Acc(\mathcal{R}, \mathcal{S}_T) \coloneqq \frac{100}{|\mathcal{S}_T|} \sum_{s \in \mathcal{S}_T} \frac{1}{|s|-1} \sum_{i=1}^{|s|-1} P^{\mathcal{R}}(s_{i+1} | s_1 s_2 \dots s_i) \tag{4}$$

The increase in accuracy between FON$_1$ and the other models justifies the use of higher-order models. For example we can correctly predict almost half of the ports visited by ships in the PORTS dataset using D_{KL}-VON or MC-VON. This score drops to 13% for the regular FON$_1$ network. This difference is less important for WIKI. Accordingly, the optimal order found using Scholtes' method [19] is 1 for this dataset.

[1] `https://pypi.org/project/honyx/`

Table 2: Comparison of the network models used.

Dataset	**Network**	$\|V\|$	**Order**	**Acc** $\pm 2sd$
PORTS	FON$_1$	909	1	13.71 ± 0.73
	FON$_2$	9,437	2	31.73 ± 1.38
	D_{KL}-VON(1.95)	9,559	6	38.56 ± 1.63
	D_{KL}-VON(1)	18K	8	46.48 ± 1.89
	MC-VON(0.001)	9,553	16	42.93 ± 2.22
	MC-VON(0.05)	18K	27	48.17 ± 2.23
AIR	FON$_1$	175	1	19.48 ± 0.09
	FON$_2$	1,716	2	27.44 ± 0.10
	D_{KL}-VON(2.85)	28K	6	36.50 ± 0.15
	D_{KL}-VON(1)	58K	6	39.37 ± 0.19
	MC-VON(0.001)	28K	6	37.11 ± 0.15
	MC-VON(0.29)	58K	6	39.19 ± 0.20
MSNBC	FON$_1$	17	1	13.82 ± 0.07
	FON$_3$	4,061	3	22.18 ± 0.16
	D_{KL}-VON(1.585)	5,774	8	22.04 ± 0.15
	D_{KL}-VON(1)	28K	11	22.29 ± 0.17
	MC-VON(0.001)	5,771	122	22.44 ± 0.17
	MC-VON(0.027)	28K	145	22.43 ± 0.16
WIKI	FON$_1$	100	1	21.48 ± 0.65
	D_{KL}-VON(3.39)	306	4	21.87 ± 0.67
	D_{KL}-VON(1)	2,260	4	23.29 ± 0.64
	MC-VON(0.001)	304	4	22.85 ± 0.65
	MC-VON(0.35)	2,257	12	23.39 ± 0.70

We now compare the networks created using D_{KL}-VON and MC-VON. For a given order, the set of contexts found relevant is different even if the parameters are tuned to have sets of similar size. This suggests that the difference between the two methods is not just a matter of parameterization. The relevant contexts occur less frequently on average in D_{KL}-VON than the contexts found using MC-VON. This effect may be similar to the situation shown in Example 1: the low-frequency contexts may have a large D_{KL} value that easily passes the test of Eq. 1. Networks constructed using MC-VON have a larger maximum order which can be very large. This is expected since the criterion used does not inherently penalize large contexts. The largest discrepancies are obtained with MSNBC and PORTS. Note, however, that such contexts are rare; the vast majority of memory nodes are of order 2 or 3.

When comparing variable-order networks of similar size, MC-VON seems to match D_{KL}-VON in terms of accuracy. For PORTS or WIKI, it clearly outperforms it. For MSNBC and AIR, the results are closer. This supports the idea that the criterion used for MC-VON helps to produce higher-order networks that are more consistent with the data. A final observation is that the computational time required to construct networks using MC-VON is several orders of magnitude higher than the time required using D_{KL}-VON (*e.g.* half an hour versus a few seconds for MSNBC). Although network analysis tasks rarely come with online computational constraints, a future challenge would be to improve the computation of MC-VON.

Table 3: Spearman correlations between higher-order PageRank rankings and Levenshtein edit distance between the Top10s.

Dataset	**Networks**	FON$_*$	D_{KL}-VON(1)	MC-VON
PORTS	FON$_1$	0.96 / 3	0.95 / 5	0.98 / 5
	FON$_*$		0.99 / 4	0.96 / 5
	D_{KL}-VON(1)			0.96 / 4
AIR	FON$_1$	0.99 / 3	0.98 / 4	0.97 / 4
	FON$_*$		0.99 / 2	0.98 / 2
	D_{KL}-VON(1)			0.99 / 0
MSNBC	FON$_1$	0.99 / 4	0.97 / 4	0.98 / 3
	FON$_*$		0.98 / 2	0.99 / 2
	D_{KL}-VON(1)			0.98 / 4
WIKI	FON$_1$	1. / 0	0.90 / 6	0.96 / 5
	D_{KL}-VON(1)			0.94 / 6

4.2 Higher-order PageRank

We now investigate how the choice of model impacts network mining algorithms results. In particular, we look at the rankings achieved using a higher-order version of the PageRank (PR) centrality measure. Since higher-order networks are still weighted graph, we can compute nodes' PRs and then define *items*' PR as the sum of their representations' PR values [23]. To be more precise, we use a corrected version of the PR metric for higher-order networks [5]. This reformulation corrects a bias due to the multiplicity and the non-normal distribution of the representation of each item. With this correction, we can compare PR items for HON of different sizes.

Table 3 reports the Spearman correlation coefficient as well as the edit distance between the top 10 found for each network. We observe strong similarities in rankings between all of the networks including FON$_1$. For all of Spearman correlations, the hypothesis of independence between the samples can be rejected. Therefore, the choice of HON model does not completely reverse the hierarchy of items. Even if sequential dependencies exist in the dataset, PR-based centrality analysis is still relevant without taking them into account. However, there are still differences between the ranking found as suggested by the difference between the top 10 most important items. These rank differences are more common when the PR values are more evenly distributed (*e.g.* for the WIKI dataset). In order to better see these differences, Fig. 2 shows the actual top 10s for the PORTS dataset.

5 Conclusion

We introduced a variable-order network model MC-VON that uses statistical significance for the identification of relevant contexts in a variable-order Markov model. Experiments have shown that we can construct sparser networks in which random walks will represent input sequences almost as well or even better than using known models. We therefore argue that our approach is a good alternative to D_{KL}-VON. On the other hand, the difference between the networks is not as important when looking at the items PageRank rankings. This suggests

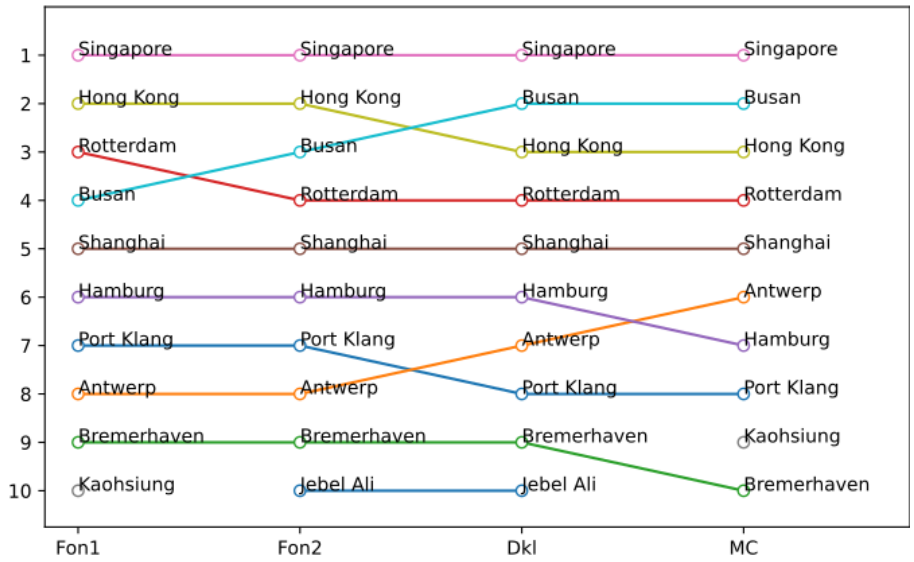


Figure 2: Differences between the top 10 ports in terms of PageRank according to the network model (PORTS dataset).

the effects of the choice of model on the information we extract from those network may be more limited. However, the choice of model may be more important when using other network mining algorithms [13].

A direction for future work is the improvement of the computation of our model p value. The method here is designed to obtain a stable solution that is not affected a lot by Monte-Carlo innate randomness. We believe that faster approximations and still stable procedures are possible, for example using Sequential Monte-Carlo techniques.

References

[1] Ron Begleiter, Ran El-Yaniv, and Golan Yona. On prediction using variable order markov models. *Journal of Artificial Intelligence Research*, 22:385–421, 2004.

[2] Irad Ben-Gal, Gail Morag, and Armin Shmilovici. Context-based statistical process control: A monitoring procedure for state-dependent processes. *Technometrics*, 45(4):293–311, 2003.

[3] Jose Borges and Mark Levene. Evaluating variable-length markov chain models for analysis of user web navigation sessions. *IEEE Transactions on Knowledge and Data Engineering*, 19(4):441–452, 2007.

[4] Brenno Caetano Troca Cabella, Márcio Júnior Sturzbecher, Walfred Tedeschi, Oswaldo Baffa Filho, Dráulio Barros de Araújo, and Ubiraci Pereira da Costa Neves. A numerical study of the kullback-leibler distance in functional magnetic resonance imaging. *Brazilian Journal of Physics*, 38:20–25, 2008.

[5] Célestin Coquidé, Julie Queiros, and François Queyroi. Pagerank computation for higher-order networks. In *International Conference on Complex Networks and Their Applications*, pages 183–193, 2021.

[6] Dong Ding, Axel Gandy, and Georg Hahn. A simple method for implementing monte carlo tests. *Computational Statistics*, 35:1373–1392, 2020.

[7] Tina Eliassi-Rad, Vito Latora, Martin Rosvall, and Ingo Scholtes. Higher-Order Graph

Models: From Theoretical Foundations to Machine Learning (Dagstuhl Seminar 21352), 2021.

[8] Axel Gandy. Sequential implementation of monte carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association*, 104(488):1504–1511, 2009. https://www.jstor.org/stable/40592357.

[9] Richard E Korf. A complete anytime algorithm for number partitioning. *Artificial Intelligence*, 106(2):181–203, 1998.

[10] Renaud Lambiotte, Martin Rosvall, and Ingo Scholtes. Understanding complex systems: From networks to optimal higher-order models. *arXiv preprint arXiv:1806.05977*, 2018.

[11] Tiago P Peixoto and Martin Rosvall. Modelling sequences and temporal networks with dynamic community structures. *Nature communications*, 8(1):1–12, 2017.

[12] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *J. Graph Algorithms Appl*, pages 191–218, 2006.

[13] Julie Queiros, Célestin Coquidé, and François Queyroi. Toward random walk-based clustering of variable-order networks. *Network Science*, 10(4):381–399, 2022.

[14] François Queyroi. Least likely sample in multivariate hypergeometric distributions. Mathematics Stack Exchange. (version: 2022-03-02).

[15] François Queyroi, Julie Queiros, and Samuel Maistre. Code and Datasets "Sampling based sequential dependencies discovery in Higher-Order Network Models", February 2024.

[16] Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.

[17] Martin Rosvall, Alcides V Esquivel, Andrea Lancichinetti, Jevin D West, and Renaud Lambiotte. Memory in network flows and its effects on spreading dynamics and community detection. *Nature communications*, 5(1):1–13, 2014.

[18] Mandana Saebi, Jian Xu, Lance M. Kaplan, Bruno Ribeiro, and Nitesh V. Chawla. Efficient modeling of higher-order dependencies in networks: from algorithm to application for anomaly detection. *EPJ Data Sci.*, 9(1):15, 2020.

[19] Ingo Scholtes. When is a network a network? multi-order graphical model selection in pathways and temporal networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1037–1046, New York, USA, 2017. ACM.

[20] Philipp Singer, Denis Helic, Behnam Taraghi, and Markus Strohmaier. Detecting memory and structure in human navigation patterns using markov chain models of varying order. *PLOS ONE*, 9(7):1–21, 07 2014.

[21] Leo Torres, Ann S Blevins, Danielle Bassett, and Tina Eliassi-Rad. The why, how, and when of representations for complex systems. *SIAM Review*, 63(3):435–485, 2021.

[22] Jian Xu, Mandana Saebi, Bruno Ribeiro, Lance M Kaplan, and Nitesh V Chawla. Detecting anomalies in sequential data with higher-order networks. *arXiv preprint arXiv:1712.09658*, 2017.

[23] Jian Xu, Thanuka L. Wickramarathne, and Nitesh V. Chawla. Representing higher-order dependencies in networks. *Science Advances*, 2(5), 2016.

[24] Hao Yin, Austin R. Benson, Jure Leskovec, and David F. Gleich. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 555–564, New York, NY, USA, 2017. Association for Computing Machinery.

FRCCS

Structify-Net: A python library for generating Random Graphs with controlled size and customized structure

Rémy Cazabet[1✓], Salvatore Citraro[2] and Giulio Rossetti[2]

[1] *Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France ; remy.cazabet@gmail.com*
[2] *Institute of Information Science and Technologies "A. Faedo" (ISTI), National Research Council (CNR), Italy*
✓ *Presenting author*

Abstract. This extended abstract introduces Structify-Net, a python library implementing a network generator. This generator has been published in a recent article[2].

Keywords. *Random Networks; Network Generator; Network Structure; Community Structure; Spatial Structure*

1 The Structify-Net Model

The mesoscale organization of networks is one of the most studied topics in network science. Some structures in particular have attracted a lot of attention, such as the block or community structure, the spatial structure, or the core-periphery structure. Many works have been published on how to detect these structures in observed networks, and how to generate random graphs having such a structure. In this work, we propose a framework to generate random graphs 1)having a desired number of nodes and edges, 2) following a desired structure –not limited to blocks and spatial structures 3) whose structure strength is controlled with a single parameter, from deterministic to fully random.

The principle is to define a **rank-score function** $f(u, v) = d, d \in \mathbb{R}$. d is used to rank pairs of nodes in their order of structure preference. A parametric function derived from Bézier curves is used to set the probability of observing each pair of edges among n nodes such as the expected number of edges equals m. m and n are chosen as objectives by the experimenter. The function itself is controlled by a structure strength parameter $\epsilon \in [0, 1]$ such as when $\epsilon = 1$, all node pairs have the same probability to be connected (ER random graph); when $\epsilon = 0$, the m edges connect the m node pairs of highest d, and the probabilities to observe an edge given the rank in d interpolates smoothly for values in-between.

The Python library contains a collection of well-known structures called the *structure zoo*, as illustrated in Fig. 1. It contains for instance block structures, spatial structures, nested structures, etc.

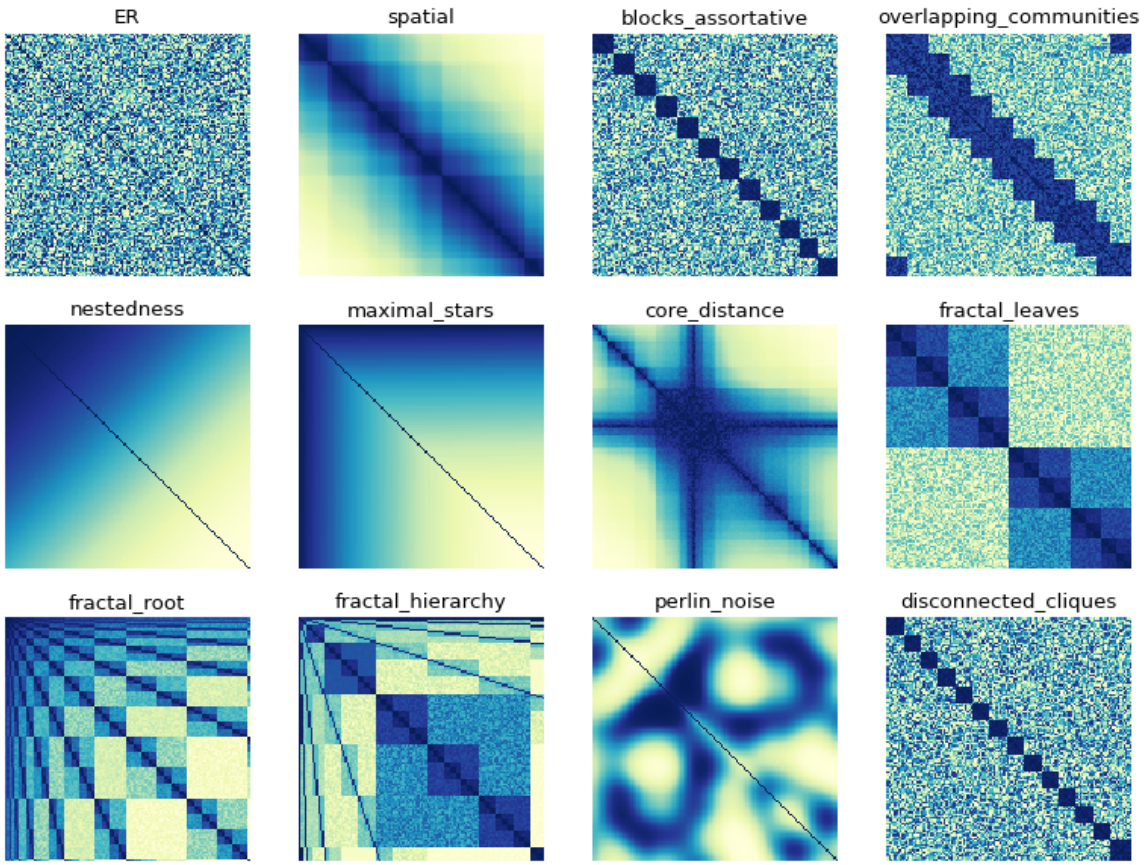


Figure 1: The Structure Zoo. Matrix of node-pairs ranks for networks with 128 nodes. Matrices can be read as adjacency matrices, with values representing edge probability instead of edge presence. Darker colors correspond to lower ranks. When involving spatial or clique positions, nodes are ordered according to this value.

2 Interest and Applications

Structify-Net can be used in a variety of contexts. Here are some illustrative examples:

- When studying propagation on networks, whether for biological viruses or information on social media, it is well known that the structure of the network (degree distribution, spatial structure, etc.) affects the diffusion. Using Structify-net, one could compare the diffusion process, its speed and amount of nodes reached, according to chosen structures, varying the degree of randomness.
- In community detection, evaluation on synthetic benchmarks such as SBM or LFR is an essential step. Using Structify-Net, one could design a large variety of network generators with community structure, having different properties. For instance, one could have a generator mixing spatial and block structure, or for which each bloc is organized as a spatial network, or hierarchically. All these variants can be done simply by providing an appropriate function.
- Link prediction benchmarks. As for community detection, link prediction algorithm are often compared on synthetic benchmark. Structify-Net could be used to compare how different methods perform differently when confronted to different type of structures. Furthermore, one could also easily include node attributes in the generation process: the rank function can depend, for instance, on node positions (spatial structure) and on node

attributes (homophily/heterophily), while also imposing some hierarchical structure. One simply needs to provide a ranking function $f(X)$, which, given the node properties X, return a higher scores for nodes that are close and/or have similar structures. Positions and attributes can be generated, or come from real data.

3 Python library

An important aspect of such a generator is to allow other researchers to use it for their own needs, whether it be to generate networks according to a structure described in the structure zoo, or to define their own. We thus release a pip installable python library [1], together with its documentation[1]. For convenience, the library is compatible with Networkx ([3]). Obtaining a rank model corresponding to one of those defined in the zoo, such as the nested structure, is as simple as calling it:

```
import structify_net.zoo as zoo
n=128
rank_model = zoo.sort_nestedness(n)
```

Generating a network as a Networkx object from it is straightforward:

```
import structify_net.zoo as zoo
n=128
m=512
generator = zoo.sort_nestedness(n).get_generator(epsilon=0.5,m=m)
g_generated = generator.generate()
```

One can also define a custom structure by providing a rank-score function:

```
import structify_net as stn
n=128
m=512

def R_nestedness(u,v,_):
    return u+v
rank_nested = stn.Rank_model(n,R_nestedness)
g = rank_nested.generate_graph(epsilon=0.1,m=m)
```

The library allows easy plotting of the rank-score matrices and node-pair probability matrices, and more generally reproduces all the content of the current article.

References

[1] Remy Cazabet. Structify-net. `https://github.com/Yquetzal/structify_net/`, 2023.
[2] Remy Cazabet, Salvatore Citraro, and Giulio Rossetti. Structify-net: Random Graph generation with controlled size and customized structure. *Peer Community Journal*, 3, 2023.
[3] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[1] `https://structify-net.readthedocs.io/en/latest/`

Economics & Finance

FRCCS

A simple model to describe an in-silico financial market populated by real traders

Michele Vodret[1✓], Damien Challet[1]

[1] *Université Paris-Saclay, CentraleSupélec, Laboratoire de Mathématiques et Informatique pour la Complexité et les Systèmes, 91192 Gif-sur-Yvette, France.;*
mvodret@gmail.com, damien.challet@centralesupelec.fr
[✓] *Presenting author*

Abstract. In a previous piece of literature, a laboratory experiment assessed that traders in an in-silico financial market were prone to excess trading. Here we propose a simple model able to rationalize many of the regularities exhibited by the associated datasets. Surprisingly, previous returns impact the price asymmetrically based on their sign. We suggest that this bias might not be caused by an idiosyncratic trader's behavior but instead can be understood as an emerging property of the collective dynamics.

Keywords. *Market Microstructure; Learning; Collective behavior;*

1 Data, price impact simplification

~ 200 humans took part in a laboratory experiment [1]: they were divided into 9 pools. They were asked to take part in two trading sessions. Here we will deal only with the first of the two session for each pool of partecipants.

Returns, i.e., log-price differences, have the following dynamics:

$$r_t = m + s\eta_t + I_t, \tag{1}$$

where m is a constant positive growth rate, η_t is a white noise, s is the associated standard deviation and I_t is the so-called price impact function which related order flows to price changes. Importantly for what follows, the history of innovations $m + s\eta_t$ is the same across all sessions and experiments. This allows us to investigate statistical regularities.

The price impact function I_t is given by:

$$I_t = \frac{N_t^{active}}{N} \frac{Buy_t - Sell_t}{Buy_t + Sell_t}, \tag{2}$$

where N is the total number of traders in each sessions (~ 20), N_t^{active} is the number of traders active at time t, Buy and $Sell$ are the orders in currency unit. All traders have structural knowledge, i.e. they know the equations above, as well as the values of m and s.

The price impact function can be substantially simplified. In fact, the impact function recorded

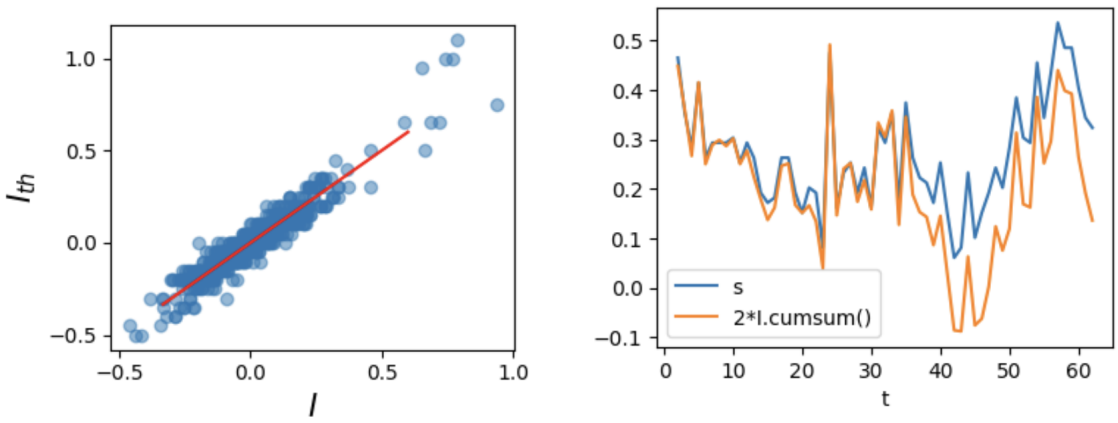


Figure 1: (Left) comparison between new proposed impact model and the one originally formulated. (Right) Correctness of Eq. (5).

in the dataset is quite well accounted for by the simpler parametrization given by

$$I_t^{th} = -\frac{1}{N}\sum_{i=1}^{N} a_t^i s_{t-1}^i, \tag{3}$$

where $a_t^i \in \{0, 1\}$ depending wether the trader i was active at time t and $s_{t-1}^i \in \{-1, 1\}$ depending whether the trader was 'in' or 'out' of the market [1]. The summation in the equation above accounts for the term $Buy_t - Sell_t$, as we can see from Fig. 1. This insight allows for quite a few more intuitions about the underlying dynamics, given its direct link with state variables a_t^i and s_{t-1}^i.

2 State dynamics

The micro-state dynamics is given by

$$s_t^i = (1 - 2a_t^i)s_{t-1}^i \tag{4}$$

so that one obtains

$$s_t := \frac{\sum_i s_t^i}{N} = s_{t-1} + 2I_t, \tag{5}$$

which again can be verified as shown in the right panel of Fig. 1.

Interestingly, we have one more insight: the price impact function is highly correlated with the previous return; that is, we found empirically the relation:

$$I_t \propto -\beta r_{t-1}, \tag{6}$$

in particular, two different proportionality factors for positive and negative returns allow for a good collapse of the three curves as can be appreciated in Fig. 2.

2.1 recap

Let me here recap what we have achieved so far:

[1] In this trading experiment traders at each step can only convert all their asset to cash, all their cash to asset, or keep their inventory unchanged. Therefore the state variable s_t^i is two-valued.

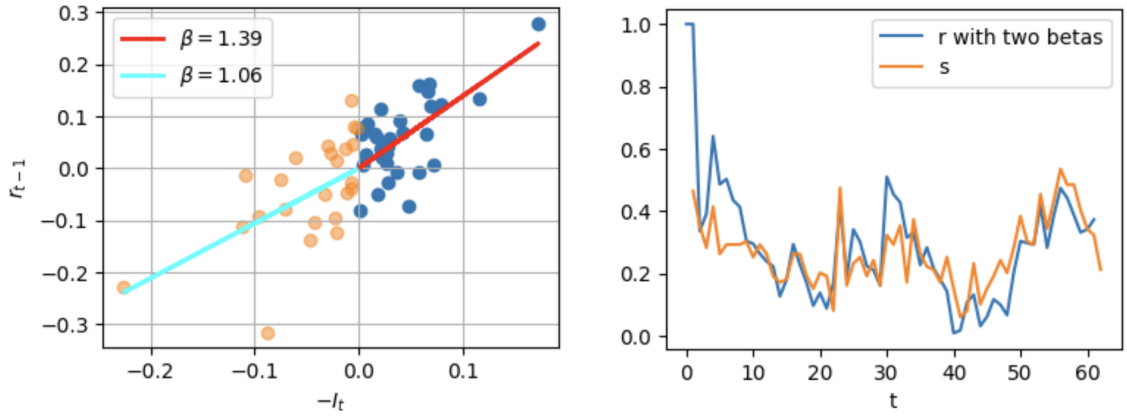


Figure 2: Two different proportionality factors β for positive and negative returns. On the left, we calibrate the βs; on the right, we superpose the average of 'predictions' and the actual mean state coming from experiments. $\sim$ Correctness of Eq. (6).

1. simplification of impact dynamics: Eq. (3). This leads to a simplification in the state dynamics given by Eq. (5).
2. impact negatively correlated with previous return: Eq. (6). Two possible rates for positive and negative returns?

If the final statement in item [2.] above is correct, then this means that positive previous return triggers more (sell) orders than those (buy) triggered by negative previous returns.

Therefore: positive returns lead to negative information, which is then over-interpreted by the market as a whole. Note that we cannot conclude this at the level of a single agent at this stage. Setting up a micro-model is therefore necessary.

Anticipation: a very simple model without positivity/negativity bias is able to reproduce these quantitative findings.

3 A model

The model is given by:

$$P(a_t^i = 1|s_{t-1}^i) = \frac{1 - s_{t-1}^i \tanh\left(\mathcal{I}_t^i|s_{t-1}^i\right)}{2}, \tag{7}$$

where $\mathcal{I}_t^i|s_{t-1}^i$ means that agents have different sources of information if they hold stock or if they do not. We assumed the following parametrization

$$\mathcal{I}_t^i|1 = LTT^i - W_t^i + \eta_t^i, \tag{8}$$

where LTT^i is an idiosincratic long-term expected gain, $W_t^i = W_t^i = -\sum_{t'=t_i^*}^{t-1} r_{t'}$ is the up-do-date gain of the current trade, while η_t^i is an idionsicratic white noise error. So, if agents hold stock, they wait until their investment reaches their long-term target LTT^i. On the other hand, if the agent is out of the market, his signal is given by

$$\mathcal{I}_t^i| - 1 = \alpha r_{t-1} + \eta_t^i. \tag{9}$$

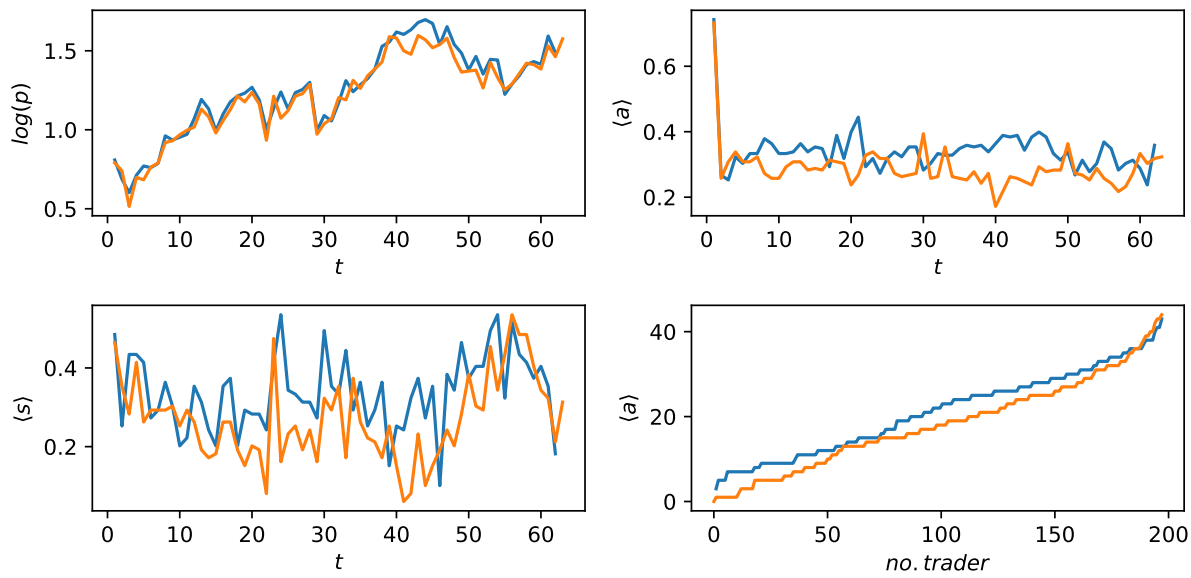


Figure 3: Comparison between the outcome of the model (blue lines) and the one recorded in the dataset (orange lines). We simulated different sessions, and we plot the average across time of the results.

With the definition given by Eq. (7) we have

$$s_t = \langle \tanh(\mathcal{I}_t^i) \rangle. \tag{10}$$

So we obtain

$$s_t = s_{t-1} + (\tanh \mathcal{I}_t^i \rangle - \langle \tanh \mathcal{I}_{t-1}^i) \tag{11}$$

and this equation reduces to Eq. (6) if i) $\tanh x \sim x$ and ii) if $\mathcal{I}_t - \mathcal{I}_{t-1} \propto -r_{t-1}$, but this is (most of the time) true in our model given Eq. (8). In fact W_t is the accumulated gain at time-step t before taking actions.

4 Results

Here, we anticipate what we found and the issues (i.e., things to do) with this section. Basically by choosing

$$P(LTT^i) = 0.5 + Uniform(-1, 1) \tag{12}$$

and by tuning the parameters one is able to reproduce several key metrics (see Fig. 3).

5 Conclusion

We have elucidated several stylized facts on a dataset from a laboratory experiment where $\sim$ 200 participants traded in an in-silico financial market. We built up a stylized model able to capture these stylized facts. Interestingly, this brief discussion shows that some 'bias' at the aggregate level can be derived from a model where participants are not exposed directly to this very same bias. In other words, biases can result from the dynamics itself; this is reminiscent of the finding in Ref. [2], where the learning dynamics induced the risk-aversion, and it was

not imposed a priori. We plan to conduct a robust calibration of the presented model on the dataset.

References

[1] Joao da Gama Batista, Domenico Massaro, Jean-Philippe Bouchaud, Damien Challet, and Cars Hommes. Do investors trade too much? a laboratory experiment. *Journal of Economic Behavior & Organization*, 140:18–34, 2017.

[2] Michele Vodret. Cognitive energy cost of informed decisions. *arXiv preprint arXiv:2310.15082*, 2023.

Economic Integration of Africa in the 21st Century: Complex Network Approach and Panel Regression Analysis (Extended Abstract)

Tekilu Tadesse Choramo[1,2], Jemal Abafita[2], Yerali Gandica[3,5], and Luis E C Rocha[1,4]

[1]Department of Economics, Ghent University, Ghent, Belgium
[2]Department of Economics, Jimma University, Jimma, Ethiopia
[3]Department of Mathematics and Master in Big Data, Universidad Internacional de Valencia (VIU), Valencia, Spain
[4]Department of Physics and Astronomy, Ghent University, Ghent, Belgium
[5]ECAM-EPMI, Research Laboratory in Industrial Eco-Innovation and Energetics LR2E, Laboratory of QUARTZ (EA 7393), Cergy 95092, France

April 20, 2024

{Email: tekilutadesse.choramo@ugent.be, jemal.abafita@ju.edu.et, ygandica@gmail.com, luis.rocha@ugent.be}

Abstract

The objective of the study is to analyze the effect of macroeconomic indicators on network-based measures of economic integration. We apply our network methodology to intra-African trade and find a sizable and significant positive relationship between our network-based integration measure and a country's economic development, institutional quality, regional trade agreements, human capital, FDI, and infrastructure quality while that of trade costs, global financial crisis, and overlapping membership were negatively associated with on network-based economic integration. These findings imply that identifying key economic and institutional factors of trade partners is critical for economic integration in Africa.

Keywords: Africa Trade, Network Model, Economic Integration, Network-based Indicators, Dynamic Panel

JEL Code: C33, F14, F15, L14

1 Introduction

Countries around the world, including Africa, have witnessed a significant increase in economic integration in recent decades, leading to trade diversification and sustainable development [4]. Empirical evidence and economic theory assert that a country's economic integration through trade and finance has a great contribution to promoting economic growth by removing frictions and barriers to trade, improving the allocation of resources efficiently, and reducing trade costs [5]. Complex networks are promising tools for understanding complex systems and analyzing how countries are interconnected with each other. The structure of the relationships between the countries can be highlighted by visualizing trade flows as a network of nodes (countries) and links (trade flows between each country pair) [3]. Many structural and topological features of trade networks can be explained by network metrics, which are crucial indicators for differentiating the competitive advantage of a given country ('[1]; [2]). However, existing measures of economic integration fail to capture the complex interactions among trading partners. Therefore, there is a need for a comprehensive analytical framework to understand economic integration in Africa holistically. In this context, the method of complex network analysis has gained popularity as an in-depth analysis of trade integration. This study aims to investigate the factors driving economic integration in Africa by utilizing network-based indicators.

2 Methodological Approach

The study adopts a combination of complex network analysis and dynamic panel regression. Complex network analysis is utilized to analyze the roles and central positions of countries in intra-African trade in Africa. In this paper, we propose different dimensions of network-based measures of economic integration able to capture different network structures and thus different indicators of economic integration mainly weighted in and out-degree, PageRank, betweenness, random-walk, and closeness centrality from network centrality measures while k-core decomposition and clustering coefficient from non-centrality measure category. Dynamic panel regression is employed to examine the effect of macroeconomic indicators, such as economic development, institutional quality, Regional trade agreements, trade cost, human capital, infrastructure quality, population, FDI inflow, overlapping membership, and global financial crisis on network-based measures of economic integration. The analysis uses data from the United Nations Conference on Trade and Development, spanning the period between 2000 and 2019 to focus on the recent development of African trade We create a balanced panel of N= 54 countries for which we have the total import and export flows from 2000 to 2019 (T= 20 years) in the USD and build the trade matrix for the African countries.Data from 2000 to 2019 is used for the analysis. Network measures are computed using the NetworkX package in Python, and econometric analysis is conducted using the R package.

3 The Main Findings

Based on UNCTAD data in 2021, although it has expanded from around 10.4% in 2000 to roughly 17.8% in 2017, the percentage of intra-African trade as a share of all African exports and imports is still low when compared to the regions of North America (47%), Asia (59%), and Europe (69%), demonstrating how Africa's trade agreement development lags well behind that of the rest of the world. Regarding the network perspectives, network analysis shows that countries have diverse

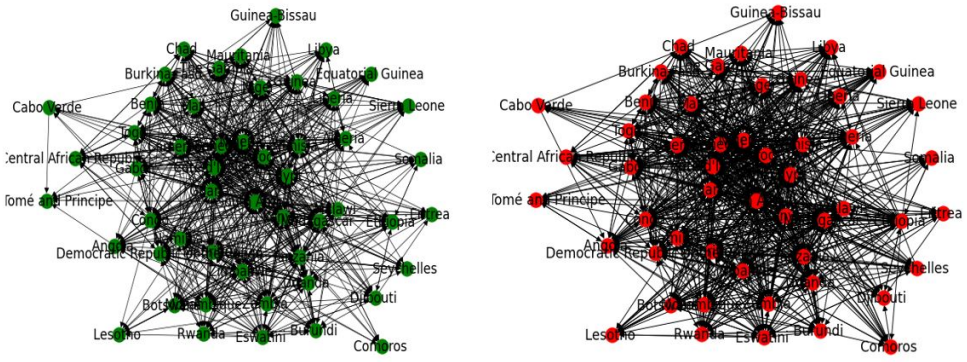


Figure 1: African trade network in 2000 and 2019

roles and capture various central positions in continuously growing intra-African trade. Moreover, African trade networks possess core-periphery structures, and the trade agglomeration effect of the core countries has increased over time. The countries with the highest k-core, forming the core of the networks, are more exposed and more likely to trigger system-wide economic shocks. Similarly, the clustering coefficient indicates an increasing pattern with African countries building more relationships with countries that also trade with each other. The Econometric model reveals there is a heterogenous effect of macroeconomic variables on different dimensions of network centrality indicators. Accordingly, economic development, institutional quality, regional trade agreements, human capital, and infrastructure quality had a significant positive impact on network-based economic integration. Conversely, trade costs, global financial crisis, and overlapping membership show a negative impact on network-based economic integration. Using an external instrument for robustness checking, our estimated model is robust for all network-based measures considered in this study.

4 Theoretical importance

This study contributes to the existing literature on economic integration by utilizing network metrics and econometric models in combination. It provides a comprehensive understanding of economic integration in Africa from network perspectives where traditional measures failed to capture complex interdependence, shedding light on the factors that drive network-based measures integration. Network perspectives can shed light on the dynamics of economic integration over time. By analyzing changes in the network structure over time, researchers can identify how economic integration is evolving and what factors are driving those changes. Moreover, network perspectives can help identify key players and hubs in the economic integration system. By analyzing the centrality of different countries in the network, researchers can identify which actors are most important for

[1]Out of Africa's 54 countries, these countries will account for more than 45% of the continent's GDP and 44% of its population in 2022. These six countries all have thriving economies, and we think that their network-based integration would have a significant impact on the integration of all 54 African countries. Our justification for this decision is that if the continent's largest economies have consistently seen increasing integration trends in their network centrality metrics, the smaller economies will eventually make up with them, and we can therefore claim that the economic conditions are favorable for a future full continental level of economic integration.

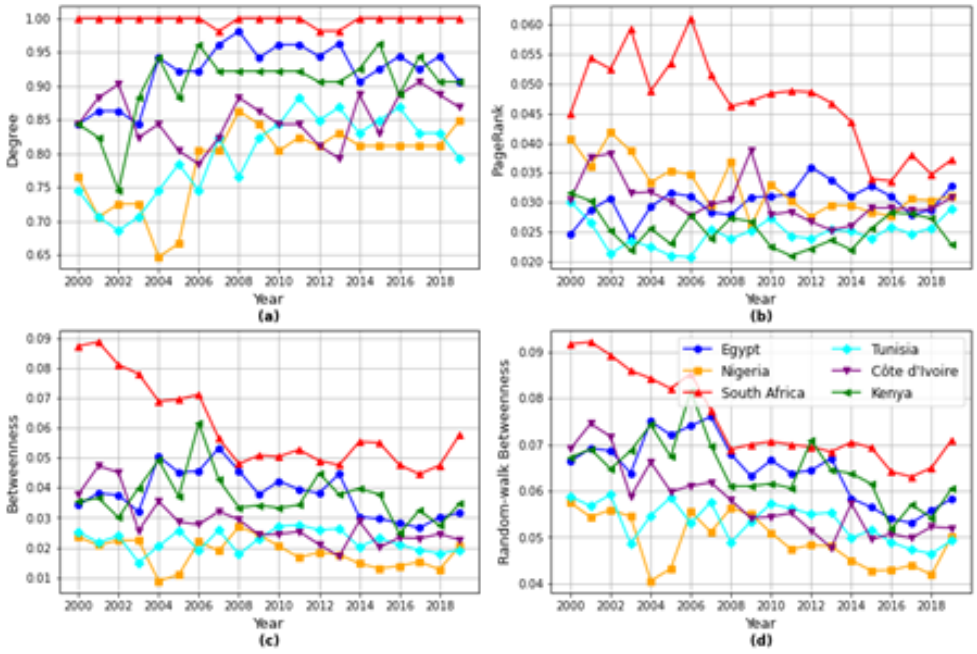


Figure 2: Network centrality measures of six selected African countries.[1]

maintaining and developing economic integration, and which ones are most vulnerable to disruption.

5 Conclusion

The findings of this study highlight the importance of identifying key economic and institutional factors for achieving broader economic integration in Africa. The study demonstrates the usefulness of complex network analysis in understanding the roles and central positions of countries in intra-African trade networks. Additionally, the study emphasizes the positive impact of economic development, weighted version of regional trade agreements, institutional quality, infrastructure, FDI inflows, and human capital on different dimensions of network-based economic integration, while noting the negative impact of trade cost measured by tariff imposed and overlapping frequency ratio proxy for overlapping membership in regional trade blocs. This research provides valuable insights for policymakers and researchers interested in promoting economic integration in Africa through identifying not only trade partners but also trade partner's partners and strengthening the economies of African countries.

References

[1] U. Brandes, P. Kenis, and D. Wagner. "Communicating centrality in policy network drawings". In: *IEEE Transactions on Visualization and Computer Graphics* 9.2 (2003), pp. 241–253. DOI: 10.1109/TVCG.2003.1196010.

[2] Matthew O Jackson et al. *Social and economic networks*. Vol. 3. Princeton university press Princeton, 2008.

Table 1: System GMM estimation for network-based integration measures

Variable	Sin	Sout	PR	B	RWB	C	Cc	k-core
RGDPc	0.132***	0.125***	0.015**	0.064**	0.0042**	0.0063**	0.0024**	0.739***
	(0.029)	(0.044)	(0.006)	(0.009)	(0.0017)	(0.0026)	(0.001)	(0.055)
Human capital	0.109***	0.077*	0.091***	-0.084	0.0075**	0.0063**	-0.0020	0.859***
	(0.039)	(0.043)	(0.030)	(0.057)	(0.0027)	(0.0026)	(0.009)	(0.137)
Population	0.020	-0.0194	-0.0035	0.065***	-0.0045	0.0081**	0.0043**	-1.146**
	(0.016)	(0.034)	(0.007)	(0.020)	(0.007)	(0.0039)	(0.0015)	(0.099)
Trade cost	-	-	-0.0076	-0.093**	0.0025	0.0029***	0.0069***	-
	0.188***	0.328***	(0.0277)	(0.033)	(0.042)	(0.007)	(0.0024)	1.288***
	(0.032)	(0.108)						(0.234)
Infrastructure	0.213***	0.222***	0.0076***	0.056**	-0.0027	-0.0019	0.0084***	1.279***
	(0.056)	(0.099)	(0.020)	(0.021)	(0.0044)	(0.0044)	(0.0021)	(0.451)
Institutions	0.399***	0.247***	-0.0059	0.079**	0.0078**	0.0096**	0.0074**	1.346***
	(0.088)	(0.066)	(0.0057)	(0.035)	(0.0036)	(0.0041)	(0.0032)	(0.488)
Regional trade	0.233***	0.219***	0.0075***	0.045	0.0043***	0.0053**	-0.0039	1.668***
agreement	(0.045)	(0.087)	(0.002)	(0.034)	(0.0018)	(0.0022)	(0.0025)	(0.536)
FDI inflows	0.644***	0.475***	0.409**	0.316	0.355	0.523***	0.415*	2.542***
	(0.234)	(0.193)	(0.188)	(0.248)	(0.277)	(0.217)	(0.265)	(0.438)
Overlap	-	0.145	-	-0.215**	-0.177**	-0.205**	0.010*	-1.558**
frequency	0.219***	(0.099)	0.036***	(0.112)	(0.101)	(0.091)	(0.005)	(0.455)
ratio	(0.079)		(0.009)					
Financial	-0.0022	-0.030	-0.009	-0.044	-	0.005***	-	-
crisis	(0.021)	(0.027)	(0.005)	(0.032)	0.006***	(0.002)	0.005***	2.780***
					(0.002)		(0.002)	(0.734)
Lag of Dep.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hansen	0.267	0.356	0.376	0.456	0.542	0.266	0.389	0.458
AR(1) p-value	0	0	0	0	0	0	0	0
AR(2) p-value	0.822	0.874	0.184	0.849	0.118	0.108	0.099	0.904
Obs	720	720	720	720	720	720	720	720

Note: Clustered Robust standard errors are included in parentheses. *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Also note that Sin stands for weighted in-degree, Sout represents weighted out-degree, PR stands for PageRank, B stands for betweenness centrality, RWB represents random walk betweenness, C stands for closeness centrality and cc stands for clustering coefficient.

[3] Luca De Benedictis and Lucia Tajoli. "The World Trade Network". In: *The World Economy* 34.8 (2011), pp. 1417–1454.

[4] Jin Zhang, Zhiwei Cui, and Lei Zu. "The evolution of free trade networks". In: *Journal of Economic Dynamics and Control* 38 (2014), pp. 72–86.

[5] Mita Bhattacharya, John Nkwoma Inekwe, and Maria Rebecca Valenzuela. "Financial integration in Africa: New evidence using network approach". In: *Economic Modelling* 72 (2018), pp. 379–390. ISSN: 0264-9993.

Geometric and Topological Approach to Market Critical Points

Lucas Carvalho[1✓], Tanya Araujo[12]

[1] *ISEG - Instituto Superior de Economia e Gestão - Universidade de Lisboa; lucas_carvalho@aln.iseg.ulisboa.pt*
[2] *Research Unit on Complexity in Economics (REM/UECE); tanya@iseg.ulisboa.pt*
[✓] *Presenting author*

Within the classical frameworks of economic and financial theories, particularly those underpinned by axiomatic methods such as the Efficient Market Hypothesis (EMH) [1, 2], a paradigm exists wherein market prices are assumed to integrate all available information, reflecting a presumed rationality among investors. This traditional perspective is progressively challenged by empirical evidence, which suggests the emergence of systematic patterns in market prices. This is viewed in Econophysics as emergent characteristics of a complex system[4]. To address those emerging patterns in financial markets, researchers have adopted alternative approaches in recent decades [5, 6, 7, 8].

This study follows this path by employing both the Stochastic Geometry Technique (SGT) and the Minimal Spanning Tree (MST) to investigate the interconnectedness of financial markets. SGT, as detailed by Araújo and Louçã [11] and in [9, 10, 11], reduces the complex market data to a simpler geometric form. This method helps in identifying patterns in the market by mapping relationships between assets onto a space with fewer dimensions. Complementary, the MST method, following the seminal paper by Mantegna [3], streamlines the web of market connections into a tree structure. This approach highlights the most significant connections between assets, making it easier to see how different market components are linked.

In a departure from conventional analyses that predominantly scrutinize single asset classes, like equities[13, 14, 15, 16, 18], FX [17] or even countries liabilities[12] this research extends the SGT and MST methodologies to the use of Exchange-Traded Funds (ETFs). By Focusing on ETFs, serving as proxies for major asset classes, our approach allowed for an asset class-specific investigation of market behavior, affording insights into the individual and collective responses to market stresses.

To differentiate business-as-usual to turbulent periods, it was elected a rule-based algorithm designed by Zegadlo [19] due to it's systematic and simplicity identification of critical market periods.

The dataset, which spans from January 02, 2006, to December 30, 2022, covers 85 securities representing 11 distinct asset classes. The study identifies key moments of heightened market activity across various classes. The Global Financial Crisis (GFC) serves as a good example, where a discernible sequence of peaks emerged: starting with Real Estate, followed by the other classes.

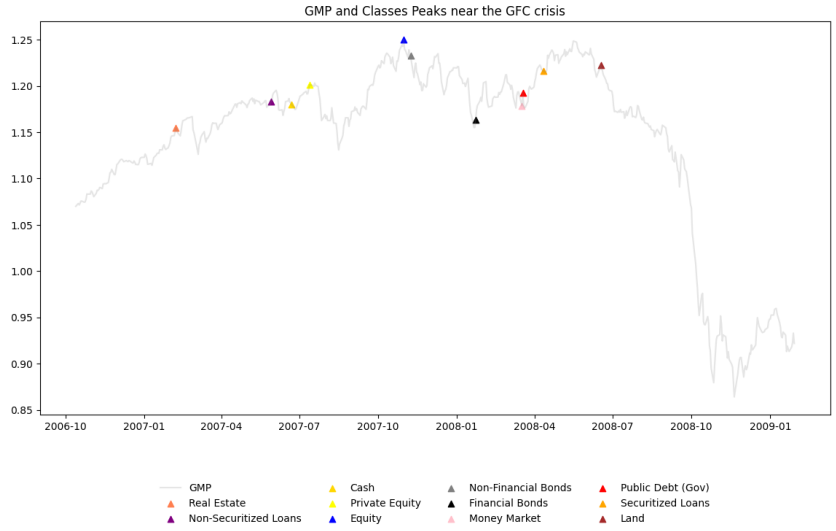


Figure 1: Asset Classes Peaks and Valleys Near the GFC

These peaks not only corroborate the initial impact within the Real Estate sector but can also illustrate the sequential distress propagated through different asset classes. The granularity of the data provides a picture of the systemic risks and interdependencies that characterize financial markets.

Figures 2, 3, 4 and 5 provide a comparative spatial and network analysis of asset classes during two different periods - Real Estate (RE) peak at 2007-02-07 and Public Debt (GOV) peak at 2008-03-18 - preceding the Global Financial Crisis. Utilizing a 66-day correlation window, these figures capture the temporal evolution of market structures through both a three-dimensional subspace, derived by Stochastic Geometry Technique (SGT), and connectivity patterns, revealed by Minimum Spanning Tree (MST).

During the first period, RE is observed to be closely aligned with multiple asset classes in the 3D subspace, suggesting its potential role in shaping market sentiment, possibly as a gauge of broader economic health. However, in the MST, RE peripheral position implies its secondary role as a direct transmission vector among assets.

In the second period, towards the Government Debt (GOV) class peak, the proximity and clustering of RE relative to other assets in the 3D subspace diminish, coinciding with the escalation of the GFC. This transition may reflect a reallocation of market focus and a reassessment of risk. In this period the MST shows an increase in direct links to RE, which may illustrate RE change of role from influencing market sentiment to becoming an intermediary in the dissemination of risk.

This apparently dichotomy presented by the 3D subspace and MST approaches shows signs of the complexity of financial markets, where an asset class can simultaneously be a driver of sentiment and not a direct transmission channel in the network. It could make sense to argue that RE might be a driver of sentiment due to its broad impact on the economy and investor psychology, even if it isn't the most direct pathway for risk or information flow in the financial network during the period analyzed.

From this analysis, we can infer that the influence of asset classes on the market is dynamic, changing as external conditions evolve. The disposition of an asset class in the 3D subspace and its connectedness in the MST during different periods can reveal how certain asset classes serve as indicative for the market's health.

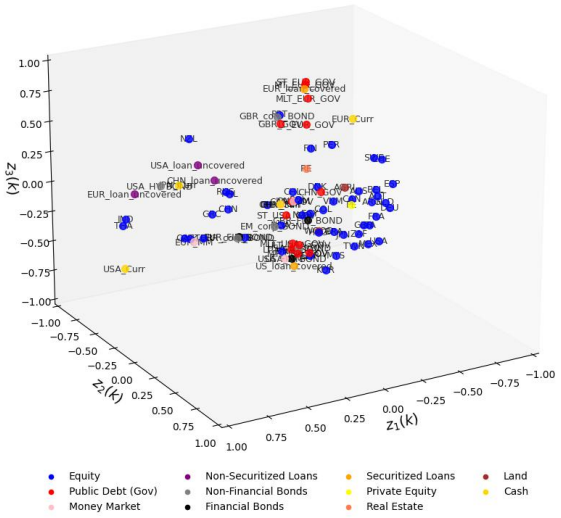


Figure 2: 3D RE Peak (66 days corr) - 2007-02-07

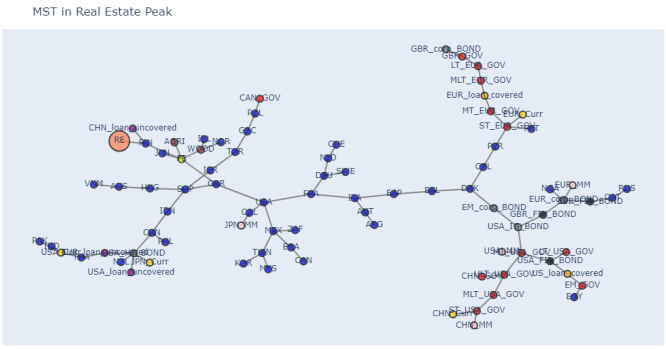


Figure 3: MST RE Peak (66 days corr) - 2007-02-07

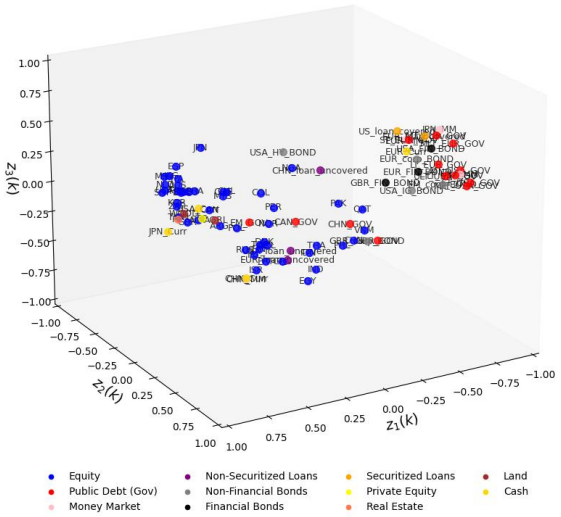


Figure 4: 3D GOV Peak (66 days corr) - 2008-03-18

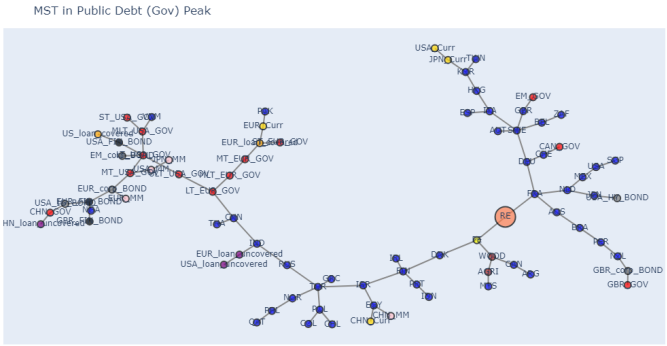


Figure 5: MST GOV Peak (66 days corr) - 2008-03-18

References

[1] Martin Sewell, "History of the efficient market hypothesis," *Rn*, vol. 11, no. 04, pp. 04, 2011.

[2] Eugene F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of Finance*, vol. 25, no. 2, pp. 383–417, 1969.

[3] Rosario N. Mantegna, "Hirarchical structure in financial markets," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 11, pp. 193–197, 1999, Springer.

[4] Dean Rickles, "Econophysics and the complexity of financial markets," in *Philosophy of complex systems*, pp. 531–565, 2011, Elsevier.

[5] R. Vilela Mendes, R. Lima, and Tanya Araújo, "A process-reconstruction analysis of market fluctuations," *International Journal of Theoretical and Applied Finance*, vol. 5, no. 08, pp. 797–821, 2002, World Scientific.

[6] Giovanni Bonanno, Guido Caldarelli, Fabrizio Lillo, and Rosario N. Mantegna, "Topology of correlation-based minimal spanning trees in real and model markets," *Physical Review E*, vol. 68, no. 4, 046130, 2003, APS.

[7] Paolo Giudici and Alessandro Spelta, "Graphical network models for international financial flows," *Journal of Business & Economic Statistics*, vol. 34, no. 1, pp. 128–138, 2016, Taylor & Francis.

[8] Paolo Giudici, Peter Sarlin, and Alessandro Spelta, "The interconnected nature of financial systems: Direct and common exposures," *Journal of Banking & Finance*, vol. 112, 105149, 2020, Elsevier.

[9] R. Vilela Mendes, Tanya Araújo, and Francisco Louçã, "Reconstructing an economic space from a market metric," *Physica A: Statistical Mechanics and its Applications*, vol. 323, pp. 635–650, 2003, Elsevier.

[10] Tanya Araújo and Francisco Louçã, "The seismography of crashes in financial markets," *Physics Letters A*, vol. 372, no. 4, pp. 429–434, 2008, Elsevier.

[11] Tanya Araújo and Francisco Louçã, "Trouble ahead-the subprime crisis as evidence of a new regime in the stock market," 2008, ISEG–Departamento de Economia.

[12] Alessandro Spelta and Tanya Araújo, "The topology of cross-border exposures: beyond the minimal spanning tree approach," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 22, pp. 5572–5583, 2012, Elsevier.

[13] Akbar Esfahanipour and SE Zamanzadeh, "A stock market filtering model based on minimum spanning tree in financial networks," *AUT Journal of Modeling and Simulation*, vol. 45, no. 1, pp. 67–75, 2015.

[14] Maman Abdurachman Djauhari and Siew Lee Gan, "Minimal spanning tree problem in stock networks analysis: An efficient algorithm," *Physica A: Statistical mechanics and its applications*, vol. 392, no. 9, pp. 2226–2234, 2013.

[15] Tristan Millington and Mahesan Niranjan, "Construction of minimum span-

ning trees from financial returns using rank correlation," *Physica A: Statistical Mechanics and its Applications*, vol. 566, 125605, 2021.

[16] Artur F Tomeczek, "A minimum spanning tree analysis of the Polish stock market," *Journal of Economics and Management*, vol. 44, no. 1, pp. 420–445, 2022.

[17] Gang-Jin Wang, Chi Xie, Feng Han, and Bo Sun, "Similarity measure and topology evolution of foreign exchange markets using dynamic time warping method: Evidence from minimal spanning tree," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 16, pp. 4136–4146, 2012.

[18] Yiting Zhang, Gladys Hui Ting Lee, Jian Cheng Wong, Jun Liang Kok, Manamohan Prusty, and Siew Ann Cheong, "Will the US economy recover in 2010? A minimal spanning tree study," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 11, pp. 2020–2050, 2011.

[19] Piotr Zegadło, "Identifying bull and bear market regimes with a robust rule-based method," *Research in International Business and Finance*, vol. 60, 101603, 2022, Elsevier.

FRCCS

Using complex networks for the analysis of the global land trade market

Marie Gradeler[1✓], Roberto Interdonato[2], Jeremy Bourgoin[2,1] and Ward Anseeuw[3]

[1] *International Land Coalition, Rome, Italy ; m.gradeler@ifad.org*
[2] *CIRAD, UMR TETIS, Montpellier, France ; {roberto.interdonato,jeremy.bourgoin}@cirad.fr*
[3] *Food and Agriculture Organization, Rome, Italy*
[✓] *Presenting author*

Abstract. Land deals around the world continue to reinforce the dispossession of traditional communities, foment the illegal transfer of public lands to private actors and generate violence. While accountability measures are widely promoted, they lack enforcement as investors responsible for land deals remain elusive. Overall, control over land remains opaque: the common trend for most of land investment is that the ultimate beneficiaries and major investors in these corporate and financial firms, especially investment funds, are often unknown. In this work, we aim at exploiting two linked data sources in order to shed light on this phenomenon: open data about worldwide Large Scale Land Acquisitions (LSLAs) provided by the Land Matrix Initiative, and detailed firm-level financial data provided by the Orbis database. Investors involved in LSLA deals are linked to the entries in Orbis, in order to retrieve information about their ownership chain (e.g., global ultimate owner, controlling shareholders, investor type). The resulting data ecosystem is then used to model complex networks at different scale, that represent relations among countries (e.g., location of the investor vs target country) and among investors (e.g., ownership/shareholding relations). We then apply state of the art complex network analysis techniques on these data structures in order to get new insights on the phenomenon of LSLAs and its global dynamics.

Keywords. *Large-scale land acquisitions; Investment networks; Land trade market*

1 Introduction

Limiting the impacts of global warming is one of the major issues that governments and intergovernmental organization are facing nowadays. Climate change is induced by a combination of drivers, including land use and land use change: the Intergovernmental Panel on Climate Change (IPCC) recently reported that Agriculture, Forestry and other land use (AFOLU) accounted for 22% of global greenhouse gases emissions [3]. As stated in 2020 by the Land Inequality report, control over land remains opaque: shareholdings in agricultural assets, particularly land, are not made public. Land and natural resources have become attractive assets to public and private investors, including actors who were formerly strangers to the rural sector. For instance, there is an increasing interest of financial capital (pension funds, private equity, and hedge funds) in non-conventional investment options; these are known as “alter-

native assets", and include commodities, land, and agricultural infrastructure. The common trend for most of land investment are that the ultimate beneficiaries and major investors in these corporate and financial firms, especially investment funds, are often unknown.

In the aim to promote transparency towards land transactions happening worldwide, the Land Matrix Initiative (LMI), a consortium of research and development partners, has been collecting data about LSLAs since 2012. The LMI database, which is provided in open access[1], includes data on large-scale land acquisitions (exceeding 200 hectares) for different investment sectors (e.g. agriculture, forestry, mining), and from diverse sources and methodologies (e.g. company reports, contracts, analytical and research reports, and press articles).

In order to study the global dynamics behind the land trade market, such data can be conveniently organized into complex networks. Previous research works by our team [2, 1] have proved the benefits of adopting this interdisciplinary approach, by focusing on the study of country-to-country land trade networks. These networks model the relations between target countries and the ones where the investing companies are located, by also taking into account the entity of the deals (e.g., size in hectares). This allowed to characterize the phenomenon of LSLAs, identify behavioral profiles of the different countries, and identify and rank anomalies in the land trade market (e.g., countries with a double investor/target role).

Nevertheless, despite all the efforts dedicated to data collection endeavors, the LMI database continues to exhibit gaps in detailed information pertaining to numerous transactions, that may raise concerns about the efficacy of current transparency measures. In this work, with the aim to complete the information about investors included in the Land Matrix database, we resort to detailed firm-level database provided by the Orbis database[2], provided by Bureau van Dijk—a subsidiary of Moody's Analytics. The linking between the Land Matrix and Orbis datasets allows us to add a further level of detail to the knowledge about the ownership chains behind transnational deals, resulting in a more complex modeling of the land trade network that, in turn, opens to advanced analyses and research questions on the subject.

More specifically, we are interested in discovering which sectors are involved in large land transactions and which actors can be made accountable for the recorded environmental and human-rights infringements. The resulting data ecosystem interconnecting a large network of entities and processes allows understanding and acknowledging patterns of land concentration and the sectors and actors directly or indirectly responsible for pushing our planet beyond its natural limits, especially through climate change, biodiversity loss, freshwater use and land system change.

2 Data and results

At the time of writing (March 2024), the Land Matrix provides open access data about nearly $7,000$ LSLAs, around $4,000$ of which are transnational (i.e., investor is located in a country other than the target one). These deals involve more than $10,000$ different investors (a single land acquisition deal can be associated to more than one investor). Beyond the fact that in Land Matrix location can be missing for some investors, the main issue is that ultimate ownership of the investing company may not correspond to the reported one. The hypothesis is that, in many cases, global ultimate owners and/or controlling shareholders of a given Land Matrix investor can be retrieved from the Orbis data, as well as their location, thus allowing

[1] https://landmatrix.org/

[2] https://www.moodys.com/web/en/us/capabilities/company-reference-data/orbis.html

Node Source	Node Target	Weight
Cyprus	Russian Federation	64
Cyprus	Ukraine	63
Ukraine	Cyprus	46
United Kingdom	Indonesia	40
Malaysia	Indonesia	34
Germany	Russian Federation	33
Italy	Romania	29
Belgium	Indonesia	24
Panama	Peru	24
United Kingdom	Mozambique	20
Netherlands	Russian Federation	19
Luxembourg	Ukraine	18

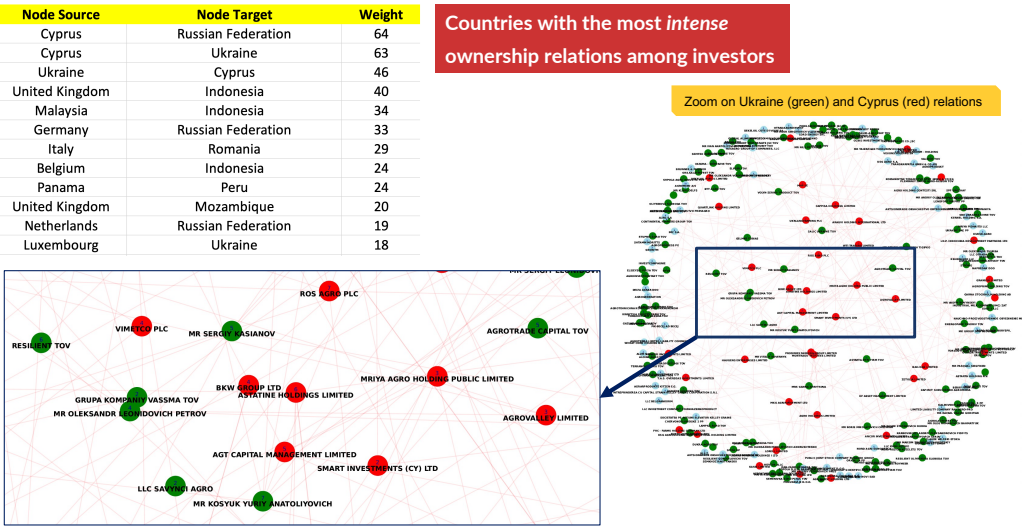


Figure 1: Countries having the most intense ownership relationship among investors, with a focus on the relation between Cyprus and Ukraine.

for a more detailed and accurate modeling of country-to-country relations, and even to the identificaton of more fine-grained relations. Moreover, Orbis provides detailed information about other actors related to a company, such as minor shareholders and subsidiaries. Data contained in the Orbis database includes detailed information about companies, including, among others detailed information about ownership (i.e., details about the corporate group, who-owns-whom information, global ultimate owner, hierarchical ownership graph) and detailed information about stakeholders, with percentages of investment. Through an iterative process, we succeeded in matching investing and operating companies, and their shareholdings, with grounded observations on land investments. Natural Language Processing techniques were used to match Land Matrix and Orbis investing company names based on their names and locations. Then the automatically retrieved matches were carefully verified by a team of domain experts, in order to avoid false positives. As a result, we successfully matched 83% of the Land Matrix deals, corresponding to a network of $8,514$ unique investors. Figure 1 shows couples of countries having the most intense ownership relationship among investors (e.g., ownership or shareholding relations), with a focus on the relation between Ukraine and Cyprus, two countries that show a notable *bidirectional* intense ownership relation (i.e., up to 63 entities located in Cyprus controlling Ukrainian companies, and up to 46 entities located in Ukraine controlling companies in Cyprus). Finally, Figure 2 reports some preliminary insights on the obtained ownership networks, focusing on *in operation* (i.e., currently active) land acquisition deals.

References

[1] R. Interdonato, J. Bourgoin, Q. Grislain, and A. Tagarelli. The new abnormal: Identifying and ranking anomalies in the land trade market. *PLOS ONE*, 17(12):1–18, 12 2022.

[2] R. Interdonato, J. Bourgoin, Q. Grislain, M. Zignani, S. Gaito, and M. Giger. The parable of arable land: Characterizing large scale land acquisitions through network analysis. *PLOS ONE*, 15(10):1–31, 10 2020.

[3] Intergovernmental Panel on Climate Change (IPCC). Summary for policymakers. in: Climate change 2023: Synthesis report. ipcc, geneva, switzerland, pp. 1-34., 2023.

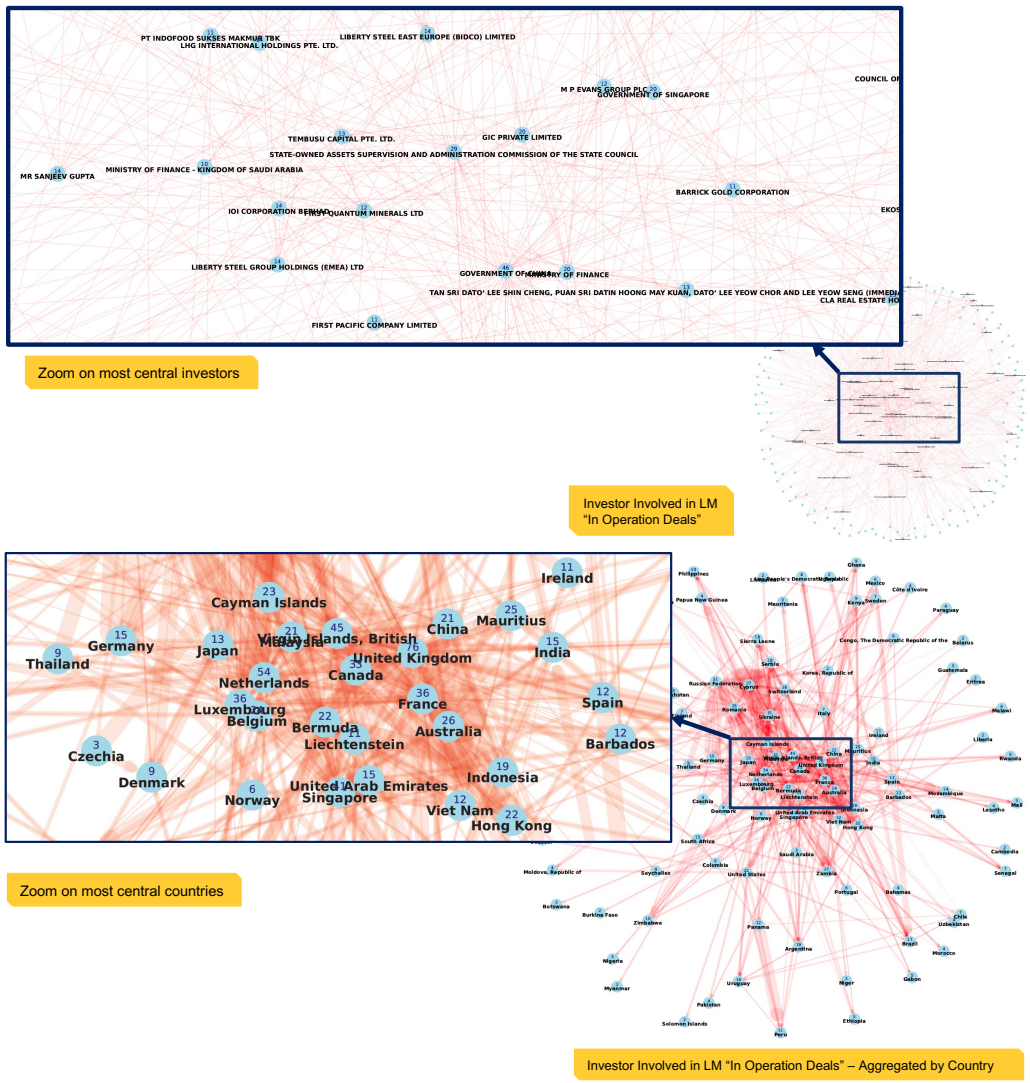


Figure 2: The obtained ownership networks, focusing on *in operation* (i.e., currently active) land acquisition deals, with a zoom on most central nodes.

Complex Networks: Communities

FRCCS

An Evaluation of Graph Based Approaches for Clustering: a Case Study in Chronic Pain Categories

Iris Ho[1✓], Paul Anderson[1], Jean Davidson[2], Jeffrey Lotz[3] and Theresa Migler[1]

[1] *Department of Computer Science and Software Engineering, California Polytechnic State University San Luis Obispo, CA, USA ; iwho@calpoly.edu, pander14@calpoly.edu, tmigler@calpoly.edu.*
[2] *Department of Biological Sciences, California Polytechnic State University San Luis Obispo, CA, USA ; jdavid06@calpoly.edu.*
[3] *Department of Orthopaedic Surgery, School of Medicine, University of California San Francisco, CA, USA ; jeffrey.lotz@ucsf.edu.*
[✓] *Presenting author*

Abstract. Chronic pain treatment varies with each person's experience. Leveraging node embedding algorithms and k-means clustering, we explore the potential of graph-based clustering methods to group patients based on pain characteristics, pain, and beliefs. We aim to suggest potential patient subgroups, offering valuable insights into tailored treatment strategies. Our findings highlight the promise of graph-based approaches in informing tailored pain management practices.

Keywords. *Patient Clustering; Patient Similarity; Chronic Pain; Node Embeddings*

1 Introduction

Chronic pain, characterized by its persistence for longer than three months, can have a number of causes ranging from severe injuries to genetic or psychological factors, along with several treatments ranging from physical therapy or diet change to pharmacological treatments. A 2019 study conducted by the National Center for Health Statistics revealed that 20.4% of adults experienced chronic pain, with 7.4% enduring high-impact chronic pain which severely limits daily life and work activities in the past 3 months [13].

While chronic pain is a deeply personal issue that impacts everyone differently, it is helpful to take a step back and notice larger trends relating to chronic pain patients. Namely, it would be beneficial to be able to group, or *cluster*, chronic pain patients into various groupings based on their treatment needs. For example, a group of chronic pain patients will benefit significantly from group therapy sessions while another group of chronic pain patients benefit more from physical therapy. While several methodologies have been proposed for patient grouping in recent years, using graphs to produce or evaluate such groupings remains largely unexplored. Recently, the use of knowledge graphs have been integrated into the study of patient similarity in various biological domains [8, 14]. Moreover, there is growing interest in exploring patient similarity [5]. Graphs can provide a transparent and easily interpretable framework, yielding insightful observations regarding proposed patient classifications.

In this study, we propose leveraging graph theory to address patient clustering. By constructing a graph of patients and utilizing a clustering algorithm, we aim to find commonalities among chronic pain patients that when further refined may aid in creating effective treatment strategies.

2 Background

Clustering in a graph context is the idea of partitioning the nodes in a graph into groups based on their properties and connections. For example, we might have a graph where the nodes are patients and two patients are connected if they experience the same symptoms. A clustering algorithm might cluster these patients based on the severity of their symptoms, but an equally valid grouping would be by their response to treatment.

There are various methods to approach clustering nodes in a graph, here we discuss k-means clustering.

2.1 k-means clustering

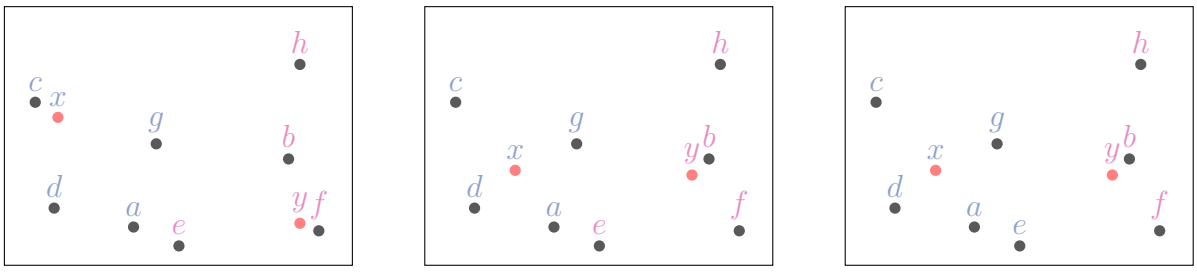
The k-means clustering algorithm, also referred to as the Lloyd–Forgy algorithm, aims to divide the data set into k clusters, where k is a parameter for the algorithm.

Take n points in a m-dimensional space. The algorithm will start with k initial (and perhaps not optimal) clusters of these points. The center of a cluster is defined by the mean of all its member's points. On each iteration:

- Every point will be relocated into its nearest center's cluster
- The cluster will then update its center by recalculating the mean of all its members points

This process of relocation and recalculation of the mean will repeat until either (a) little or no relocation occurs or (b) a predefined number of iterations has occurred [9, 10, 4].

For example, in Figure 1, there are points a through h in a two-dimensional space. Let $k = 2$ and let x and y be the randomly placed centers (shown with a black dot). For every point z, if z is closer to x, then z will join x's cluster, otherwise z must be closer to y and will therefore join y's cluster. Points in x's cluster will have blue text and points in y's cluster will have red text. Now, the center will be recalculated based on all the points in the cluster, shown in Figure 1b. Now, each point will recalculate its distance to the new centers and join the associated cluster, shown in Figure 1c. Notice how e which was originally in y's cluster, is now closer to x and therefore joins x's cluster.



(a) Let x and y be the initial centers. All other points are part of x's cluster or y's cluster.

(b) Each cluster re-calculates their own center.

(c) Each point calculates its distance to the new centers and joins the cluster associated with that center.

Figure 1: An example of k-means clustering with $k = 2$ on 2-dimensional data points.

Note that depending on where the initial centers are placed, the clusters can vary slightly, but overall should converge to roughly the same clusters.

One way to measure the validity of the clustering is through the average *silhouette score* of every point, which is a measure of a how well clustered a node is and is calculated for a single node as follows

$$s = \frac{b - a}{\max(a, b)} \tag{1}$$

where a is the mean intra-cluster distance (or the mean distance between this node and all the other nodes in the same cluster) and b is the mean nearest cluster distance (or the average distance between this node and the nearest cluster that the node is not a part of). The silhouette score has a range of -1 to 1, where a higher value is associated with the node being closer to values in their cluster as opposed to neighboring clusters. Simply put, a silhouette score of -1 for a node implies that the node is likely in the wrong cluster and a score of 1 means that the node is in the best cluster it could possibly be in.

Note that the input of the k-means algorithm is a specific value for k and a set of vectors (importantly not a graph). Users of the k-means algorithm often try a range of values for k and pick the best k value based on the silhouette score and/or the clusters that make the most sense based on the domain of objects that are being clustered. Since k-means doesn't run on graphs, and instead takes in vectors, each node in the graph will be represented as a vector through the use of node embedding algorithms.

2.2 Node Embeddings

Algorithms that transform graphs into vectors are called node embedding algorithms.

Node embedding algorithms aim to represent nodes and their relationships (edges) in the form of a vector. There are various algorithms that attempt to do so. We will discuss Node2Vec and Fast Random Projection (FastRP).

Node2Vec Presented in 2016, Node2Vec is a common and flexible node embedding algorithm that utilizes random walks to generate a vector for every node [6].

Fast Random Projection Later in 2019, Chen et al. presented Fast Random Projection (FastRP) [3]. FastRP is 4,000 times faster than Node2Vec by explicitly constructing a node similarity matrix and by utilizing sparse random projection. Random projection is a dimension reduction method that preserves pairwise distances between data points.

Both Node2Vec and FastRP allow the user to determine the number of dimensions to represent each node.

3 Related Works

While to our knowledge this paper is the first to explore utilizing graphs to cluster chronic pain patients, it is not the first to explore clustering chronic pain patients in general. Tagliaferri et al. used data from the UKBioBank and found five groups of chronic back pain patients using machine learning mainly characterized by a varying combination of social isolation and depressive symptoms [12]. Another study by Larsson et al. identified four groups of chronic pain among Swedish older adults distinguished by varing degrees of pain and psychological symptoms [7]. Larsson et al. utilized two-step cluster analysis (TSCA), which consisted of pre-clustering and then hierarchical methods on the basis of best fit.

Lastly a 2018 study by Bäckryd et al. using psychometric data from the Swedish quality registry for pain rehabilitation (SQRP), identified that chronic pain patients belong to one of four groups [2]: (a) low "psychological strain", (b) high "psychological strain", (c) high "social distress" and high "psychological strain", low "social distress", and high pain intensity. Individuals in group (a) low "psychological strain" was comparatively in the best situation with respect to pain characteristics such as intensity and spreading. They had the lowest frequency of fibromyalgia (chronic widespread pain), and were slightly older in age. This contrasts individuals in group (b) high "psychological strain" where patients were comparatively in the most negative situation with respect to pain characteristics. Group (c) high "social distress" had the longest pain duration and statistically consisted more females. The frequency of three neuropathic pain conditions was also generally lower in this group. Bäckryd et al. used statistical methods, specifically principal component analysis (PCA) and hierarchical clustering to reveal the four groups described above.

We will compare the results of our graphical clustering method to the established statistical method of clustering by Bäckryd et al.

4 Methods

4.1 Data

The data utilized in this paper is based on The Longitudinal Clinical Cohort for Comprehensive Deep Phenotyping of Chronic Low-Back Pain (cLBP) Adults Study (comeBACK). This longitudinal multicenter observational study was designed to perform comprehensive deep phenotyping in patients with cLBP and is being conducted at 4 clinical sites in the United States (U.S.) with a coordinating center at UCSF. The comeBACK clinical sites are located at four of the University of California campuses, including UC San Francisco (UCSF), UC Davis, UC Irvine, and UC San Diego. Recruitment for the study commenced in March 2021 and was completed in June 2023 with a total of 450 participants enrolled and to be followed for up to 2 years. Participants attend in-clinic baseline and annual visits (on month 12 and 24). Remote (via online surveys with a link sent by email and/or phone) visits occur at Months 1, 2, 3, 4, 5, 6, 18, and at months 7-8, if necessary.

Table 1: Data Variables

Data	Meaning
Age	Age
Sex	Sex
Gender	Gender
Ethnicity	Ethnicity
Race	Race (may identify as more than one)
Household Size	Household Size
Household Income	Household Income
Education Level	Education Level
PEG-3	Pain, Enjoyment of Life and General Activity scale
PROMIS	PROMIS Pain Interference
Duration	Length of low back pain (LBP)
Frequency	LBP frequency
Unemployment	LBP-related unemployment
Intensity	LBP intensity

Data	Meaning
QoL	PROMIS Physical Functioning Score
Depression	PROMIS Depression Score
PAS	Pain Anxiety Symptom Scale Short Form 20
Not Distracting	MAIA-SF – Not-Distracting Sub-Score
Emotion Aware	MAIA-SF – Emotional Awareness Sub-Score
Self Regulation	MAIA-SF – Self Regulation Sub-Score
Interoception	MAIA-SF – Interoceptive Attention Sub-Score
Fear Avoidance	Fear Avoidance Beliefs with Physical Activity Score
CP Acceptance	Chronic Pain Acceptance Score
Pain Catas.	Pain Catastrophizing Scale SF Score
Self Efficacy	Pain Self-Efficacy Score

This paper utilizes a subset of the data collected in comeBACK. Table 1 presents the columns and potential values for each variable within the dataset used for this study.

In general, the data we use encompass the following categories (For detailed insights on the the meanings of acronyms, refer to Appendix A.):

- Demographic - such as age, gender, sex, race, household size, household income, and education level
- Anthropometrics (measurements and proportions of the human body) - such as Body Mass Index (BMI)
- Pain Interference - such as the PEG-3 score and PROMIS score
- Pain Characteristicts - such as duration, frequency, and impact on employment, and intensity
- Quality of Life - based on the PROMIS Physical Functioning (SF 6b) Score
- Mood - such as depression and the Pain Anxiety Symptom (PAS) score
- Thoughts and Beliefs - such as non-distracting (the tendency not to ignore or distract oneself from sensations of pain or discomfort), emotional awareness (one's awareness of the connection between body sensations and emotional states), self-regulation (one's ability to regulate distress by attention to body sensations), Interoceptive Attention, fear avoidance, chronic pain acceptance, Pain Catastrophizing, and Pain Self-Efficacy from the MAIA-SF (Multi-dimensional Assessment of Interoceptive Awareness version 2)questionnaire, FABQ-P (Fear Avoidance Beliefs with Physical Activity) Score, CPAQ-8 (Chronic Pain Acceptance) Score, PCS-6 (Pain Catastrophizing Scale SF) Score, PSEQ-4 (Pain Self-Efficacy) Score.

Since the study is still an ongoing processes, comeBACK is not yet published. Thus, we used CTGAN (Conditional Tabular Generative Adversarial Network) to generate synthetic data that mimics the characteristics of the original dataset. Specifically, data for 1000 patients was synthesized.

4.2 Graph Construction

We constructed our graph as follows: nodes represent patients and two patients are connected by an edge if they share a similar answer/score to a pain-related metric. Allowing "similar" numerical scores in a category to share an edge accounts for varying interpretations of potentially the same levels of pain. For example, a pain intensity score of 4 is likely similar to a pain intensity of 5 (on a scale of 1-10). Since each metric's range of values varies, "similar" means different things for each metric. Table 2 specifies what 'similar' means for each metric in the column 'Similar If'.

For example, if patient a had a pain duration value of 1 and patient b had a pain duration value of 2, they would not share an edge since their similarity is within ± 1 and not ± 0. In other words a ± 0 value means an edge will only be added between two patients if they have the exact same value. However, if patient a had an intensity value of 3 and patient b had an intensity value of 4, then they would be connected by an intensity edge since they are within 1 numerical value from each other.

Table 2: Edge Types

Edge Type	Possible Values	Similar If...
PEG-3	0-10	±1
PROMIS	4-20	±4
Duration	1-5	±0
Frequency	1-3	±0
Unemployment	0-2	±0
Intensity	0-10	±1
QoL	6-30	±4
Depression	4-20	±4
PAS	0-100	±10

Edge Type	Possible Values	Similar If...
Not Distracting	0-5	±0
Emotion Aware	0-5	±0
Self Regulation	0-5	±0
Interoception	0-5	±0
Fear Avoidance	0-24	±4
CP Acceptance	0-48	±6
Pain Catas.	0-12	±2
Self Efficacy	0-24	±4

Notably, edge types are only constructed based on pain-related metrics and not demographic data. We decided that while demographic data is important and has the potential to influence pain, it should not impact the creation of clusters and even has the potential to skew the graph. For example, if all individuals who identified as female were connected to each other, it would skew the connectedness of the graph, and our clusters might only reflect gender.

We implemented our undirected, unweighted graph in Neo4j, a graphical database that provides an implementation of Node2Vec and FastRP [1]. Our graph had 1000 nodes and 495,650 edges. Given the edge construction, it should then not be surprising that each patient has some relation with another patient, causing the graph to be one large connected component. The degree distribution of the graph is shown in Figure 2 and Figure 3 shows the number of edges for each edge type.

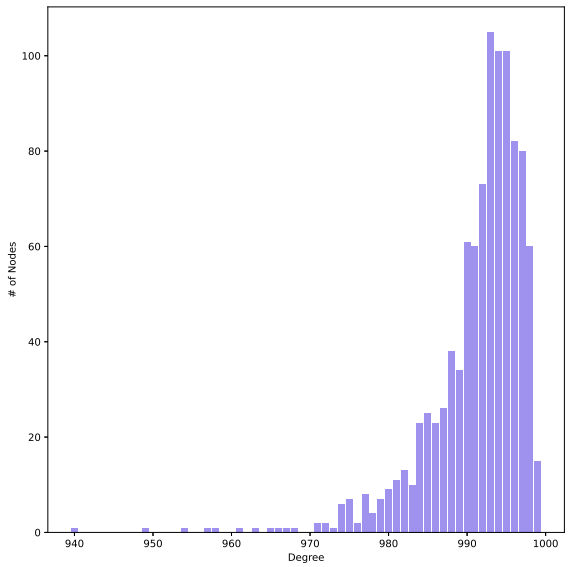


Figure 2: Degree Distribution

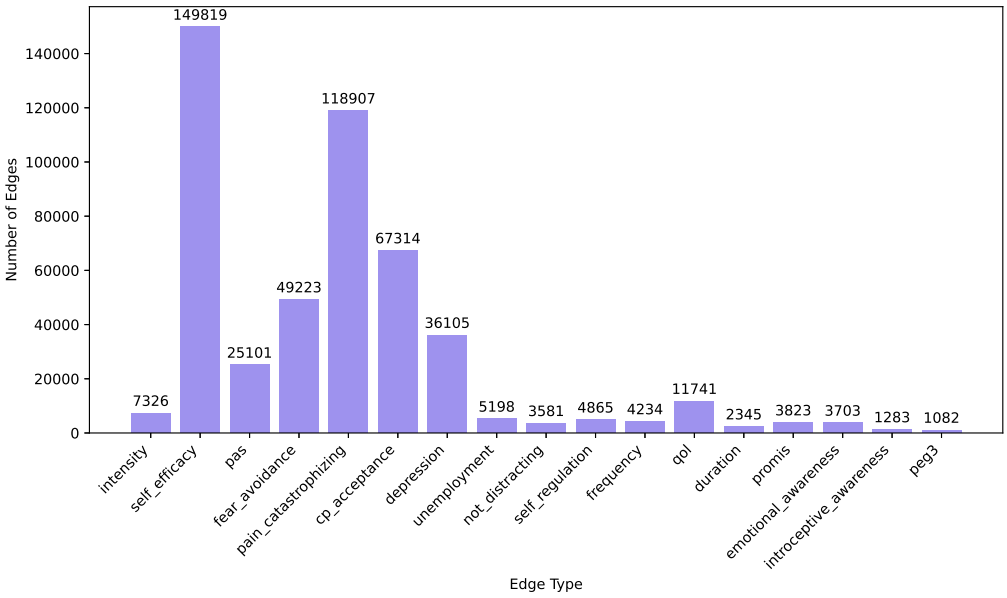


Figure 3: Number of Edges for Each Edge Type

Once the graph was constructed, we ran k-means clustering on vectors produced by Node2Vec and FastRP.

5 Results

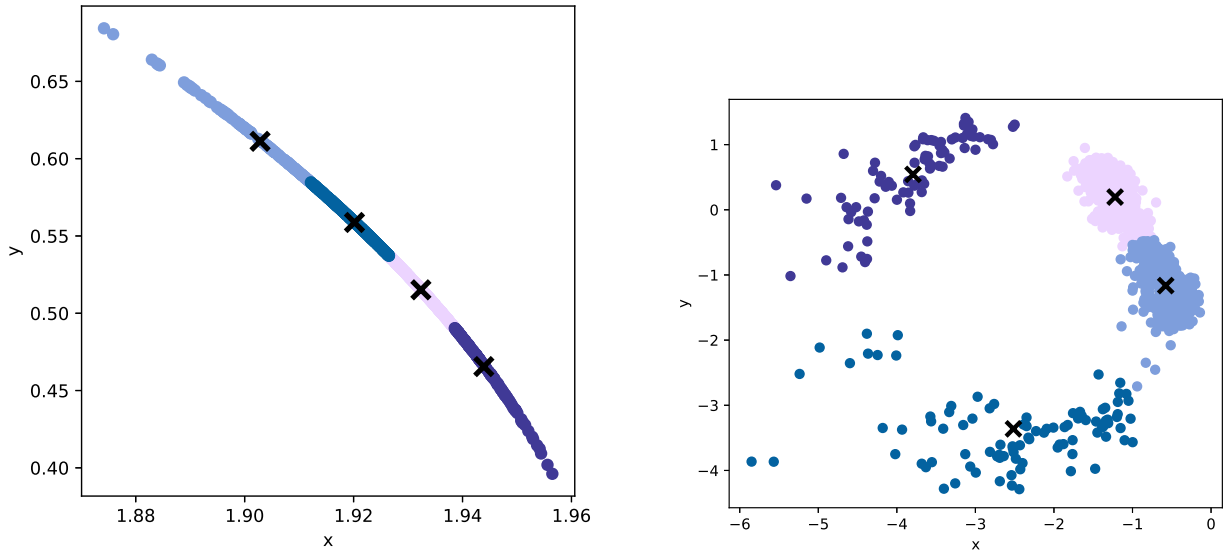
5.1 k-means Clusters

We chose to use skikit-learn's implementation of k-means with values $k = 2, 3, \ldots, 9, 10$ [11]. Then for each of the possible k values, we ran the k-means algorithm 20 times: 10 times with each of the two node-embedding algorithms, Node2Vec and FastRP, (both of which are configured to produce 2-dimensional embeddings) yielding similar silhouette scores. Recall that the silhouette score is a measure of how cohesive the clusters are. The highest average silhouette score from k-means ranged from $0.62 - 0.67$ for Node2Vec and $0.53 - 0.54$ for FastRP. This average silhouette score means the nodes were generally well-positioned within their clusters regardless of the node embedding algorithm chosen. The optimal k value (number of clusters) associated with the highest average silhouette scores was either 3 or 4 for Node2Vec and consistently 4 for FastRP. Table 3 shows the results of running k-means on both node embeddings once for each value of k. Since Node2Vec, FastRP, and k-means all utilize pseudo-randomness (for example k-means randomly selects where the initial centers are), the results discussed below were produced with a seed of 0, which is a number used to initialize the pseudorandom number generator. This results in multiple runs of the algorithm on the same input to produce consistent results.

Table 3: Average Silhouette Scores by Node Embedding algorithm using a seed of 0

k	Node2Vec	FastRP	k	Node2Vec	FastRP	k	Node2Vec	FastRP
2	0.62722	0.55798	5	0.63462	0.52088	8	0.48065	0.51326
3	0.54622	0.51924	6	0.62519	0.53089	9	0.48246	0.52125
4	0.63784	0.53650	7	0.52102	0.52044	10	0.48412	0.53727

We've visualized the node embeddings produced by FastRP and Node2Vec in Figure 4. The colors of each point (representing a patient) represent the cluster that they are in as a result of running k-means with $k = 4$. Of the four clusters that emerged as a result of the FastRP embeddings, there were two larger clusters (C_1 of size 371 and C_3 of size 331) along with two smaller groups (C_2 of size 137 and C_4 of size 161). Similarly, Node2Vec had two larger groups (C_1 of size 430 and C_2 395) and two smaller groups (C_3 of size 93 and C_4 of size 84).



(a) Fast RP

(b) Node2Vec

Figure 4: Node Embeddings produced by (a) Fast RP and (b) Node2Vec

Aside from the silhouette score, we can utilize statistical measures such as a one-way ANOVA test, Mann Whitney U Test, and Pearson's chi-squared test to evaluate and compare the clusters produced with FastRP embeddings and Node2Vec embeddings. For all statistical tests used, a p-value of $< 0.05(5\%)$ was considered significant. The lower the p-value, the more likely that the difference measured was not pure chance and is actually a meaningful difference.

Clusters created from Node2Vec proved to be more homogenous and showed fewer statistical differences from each other. However, clusters created as a result of FastRP embeddings proved to be more statistically different from each other. Thus, cluster analysis was done on clusters created by the FastRP node embedding algorithm. The results of which are shown in Table 4. Some variables from the data were not statistically different between any two groups and therefore are omitted from the table. It is not surprising that there are statistical differences between the pain characteristics of each cluster because these variables are reflected in the edges of the graph. However, demographic variables such as sex and race suggest some potential relationships between demographics and specific clusters. For example, between cluster 1 (C_1) and cluster 3 (C_3), there were statistical differences between the sex of the individuals. Pairing this with Figure 5, we can see that C_1 consists of more females than C_3. We've chosen to represent these variables per cluster by percent since clusters aren't of the same size, but the size of the cluster as well as the percent of individuals with a particular value per variable could affect the statistical significance. (See Appendix B for more figures showing significant variables across the four clusters.)

Table 4: p-values for variables between k-means clusters (* means that the p-value was < 0.001 whereas an X indicates a p-value > 0.05) computed by Mann Whitney U Test, except for Sex, Black or African American, White, Duration, Frequency, and Unemployment (Chi-Square)

Variable	C_1 vs C_2	C_1 vs C_3	C_1 vs C_4	C_2 vs C_3	C_2 vs C_4	C_3 vs C_4
Sex	X	0.0126	X	X	X	X
Black or African American	X	X	0.0373	X	X	0.0356
White	0.0116	X	X	0.0292	X	X
PROMIS	*	*	*	*	*	*
Duration	*	0.0012	*	0.0030	*	*
Frequency	*	X	*	0.0024	*	*
Unemployment	*	*	0.0159	0.0111	X	X
Intensity	*	*	0.0049	X	*	*
QoL	X	X	0.0015	X	0.0391	*
Depression	*	*	0.0279	*	*	*
PAS	*	0.0018	X	0.0013	*	*
Not Distracting	X	0.0408	X	X	X	0.0086
Fear Avoidance	0.0350	0.0117	X	X	X	0.0349
CP Acceptance	*	X	X	0.0041	0.0060	X
Pain Catas.	X	X	X	X	0.0151	0.0199

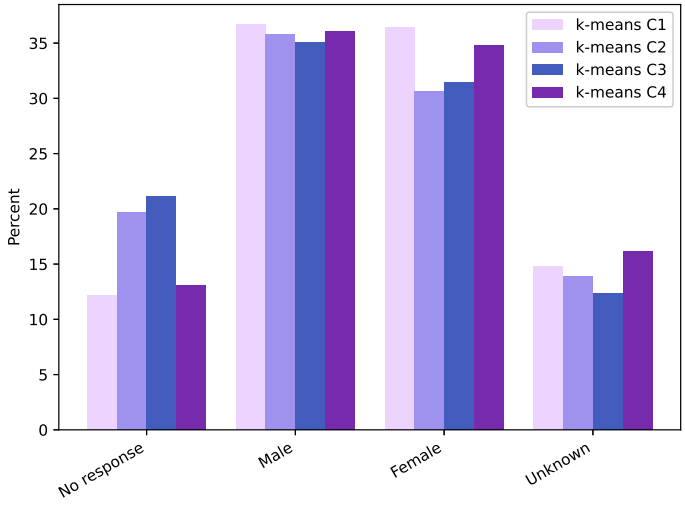


Figure 5: Sex of k-means clusters by percent

5.2 Comparison with HCA clusters

Since there is no "true" clustering of chronic pain patients, we will evaluate our method and the resulting clusters by utilizing the method's described in the work of Bäckryd ed al. (which we will refer to as "hierarchical clustering" or HCA). Utilizing these methods on our data resulted in produced 4 clusters. The dendrogram from hierarchical clustering is shown in Figure 6. The same statistical measures were performed on their clusters and the results are shown in Table 5. Similar to Table 4, if variables from the data were not statistically different between any two groups, the variable is omitted from the table.

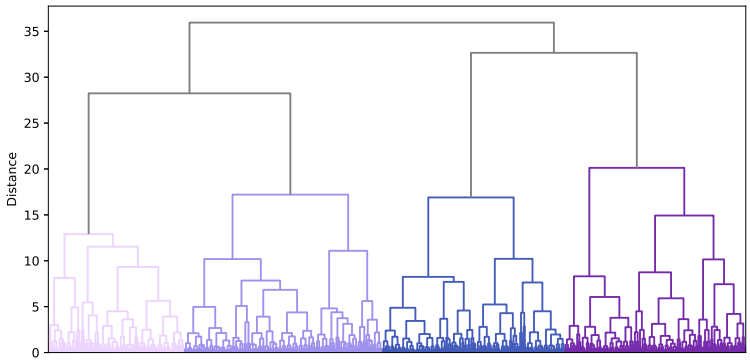


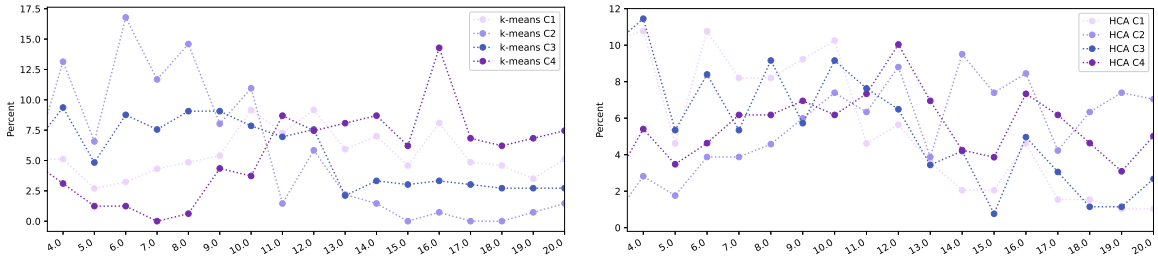
Figure 6: Dendrogram showing four clusters from Hierarchical Clustering. The y-axis shows similarity/dissimilarity and the x-axis shows patients.

Table 5: p-values for variables between HCA clusters (* means that the p-value was < 0.001 whereas an X indicates a p-value > 0.05) computed by Mann Whitney U Test, except for Gender, American Indian or Alaska Native, Native Hawaiian or Pacific Islander, Race unknown or not reported, Frequency, and Unemployment (Chi-Square)

Variable	C_1 vs C_2	C_1 vs C_3	C_1 vs C_4	C_2 vs C_3	C_2 vs C_4	C_3 vs C_4
Gender	X	X	X	X	X	0.0169
American Indian or Alaska Native	X	X	X	0.0478	X	X
Native Hawaiian or Pacific Islander	X	X	0.0146	X	X	X
Race unknown or not reported	X	X	X	X	X	0.0088
Household Size	X	X	X	X	0.0454	X
PEG-3	*	X	X	0.0099	0.0334	X
PROMIS	*	X	*	*	*	*
Frequency	X	X	X	*	X	0.0047
Unemployment	X	0.0199	X	0.0148	X	0.0249
Intensity	*	0.0106	X	*	*	0.0041
QoL	*	*	*	*	*	X
Depression	0.0033	X	*	*	*	*
PAS	X	X	*	X	*	*
Not Distracting	*	*	*	X	X	0.0267
Emotion Aware	*	0.0239	X	0.0319	0.0011	X
Self Regulation	*	*	*	X	*	*
Interoception	*	*	*	*	0.0426	*
Fear Avoidance	X	0.0222	*	*	*	*
CP Acceptance	0.0307	X	X	*	0.0018	X
Pain Catas.	0.0251	*	*	*	*	X
Self Efficacy	X	*	*	*	*	*

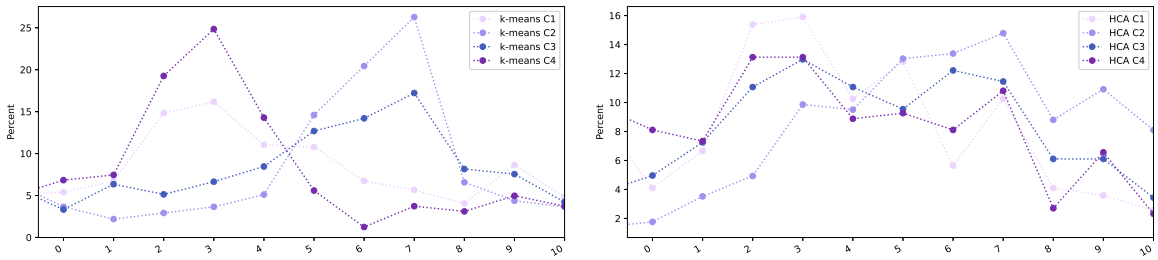
For the variables with statistical significance, we've provided a line graph comparing the percent of individuals with each variable value for each cluster. (See Appendix B for all figures.) Notably PROMIS was the only variable that had statistical differences across every pair of clusters produced by k-means. Figure 7 visualizes the PROMIS scores across the four k-means and

HCA clusters compared to each other. For k-means, we can see that cluster 2 (k-means C_2) and cluster 3 (k-means C_3) generally have lower PROMIS scores, cluster 1 (k-means C_1) is generally evenly spread across PROMIS scores and cluster 4 (k-means C_4) contains more individuals with higher PROMIS scores. For HCA, we can see that cluster 1 (HCA C_1) and cluster 3 (HCA C_3) had generally lower PROMIS scores whereas cluster 2 (HCA C_2) has generally higher PROMIS scores and cluster 4 (HCA C_4) consists mostly of mid-ranged PROMIS scores.



(a) k-means (b) HCA

Figure 7: PROMIS across Four Clusters



(a) k-means (b) HCA

Figure 8: Intensity across Four Clusters

We also want to highlight how for some variables, there were dramatic differences between some groups but not between all possible pairs of groups. For example, Figure 8 shows pain intensity across the four clusters. In the clusters produced by k-means, cluster 4 (k-means C_4) contained individuals with lower intensity values whereas cluster 2 (k-means C_2) and cluster 3 (k-means C_3) contained individuals with higher pain intensity values. Thus cluster 2 (k-means C_2) and cluster 4 (k-means C_4) are very different, but cluster 2 (k-means C_2) and cluster 3 (k-means C_3) are considered similar with respect to pain intensity.

In general, clusters produced with Bäckryd et al.'s methods were more statistically different from each other. However, while more work can be done to refine our methods, these initial results show that graphs can produce meaningful patient clusters.

5.3 Other clustering methods

Given the same graph construction, we experimented with a few other clustering algorithms: Label Propagation, DBSCAN (`eps`=.5 and `min_samples`=5, and Louvain. Both Label Propagation and DBSCAN resulted in one cluster. Contrasting this, Louvain revealed 3 clusters, of size 359, 353, and 288. These three clusters showed to have less variables with statistical differences. Due to how k-means preformed better, we chose to not conduct a full analysis of the Louvain communities.

6 Conclusions and Future Directions

In this paper, we've explored clustering chronic pain patients using graphs due to the transparent and intuitive nature of graphs. Thus we proposed using well known algorithms such as k-means clustering and Louvain community detection to identify clusters in a graph. The conclusions drawn from the resulting clusters on the synthetic data is limited but once the real data is publishable, we hope to use these methods on the real data to find meaningful clusters.

In our brief comparison to an established method, it would have been ideal to test our method on the dataset used by Bäckryd et al., however we did not have access to that dataset. We recognize that the scope of this comparison is constrained by the limited access of data.

Graphs can provide a novel approach to patient clustering. While this paper proposes a new method, more work can be done to further improve the results. Namely, we can consider other node embedding and clustering methods. Further, we could consider utilizing a one-hot encoder for categorical variables with no gradation and then using spectral clustering. We can also consider other methods of graph construction and incorporating more data into the graph.

7 Acknowledgements

We extend our gratitude to members of the Cal Poly Bioinformatics Research Group, especially Sam Kaplan and Madeline Bittner, for their invaluable support throughout this study. Additionally, we would like to thank Felicia Korengel and Jeffrey Sklar for generously sharing their time and expertise, and aiding in the comprehension of statistical concepts. This research was supported by "UCSF Core Center for Patient-centric Mechanistic Phenotyping in Chronic Low Back Pain (UCSF REACH)" funded by NIH. Support was provided for the Computational Molecular Sciences Center by the Frost Fund at the Cal Poly Bailey College of Science and Math.

Appendices

A Data

Below, we will provide a more detailed explanation of the variables used in this study.

Age This is the patient's age rounded down to the nearest year.

Sex Patients were asked "What was your sex at birth?" and were given the following options:
1: Male
2: Female
3: Unknown
4: Intersex

Gender Patients were asked "What was your gender identity?" and were given the following options:

1: Male
2: Female
3: Transgender man
4: Transgender woman
5: Genderqueer or gender nonconforming
6: Unknown
7: Other

Ethnicity Patients were asked "What was your ethnicity?" and were given the following options:

1: Hispanic or Latino
2: Not Hispanic or Latino
3: Unknown
4: Prefer not to answer

Race Patients were asked "What is your race?" and could select any of the following options that applied:

- American Indian or Alaska Native
- Asian
- Black or African American
- Native Hawaiian or Pacific Islander
- White
- More than one race
- Unknown or not reported

Household Size Patients were asked "Including yourself, how many people live in your household?"

Household Income Patients were asked "What is your annual household income from all sources?" and given the following options:

1: Less than $10,000
2: $10,000 to $24,999
3: $25,000 to $34,999
4: $35,000 to $49,999
5: $50,000 to $74,999
6: $75,000 to $99,999
7: $100,000 to $149,999
8: $150,000 to $199,999
9: $200,000 or more
10: Prefer not to answer

Education Level Patients were asked "What is the highest level of education you have completed?" and given the following options:

1: Did not complete Secondary School (or less than High School)
2: Some Secondary School (or High School) education
3: High School (or Secondary School) Degree complete
4: Associate's or Technical Degree complete
5: College or Baccalaureate Degree complete
6: Doctoral or Postgraduate education

PEG-3 The PEG-3 Score is calculated from the mean of the following three questions each range 0-10:

- What number best describes your pain on average, in the past week?
- What number best describes how, during the past week, pain has interfered with your enjoyment of life?
- What number best describes how, during the past week, pain has interfered with your general activity?

... where 0 referred to no pain and 10 is the worst pain you can imagine.

PROMIS The PROMIS Pain Interference Score is calculated as the sum of the following 4 values each ranged 1-5: In the past 7 days, how much did pain interfere with your...

- day to day activities?
- work around the home?
- ability to participate in social activities?
- household chores?

... where 1 corresponds to 'not at all', 2 corresponds to 'a little bit', 3 corresponds to 'somewhat', 4 corresponds to 'quite a bit' and 5 corresponds to 'very much'.

Duration This describes the length of low back pain (LBP), where:

1: $<$ 3 months (ineligible for study)
2: 3 - 6 months
3: 6.1 months - 1 year
4: 1.1 - 5 years
5: $>$ 5 years

Frequency This refers to the frequency at which patients experience chronic back pain. Patients were asked to answer the following question: "How often has low-back pain been an ongoing problem for you over the past 6 months?" Their answers were on a scale of 1-3 where:

1: Every day or nearly every day in the past 6 months
2: At least half of the days in the past 6 months
3: Less than half of the days in the past 6 months

Unemployment This asks patients to answer the following question related to LBP related unemployment: "Have you been off work or unemployed for 1 month or more due to your low back pain?" Their answers are on a scale of 0-1 where:

1: yes
0: no
2: Does not apply

Intensity This refers to the intensity of patient's lower back pain in the past 7 days on a scale of 0-10.

QoL (Quality of Life) This is the PROMIS Physical Functioning (SF 6b) Score which ranges from 6-30, where a higher score means better function. It is calculated as the sum of the answer to following six questions, each on a scale of 1-5:

1. Are you able to do chores such as vacuuming or yard work?
2. Are you able to go up and down stairs at a normal pace?
3. Are you able to go for a walk of at least 15 minutes?
4. Are you able to run errands and shop?
5. Does your health now limit you in doing two hours of physical labor?
6. Does your health now limit you in doing moderate work around the house like vacuuming, sweeping floors, or carrying in groceries?

where for questions 1-4, the scores have the following meaning:

5: Without any difficulty
4: With a little difficulty
3: With some difficulty
2: With much difficulty
1: Unable to do

and for questions 5-6, the scores have the following meaning:

5: Not at all
4: Very little
3: Somewhat
2: Quite a bit
1: Cannot do

Depression The PROMIS Depression score is calculated as the sum of a patient's answers to the following 4 questions, where each question has a range of 1-5, for a overall range of 4-20.

In the past 7 days...

- I felt worthless
- I felt helpless
- I felt depressed

- I felt hopeless

where:

1: Never
2: Rarely
3: Sometimes
4: Often
5: Always

PAS This is the PASS-20, or the Pain Anxiety Symptom Scale Short Form 20, is calculated based on the Avoidance sub-score, Physiological anxiety sub-score. Each sub-score is calculated by taking 5 times the mean of the responses to questions.

The avoidance sub-score is composed of the following 5 questions, each on a scale of 0-5:

1. I will stop any activity as soon as I sense pain coming on
2. Pain seems to cause my heart to pound or race
3. As soon as pain comes on, I take medication to reduce it
4. When I feel pain I think that I might be seriously ill
5. During painful episodes it is difficult for me to think of anything besides the pain

The avoidance sub-score is then $5 * \text{mean}$ of the responses to questions 1-5.

The Physiological anxiety sub-score is composed of the following 5 questions, each on a scale of 0-5:

6. When pain comes on strong, I think that I might become paralyzed or more disabled
7. I find it hard to concentrate when I hurt
8. I find it difficult to calm my body down after periods of pain
9. I worry when I am in pain
10. I try to avoid activities that cause pain

The physiological anxiety sub-score is then $5 * \text{mean}$ of the responses to questions 6-10.

For all questions above, 0 is associated with Never and 5 is Always.

The overall PASS score is then calculated by taking the sum of both sub-scale scores and multiplying it by 2.

Not Distracting This is the MAIA-SF (Multi-dimensional Assessment of Interoceptive Awareness, v2) – Not- Distracting Sub-Score. Participants are asked to answer the following question:

Please indicate how often each statement applies to you generally in daily life.

- I ignore physical tension or discomfort until they become more severe.
- I distract myself from sensations of discomfort.
- When I feel pain or discomfort, I try to power through it.
- I try to ignore pain.
- I push feelings of discomfort away by focusing on something.
- When I feel unpleasant body sensations, I occupy myself with something else so I don't have to feel them.

where responses are on a scale from 0-5 where 0 means never and 5 means always.

The non-distracting sub-score is then calculated as the mean of 5 minus each of their responses, thus a higher score means less distracting.

Emotion Aware This is the MAIA-SF Emotional Awareness Sub-Scale score which is the mean of their answers to the following questions:

- I notice how my body changes when I am angry.
- When something is wrong in my life I can feel it in my body.
- I notice that my body feels different after a peaceful experience.
- I notice that my breathing becomes free and easy when I feel comfortable.
- I notice how my body changes when I feel happy / joyful.

where responses are on a scale from 0-5 where 0 means never and 5 means always, thus a higher score means more awareness.

Self Regulation This is the MAIA-SF Self-Regulation Sub-Scale which is the mean of their answers to the following questions:

- When I feel overwhelmed I can find a calm place inside.
- When I bring awareness to my body I feel a sense of calm.
- I can use my breath to reduce tension.
- When I am caught up in thoughts, I can calm my mind by focusing on my body/breathing.

where responses are on a scale from 0-5 where 0 means never and 5 means always, thus a higher score means more self-regulation.

Interoception This is the MAIA-SF Interoceptive Attention Sub-Scale which is the mean of their answers to the following questions:

- I can pay attention to my breath without being distracted by things happening around me
- I can maintain awareness of my inner bodily sensations even when there is a lot going on around me
- When I am in conversation with someone, I can pay attention to my posture.
- I can return awareness to my body if I am distracted
- I can refocus my attention from thinking to sensing my body.
- I can maintain awareness of my whole body even when a part of me is in pain or discomfort.
- I am able to consciously focus on my body as a whole.
- I notice how my body changes when I am angry.
- When something is wrong in my life I can feel it in my body.
- I notice that my body feels different after a peaceful experience.
- I notice that my breathing becomes free and easy when I feel comfortable.
- I notice how my body changes when I feel happy / joyful.
- When I feel overwhelmed I can find a calm place inside.
- When I bring awareness to my body I feel a sense of calm.
- I can use my breath to reduce tension.
- When I am caught up in thoughts, I can calm my mind by focusing on my body/breathing.

where responses are on a scale from 0-5 where 0 means never and 5 means always, thus a higher score means more attention.

Fear Avoidance This is the FABQ-P (Fear Avoidance Beliefs with Physical Activity) score, which is calculated as 4 times the mean of their answers to the following questions

- Physical activity makes my pain worse
- Physical activity might harm my back
- I should not do physical activities which (might) make my pain worse
- I cannot do physical activities which might make my pain worse

where responses are on a scale from 0-6 where 0 means completely disagree and 6 is completely agree, thus a higher score means more fear.

CP Acceptance This is the CPAQ-8 (Chronic Pain Acceptance) score, which is calculated as 2 times the mean of willingness to accept pain and activity engagement despite pain sub-score.

The willingness to accept pain is calculated as 4 times the mean of their answers to the following questions:

- I am getting on with the business of living no matter what my level of pain is
- Keeping my pain level under control takes first priority whenever I am doing something
- Although things have changed, I am living a normal life despite my chronic pain
- Before I can make any serious plans, I have to get some control over my pain

The activity engagement despite pain sub-score is calculated as 4 times the mean of their answers to the following questions:

- I lead a full life even though I have chronic pain
- When my pain increases, I can still take care of my responsibilities
- I avoid putting myself in situations where my pain might increase
- My worries and fears about what pain will do to me are true

For all questions above, responses are on a scale from 0-6 where:

0: Never true
1: Very rarely true
2: Seldom true
3: Sometimes true

Thus the overall range of the Chronic Pain Acceptance score is 0-48 where a higher value means more pain acceptance.

Pain Catas. This is the PCS-6 (Pain Catastrophizing Scale SF) score, which is calculated as 3 times the mean of helplessness sub-score, magnification sub-score, and rumination sub-score.

The helplessness sub-score is calculated as the mean of their answers to the following questions:

- It's awful and I feel that it overwhelms me
- I feel I can't stand it anymore

The magnification sub-score is calculated as the mean of their answers to the following questions:

- I become afraid that the pain will get worse
- I keep thinking about how much it hurts

The rumination sub-score is calculated as the mean of their answers to the following questions:

- I keep thinking about how badly I want the pain to stop
- I wonder whether something serious may happen

For all questions, responses are on a scale of 0-4 where:

0: Not at all

1: To a slight degree
2: To a moderate degree
3: To a great degree
4: All the time

Thus a higher score means more catastrophizing.

Self Efficacy This is the PSEQ-4 (Pain Self-Efficacy) score, which is calculated as 4 times the mean of their answers to the following questions:

- I can cope with my pain in most situations.
- I can still do many of the things I enjoy doing, such as hobbies or leisure activity, despite pain
- I can still accomplish most of my goals in life, despite the pain.
- I can live a normal lifestyle, despite the pain.

where responses were originally on a scale of 1-6 with 1 meaning not at all confident and 6 meaning completely confident. These responses were then mapped to the following values for calculation:

1: 0
2: 1.2
3: 2.4
4: 3.6
5: 4.8
6: 6

B Cluster Figures

Here we have figures for the remaining statistically significant variables. Categorical variables are represented with bar graphs whereas numerical variables are represented with line graphs. If a variable is significant in both the k-means clusters and hierarchical clustering clusters, they are in the same figure.

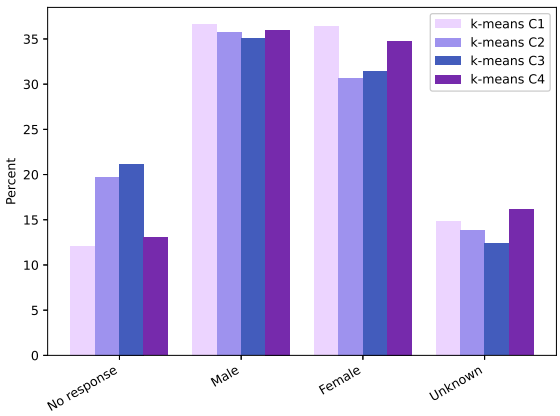


Figure 9: Sex across Four k-means Clusters

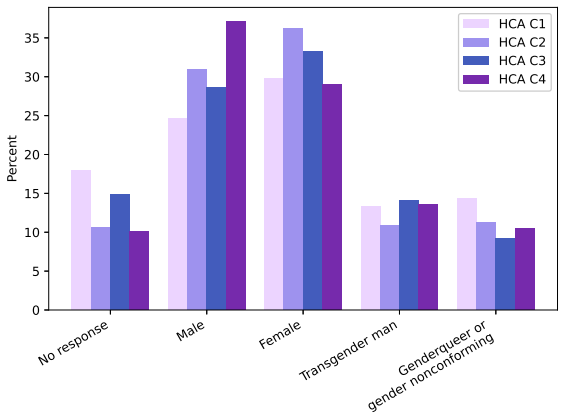


Figure 10: Gender across Four HCA Clusters

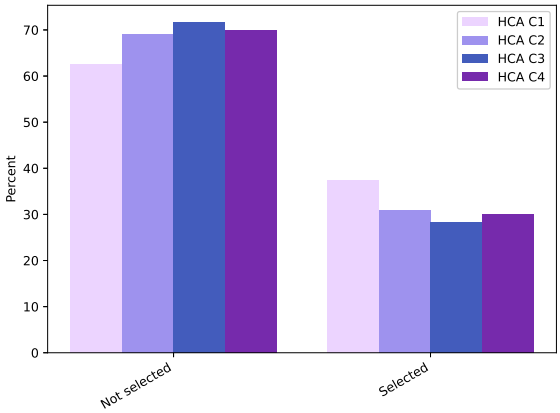


Figure 11: American Indian or Alaska Native identifying individuals across Four HCA Clusters

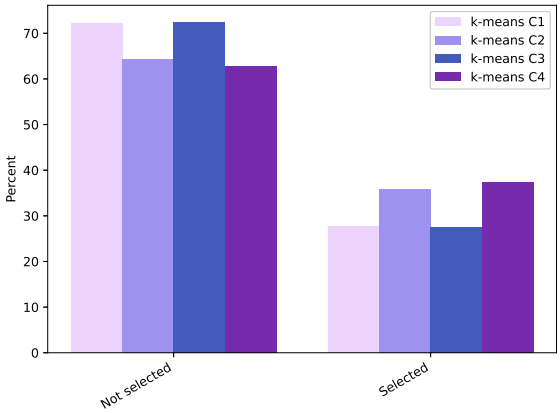


Figure 12: Black or African American Identifying individuals across Four k-means Clusters

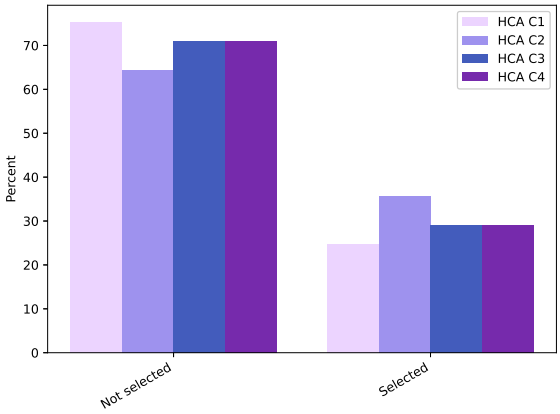


Figure 13: Native Hawaiian or Pacific Islander individuals across Four HCA Clusters

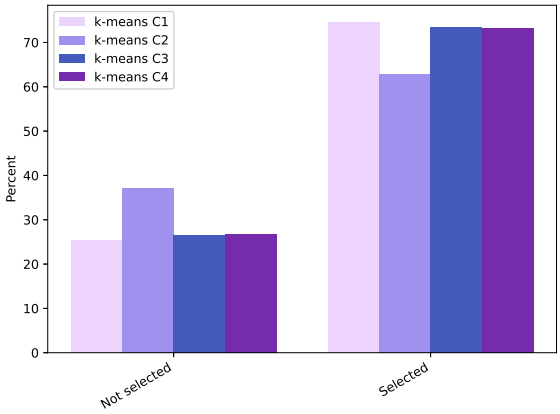


Figure 14: White identifying individuals across Four k-means Clusters

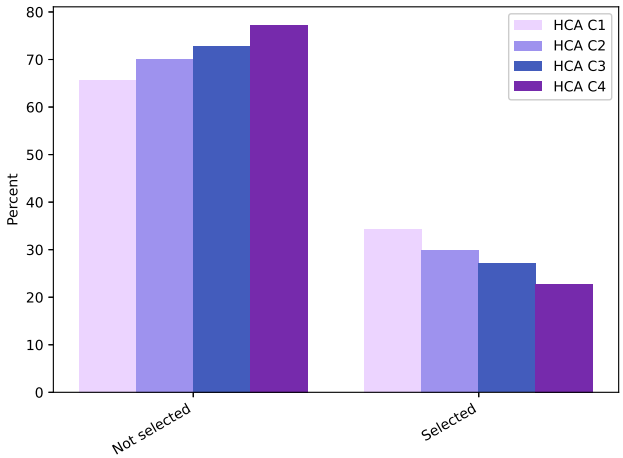


Figure 15: Race unknown or not reported individuals across Four HCA Clusters

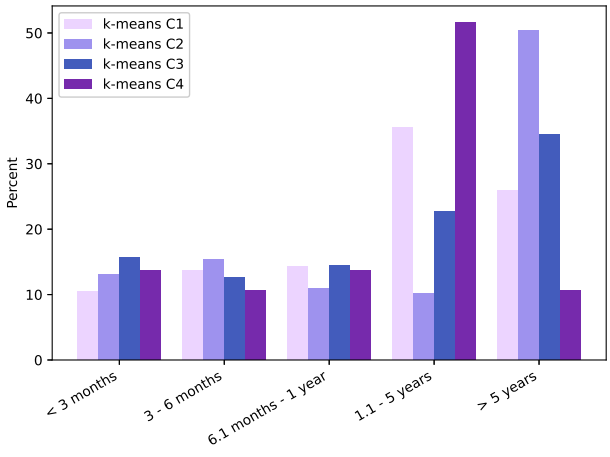


Figure 16: Duration across Four Clusters created by k-means

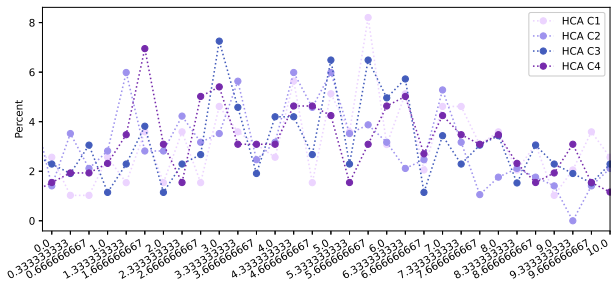


Figure 17: PEG-3 across Four Clusters created by Hierarchical Clustering

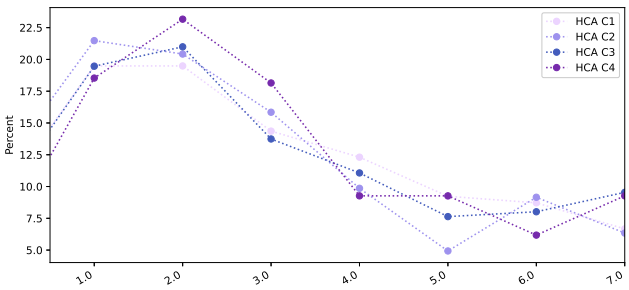


Figure 18: Household Size across Four Clusters created by Hierarchical Clustering

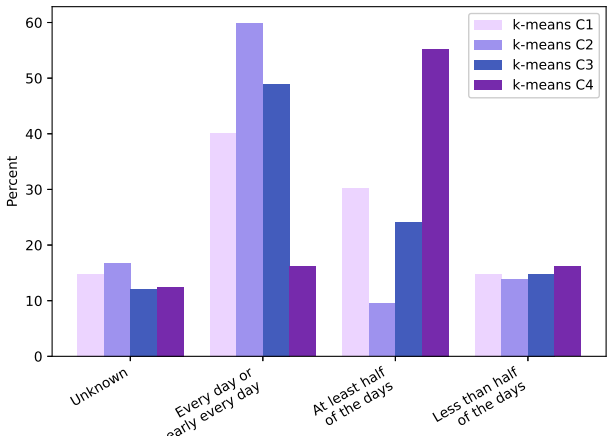


Figure 19: Frequency (in the past 6 months) across Four Clusters created by *k*-means

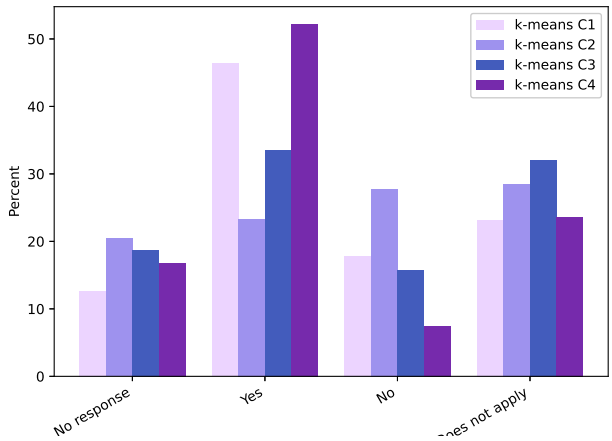
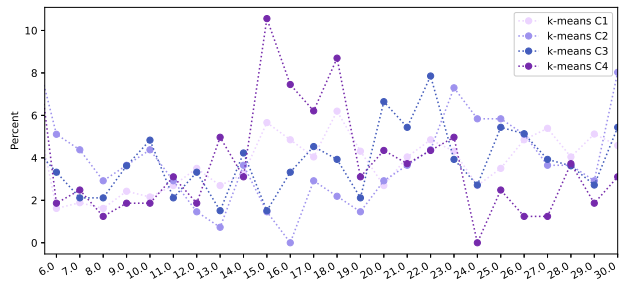
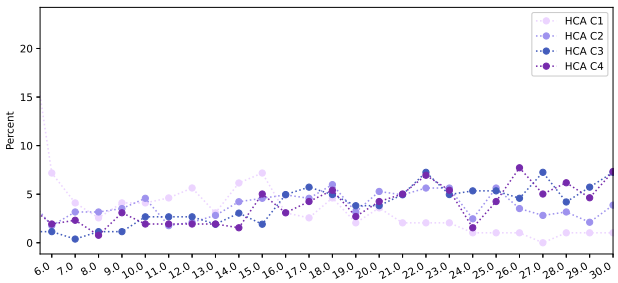


Figure 20: Unemployment across Four Clusters created by *k*-means

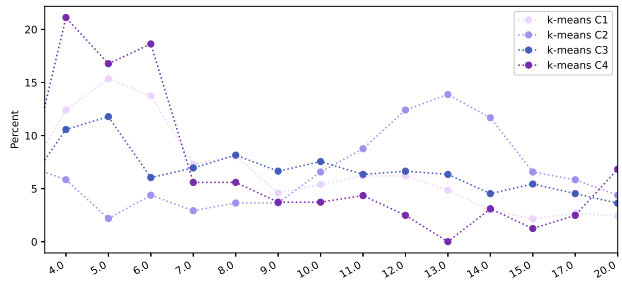


(a) *k*-means

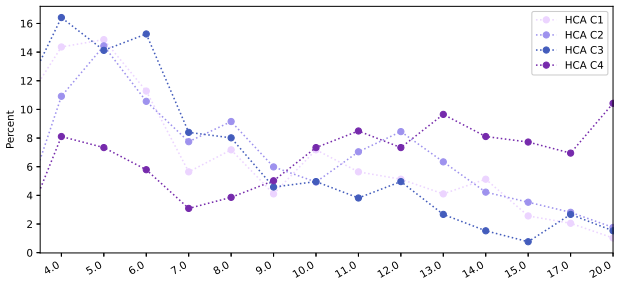


(b) HCA

Figure 21: QoL across Four Clusters

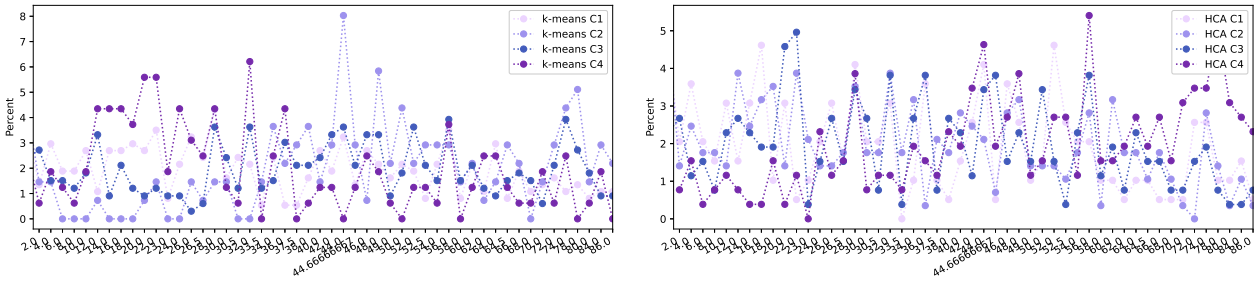


(a) *k*-means



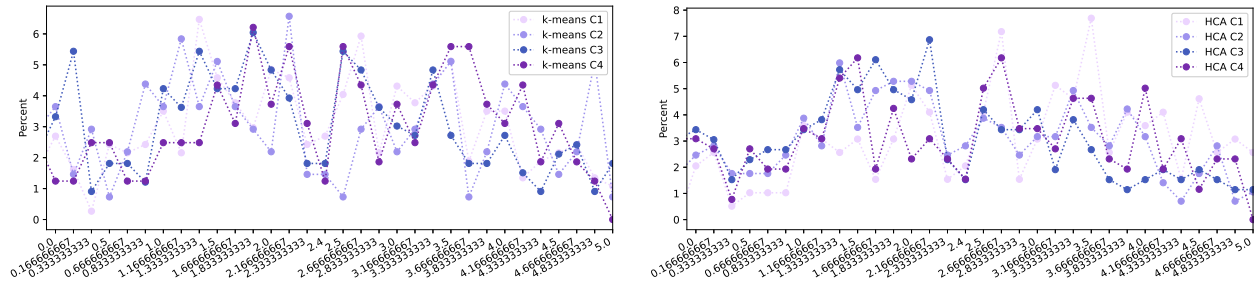
(b) HCA

Figure 22: Depression across Four Clusters



(a) k-means (b) HCA

Figure 23: PAS across Four Clusters



(a) k-means (b) HCA

Figure 24: Not Distracting across Four Clusters

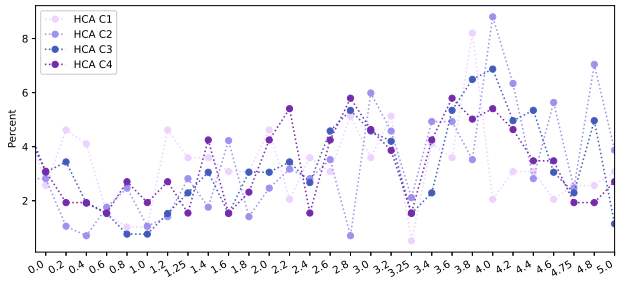


Figure 25: Emotional Awareness across Four Clusters created by Hierarchical Clustering

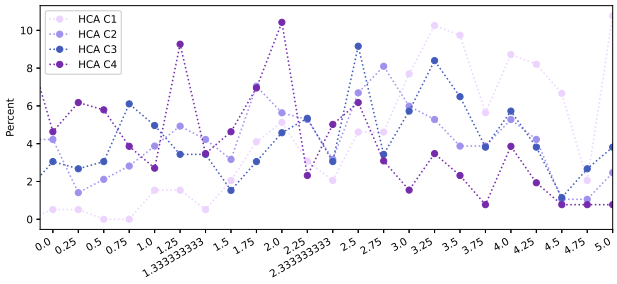


Figure 26: Self Regulation across Four Clusters created by Hierarchical Clustering

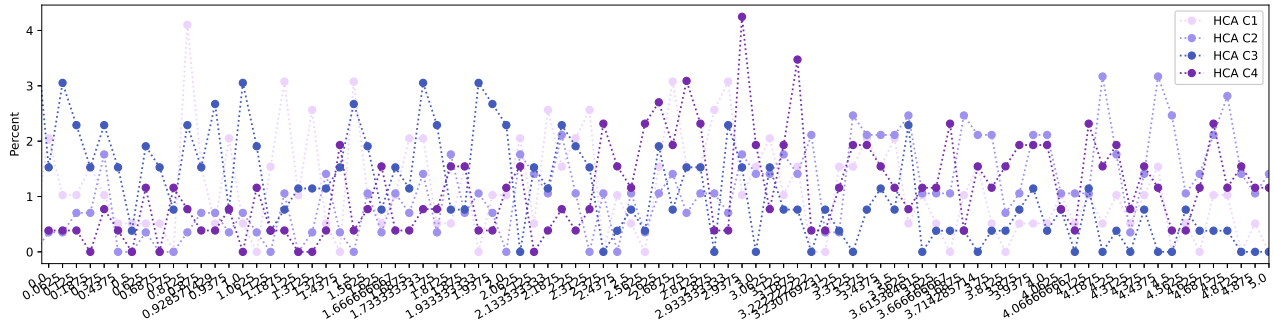
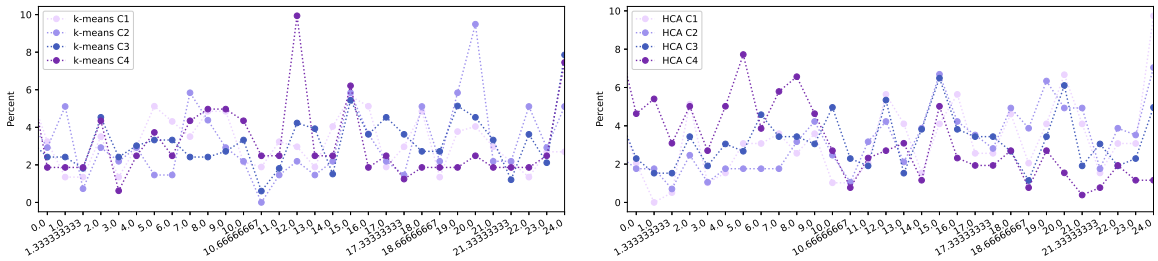
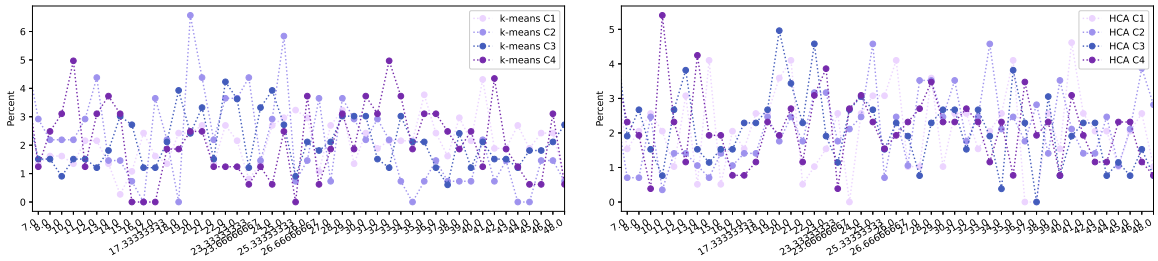


Figure 27: Interoceptive Awareness across Four Clusters created by Hierarchical Clustering



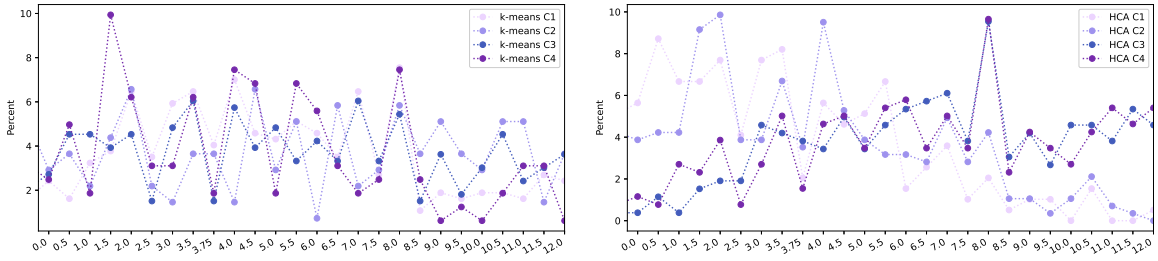
(a) *k*-means (b) HCA

Figure 28: Fear Avoidance across Four Clusters



(a) *k*-means (b) HCA

Figure 29: Chronic Pain Acceptance across Four Clusters



(a) *k*-means (b) HCA

Figure 30: Pain Catastrophizing across Four Clusters

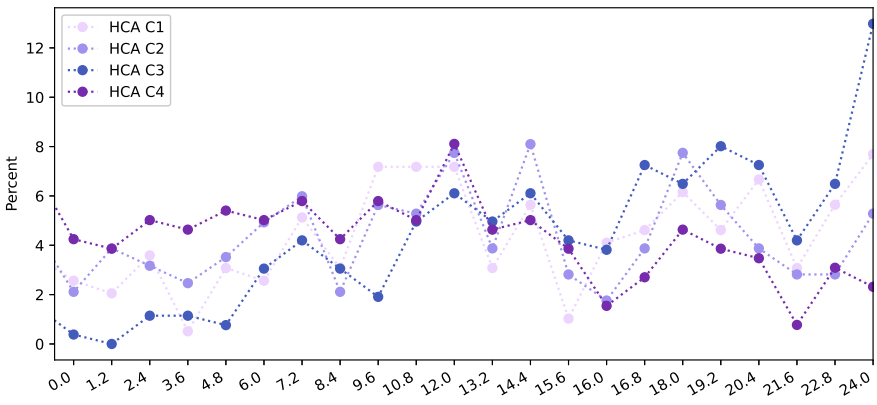


Figure 31: Self Efficacy across Four Clusters created by Hierarchical Clustering

References

[1] Neo4j Graph Database & Analytics – The Leader in Graph Databases.

[2] Emmanuel Bäckryd, Elisabeth Persson, Annelie Inghilesi Larsson, Marcelo Rivano Fischer, and Björn Gerdle. Chronic pain patients can be classified into four groups: Clustering-based discriminant analysis of psychometric data from 4665 patients referred to a multidisciplinary pain centre (a SQRP study). *PLOS ONE*, 2018.

[3] Haochen Chen, Syed Fahad Sultan, Yingtao Tian, Muhao Chen, and Steven Skiena. Fast and Accurate Network Embeddings via Very Sparse Random Projection. *International Conference on Information and Knowledge Management*, 2019.

[4] Geert De Soete and J. Douglas Carroll. K-means clustering in a low-dimensional Euclidean space. In Edwin Diday, Yves Lechevallier, Martin Schader, Patrice Bertrand, and Bernard Burtschy, editors, *New Approaches in Classification and Data Analysis*, pages 212–219, Berlin, Heidelberg, 1994. Springer Berlin Heidelberg.

[5] Jessica Gliozzo, Marco Mesiti, Marco Notaro, Alessandro Petrini, Alex Patak, Antonio Puertas-Gallardo, Alberto Paccanaro, Giorgio Valentini, and Elena Casiraghi. Heterogeneous data integration methods for patient similarity networks. *Briefings in Bioinformatics*, 23(4):bbac207, 06 2022.

[6] Aditya Grover and Jure Leskovec. Node2vec: Scalable Feature Learning for Networks. pages 855–864, 2016.

[7] Björn Larsson, Björn Gerdle, Lars Bernfort, Lars-Åke Levin, and Elena Dragioti. Distinctive subgroups derived by cluster analysis based on pain and psychological symptoms in Swedish older adults with chronic pain - a population study (PainS65+). *BMC Geriatrics*, 2017.

[8] Zhihuang Lin, Dan Yang, and Xiaochun Yin. Patient similarity via joint embeddings of medical knowledge graph and medical entity descriptions. *IEEE Access*, 8:156663–156676, 2020.

[9] Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[10] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[12] Scott D. Tagliaferri, Tim Wilkin, Maia Angelova, Bernadette M. Fitzgibbon, Patrick J. Owen, Clint T. Miller, and Daniel L. Belavy. Chronic back pain sub-grouped via psychosocial, brain and physical factors using machine learning. *Scientific Reports*, 2022.

[13] Carla E Zelaya, James M Dahlhamer, Jacqueline W Lucas, and Eric M Connor. Chronic pain and high-impact chronic pain among us adults, 2019. *NCHS data brief*, 11 2020.

[14] Rukui Zhang, Zhaorui Liu, Chaoyu Zhu, Hui Cai, Kai Yin, Fan Zhong, and Lei Liu. Constructing cancer-specific patient similarity network with clinical significance. *medRxiv*, 2023.

FRCCS

Evaluating Community Structure Preservation of Network Embedding Algorithms

Jason Barbour[1✓], Stephany Rajeh[2], Sara Najem[3] and Hocine Cherifi[4]

[1] *Graduate Program in Computational Science, American University of Beirut, Beirut, Lebanon ; jgb21@mail.aub.edu.lb*
[2] *Efrei Research Lab, EFREI Paris, Villejuif, France ; stephany.rajeh@efrei.fr*
[3] *Center for Advanced Mathematical Sciences, American University of Beirut, Beirut, Lebanon; sn62@aub.edu.lb*
[4] *ICB UMR 6303 CNRS, University of Burgundy, Dijon, France ; hocine.cherifi@u-bourgogne.fr*

[✓] *Presenting author*

Abstract. Network embedding compresses network information into low-dimensional vectors while retaining structural and semantic details. Preserving community structure is vital. Existing evaluation metrics often overlook community structure. This study assesses network embedding algorithms across various community strengths, demonstrating performance variation in mesoscopic quality metrics.

Keywords. *Network Embedding, Community Structure, Evaluation Metrics, Quality Metrics*

1 Introduction

Networks often exhibit a modular structure, where nodes cluster into communities with shared characteristics or functions [3]. Understanding these community structures is crucial for various applications, from recommendation systems to the optimal spread of information and disease control [2, 7, 8]. With network sizes increasingly increasing, generating lower order representation, known as network embedding, has gained significant attention in recent years [4]. This technique transforms networks into low-dimensional vector representations.

While certain techniques are designed to explicitly maintain or enhance the community structure through the embedding process, others may not consider community structure preservation a primary objective. Nonetheless, one of the fundamental goals of all network embedding techniques is to project the similarity of the nodes of the original network onto the lower-dimensional space.

Further, network embedding techniques are commonly evaluated through classification metrics [4]. Nonetheless, these metrics are agnostic about the community structure: they do not indicate whether it is well preserved after the embedding process. In other words, they offer information about the overall quality of results but do not reveal the fine-grained details of community structure within a network.

Consequently, there is a need for a comprehensive comparative analysis of network embedding algorithms from a modular perspective. This paper analyzes the performance of the most prominent network embedding algorithms on controlled synthetic networks.

Note that the full paper was published in Complex Networks & Their Applications XII [1]

2 Experimental Setup

This section describes the fundamental steps of the experimental setup employed to evaluate the network embedding algorithms' efficacy in maintaining the community structure. Seven main steps are described below to evaluate the network embedding algorithms. A bird's-eye view of the experimental setup is illustrated in Figure 1.

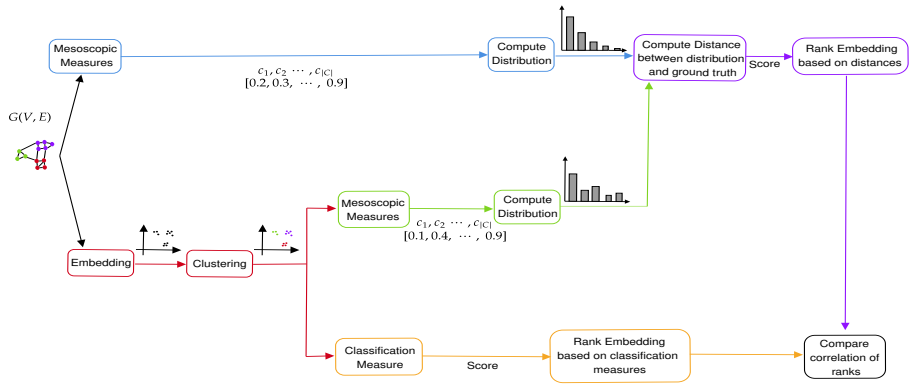


Figure 1: Flowchart of the experimental setup to evaluate the performance of the network embedding algorithms. Mesoscopic metrics are calculated individually for each of the network's communities.

Network Generation:

Synthetic networks are generated with the help of the ABCD network generator as to know the ground truth with certainty. ABCD offers the ability to change multiple parameters of the network generation process, namely: number of nodes (N), power-law exponent for degree distribution (τ_1), minimum degree (d_{min}), maximum degree (d_{max}), power-law exponent for community size distribution (τ_2), minimum community size (c_{min}), maximum community size (c_{max}) and the mixing parameter (μ) [5]. Note that in the study, we fix all parameters and just vary the mixing parameter.

Algorithms

The study employs ten embedding algorithms to embed generated networks into a 128-dimensional space: DeepWalk, Node2Vec, Diff2Vec, Walkets, Modularity-Normalized Matrix Factorization (M-NMF), Laplacian Eigenmaps (LEM), Randomized Network Embedding (RandNE), Boosted Network Embedding (BoostNE) and Network Matrix Factorization (NetMF) [1].

Classification Metrics

Classification metrics such as adjusted mutual information score (AMI), normalized mutual information score (NMI), adjusted random score (ARI), Micro-F1 score, and Macro-F1 score, which are not necessarily community-aware evaluators, are then calculated. These are the metrics that are usually used in the literature.

Quality Measures

Multiple quality metrics are used here. We check these metrics for each community in the graph, and we can get a distribution of these metrics for a given graph[1].

1. **Internal distance:** is the average shortest distance of nodes inside a given community
2. **Internal density:** is the edge density inside a given community
3. **Maximum-out degree fraction (Max-ODF):** is the maximum of the ratio of inter-community links vs. intra-community links.
4. **Average-out degree fraction (Average-ODF):** same as Max-ODF, but averaging.
5. **Hub dominance:** is based on the intra-community links of a node that has the highest intra-community links in its community
6. **Flake-Out degree fraction (Flake-ODF):** is the percentage of the out degree fraction.
7. **Embeddedness:** quantifies intra-community links. It is the opposite of Average-ODF.
8. **Hub dominance:** is based on the intra-community links of the node that has the highest intra-community links in its community

After embedding, K-means clustering is applied to group embeddings into the same number of clusters as the ground truth. The Kullback-Leibler divergence is computed between the ground truth and embedded distributions. The distance between the two distributions gives a measure of how good the embedding is. The smaller the distance, the better the algorithm. For classification Metrics, we can just look at the performance of each algorithm.

Voting Model

A ranking scheme utilizing Schulze's voting model is employed to evaluate the overall quality of the algorithms based on performance metrics[14]. KL-divergence scores and classification metrics serve as voters, while algorithm ranks act as candidates. The model compares candidates in head-to-head matchups to identify the algorithm with the broadest preference, aiming for a consensus winner. The final ranks represent a consensus across all metrics, providing a comprehensive assessment of network embedding algorithm performance.

3 Performance of Embedding Algorithms based on Quality Metrics

As shown in Table 1, LEM demonstrates outstanding performance within a robust community structure, excelling in community awareness and classification metrics. However, as the community structure strength diminishes, its effectiveness prominently declines with classification metrics. The opposite behavior is seen with NetMF. In contrast, M-GAE maintains outperformance across both community-aware and classification metrics, regardless of the community structure strength, by ranking either first or second.

4 Conclusion

Preserving network community structure is crucial, and network embedding techniques offer significant potential. However, the evaluation metrics commonly used in the literature fail to capture this preservation effectively. This study highlights the need for a comprehensive comparison of network embedding algorithms from a modular perspective. Our work is limited to evaluating the effect of the mixing parameter on the embedding quality. Our study specif-

		Ranks									
μ	Metrics	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}
$\mu \leq 0.4$	Meso	LEM	M-GAE	RandNE	NetMF	Deepwalk	Node2Vec	Diff2Vec	Walklets	M-NMF	BoostNE
	CL	LEM	M-GAE	RandNE	Walklets	NetMF	M-NMF Node2Vec		Deepwalk	Diff2Vec	BoostNE
$\mu > 0.4$	Meso	M-GAE	NetMF	RandNE LEM		M-NMF	Deepwalk	Diff2Vec	Node2Vec	BoostNE	Walklets
	CL	M-GAE	NetMF	M-NMF	BoostNE	LEM	Deepwalk	Node2Vec RandNE		Diff2Vec	Walklets
Total	Meso	LEM	M-GAE	RandNE	NetMF	Deepwalk	Node2Vec	Diff2Vec	M-NMF	Walklets	BoostNE
	CL	M-GAE	LEM	NetMF	M-NMF	RandNE	Node2Vec	Deepwalk	BoostNE	Walklets	Diff2Vec

Table 1: The ranking of the embedding algorithms based on Schulze's method for mesoscopic (Meso) and classification metrics (CL) with respect to the mixing parameter(μ)

ically aims to determine the adequacy of classification metrics employed in the literature to comprehend the effectiveness of network embeddings. Results reveal that these metrics do not comprehensively reflect the network's community structure. Furthermore, the efficacy of certain embedding techniques, such as LEM, M-GAE, and NetMF, is influenced by the strength of the community structure. These findings underscore the need for a more attentive approach in evaluating embedding techniques tailored to the specific application. Finally, it is worth mentioning that we only considered the effect of the mixing parameter in this study, further research can be done to check for other parameters' influence such as the dimensionality of the embedding, and other coefficients that describe community structure.

References

[1] Jason Barbour, Stephany Rajeh, Sara Najem, and Hocine Cherifi. Evaluating network embeddings through the lens of community structure. In Hocine Cherifi, Luis M. Rocha, Chantal Cherifi, and Murat Donduran, editors, *Complex Networks & Their Applications XII*, pages 440–451, Cham, 2024. Springer Nature Switzerland.

[2] Debayan Chakraborty, Anurag Singh, and Hocine Cherifi. Immunization strategies based on the overlapping nodes in networks with community structure. In *Computational Social Networks: 5th International Conference, CSoNet 2016, Ho Chi Minh City, Vietnam, August 2-4, 2016, Proceedings 5*, pages 62–73. Springer International Publishing, 2016.

[3] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.

[4] Mingliang Hou, Jing Ren, Da Zhang, Xiangjie Kong, Dongyu Zhang, and Feng Xia. Network embedding: Taxonomies, frameworks and applications. *Computer Science Review*, 38:100296, 2020.

[5] Bogumił Kamiński, Paweł Prałat, and François Théberge. Artificial benchmark for community detection (abcd)—fast random graph model with community structure. *Network Science*, 9(2):153–178, 2021.

[6] Solomon Kullback and Richard A Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22(1):79–86, 1951.

[7] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Characterizing the interactions between classical and community-aware centrality measures in complex networks. *Scientific reports*, 11(1):10088, 2021.

[8] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Comparative evaluation of community-aware centrality measures. *Quality & Quantity*, 57(2):1273–1302, 2023.

Complex Networks: Structure & Dynamics II

FRCCS

Emergence of dynamical networks in termites

Louis E. Devers[1✓], Perrine Bonavita[2] and Christian Jost[2]

[1] *Institut de Mathématiques de Toulouse (IMT), Université Paul Sabatier Toulouse III ; devers.louis@gmail.com*
[2] *Centre de Recherche sur la Cognition Animale (CRCA-CBI), Université Paul Sabatier Toulouse III ; perrine.bonavita@univ-tlse3.fr et christian.jost@univ-tlse3.fr*
[✓] *Presenting author*

Abstract. Termites form complex dynamical trail networks from simple individual rules when exploring their environment. To help identify those simple rules, we reconstructed trail networks from time-lapse images of roaming termites. We quantified the trails' frequentations over time and compared them to the ones obtained by a null model. Arena borders were preferred in both simulated and observed data. Yet, the amplification phenomenon was higher with real termites, underlining the role of pheromones.

Keywords. *Dynamical Networks; Social Insects; Network Reconstruction; Termites; Biological Networks*

1 Introduction

In social insects, one can consider that the whole is more than the sum of its parts. Colony-level properties emerge from simple individual-based rules. For instance, in ants, it has been shown that the pheromones deposited by individuals allowed the colony to better exploit food sources [7, 8, 13, 9]. Termites, similarly to ants, build nests, forage, form tunnel networks or even cultivate fungi[1]. If pheromone trail emergence has been well studied in ants [21], it is not the case of termites. Studies on termites mainly focused on the tunnelling network [15], their dynamics [19, 16] and their nest architecture [23, 22, 14]. Unlike tunnelling behaviours, trail networks can be more difficult to observe. The movements of termites on surfaces without any building material, artificial galleries, or nest-oriented behaviours are not that well documented. Just like ants, termites' movements might be influenced by: other individuals [20], angles [17, 25] and pheromones [1, 27, 28].

This paper aims to investigate individual behaviours that are sufficient and necessary to reproduce higher-level properties. In our case, we will focus on the trail network formed by freely roaming termites without any stimuli (nest, gallery, building material). Which are the individual rules reproducing such networks? and how to describe such networks? We detail how we can extract a trail network of invisible pheromones through image processing. And we detail a method to follow the network's dynamical properties over time. To further explore this network, we developed a null model based on simple individual and voluntarily naive local rules. By comparing our observations to our null model, we can further understand how those

networks are formed.

2 Methods

2.1 Setup and species

The experiment consists of filming termites roaming freely in a circular box and analysing the network they form in 20 minutes (Fig.1). Termites will seach a shelter in such unfamiliar environment. The termites used: *Procornitermes araujoi* measure about 5-6mm and originate from South America [11]. Experiments were run in 2012 by Christian Jost and Christine Lauzeral in Rio Claro, Brazil. The experiment was replicated 15 times. For each experiment, 106 *Procornitermes araujoi* were extracted from the same nest on the university campus of UNESP Rio Claro. To maintain the polymorphism in natural populations, 100 of them were workers (smaller termites), and 6 of them were soldiers (bigger termites). Termites were contained in a 3cm diameter zone before being let free in arenas of 24 or 40cm diameter (respectively 6 and 9 replicates). Experiments were filmed at 25fps, and one picture in ten was kept (one every 0.4s). Thomas Colin segmented the termites into 3000 binary pictures for each experiment.

2.2 Network reconstruction

For each experiment, we want to form a network from these segmented images and get the termite flow observed on each edge over time. We thus treated the images obtained using Matlab [18] (Fig.1).

We subtracted each frame from the previous one to obtain a binary mask of (only) moving termites. We then summed the binarized differentiated images before applying a log transformation to it (Fig.1D) over the first 20 minutes (3000 pictures). The brighter the image, the more frequented it is. We can already observe that some paths are more frequented than others. To segment the network, we detected vessel-like objects using a Frangi filter [12]. As it does not detect intersections, we obtained the whole network binary mask through additional morphological operations (Fig.1E). We spurred this mask to form edges and split them by placing nodes on intersections and extremities (Fig.1F). Some nodes were regrouped if too close to each other, thus forming nodes of degrees higher than 2 or 3. This process can be applied to any image (or stacked image) of a network, feel free to approach authors for more information.

We used the termites' binary masks to compute the dynamics of termite fraction on edges. If a termite is within the binary mask in Fig.1E, its pixels get assigned to the nearest edge, thus giving non-directional data of all termite fraction on edges over time $N_{ij}(t)$. Notably, the sum of $N_{ij}(t)$ over all edges equals 1 for all time-steps.

2.3 Null model

There is little information about freely roaming insects network properties in the literature. Insects networks are usually studied in foraging (when they form networks around their nest) or nest-building context. To better compare the properties of this dynamical network we needed a null model [4]. We propose here a freely roaming termite deterministic null model. The termites can move freely within all possible edges in the observed network (detected in the first 20 minutes of experiment). The model functions as follows:

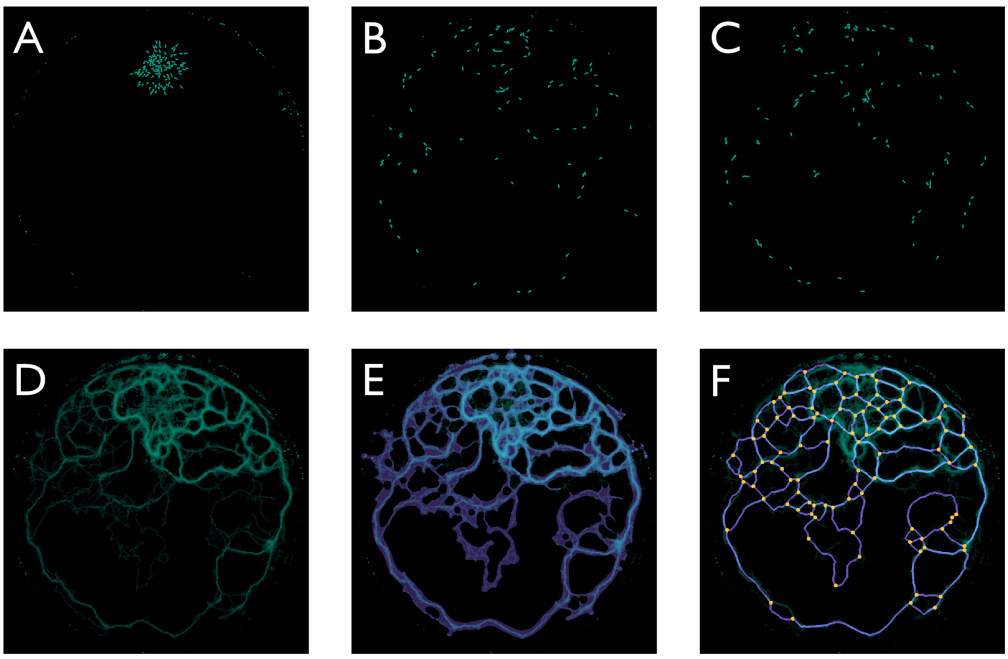
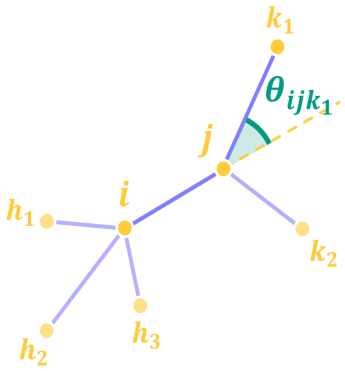


Figure 1: Method of network reconstruction from time-lapse images (arena diameter: 40cm). (A-C) are images of the binarised termites spreading and exploring the arena at time $t = 0, 5, 10$ minutes. (D) is the cumulative image of moving termites' presence. (E) is the segmentation of the previous cumulated image (obtained using Frangi filters [12]). (F) is the obtained network overlapped with the cumulated image for reference.



$$p_{ijk} = \frac{|\cos\left(\frac{\theta_{ijk}}{2}\right)|}{\sum_l |\cos\left(\frac{\theta_{ijl}}{2}\right)|} \tag{1}$$

We noted p_{ijk} the probability of joining the edge jk (going from node j to node k) for an agent present in the edge ij. The probability of joining jk is computed as a ratio of a preference score over all the possible jl edges accessible from node j. That preference score is computed as $|\cos\left(\frac{\theta_{ijk}}{2}\right)|$ where θ_{ijk} is the turning angle of a termite moving from edge ij and edge jk. The preference score equals one if i, j and k are aligned ($\theta_{ijk} = 0$), and 0 if going backwards (if $i = k, \theta_{ijk} = \pi$ or $-\pi$). In the illustration of Eq. (1), $\theta_{ijk_1} = \pi/6$ and $\theta_{ijk_2} = -\pi/3$. The preference scores $|\cos\left(\frac{\theta_{ijk}}{2}\right)|$ are respectively 0.9659 and 0.8660 for k_1 and k_2. Note that the preference score of going to i from ij is 0. Thus by Eq. (1): $p_{ijk_1} = 0.5273$ and $p_{ijk_2} = 0.4727$.

We can then write that $N_{ij}(t)$, the termite fraction on an edge ij at time t, fluctuates as :

$$\frac{dN_{ij}(t)}{dt} = v \times \left(\sum_h \frac{N_{hi}(t) p_{hij}}{L_{hi}} - \sum_k \frac{N_{ij}(t) p_{ijk}}{L_{ij}} \right) \tag{2}$$

In Eq. (2), $N_{ij}(t)$ is the termites fraction on the edge going from nodes i to j at time t. v is a single termite velocity (1 cm.s^{-1}). $N_{ij}(t)$ evolves positively with incoming termites coming from all possible nodes h, connected to i. The incoming flux is averaged to the termite fraction in hi times the probability to join ij from hi (p_{hij} in Eq. (1)). The incoming flux must be divided by the length of said edge, L_{hi}, while the termite goes at a velocity v. Similar reasoning is made for leaving fluxes: $N_{ij}(t)$ evolves negatively with leaving termites going to all possible nodes k, connected to j. The leaving flux is averaged to the termite fraction in ij times the probability to leave ij to jk (p_{ijk} in Eq. (1)). The leaving flux must be divided by the length of said edge, L_{ij}, while the termite goes at a velocity v.

To determine initial conditions, we identified the node i closest to the termites' experimental release point. We evenly distributed termites in all out-going edges connected to node i. Similarly as the observed data, the sum of $N_{ij}(t)$ over all edges equals 1 for all time-steps.

These rules are simple, local, and only based on angle preferences. They roughly match termite angle preferences observed in tunnels [17, 25]. Authors argue that the preference function can be any function returning one if edges are aligned ($\theta_{ijk} = 0$) and returning 0 if going backward ($\theta_{ijk} = \pi$ or $-\pi$).

3 Results

3.1 Final states Networks

To first describe the networks obtained, we will focus in this section on "final states networks". These networks include all the edges, and all the nodes extracted at t=20min. Each edge is associated with its termite fraction $N_{ij}(t)$ over the course of the experiment. For observed networks, we plotted the mean termite fraction considering all previous time-steps. For simulated networks, it consists of the final termite fraction (as the simulation reaches equilibrium corresponding to the mean termite fraction). We can thus compare simulated and observed termite fraction on edges.

We can observe on Fig.2 the obtained final network for both observation and simulation (Fig.2A and C respectively). The colour intensity corresponds to the final termite fraction on an edge. We can note that in the observed networks, termites are not uniformly distributed, especially compared to the simulated case. Indeed, some edges are highly preferred to others over time (Fig.2B). In the simulated case, termites do not exhibit strong preferences and rapidly reach a stable state (Fig.2D). Termite fraction on edges are not distributed in the same way (Fig.2F).

To identify edges that are over-frequented in the observed network, we subtract observed and simulated termite densities (Fig.2E). Edges preferred by termites are in green, and edges preferred by the null model are in purple. Edges are white if densities are equivalent for the termites and the null model. In this specific network, edges on the border are over-frequented compared to our null model. Conjointly, most edges in the middle of the network are slightly preferred by the model.

Can this observation be generalised to other networks ? In Fig.3 we compare observed simulated edges' termite fraction for all networks treated (12 out of 15). To visualise edges relative position in the arena, edges close to the border are represented in bright blue, while edges close to the centre of the area are represented in pink. In Fig.3A, we plotted all observed edges' fraction against simulated ones. Over-represented edges compared to a null model are present over the dashed diagonal line (and respectively, under-represented ones bellow the line). Most edges are

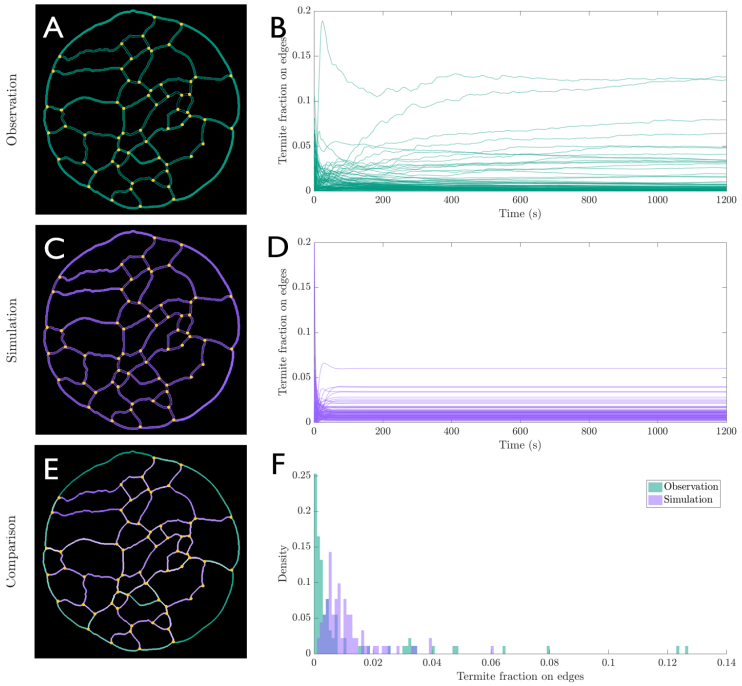


Figure 2: Termite fraction on edges over time in a single network (arena diameter: 24cm). (A) Extracted network and observed mean fraction. The filling of edges represents the fraction (dark to green for respectively low to high fraction). (B) Mean termite fraction over time. (C) Extracted network and simulated final termite fraction. The filling of edges represents the fraction (dark to purple for respectively low to high fraction). (D) Edge fraction over time (Eq. (2)) (E) Difference of termite fraction on edges between observed and simulated data. Green edges are over-represented in the observed data, and purple edges are under-represented in observed data. White edges are equivalently dense in both. (F) Density distribution of observed mean termite fraction (green) and simulated final termite fraction (purple).

under-represented observations, meaning that termites prefer to focus on a few edges with a high activity. Bright blue points following the diagonal line in Fig.3A show that edges located at the border of the arena are preferred in both models. However, the preference is way higher in actual termites' networks. This common preferences is also visible by plotting percentile rank of fraction of simulated vs observed edges (B). We note in the top-right corner that frequented edges are common in simulations and observation and correspond to border edges. However, other edges show few to no correspondence.

Which are the edges preferred in termites' networks ? From observation of Fig.2, we hypothesised that edges that are far from the middle of the arena are over-represented compared to a null model. We represented the difference of fraction (Observed - Simulated) against the edge position in the arena (Fig.3C). Indeed, edges far from the centre of the arena (close to the border) are over-represented. We also plotted the difference of fraction against edge orientation with regard to the arena's radius (Fig.3D). We note that edges perpendicular to the radius are over-represented compared to our null model. Both these observations support the fact that the network in Fig.2 is representative of that phenomenon.

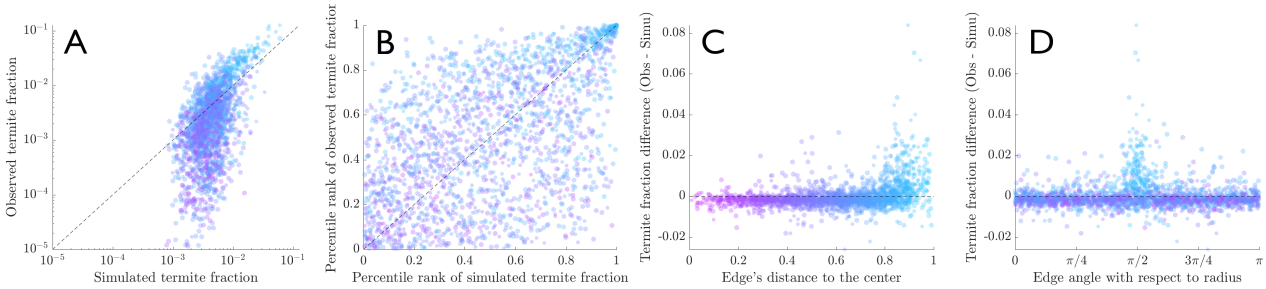


Figure 3: Comparison of simulated and observed termite fraction for each edge. (A) observed vs simulated termite densities (log-log scale). The diagonal dashed line visually supports testing for equality. (B) Percentile rankings of termite fraction (Simulated vs Observed) (C) Difference of termite fraction (Observed minus Simulated) as a function of edge position. Edge position was computed as its distance from the centre of the arena divided by the radius of the arena. The horizontal dashed line visually supports testing for equality. (D) Difference of termite fraction (Observed minus Simulated) as a function of edge orientation. Edge orientation was computed as its angle with the radius of the arena. The horizontal dashed line visually supports testing for equality. Colour is function to the edge's distance to the centre (pink for centre, bright blue for border edges) and marker size depends the diameter of the experimental arena.

3.2 Dynamical Networks

We showed in the previous section that the termite fraction on edges varies over time for both observed and null model networks. However, if an edge had a low termite fraction, meaning that a path was rarely frequented, it remained in the network. In this section, we propose a method to dynamically modify network topology as a function of edge fraction. Low fraction edges will be discarded and can be added back to the network later on. The structure of the network thus changes over time, and with it, its properties.

As seen in Fig.2B and D, the observed and simulated termite densities are not distributed in the same way. So, an absolute filter above which an edge is considered "active" will not suffice. To discriminate active and non-active edges, we propose a method inspired by social insects like ants and termites: pheromones. The amount of pheromones on a given edge increases with passing termites but decreases through evaporation at a constant rate μ. Pheromones are usually key to understanding routing problems and path selection in social insects [26, 9]. Here, we computed the amount of pheromones on each edge Ph_{ij} for each time step as follows :

$$\frac{dPh_{ij}(t)}{dt} = -\mu Ph_{ij}(t) + \frac{N_{ij}}{L_{ij}} \tag{3}$$

In Eq. (3), the concentration of pheromones $Ph_{ij}(t)$ on edge ij evaporates at rate μ. Previous work estimated the half life of *Procornitermes araujoi* of being 16 minutes [11]. Implying a rate of evaporation of $\mu = 7.26 \times 10^{-4}\text{s}^{-1}$. The concentration of pheromones increases with the number of individuals present in edge ij. We need to divide by the length of the edge L_{ij} to obtain concentrations of pheromones per cm.

From there, we conserved the edges with the higher amount of pheromones that totalled p_{thresh} per cent of all the pheromones at time t. In our case,$p_{thresh} = 0.8$ meant that active edges were the biggest ones representing a total of 80% of all pheromones. Such criterion allows easy comparison between the observed and simulated networks.

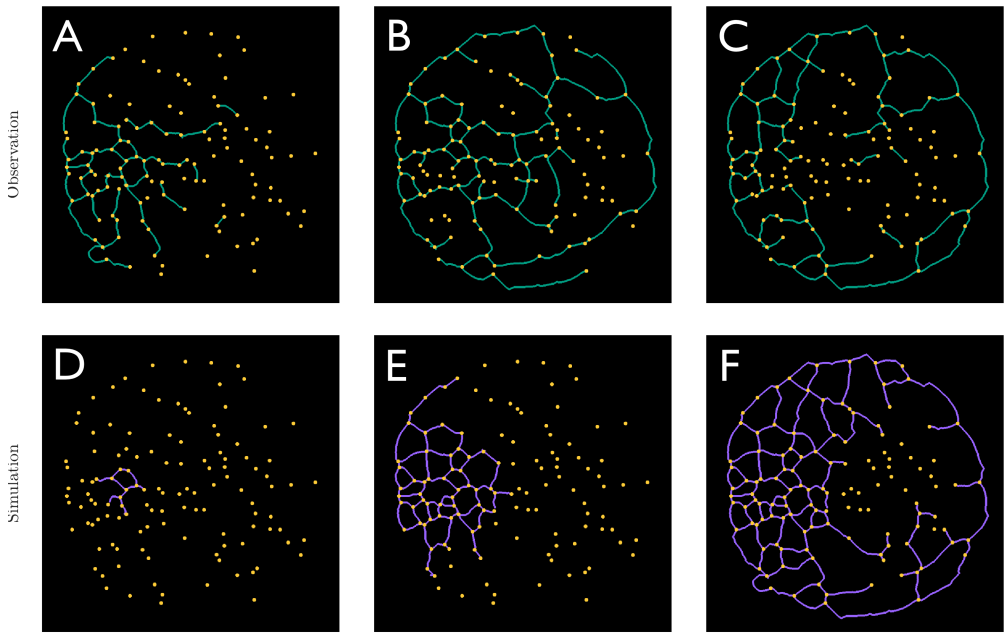


Figure 4: Examples of dynamical networks (arena diameter: 24cm). (A-C) Observed networks ($t = 10, 100, 1000$ s). (D-F) Simulated networks ($t = 10, 100, 1000$ s).

In Fig.4, is represented both observed (A-C) and simulated (D-F) networks over time ($t = 10, 100, 1000s$). Concerning the observed network (A-C), we first observe a spread of the termites through the whole arena, followed by a selection of edges. The edges on the border are mainly selected. Concerning the networks simulated by our null model, we also observe a spread, but not followed by a drastic edge selection. However, border edges seem to be preferred as well. The main difference thus lies in the intensity of the filtering, rather than the edges being filtered.

The dynamics of the formed networks properties can be extensively studied. We propose here preliminary results concerning the total length of the networks and the number of conserved edges over time. Future work will be needed to focus on metrics like efficiency, robustness or meshedness for instance [2, 4, 3]. In Fig.5, we represented (A) the total number of edges and (B) the total length of the network in cm over time. Both observed (green) and simulated (purple) networks are shown. We note that the number of edges and total length of the networks are different between the observed and simulated networks. However, we observe no differences in edge number and total length between 24 (light green) and 40cm (dark green) arenas. It could mean that 106 termites can only sustain a pheromone track of about 200cm independently of the arena's diameter.

4 Discussion

This paper investigated individual termites roaming behaviour by studying the network they collectively form dynamically. To do so, we observed termites forming networks while exploring a circular arena. We extracted its nodes and edges using image processing and Frangi filter [12]. We measured the termite fraction of each edge over time, and underlined a preference for the border of the arena. Thigmotaxis, the preferences of animals for borders and contacts, is well known, especially in stressful situations [24, 6, 5]. However, the preference for borders could

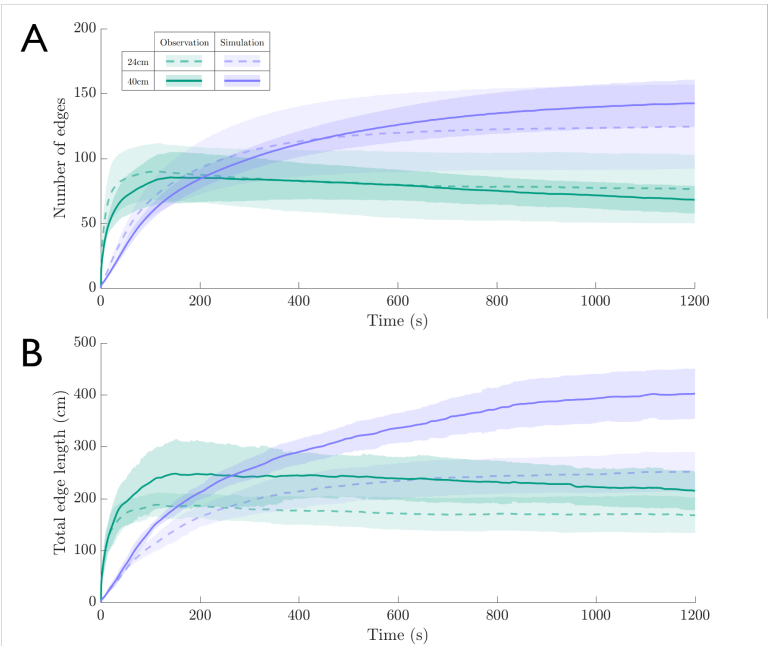


Figure 5: Mean length and number of edges (6 networks for each condition, confidence interval of one standard deviation). Observations in green, and simulations in purple. 24cm arenas with dashed lines and 40cm with solid lines.

also be emerging from the geometry of the arena. So, to assess whether that preference was due to behavioural biases or to the arena's topology, we established a null model simulating termites movements based on turning angles in the existing network. Our null model managed to explain the preference for the borders, without explicitly implementing it. However, the intensity of termites' edge selection was not reproduced. The individual rules we implemented in our null model were not sufficient to reproduce such collective behaviour.

In our null model, agents prefer lower turning angles. Additionally, one can expect that pheromones drive an important role in the turning decisions [21]. Our model is missing the amplification some edge benefits, and pheromones play a key role in the amplification of an individual decision to a collective one [10]. We also showed that the termites' networks total lengths stabilise around 200cm independently of the size of the arena. This fact supports the hypothesis of pheromone trails, as 106 individuals may only sustain a 200cm long pheromone track (considering evaporation rates). The future work should focus on improving the null model with a turning preference based on both angle and pheromone quantity. This next step will be straightforward from our data, as we already implemented pheromones in our model to discriminate between active and inactive edges. Our work would benefit from more pertinent network metrics especially suitable for planar network efficiency. New metrics will allow us to better differentiate our null models from observed collective behaviours over time. The future work should also focus on the survival analysis of edge activation depending on their location, branching, or orientation for instance [29].

References

[1] David Edward Bignell, Yves Roisin, and Nathan Lo, editors. *Biology of Termites: a Modern Synthesis*. Springer Netherlands, Dordrecht, 2011.

[2] J. Buhl, J. Gautrais, R. V. Solé, P. Kuntz, S. Valverde, J. L. Deneubourg, and G. Theraulaz. Efficiency and robustness in ant networks of galleries. *The European Physical Journal B*, 42(1):123–129, November 2004.

[3] Jerome Buhl, Kerri Hicks, Esther R. Miller, Sophie Persey, Ola Alinvi, and David J. T. Sumpter. Shape and efficiency of wood ant foraging networks. *Behavioral Ecology and Sociobiology*, 63(3):451–460, January 2009.

[4] Jérôme Buhl, Jacques Gautrais, Jean Louis Deneubourg, Pascale Kuntz, and Guy Theraulaz. The growth and form of tunnelling networks in ants. *Journal of Theoretical Biology*, 243(3):287–298, December 2006.

[5] E. Casellas, J. Gautrais, R. Fournier, S. Blanco, M. Combe, V. Fourcassié, G. Theraulaz, and C. Jost. From individual to collective displacements in heterogeneous environments. *Journal of Theoretical Biology*, 250(3):424–434, February 2008.

[6] R. P. Creed and J. R. Miller. Interpreting animal wall-following behavior. *Experientia*, 46(7):758–761, July 1990.

[7] J.L. Deneubourg and S. Goss. Collective patterns and decision-making. *Ethology Ecology & Evolution*, 1(4):295–311, December 1989.

[8] J.L. Deneubourg, J.M. Pasteels, and J.C. Verhaeghe. Probabilistic behaviour in ants: A strategy of errors? *Journal of Theoretical Biology*, 105(2):259–271, January 1983.

[9] Marco Dorigo, Vittorio Maniezzo, and Alberto Colorni. The Ant System: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics*, 96(1):29–41, 1996.

[10] Audrey Dussutour, Jean-Louis Deneubourg, and Vincent Fourcassié. Amplification of individual preferences in a social context: the case of wall-following in ants. *Proceedings of the Royal Society B: Biological Sciences*, 272(1564):705–714, April 2005.

[11] Diane Fouquet and Christian Jost. Construction du nid des termites Procornitermes araujoi. Master 2 (Neurosciences, Comportement, Cognition), Centre de Recherche sur la Cognition Animale, Toulouse, 2011.

[12] Alejandro F. Frangi, Wiro J. Niessen, Koen L. Vincken, and Max A. Viergever. Multiscale vessel enhancement filtering. In William M. Wells, Alan Colchester, and Scott Delp, editors, *Medical Image Computing and Computer-Assisted Intervention — MICCAI'98*, Lecture Notes in Computer Science, pages 130–137. Springer Berlin Heidelberg, 1998.

[13] S. Goss, R. Beckers, J. L. Deneubourg, S. Aron, and J. M. Pasteels. How trail laying and trail following can solve foraging problems for ant colonies. In Roger N. Hughes, editor, *Behavioural Mechanisms of Food Selection*, pages 661–678. Springer Berlin Heidelberg, Berlin, Heidelberg, 1990.

[14] Alexander Heyde, Lijie Guo, Christian Jost, Guy Theraulaz, and L. Mahadevan. Self-organized biotectonics of termite nests. *Proceedings of the National Academy of Sciences*, 118(5):e2006985118, February 2021.

[15] P. Jmhasly and R. H. Leuthold. The system of underground passages in Macrotermes subhyalinus and comparison of laboratory bioassays to field evidence of intraspecific encounters in M. subhyalinus and M. bellicosus (Isoptera, Termitidae). *Insectes Sociaux*, 46(4):332–340, November 1999.

[16] C. Jost, I. Haifig, C. R. R. de Camargo-Dietrich, and A. M. Costa-Leonardo. A comparative tunnelling network approach to assess interspecific competition effects in termites. *Insectes Sociaux*, 59(3):369–379, August 2012.

[17] Sang-Hee Lee, Seungwoo Sim, and Hark-Soo Song. Path selection by the termite Reticulitermes speratus kyushuensis (Isoptera: Rhinotermitidae) at a bifurcation node of Y-shaped tunnel. *Journal of Asia-Pacific Entomology*, 19(2):497–501, June 2016.

[18] MATLAB. Matlab 2016a, 2016.

[19] Nobuaki Mizumoto, Paul M. Bardunias, and Stephen C. Pratt. Complex Relationship between Tunneling Patterns and Individual Behaviors in Termites. *The American Naturalist*, 196(5):555–565, November 2020.

[20] Leticia R. Paiva, Alessandra Marins, Paulo F. Cristaldo, Danilo Miranda Ribeiro, Sidiney G. Alves, Andy M. Reynolds, Og DeSouza, and Octavio Miramontes. Scale-free movement patterns in termites emerge from social interactions and preferential attachments. *Proceedings of the National Academy of Sciences*, 118(20):e2004369118, May 2021.

[21] Andrea Perna, Boris Granovskiy, Simon Garnier, Stamatios C. Nicolis, Marjorie Labédan, Guy Theraulaz, Vincent Fourcassié, and David J. T. Sumpter. Individual rules for trail pattern formation in Argentine ants (*Linepithema humile*). *PLoS Computational Biology*, 8(7):e1002592, July 2012.

[22] Andrea Perna, Christian Jost, Etienne Couturier, Sergi Valverde, Stéphane Douady, and Guy Theraulaz. The structure of gallery networks in the nests of termite Cubitermes spp. revealed by X-ray tomography. *Naturwissenschaften*, 95(9):877–884, September 2008.

[23] Andrea Perna, Sergi Valverde, Jacques Gautrais, Christian Jost, Ricard Solé, Pascale Kuntz, and Guy Theraulaz. Topological efficiency in three-dimensional gallery networks of termite nests. *Physica A: Statistical Mechanics and its Applications*, 387(24):6235–6244, October 2008.

[24] Hermann Schöne and Hermann Schöne. *Spatial orientation: the spatial control of behaviour in animals and man*. Princeton series in neurobiology and behavior. Princeton University Press, Princeton, 1984.

[25] Seungwoo Sim and Sang-Hee Lee. Direction selection of termites at a skewed T-shaped tunnel junction. *Journal of Asia-Pacific Entomology*, 20(1):61–64, March 2017.

[26] Guy Theraulaz and Eric Bonabeau. A brief history of stigmergy. *Artificial Life*, 5(2):97–116, April 1999.

[27] J. F. A. Traniello. Recruitment and orientation components in a termite trail pheromone. *Naturwissenschaften*, 69(7):343–345, July 1982.

[28] J F A Traniello. Foraging strategies of ants. *Annual Review of Entomology*, 34(1):191–210, 1989.

[29] Mathilde Vernet, Yoann Pigné, and Éric Sanlaville. A study of connectivity on dynamic graphs: computing persistent connected components. *4OR*, 21(2):205–233, June 2023.

FRCCS

Persistence of Information in Dynamic Graphs

Vincent Bridonneau[1✓], Frédéric Guinand[1] and Yoann Pigné[1]

[1] *LITIS Laboratory, Le Havre Normandy University (France). firstname.lastname@univ-lehavre.fr*
[✓] *Presenting author*

Abstract. An information is present in a dynamic graph at time 0. According to a chosen communication scheme, is this information still present at time T in the graph despite a change in the set of vertices over time? This preliminary work proposes a formal description of this problem and some of its variants. It is shown that for simple cases a polynomial time algorithm can answer the question while remaining open for more complex variants.

Keywords. *Dynamic Graphs, Broadcasting Algorithms, Persistent Information, Information Coverage*

1 Introduction

In recent years, the study of temporal graphs has seen significant advancements. While prior work has laid a foundation for understanding the temporal evolution of graphs [2], this article takes a distinctive approach, centering on the profound challenge of information persistence within dynamic networks. The *information persistence* problem addressed here focuses on how information remains present in a dynamic graph when the set of vertices changes over time. To the best of our knowledge this is the first time this problem is formally defined and addressed. While they may appear similar in appearance, the broadcasting and epidemic spreading problems [5, 6, 7] are, to a large extent, different from the problem described in this work. It can be considered as an extension of the reachability property in temporal graphs.

In the context of a dynamic graph wherein the set of vertices evolves over time, along with a defined communication strategy, the aim is to check whether information persists within the graph at a specified step T, and potentially quantify the proportion of vertices retaining this information.

The objective of this study is to introduce two novel problems relating to this subject. Firstly, in Section 2, we delve into the formulation of the problem. Concurrently, we provide clear definitions of dynamic graphs and broadcasting strategies to mitigate any potential ambiguity. Additionally, we introduce concepts relevant to dynamic graphs, such as dynamic graph generation, to broaden the scope of information persistence study. This section culminates in the formal definition of the *information persistence problem* and the *information covering problem*. Section 3, we propose some preliminary results for a restricted set of instances. In particular we provide algorithms solving the defined problems and show that their time complexity are polynomial. After dealing with these points, a last section (Section 4) is dedicated to open problems this new paradigm offers.

2 Problem Formulation

The *persistence-of-information* problem addressed here focused on how information may spread and remain present in a dynamic graph, while the set of vertices changes over time.

The problem is defined using two parameters: a dynamic graph $G = (G_t)_{t\in\mathbb{N}}$ and a communication strategy A specifying how information spread. Knowing these two information, the question is whether or not there exists a couple of vertices (u, v) such that $u \in V_0$, $v \in V_T$ and u can reach v through a temporal path (a.k.a. time-respecting path TRP) according to the communication strategy.

2.1 Graphs

The parameter G concerns the instance of dynamic graph in which the information will spread. There are no constraints on the choice of an instance of dynamic graph. Note that vertices may appear and/or disappear over time. In terms of set, it means that V_t may be different from V_{t+1} for any $t \geq 0$. Moreover, the information definitively disappears from the graph as soon as no vertex holds it anymore.

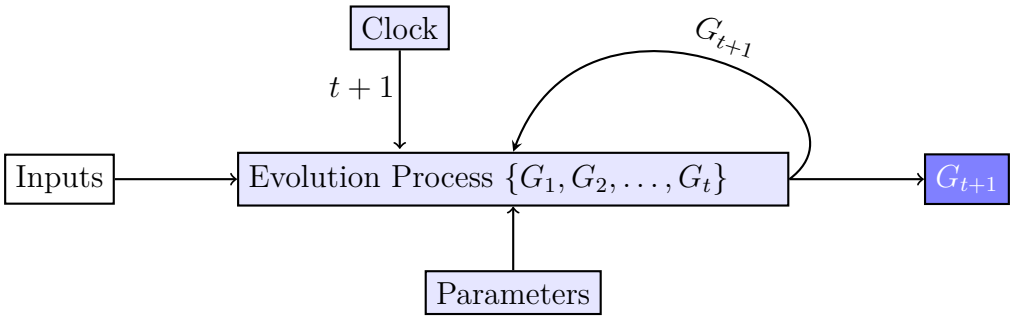
A dynamic graph is defined as a sequence of static graphs ordered by a timestamp :

Dynamic Graph: Let $G_0 = (V_0, E_0), G_1 = (V_1, E_1), \cdots, G_T = (V_T, E_T)$ be a sequence of graphs, then $\mathcal{G} = (G_t)_{0\leq t\leq T}$ is a dynamic graph.

The addressed problem is only significant if the set of vertices changes over time, thus, if there exist $t \in [0, T-1]$ such that $V_t \neq V_{t+1}$. This might be the case either because vertices have been added or removed. From here, it is possible to state two first results. First, if there exists $t \in [0, T-1]$ such that $V_t \cap V_{t+1} = \emptyset$, then the graph is not information persistent as the information will disappear with the loss of all vertices present at time t. Second, if the set of vertices never changes (i.e., if $V_t = V_0$ for all $t \in [0, T]$), then the information never disappear and the graph is *information-persistent*. The information-persistent problem is therefore relevant only if those two cases do not appear. In addition with these two conditions, we also consider the set of initial vertices V_0 and the set of final vertices V_T does not share any vertex ($V_0 \cap V_T = \emptyset$), otherwise the information persistence problem would be obviously solved by looking at $V_0 \cap V_T$.

2.1.1 Generation of Dynamic Graph

We also consider the possibility for dynamic graphs to be produced by a dynamic graph generator. From a general point of view, a dynamic graph generator can be defined as a process with input data, that produces at each time step $t + 1$ a new static graph G_{t+1} from already generated static graphs $\{G_1, \ldots, G_t\}$ and possibly additional information (the parameters of the generator). Thus, the output of a dynamic graph generator is a flow of static graphs identified by time stamps. The time stamps may also correspond to events, and in such a case, the time interval between two time stamps may be different. However, in this report, for sake of clarity, we consider integer time stamps. If the flow stops, for whatever reason (e.g. clock has been stopped, evolution process is finished) at step T, the set of generated static graphs $\{G_1, G_2, \ldots, G_T\}$ corresponds to a temporal network (TN).



As an illustration, the Barabasi-Albert model [1], Edge-Markovian Graphs [4] and Degree-Driven Dynamic Geometric Graph Generators (D3G3) [2] are instances of dynamic graph generators. The Barabasi-Albert model can be seen as a generator of dynamic graph as it is parameterized by an initial configuration and then, the transition rules are described by the preferential attachment mechanism. The second one is a generative model with transition rules based on Markov Chains. The last one is also a generative process. It is modeled by defining two transition rules driving the evolution of the graph between two consecutive time steps. The application of the rules in this model leads to creation and deletion of vertices in the graph.

The two first models does not consider the possibility of deleting vertices. Thus, the information persistent problem is irrelevant for these models as information never disappears once introduced. This is not the case for the last model as it allows vertices to disappear. Therefore, for this last case it is worth studying the information persistence as it is not guaranteed. In the sequel this generator will be used as our dynamic graphs instances generator. Rules are based on node degrees only and rely on a random generator for positioning new nodes on the 2D euclidean space. This leads to the name of the generator: *Degree-Driven Dynamic Geometric Graphs Generator* or D3G3.

Degree Driven Dynamic Geometric Graph Generator:
An instance of D3G3 is defined by an initial graph, a set of parameters and two rules:

- $G_0 \neq (\emptyset, \emptyset)$ the seed graph,
- parameters:
  - $d \in (0, \frac{\sqrt{2}}{2})$ (distance threshold for connection),
  - S_S a set of non-negative integers
  - S_C a set of non-negative integers
- rules applied on G_t leading to G_{t+1}:
  - if $v \in V_t$, then $v \in V_{t+1}$ if and only if $\deg(v) \in S_S$ (conservation rule)
  - if $v \in V_t$ and if $\deg(v) \in S_C$ then add a new vertex to V_{t+1} with a random position in the unit-torus (creation rule)

At a given time step, two vertices are connected if and only if their euclidean distance is lower than d. Graph evolution between two consecutive time steps t and $t+1$, is driven by two rules applied to each vertex $v \in V_t$ simultaneously. The first rule determines for a vertex $v \in V_t$ whether it is kept at step $t+1$ while the second rule concerns the possibility for a vertex $v \in V_t$ to create a new vertex in V_{t+1} according to its degree.

2.2 Sustainability of Dynamic Graphs

The study of dynamic graphs where the set of vertices may change over time raises questions about the evolution of the graph. Among others, a notion called sustainability, introduced in [2] shed the light on one of these questions. Indeed, some generator models may produce dynamic graphs that becomes static or empty. Sustainability qualifies a dynamic graph that never becomes null or periodic (which includes static) and is defined as follows:

Graph sustainability: A dynamic graph $\mathcal{G}$ is said sustainable if both Condition 1 and Condition 2 are not verified.

$$\begin{aligned} &\text{Condition 1:} \quad \exists T \in \mathbb{N}, \forall t \geqslant T, G_t = (\emptyset, \emptyset) \\ &\text{Condition 2:} \quad \exists T \in \mathbb{N}, \exists k \in \mathbb{N}^*, \forall t \geqslant T, G_t = G_{t+k} \end{aligned}$$

The question of information persistence complements the analysis of dynamic graph sustainability. Indeed, in the subsequent sections of this document (see Section 4.2), we prove that dynamic graphs can be sustainable without necessarily being information persistent. Therefore, information persistence is an additional property that can be studied alongside dynamic graph analysis, offering new perspectives.

2.3 Communication Strategy

The parameter A concerns the communication strategy. In this work, we restrict the communication policies to local broadcasting strategies only. Thus, when communicating at time t, a vertex sends the information to all its neighbors connected to it at time t. The broadcasting strategy describes the way information spread between vertices. Many works have been devoted to this problem, especially in the domain of mobile ad hoc networking [8]. But in most studies, the set of vertices remains the same all the time. However, whatever the case, changing or unchanging vertices set, the strategy must specify the conditions for sending a message to connected neighbors. For instance, a minimum delay of one time step might be mandatory on a vertex between the reception of the information and its transmission to its neighbors. Some strategies may select vertices to which transmit the information within the neighborhood, or may allow only a restricted number of transmissions. All these points have to be clearly defined. In this work we restrict our study to two algorithms and remarks are made to highlight relevant questions.

2.3.1 Constant Flooding

The first algorithm to be discussed is a variant of the flooding algorithm. We call it constant flooding. The principle is that once a vertex receives the information, it keeps transmit it to its neighbors as long as it is present in the graph. The constant flooding algorithm is defined as follows:

Constant Flooding Algorithm (CF): Let G be a dynamic graph. Let u be a vertex in this graph. Let t be the date at which u receives the information. Then as long as u remains in the graph, u transmits the information to its neighbors at every step $t' > t$.

It is important to notice that if a vertex receives information at step t it starts its transmission from the next time step, at $t + 1$.

2.3.2 Simple Flooding

The strategy discussed here is known as simple flooding algorithm. The principle is the following: once a vertex receives the information, it is allowed to send it to its neighbors only once.

A delay of one time step has to be observed between the reception and the emission. In the current work this algorithm only waits one time step before the transmission of the information to its neighbors, however it could be possible to consider other variants of this algorithm for which the transmission could be done later. The algorithm is defined as follows:

Simple Flooding Algorithm (SF): Let G be a dynamic graph. Let u be a vertex in this graph, meaning there exists t such that $u \in V_t$. Let t' be the first date such that u receives the information. Then u sends the information to its neighbors at $t' + 1$ only.

Note that the transmission might be done even if the vertex has no neighbors connected.

2.4 Studied Problems

In this part, we define notions and set formalism aiming at studying and defining the information persistence and the information coverage problems. For the rest of this part, we assume that a broadcasting strategy A and a dynamic $\mathcal{G}$ are defined. The first step is to introduce the notion of *reachability* between two vertices $u \in V_0$ and $v \in V_T$ as it plays a key role in the definition of the two problems.

Reachability: Let $u \in V_0$ and $v \in V_T$. We say that u can reach v using the broadcasting strategy A, and we note $u \xrightarrow{A} v$, if v can receive information from u according to the broadcasting strategy A.

Note that the reachability implies that there exists a time-respecting path between u and v, satisfying the condition implied by the broadcasting strategy A.

The reason why the reachability only concerns the initial and the final step is because it is not necessary to define it for intermediate steps. However, if the information is not introduced at a step later than 0, it may be relevant to define the reachability according to the date u sends the information and the date v receives it.

From the definition of reachability it is possible to define two decision problems. The first one is defined as the capability for a dynamic graph to exhibit a vertex $u \in V_0$ and another vertex $v \in V_T$ such that $u \xrightarrow{A} v$. This problem will be referred to as the *information persistence problem* in this document. Formally, this problem can be defined as follows:

Information Persistence Problem: Let $\mathcal{G} = (G_t)_{0 \leq t \leq T}$ be a dynamic graph and let A be a broadcasting strategy. Then, $\mathcal{G}$ is said to be A-persistent if it satisfies:

$$\exists u \in V_0, \exists v \in V_T, u \xrightarrow{A} v \tag{1}$$

The second problem defined in this work is called the *information coverage problem*. It is the capability for a dynamic graph to exhibit for every vertex $v \in V_T$, a vertex $u \in V_0$ such that $u \xrightarrow{A} v$. This means there exists a subset $S \subset V_0$ such that for every vertex $v \in V_T$, there exists $u \in V_0$ such that u can reach v using the strategy A. Formally, this problem is defined as follows:

Information Coverage Problem: Let $\mathcal{G} = (G_t)_{0 \leq t \leq T}$ be a dynamic graph and let A be a broadcasting strategy. Then, $\mathcal{G}$ is said to be A-coverable if it satisfies:

$$\forall v \in V_T, \exists u \in V_0, u \xrightarrow{A} v \tag{2}$$

Now that both problems have been defined, the remainder of this document is dedicated to their study. As a first step, an analysis of the two problems will be presented assuming the

dynamic graph is known. An algorithmic study will show that the problems can be solved in polynomial time by conducting a simple simulation of information spreading within the graph. Subsequently, a link will be established between the persistence of information and the processes of generating dynamic graphs. We demonstrate in this case that sustainability alone is not sufficient for information to persist within a graph and attempt to identify conditions under which the generated graphs exhibit information persistence in the case of the D3G3 model.

3 Remarks and First Results

The main result is an algorithm to solve both the information persistence problem and the information coverage problem when the broadcasting strategy is CF (constant flooding). Its time complexity is studied and is shown to be polynomial. This algorithm takes as an input a dynamic graph $\mathcal{G}$ and a set of vertices $I_0 \subset V_0$ having information at date 0. This algorithm depends on the broadcasting strategy studied and simulates the spread of information.

Algorithm 1 *Spreading*$(\mathcal{G}, I_0)$

Require: $\mathcal{G} = (G_t)_{0 \leq t \leq T}$ a dynamic graph, $I_0 \subset V_0$ set of vertices having the information at date $t = 0$.

Ensure: $I_T \subset V_T$ set of vertices receiving information from vertices in I_0 or $\emptyset$ if no such vertices exist.

```
1: I ← I_0
2: for t ← 1 to T do
3:     I ← I ∩ V_t
4:     if I = ∅ then
5:         return ∅
6:     end if
7:     for (x, y) ∈ E_t do
8:         if x ∈ I and y ∉ I then
9:             I ← I ∪ {y}
10:        else if x ∉ I and y ∈ I then
11:            I ← I ∪ {x}
12:        end if
13:    end for
14: end for
15: return I
```

With this algorithm, it is possible to answer the question of the persistence problem. Indeed, it is sufficient to apply this algorithm with the whole set of initial vertices $I_0 = V_0$. If the algorithm ends returning a non-empty set, then, $\mathcal{G}$ is A-persistent. With the same idea, it is possible to answer the information coverage problem. If the result of the spreading algorithm with $I_0 = V_0$ is V_T, then it means every vertex in V_T can be reached by at least one vertex in V_0.

3.1 Time Complexity of the Spreading Algorithm

Let us now study the time complexity of algorithm 1. The goal is to prove that the time complexity of this algorithm is $O\left(\sum_{t=1}^{T} n_t m_t\right)$, with $m_t = |E_t|$. This comes from the complexity

of the most nested *for* loop. In the worst case, the time complexity of set operations are bounded by the size of the set. Here, the considered set is $I \subset V_t$ for any given $t \in [1, T]$. Therefore, the time complexity of the lines 7–13 is $O(m_t \times n_t)$. As it is the biggest time complexity of the first *for* loop (lines 2–14), the time complexity of the whole algorithm is thus $O\left(\sum_{t=1}^{T} n_t m_t\right)$.

As mentioned above, this algorithm solves both the information persistence problem and the information coverage problem. We can therefore deduce that the complexity of these two algorithms is $O\left(n_0 + \sum_{t=1}^{T} n_t m_t\right)$, where the term n_0 comes from the construction of I_0, a copy of V_0. Thus, we have established the existence of algorithms solving the information persistence and the information coverage problems in polynomial time.

4 Questions and Open Problems

4.1 Questions Related to the Simple Flooding Algorithm

The reader may have noticed that the previously defined algorithm is not convenient if the node has no neighbor at the moment of transmission. It would be interesting to postpone the transmission at another date in order to improve the performances of the process from an information-persistence point of view. However, considering such a possibility raises many questions.

First, it could be possible to remove the delay between the reception of the information and its transmission to neighbors. This situation is similar to the notion of non-strict path in temporal graphs. In that case, given a connected component, as soon as one vertex receives the information, then all the vertices of that connected component also receive and send the information. As a consequence, all these nodes will never send the information at later dates. This means that only the vertices present at step 0, if the information is introduced at date 0, will have the information. This means that some vertices must exists both at step 0 and T to ensure the persistence of information.

In an opposed direction there are some ways that takes into account the possible future neighborhood of the node to estimate the moment when the information could be send. Indeed, another way to define the moment to send the information would be to wait until the neighborhood is not empty. This question is not treated in this document, however it may offer interesting wondering. For instance, it is possible to study questions such as:

- When should a given node send the information it owns so that information persistency is guaranteed.

A last question concerns the possibility for a node to transmit the information several times. The defined algorithm does not allow multiple transmission: once the information has been spread by one vertex, this vertex does not transmit it again. It is possible to imagine some applications where the vertex can receive the information several times. Every time the information is received, the vertex will transmit it again to its neighbors. This defines a new algorithm and the same questions as presented above may be addressed for this new strategy.

4.1.1 Connection between Simple Flooding and Constant Flooding

One final aspect to address here is the connection between simple flooding and constant flooding strategies. Specifically, it is observed that if a dynamic graph is SF-persistent, then it is also CF-persistent. To illustrate this, it suffices to note that if a dynamic graph $\mathcal{G}$ is SF-persistent,

then there exist vertices $u \in V_0$ and $v \in V_T$ such that information is transmitted from u to v via a temporal path using the simple flooding communication strategy. However, this path is also observable using the constant flooding communication strategy. Indeed, for propagation with a simple flooding strategy, information can only be transmitted once it has been received, whereas in the case of a constant flooding strategy, information is continuously transmitted after being received. Thus, in the scenario where the communication strategy is constant flooding, the information can indeed follow the same path between u and v as when the communication strategy is simple flooding.

4.2 Sustainability and Information Persistence Problem

This section focuses on studying the problem of information persistence from the perspective of dynamic graph generators. Here, we assume that graphs are the product of a generation mechanism as defined earlier. One of the initial observations is that if a generator produces A-persistent graphs for any diffusion strategy A, then the graphs do not become empty, which is a characteristic of sustainability. However, the converse is not necessarily true. There are cases where the produced graphs are sustainable without being information persistent.

For instance, considering the previously defined D3G3 generation model, it is possible to find parameter values such that the produced graphs are sustainable with high probability without being information persistent. For instance, if the sets S_S and S_C are both equal to 0 and the chosen parameter d is small enough, then the produced graphs are likely to be sustainable. This has been proved in [2]. However, if we expect the condition $V_0 \cap V_T = \emptyset$ to be satisfied (no initial vertices are still present at a given time T), then the information will have vanished from the dynamic graph.

To understand this, it is essential to understand that $S_S = S_C = 0$ implies that only isolated vertices are retained and can generate new vertices in the graph. If a new vertex u connects to an isolated vertex v, then both u and v will disappear in the subsequent time step. Furthermore, only isolated vertices can retain information in the graph, as any other vertex disappears along with the information it carries. Hence, we deduce that the produced graphs when $S_S = S_C = 0$ are not A-persistent for any considered strategy A.

5 Conclusion

This study has introduced the issue of information persistence in dynamic graphs. The aim is to determine, for a given dynamic graph or a family of such graphs, and a fixed communication strategy, whether information introduced into the graph is likely to persist within it and potentially propagating and reaching all vertices of the graph at a given date. In this article, we have proposed algorithmic solutions to address these problems. These algorithms have been demonstrated to have polynomial time complexity. We have also shown that information persistence implies sustainability (the ability of a graph to avoid becoming empty or periodic), but the converse is not necessarily true. This was evidenced by demonstrating that certain families of graphs produced by generation mechanisms do not guarantee the existence of information within the graph despite being sustainable. Furthermore, discussions and conjectures have been raised regarding communication strategies. Specifically, we observed that if a graph is SF-persistent, then it is also CF-persistent. To conclude, this work lays a foundational step in the study of information persistence within a dynamic graph.

Several perspectives for future research in the field are possible. For instance, regarding the

information coverage problem, one could explore the minimum size of the initial vertex set required to reach all final vertices. In another direction, still linked to the information coverage problem, investigating the relationship between the number of information carriers at time t and the number of vertices present at that same time could yield valuable insights.

6 Acknowledgments

The authors disclosed receiving the following financial support for the research, authorship, and/or publication of this article: Supported by the French ANR, project ANR-22-CE48-0001 (TEMPOGRAL)

References

[1] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999. arXiv:cond-mat/9910332.

[2] Vincent Bridonneau, Frédéric Guinand, and Yoann Pigné. *Dynamic Graphs Generators Analysis : an Illustrative Case Study.* report, LITIS, Le Havre Normandie University, 2022.

[3] Arnaud Casteigts, Paola Flocchini, Walter Quattrociocchi, and Nicola Santoro. Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems*, 27(5):387–408, October 2012. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17445760.2012.668546.

[4] Andrea E. F. Clementi, Claudio Macci, Angelo Monti, Francesco Pasquale, and Riccardo Silvestri. Flooding Time of Edge-Markovian Evolving Graphs. *SIAM Journal on Discrete Mathematics*, 24(4):1694–1712, January 2010.

[5] C. Gkantsidis, M. Mihail, and A. Saberi. Hybrid search schemes for unstructured peer-to-peer networks. In *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, volume 3, pages 1526–1537, Miami, FL, USA, 2005. IEEE.

[6] Marcelo Kuperman and Guillermo Abramson. Small world effect in an epidemiological model. *Physical Review Letters*, 86(13):2909–2912, March 2001. arXiv:nlin/0010012.

[7] M. E. J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):016128, July 2002.

[8] Brad Williams and Tracy Camp. Comparison of broadcasting techniques for mobile ad hoc networks. In *Proceedings of the 3rd ACM international symposium on Mobile ad hoc networking & computing*, pages 194–205, 2002.

Temporal Connectivity of Maritime Transport Networks

Théo Morel[1✓], Niels Kerné[1], Yoann Pigné[1] and Claude Duvallet[1]

[1] *Université Le Havre Normandie, LITIS UR 4108, F-76600 Le Havre ;*
{theo.morel, niels.kerne, yoann.pigne, claude.duvallet}@univ-lehavre.fr
[✓] *Presenting author*

Abstract. This abstract outlines preliminary work on reconstructing a maritime transportation network using AIS data. We propose a modeling approach using temporal graphs to represent the network's dynamics. We demonstrate the utility of the model with an analysis of connections between ports, considering arrival dates of ships, and travel durations. We show that this temporal approach better highlights the maritime network properties.

Keywords. *Maritime Transport Network; AIS; Dynamic Graphs*

1 Introduction

Maritime transportation is a crucial element of the global economy. Indeed, 90% of global trade is facilitated through maritime transport. Commercial ships serve as the primary means of exchange, carrying goods from one point to another. The routes taken by these ships, the ports they visit, and the pathways they traverse represent essential components of the global maritime network. The study of this network, including its properties and evolutions, constitutes the focus of numerous researchers.

The primary challenge in this endeavor lies in reconstructing the network using real data. Access to departure and arrival data of ships in ports is not as universally available as it is for the tracking of civilian commercial aviation, notably through the Automatic Dependent Surveillance-Broadcast (ADS-B) protocol.

Some operators, like insurers, hold this information. Another common approach is utilizing Automatic Identification System (AIS) data emitted by ships for real-time tracking. These data serve purposes from maritime safety to scientific research [5]. However, their effectiveness is limited by antenna range, being more accurate in high-traffic zones and less so in remote areas. Satellite-based AIS data exist but are costly and not always accessible to researchers. On a broader scope, satellite imagery, like Sentinel-1 from the European Space Agency, can detect ships [8] and estimate their speed and trajectory.

The significance of studying ship networks lies in the fact that various short-term socio-economic and meteorological events are immediately reflected in the data constituting the network. Therefore, there is merit in investigating these networks to extract their properties and evolutions, thereby deducing information about the events that generated them.

Network analysis is a valuable tool across various disciplines, including maritime transport. It

helps understand network structures, dynamics, and their impact on disease spread (see [4] for an extensive review).

Maritime networks evolve as ships move, port traffic shifts, and routes change. Analyzing these networks demands considering their temporal dimension. Temporal graphs, powerful tools for such analysis [7], unveil network evolutions and port visitation changes. In this abstract, we prioritize constructing temporal graphs from AIS data to extract insights into maritime transportation networks. We discuss the construction steps, models, and analyses, stressing the temporal aspects' significance. Additionally, we showcase a study on temporal connectivity between ports, highlighting its importance in understanding these networks.

2 Network Reconstruction

In this study, AIS data are utilized to reconstruct the maritime transportation network. AIS data are transmitted by ships via AIS transponders using VHF radio frequencies. These signals can be intercepted by ground-based antennas within a radius of several tens of kilometers or by satellites, and then relayed to receiving stations. While freely accessible, AIS data require interpretation. To access a broad coverage of AIS data, we participate in a network of receiving stations that share their AIS data (AISHUB [1]). Additionally, we supplement our dataset with other sources of freely available AIS data (NOAA [3] and DMA.DK [2]).

In AIS data, each message includes vessel details like identifier, position, speed, and navigation status. We focus on container ships (`Ship_Type` 70-79). Navigation status, such as "anchored" or "under way," helps understand ship movements and identify visited ports. This data aids in pinpointing ships at quays and associating them with ports. It also helps detect waiting periods offshore, in waiting areas. They refer to maritime areas where ships can wait outside ports for various reasons such as capacity issues at the port, delays in loading or unloading cargo, adverse weather conditions, security concerns or regulatory issues. Figure 1 demonstrates the detection of waiting zones using AIS data.

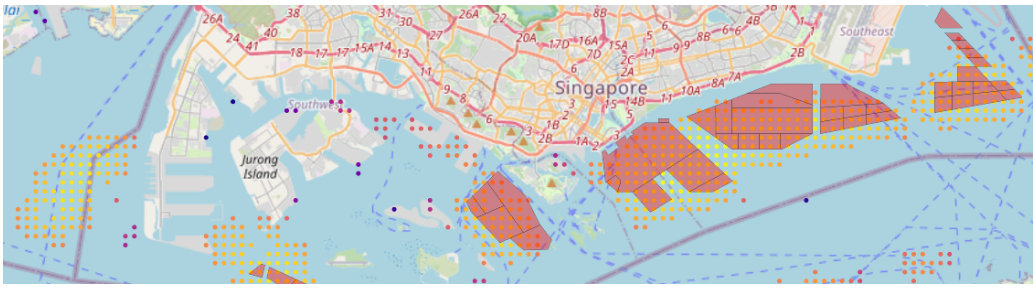


Figure 1: Detection of waiting zones from AIS data (GPS positions discretized in fractions of degrees).

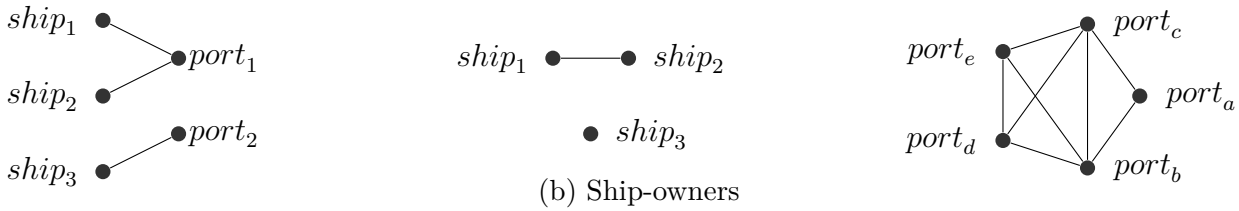
Incorporating waiting zones into the modeling of the maritime transport network is essential for understanding ship movement patterns, assessing potential waiting times, and identifying factors that can influence the maritime traffic.

Furthermore, the consideration of the different operators involved may help to analyze port performance as each operator may have its own policies, procedures, equipment and capacities, which can significantly impact how a port operates and its ability to efficiently handle maritime traffic.

An initial version [6] of the reconstructed data is freely accessible on the Zenodo platform. This

version offers tabulated data in triplets of date, ship, and port, enabling the identification of ship routes between ports.

Once the data has been reconstructed, the next step involves choosing a suitable graph model for representation. Various representations can accentuate different network properties as we can see in the Figure 2 that depicts three potential representations: a bipartite network featuring ports and ships as nodes (Fig. 2a), a ship network establishing connections between ships that have visited the same ports (Fig. 2b), and a port network linking ports visited by the sale ships (Fig. 2c). These representations complement each other, offering insights into a range of network properties.



(a) Ship-owner/port

(b) Ship-owners

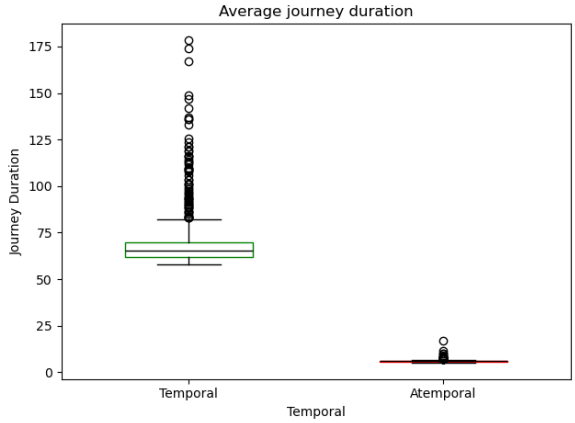
(c) Port centrality

Figure 2: Various representations of maritime transport networks.

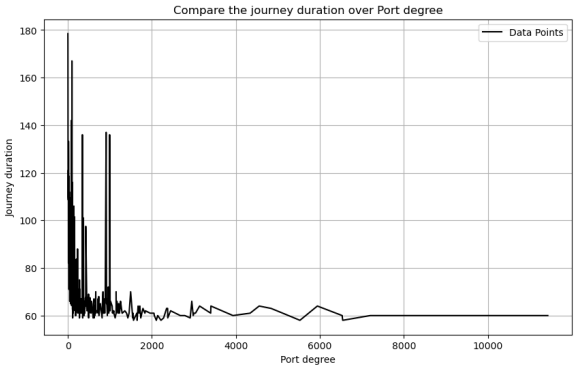
3 Port Centrality Analysis

In this analysis, we used port centrality modeling to explore its effects within both atemporal and temporal contexts. Within atemporal context, edges persist indefinitely and are characterized by the median duration of travel. Conversely, within the temporal context, edges are accessible from the source node at each departure date until reaching the median duration of travel added to the respective departure date. Paths using these edges are known as Journeys [9], which are instrumental in computing three primary objectives: Shortest (minimizing the number of edges), Fastest (minimizing duration) and Foremost (minimizing arrival date). Nodes are represented by ports while edges represent instances where at least one ship had traversed the route between two ports, additional data about the travel as departure and duration is stored as labels on the edge.

We investigated the analysis of journeys with the foremost objective, as it aligns with the notion that a container ship aims to unload its goods at the destination port at the earliest opportunity. Then, we compared these findings with a basic shortest path algorithm.



(a) Average Journey Duration



(b) Journey Duration over Port Degree

Figure 3: Port centrality over atemporal and temporal context

In the Figure 3a, we see the average journey durations in days for both the temporal and atemporal versions. The atemporal rendition encompasses journeys of significantly shorter duration, attributable to the absence of waiting time for departure, thus offering an optimistic perspective on travel time compared to the temporal version. In the Figure 3b, the average journey durations are depicted relative to a port's degree. It is observed that ports with higher connectivity exhibit shorter journey durations. This phenomenon can be elucidated by the increased likelihood of such ports being connected to a hub or serving as one themselves, thereby affording access to more favorable departure options and durations.

Acknowledgements The authors would like to thank the Normandy Region for supporting this work.

References

[1] AISHub. `https://www.aishub.net/`. Accessed: 2024-03-01.

[2] Danish Maritime Authority Navigational information. `https://dma.dk/safety-at-sea/navigational-information/download-data`. Accessed: 2024-03-01.

[3] NOAA ais data. `https://marinecadastre.gov/AIS/`. Accessed: 2024-03-01.

[4] N. G. Álvarez, B. Adenso-Díaz, and L. Calzada-Infante. Maritime Traffic as a Complex Network: A Systematic Review. *Networks and Spatial Economics*, 21(2):387–417, 2021.

[5] S. Benaïchouche, C. Le Goff, Y. Guichoux, F. Rousseau, and R. Fablet. Unsupervised Reconstruction of Sea Surface Currents from AIS Maritime Traffic Data Using Learnable Variational Models. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4100–4104, June 2021.

[6] C. Duvallet, N. Kerné, T. Morel, and Y. Pigné. Date-ship-port triplet data based on ais data. `https://zenodo.org/records/10783698`, March 2024.

[7] P. Holme and J. Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, October 2012.

[8] F. S. Paolo, D. Kroodsma, J. Raynor, T. Hochberg, P. Davis, J. Cleary, L. Marsaglia, S. Orofino, C. Thomas, and P. Halpin. Satellite mapping reveals extensive industrial activity at sea. *Nature*, 625(7993):85–91, 2024.

[9] B. B. Xuan, A. Ferreira, and A. Jarry. Computing shortest, fastest, and foremost journeys in dynamic networks. *International Journal of Foundations of Computer Science*, 14(02):267–285, April 2003.

FRCCS

Towards Balanced Information Propagation in Social Media

Mahmoudreza Babaei[1,2✓], Baharan Mirzasoleiman[3], Jungseock Joo[3], and Adrian Weller[4]

[1] *GISMA University of Applied Science, Germany, mahmoudreza.babaei@gisma.com.*
[2] *Max Planck Institute for Human Development, Germany; babaei@mpib-berlin.mpg.de.*
[3] *University of California, Los Angeles, California, United States; baharan@cs.ucla.edu, jjoo@comm.ucla.edu.*
[4] *University of Cambridge and The Alan Turing Institute, England; aw665@cam.ac.uk.*
[✓] *Presenting author*

Abstract. As people increasingly rely on social media platforms such as Twitter to consume information, there are significant concerns about the low diversity of news consumption. Users may be exposed to posts which reinforce their pre-existing views, which could lead to a more fragmented society. For example, one worry is reinforcing societally damaging misinformation prevalent in one subgroup, such as COVID-19 denial. Aiming to combat this, earlier work divided news stories into *high consensus* and *low consensus* posts, based on the degree to which reactions can be expected to be similar from users with different political views.
In this work, we propose and quantify the benefits of a strategy to *spread* high consensus news across readers with diverse political leanings. We first compile a dataset and make the following three key observations: (1) low consensus news is more likely to remain within subgroups of users with similar political leanings, whereas high consensus news spreads broadly to many users across subgroups; (2) high consensus news posted by neutral publishers spreads widely and more equally across subgroups; and (3) users that get the information from other users instead of the publishers, get an even more biased exposure to news.
We propose a strategy that spreads high consensus news through neutral publishers, and quantify the significant decrease in the disparity of users' news exposure.
Our extensive experiments on Twitter shows that seeding high consensus information with neutral publishers is an effective way to achieve high spread to many users with little disparity regarding political leaning.

Keywords. *Consensus; News consumption in social media; Filter bubble.*

1 Introduction

People increasingly rely on social media platforms, such as Facebook and Twitter, to receive news and information [31, 47]. Many news stories are divisive, often posted by polarized publishers, eliciting different reactions from users with different political leanings or pre-existing views, e.g., conservatives or liberals. While various news sources publish high and low consensus news that cover a given story, users often limit themselves to the low consensus divisive stories which can reinforce their prior views [27, 4, 16]. This selective exposure and consumption of divisive information may lead to a more politically fragmented, less cohesive society [13] and

the formation of filter bubbles or echo chambers [8, 11, 19, 36].

For example, only 12% of Conservative Fox News viewers believe that climate change is man-made. This compares with 60% of Americans overall, recognizing that 28% of Conservative don't watch Fox News.[1] Americans who relied on Fox News, or similar right-wing sources, often do not believe the coronavirus threat is serious.[2] Social media platforms such as Facebook and Twitter can amplify segregation in information consumption with recommendation algorithms that incentivise people towards more extreme positions [7].

Due to concerns about information segregation and societal polarization [45, 42], prior works have proposed exposing users to diverse news stories by nudging users to read other views [34, 37]. While this could be useful for encouraging debate in society, such approaches have been shown to *increase* the chance that users reject stories from other perspectives – perhaps because they believe other publishers and their stories are biased – thereby defeating the purpose [5, 33, 35, 46, 53, 6].

However, social media platforms receive so many posts that some form of moderation is inevitable.[3] We present one proposal to be considered: highlighting high consensus news that elicits similar responses from both sides could act as a soothing balm, helping people to connect with others, even if they have ideological differences. Babaei et al. [4] proposed such a complementary approach to increase diversity in users' information consumption by identifying high consensus, yet interesting information. Their system recommends high consensus "purple" posts to both "red" (conservatives) and "blue" (liberals) users, hoping to increase users' exposure to cross-cutting news posts, leading to lower societal polarization and lower *segregation* in information consumption [15]. Nevertheless, it still remains unclear how such information is spread across users in a network and how individuals choose to react to it. Vraga and Bode [49] claim that if users receive even one piece of news from reliable news sources about health, it has a significant effect on them. Therefore, exposing people to high consensus news related to important events such as COVID-19 may improve health and harmony across society.

In this paper, we investigate users' willingness to share and spread such posts, along with the reach of high and low consensus news stories across a diverse audience. We also examine the newsworthiness [20, 51] of both high and low consensus news. Fundamentally, we ask if we can help to propagate high consensus stories broadly across society? We hope this can help to break undesirable echo chambers, leading to a less polarized, healthier society.

Note that due to the huge number of posts, social media platforms must inevitably engage in some form of story curation or promotion. We propose one approach which we believe can benefit society. The problem of *identifying* high consensus news automatically has already been addressed by Babaei et al. [4]. We assume that platforms are good at selecting stories which are likely to prove popular, hence we focus on identifying a good set of stories from which to select. We highlight the following contributions:

I. We compile a novel dataset, which reveals how Twitter users with similar or different political leanings are connected to each other. To do so, we consider a dataset of 400 news tweets posted by 10 publishers containing 100 high and 100 low consensus posts

[1] https://www.insider.com/fox-news-republican-viewers-reject-climate-change-science-2019-3

[2] `https://www.washingtonpost.com/lifestyle/media/the-data-is-in-fox-news-may-have-kept-millions-from-taking-the-coronavirus-threat-seriously/2020/06/26/60d88aa2-b7c3-11ea-a8da-693df3d7674a_story.html`

[3] https://onezero.medium.com/the-moderation-war-is-coming-to-spotify-substack-and-clubhouse-9fe00672091b

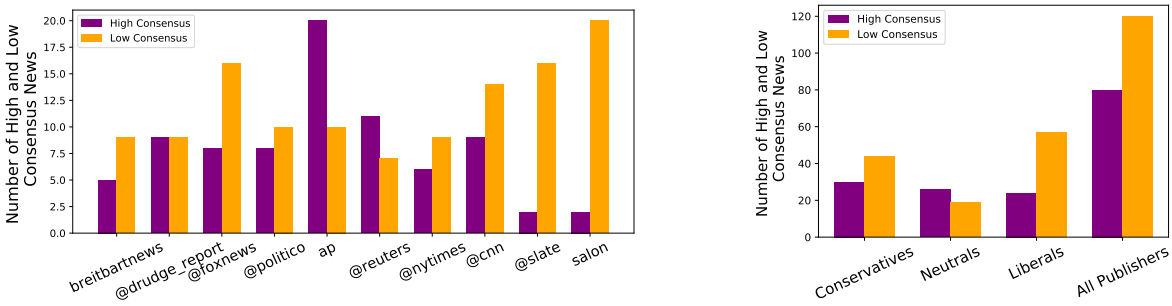
[4]. For every high or low consensus news post, we collected a subset of its 100 random retweeters and for each retweeter we collected a random set of their 100 followers. We compute the political leaning of the 1,616,000 users who either retweeted a high or low consensus news story or were exposed to it. We collected 40 high and low consensus news related to COVID-19 and global warming. Moreover, to simulate the spread of news in Twitter, we crawl a network of more than 100 million Twitter users. This allows us to compute the political leanings of 69,687 users connected by 2,907,026 links.

II. Using our dataset, we study how individuals with different political leanings get exposed to and retweet high and low consensus news posted by users from various political perspectives. We observe that low consensus news tends to proliferate primarily only amongst users with a particular political learning. Gruzd and Mai [24] show how one single tweet on COVID-19 denialism circulated among many conservatives. In contrast, high consensus news has a higher chance of spreading through the entire network. Importantly, high consensus news posted by a set of neutral publishers spreads more equally across liberal and conservative users than if posted by the same number of a mix of non-neutral publishers.

III. Based on the above observations, we propose a strategy that seeds neutral publishers to expose roughly equal fractions of people with different political leaning to high consensus news with the minimum cost (hoping this may help to break echo chambers which can trap users). We show that our proposed strategy is more effective than seeding the most influential nodes without taking the political leanings into account.

Our work provides new insights and a complementary tool which may help to reduce echo chamber, encourage healthier interaction between population subgroups, and lead to a more cohesive society.

2 Related Work



(A) (B)

Figure 1: Number of high and low consensus news posted on Twitter during 9th-15th May, 2017. (A) shows the number of high and low consensus news for 10 selected publishers, and (B) shows the aggregated result for conservative, liberals, and neutral publishers.

We review related work on diverse and polarized news dissemination, and information propagation in social networks.

2.1 News Consumption Polarization on Social Media

Several recent studies have investigated the dissemination of news in social networks [9], focusing on biases [14], political news [2], and the characteristics of spreaders [28].

Traditionally, professional news organizations played a major role in spreading news by selectively presenting news stories to citizens [43]. Accordingly, news media had a high impact on political issues and public opinions [23, 16]. Several works have focused on understanding how and to what extent news media outlets can impact people and society, such as the White Helmets in Syria [44] and the 2016 US presidential campaign [41, 39].

By examining cross-ideological exposure through content and network analysis, Himelboim et al. [27] showed that political talk on Twitter is highly partisan and users are unlikely to be exposed to cross-ideological content through their friendship network. Other studies also report similar findings such as users' higher willingness to communicate with other like-minded social media users [32].

To understand the political bias in social media better, many researchers have studied political polarization on Twitter by analyzing different groups' behavior. Conover et al. [17] showed that Twitter users usually retweet the users who have the same political ideology as themselves, making the retweeting network structure highly partitioned into left- and right-leaning groups with limited connections between them.

Previous work have mostly investigated news media political bias, and the bias introduced in the content of the news, by different methods such as crowdsourcing and machine learning [12, 21, 4]. Recently, a complementary approach was proposed by Babaei et al. [4], in which the goal is to inject diversity in users' information consumption by identifying high consensus yet informative news, based on using features such as the publishers' political leaning.

Babaei et al. [4] showed that high and low consensus posts are equally popular and cover broadly similar topics. However, their study did not investigate how low and high consensus news spread through social media, and their potential impacts on readers biased exposure, which are the main concerns of our paper.

In this work, we first investigate how low and high consensus news spread through Twitter. Then we study the effect of spreading high consensus news through users with different political leanings on decreasing the disparity in users' exposure. We show that seeding high consensus information with neutral individuals is the most effective way to achieve high spread with little disparity regarding political leaning.

2.2 Information Propagation in Social Networks

The process of increasing information propagation and network diffusion by identifying and choosing the optimal set of individuals that utilize social influences to maximize adoption or reception of information in society has been studied widely [22, 40, 29, 25]. The effectiveness of these strategies is studied by Kempe et al. [29] under different social contagion models such as Linear Threshold (LT) and Independent Cascade (IC) models.

The goal of our paper is to propagate important and reliable news with less disparity amongst users with different political leanings. Two studies on fair influence maximization [1, 48] are the most related to our work. However, their approaches may not be directly applicable to online networks such as Twitter. We show that in social media, the political leaning of the seeds can make a considerable bias in users' exposure to news. In particular, we observe that high consensus news posted by neutral publishers has the lowest disparity for spreading among all users (liberal, conservative and neutral). We use the fair influence maximization method proposed by [1] as a baseline.

In the following sections, we explain our dataset and research design, and discuss our findings.

3 Dataset

In this work, we consider the dataset of 400 news tweets posted by 10 publishers collected in [4] between 9th to 15th May, 2017. The dataset contains 80 low (20% of news with lowest consensus value) and 80 high consensus (20% of news with highest consensus value) news posts. Furthermore, we collected an additional 10 high and 10 low consensus recent news posts for each of COVID-19 and global warming, including 40 tweets in total.

To obtain the political leanings of the users who either retweeted a high or low consensus news or were exposed to it, for every high or low consensus news post in the dataset, we collected a random set of its 100 retweeters. Then for each retweeter, we collected a random set of his 100 followers. Finally, for each of these 1,616,000 users we collected their followees to compute their political leaning.

3.1 High and Low Consensus News Posts

Our news dataset consists of 10 news publishers with different political leanings varying from liberals to neutrals to conservatives: Slate, Salon, New York Times, CNN, AP, Reuters, Politico, Fox News, Drudge Report, and Breitbart News. From each publisher, 40 tweeted news posts are collected during the one week period of 9th to 15th May, 2017. For each news post, the authors in [4] set up a survey in Amazon Mechanical Turk. They asked US AMT workers about their reaction to the post by selecting one out of three options: agreement, neutral, or disagreement. At the end of the experiment, they asked about AMT workers' political leanings: liberal, conservative, or neutral. Based on the above intuition, Babaei et al. [4] proposed to capture the degree of consensus that a social media post is likely to have based on the distribution of the political leanings of the retweeters and repliers of a post, as follows.

$$\text{consensus} = 1 - \left| \frac{\#D_{disagree}}{\#D} - \frac{\#R_{disagree}}{\#R} \right|, \tag{1}$$

where $\#D_{disagree}$ and $\#R_{disagree}$ respectively denote the number of democrats and republicans who disagree with the post, while $\#D$ and $\#R$ are the total number of democrats and republicans. Authors also proposed a method to detect high and low consensus news automatically, which we use in our work to find an additional set of 40 high and low consensus tweets from the above mentioned publishers related to COVID-19 and global warming.

Note that high consensus news is distinct from low attention news. All stories used in our experiments discuss salient health, social, and political news – not lightweight gossip which may not generate much attention from or disagreement between users due to the chosen topics (see Section 7 for more discussion about the newsworthiness of our news stories). Table 1 shows random sample news that are labeled as low consensus with conservative leaning, low consensus with liberal leaning, and high consensus news respectively.

Figure 1(A) shows the number of high consensus (purple bars) and low consensus (orange bars) news posted by 10 publishers with various political leanings. Figure 1(B) shows the total number of high consensus and low consensus news posted by the same 10 publishers, grouped into liberal, conservative, and neutral categories. It also shows the total number of high consensus and low consensus news posted by all the 10 publishers. We can see that the total number of low consensus posts are considerably higher than the total number of high consensus posts.

Low Consensus News by Conservative Publishers
FoxNews: @DennisDMZ on global warming: "I think it's hot out there 'cause the sun is hot." #OReillyFactor, 278 Retweets, 122 Replies, 642 Likes
FoxNews: CNN doesn't sound alarm of COVID 'superspreaders' as thousands celebrate Biden win in the streets. 5.4 Retweets, 11.5k Replies, 14.1k Likes
New York Post: 29 COVID-19 vaccine recipients had serious allergic reactions: CDC https://trib.al/IZeEbyi: 76 Retweets, 29 Replies, 81 Likes
Fox News: Schieffer Slams Trump: Comey Firing Reminds Me of JFK-Oswald Conspiracies. 55 Retweets, 510 Replies, 149Likes.
Low Consensus News by Liberal Publishers
CNN: The Trump administration is loosening restrictions on the coal industry just days after a government report warned that aggressive action is needed to ease the impact of global warming. 307 Retweets, 105 Replies, 259 Likes
CNN: Exclusive: President-elect Joe Biden will aim to release every available dose of the coronavirus vaccine, a break with President Trump's strategy of holding back second shots. 262 Retweets, 49 Replies, 1.7k Likes
Salon: "We want them infected": Shocking email reveals top Trump appointee's plan to spread COVID-19. 36 Retweets, 39 Replies, 33 Likes
CNN: Report: Trump "revealed more information to the Russian ambassador than we have shared with our own allies" 51 Retweets, 14 Replies, 34 Likes
High Consensus News
CNN: Greenland's ice sheet has melted to a point of no return, and efforts to slow global warming will not stop it from disintegrating, according to a new study. 4.6k Retweets, 381 Replies, 5.2k Likes
POLITICO @politico · Mar 11, 2020 Anthony Fauci, the director of the National Institute of Allergy and Infectious Diseases, said #COVID2019 is 10x times more lethal than the flu, which kills nearly 0.1 percent of Americans who get it each year https://politi.co/2IEQTS5. 473 retweets, 39 replies, 573 likes
CNBC: Wearing a mask can reduce coronavirus transmission by 75%, new study claims. 9.3k Retweets, 647 Replies, 14.8k Likes
Fox News: @johnrobertsFox on firing of James Comey: "This came as a shock to literally everyone, including the @FBI Director." TheFiveSource. 156 Retweets, 259 Replies,574 Likes

Table 1: Samples of high and low consensus news posts.

As we discuss later, low consensus posts have a much smaller chance of being received by users with different political leanings, which leads to a more politically fragmented society. On the other hand, high consensus posts have a better chance of spreading through communities with various political leanings, and can be used to help break echo chambers.

3.2 Collecting Users' Political Leanings

For each post in our set of 80 high and 80 low consensus news posts, we collected a random set of its 100 retweeters. Then for each retweeter, we collected a random set of its 100 followers. Thus we have 1,616,000 twitter users. To infer a user's political leaning, one popular way is to apply NLP methods on the contexts that the user posts. However, tweets' context length is small, so we must look at users' neighbors. We inferred every user's political leaning, as a score between -1, +1, using the method of Kulshrestha et al. [30], for which we needed to collect their followees. Inferring the political leaning of a given Twitter user u is based on the following steps – (*i*) generating two representative sets of users who are known to have a liberal or conservative bias, (*ii*) inferring the topical interests of u by looking at her followees, and (*iii*) examining how closely u's interests match with the interests of the representative sets of liberal and conservative users. Formally,

$$leaning(u) = cos\_sim(I_u, I_D) - cos\_sim(I_u, I_R), \tag{2}$$

where I_u is the interest vectors of user u, and I_D, I_R are normalized aggregate interest vector for the liberal seed set (I_D) and the conservative seed set (I_R). Similarity between interest vectors are measured by cosine similarity.

For retweeters with certain political leaning, we calculated the expected fraction of their liberal, conservative, and neutral followers, as is shown in Table 2.

We also estimated the conditional probability that users with different political leanings retweet high consensus and low consensus news post from liberal, conservative, and neutral publishers (given that they retweet) in Table 3. It can be seen that users with a certain political leaning retweet low consensus posts from the publishers with the same political leaning with a very high probability. Interestingly, users retweet high consensus news posts from the publishers with the same political leaning with a smaller probability. On the other hand, there is a very

	Liberal	**Conservative**	**Neutral**
Liberal	0.76	0.04	0.2
Conservative	0.045	0.85	0.1
Neutral	0.3	0.27	0.43

Table 2: Expected fraction of liberal, conservative, and neutral followers of retweeters with various political leanings. Rows and columns correspond to retweeters and followers.

		Retweeters		
		Liberal	**Conservative**	**Neutral**
Publishers	**Liberal**	H: 0.65 L: 0.85	H: 0.08 L: 0.04	H: 0.27 L: 0.11
	Conservative	H: 0.12 L: 0.08	H: 0.68 L: 0.85	H: 0.2 L: 0.07
	Neutral	H: 0.34 L: 0.38	H: 0.33 L: 0.37	H: 0.33 L: 0.25

Table 3: Conditional probability of retweeting a high and low consensus news post (indicated H and L in the table) by users from various political leanings (given that they retweet). Rows and columns correspond to publishers and retweeters. For instance, in the first cell (first row and column), the probability that liberal users retweet high/low consensus news posts published by liberals publishers is 0.65/0.85.

small chance that users with a certain political leaning retweet low consensus posts from the publishers with different political leanings. For high consensus news this probability is larger.

4 The Gap between Proliferation of High and Low Consensus News

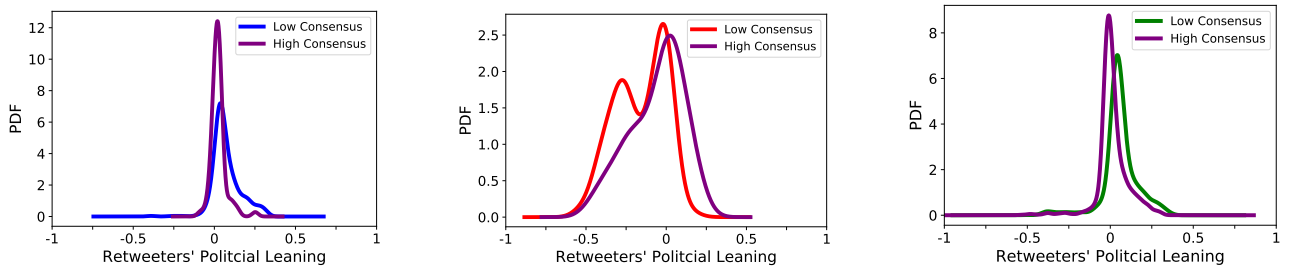
In this section, we investigate how high and low consensus news posts spread among users with various political leanings in Twitter. In particular, our goal is to answer the following key question:

How do individuals with certain political leaning (liberal, conservative, and neutral) get exposed to high and low consensus news posts?

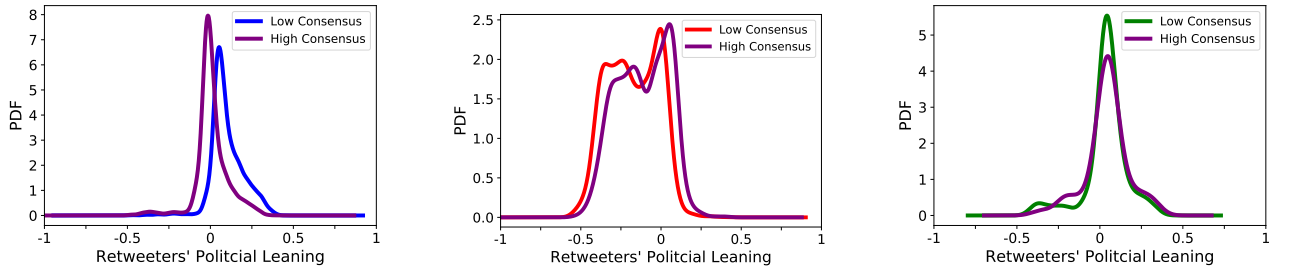
Studying the above key question allows to understand the gap between proliferation of high and low consensus news, and develop strategies to decrease the polarization in the society by breaking the echo chambers that trap users. We start by investigating users' behavior in retweeting high and low consensus news posts. Then, we discuss how the confirmation bias in retweeting behavior makes the echo chambers grow larger and promote social polarization.

4.1 Confirmation Bias in Retweeting Behavior

First, we study how users with different political leanings *share* high and low consensus news post. Specifically, we compare how users with different political leanings retweet low and high consensus news posts from publishers with different political perspectives.



(A) Liberals (B) Conservatives (C) Neutrals



(D) Liberals (E) Conservatives (F) Neutrals

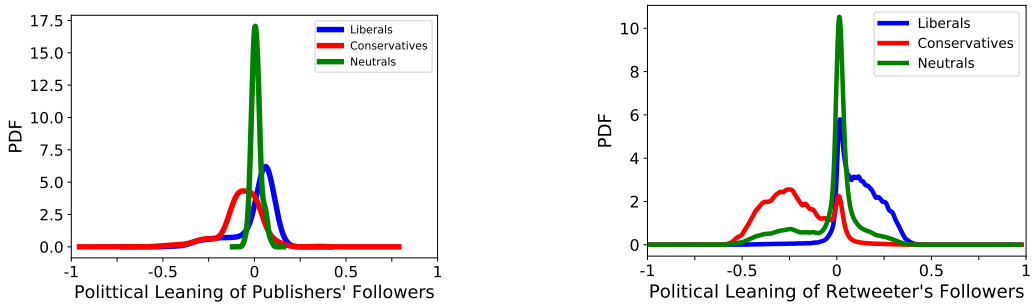
Figure 2: Distribution of retweeters' political leanings for low and high consensus news posted by publishers with different political perspectives. Distribution of political leanings for a random high and a random low consensus news posted by (A) liberal, (C) conservative, and (E) neutral publishers. Distribution of average political leanings for 100 high and low consensus news posted by (B) liberal, (D) conservative, and (F) neutral publishers. Distribution of political leanings for high consensus news (purple) is more symmetric and centered around 0.

Figures 2(A), 2(B), 2(C) show the distribution of the political leanings of all retweeters for one random low consensus and one random high consensus news posted by CNN (liberal publisher), FoxNews (conservative publisher), and Reuters (neutral publisher)[4]. Notice that the distribution of retweeters' political leanings in Figure 2(A), 2(C) has more density in the right (liberal leaning) for the low consensus news. On the other hand, the distribution of retweeters' political leanings in Figure 2(B) has considerably more density in the left (conservative leaning) for the low consensus news. Importantly, the distribution of retweeters' political leanings for high consensus news (purple curve) is more symmetric in all the Figures. Moreover, the mean of the distribution for high consensus news is close to 0.

Next, we consider 80 low consensus and 80 high consensus news posted by the 10 publishers, ranging from liberals to neutrals to conservatives: Slate, Salon, New York Times, CNN, AP, Reuters, Politico, Fox News, Drudge Report, and Breitbart News. For each news post, we consider a set of its 100 retweeters chosen at random. Figures 2(D), 2(E), 2(F) show the distribution of the *expected* political leanings of retweeters of all the low consensus and high consensus news posted by liberal, conservative, and neutral publishers, respectively. Again, the distribution of retweeters' political leanings for high consensus news (purple curve) is more symmetric, and is centered around 0 in all the Figures. In particular, the distribution of retweeters' political leanings for high consensus news posted by neutral publishers has the most symmetric shape around 0.

We summarize our key observations as follows:

[4]The PDFs have been empirically estimated using kernel density estimation [10]



(A) Publisher's followers (B) Retweeter's followers

Figure 3: Distribution of political leanings for (A) followers of liberal, conservative, and neutral publishers, and (B) followers of retweeters of liberal, conservative, and neutral publishers. As we get farther away from the publishers, the distribution of liberal and conservative followers becomes significantly more skewed (echo chambers grow larger).

- I. Low consensus news posted by publishers with a specific political leaning (liberal/conservative) are mostly retweeted by users with similar political leanings (Figures 2(D), 2(E)).
- II. High consensus news posted by publishers with a specific political leaning (liberal/conservative) are retweeted by users with various political leanings (liberal/conservative/neutral) (Figures 2(D), 2(E)).
- III. While low and high consensus news posted by neutral publishers spread with lower disparity among users with different political leanings, high consensus news posted by neutral publishers have the highest probability to be spread with minimum disparity among users (Figure 2(F)).

4.2 The Growth of Echo Chambers in Twitter

Next, we investigate how individuals with different political leanings get *exposed* to high and low consensus news posted by liberal, conservative, and neutral publishers.

Figure 3(A) depicts the distribution of political leanings for followers (level 1) of liberal, conservative, and neutral publishers. We observe that users with conservative or liberal leanings are mostly exposed to news posted by publishers with the same political leaning (the chamber effect). Therefore, the distribution of political leaning for followers of liberal and conservative publishers are skewed to the left and right, respectively. Nevertheless, followers of conservative publishers has a more skewed distribution. This is resulted from the fact that the conservative community is denser, and has fewer connections to liberals and neutrals in Twitter (*c.f.* Table 2). On the other hand, the distribution of political leanings for neutral users is very symmetric and is centered at 0. Hence, neutral users get similar exposure to liberal and conservative view points.

Figure 3(B) shows the distribution of political leanings for followers of retweeters (level 2) of liberal, conservative, and neutral publishers. We observe that while the distribution of political leanings for followers of retweeters of neutral publishers is symmetric and centered around 0, the distribution of political leanings for followers of retweeters of liberal or conservative publishers are extremely skewed. As expected, followers of retweeters of conservative publishers have a more skewed distribution. Interestingly, the skewness of the distributions for followers of retweeters (level 2) is much larger compared to the skewness of distributions for followers of publishers (level 1). This means that echo chambers in level 2 are larger than those in level 1.

Our experiments show that as we get farther away from the publishers, echo chambers grow even larger (Figure 7).

We summarize our key observations as follows:

- I. Conservatives and liberals get a biased exposure to the news posted in Twitter, while neutrals get similar exposure to liberal and conservative view points (Figure 3(A)).
- II. Users who get the news from retweeters get a significantly more biased exposure, compare to users who get the news from publishers. In other words, as we get farther away from the news publishers, the echo chambers grow larger (Figure 3(B)).

4.3 Breaking Echo Chambers

To break echo chambers, we aim for all individuals to get similar exposure to news stories. Our proposed strategy that we hope to break the chamber effect is based on the three key observations discussed earlier: (1) while low consensus news are more likely to proliferate amongst the users with a particular political leaning, high consensus news has a much higher chance of spreading among users with different political leanings; (2) high consensus news posted by neutral publishers has the lowest disparity for spreading among liberal and conservative users; and (3) as users get farther away from publishers, they get a more biased exposure to news. Based on the above observations, we conjecture:

High consensus news posted by neutral users may help to break echo chambers.

High consensus news posted by neutral users achieve high spread with little disparity regarding political leaning. We confirm our conjecture and show the effectiveness of our proposed strategy through an extensive set of experiments later in the paper.

In the following section, we first formulate the problem of information diffusion in social networks. Then, we discuss the problem of finding a near-optimal set of neutral users to seed spreading high consensus news and break the echo chamber.

5 Problem Formulation: Information Diffusion

We start by formulating the information diffusion problem to model the spread of news among individuals with various political leanings in Twitter. We simulate the proliferation of news by assuming that each user can be a publisher. Then we select a set of users and involve them to post news. We represent Twitter by a directed graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of directed edges between the nodes. The nodes in the network are partitioned into three disjoint groups $\mathcal{V} = \{_{d}, _{r}, _{n}\}$, where $_{d}, _{r}, _{n}$ represent users with liberal, conservative, and neutral leanings, respectively. A directed edge (v, u) exists if user v follows user u. When users post tweets, their followers can retweet and spread the tweets in the network. To model spread of information, e.g. news or tweets in Twitter, two well-known classical diffusion models are introduced in the literature [38]: (1) Independent Cascade model (IC) and (2) Linear Threshold (LT) model. In this work, we consider the IC model.

5.1 Independent Cascade model (IC)

In the IC model, information propagates through every edge (v, w) with probability p_{vw}. We have a set of discrete time steps which we denote with $t = \{0, 1, 2, \cdots\}$. At $t = 0$, the initial

seed set $S \subseteq \mathcal{V}$ is activated. At every time step $t > 0$, a node $v \in V$ which was activated at time $t-1$ can activate its inactivated neighbors w with probability p_{vw}. The model assumes that once a node is activated, it stays active throughout the whole process and each node has only one chance to activate its neighbors. The described process stops at time $t > 0$ if no new node gets activated at this time. We note that the IC model is a stochastic process, in which a node u can influence its neighbors w based on the Bernoulli distribution with success probability p_{uw}. A possible outcome of the process can be denoted via a set of timestamps $\{t_v \geq 0 : v \in \mathcal{V}\}$, where t_v represents the time at which a node $v \in V$ is activated.

5.2 Information Diffusion with Low Disparity

Our goal is to find the smallest seed set of users that when post a tweet, it spreads through at least a fraction $Q_p \in [0,1]$ of liberals $(_l)$, conservatives $(_c)$, and neutrals $(_n)$ in Twitter, where $p \in \{l, c, n\}$. We formulate the problem as follows:

$$\min_{S\subseteq} |S| \quad \text{subject to} \qquad (3)$$
$$\sum_{p\in\{l,c,n\}} \min\Big(f_p(S),\ Q_p \cdot |_p|\Big) \geq \sum_{p\in\{l,c,n\}} Q_p,$$

where $f_l(.), f_c(.), f_n(.)$ determine the total number users among liberals $(_l)$, conservatives $(_c)$, and neutrals $(_n)$ that are activated as a result of selecting the seed set S. We call $_p$ "saturated" by S when $\min\Big(f_p(S),\ Q_p \cdot |_p|\Big) = Q_p \cdot |_p|$. When a certain fraction Q_p of individuals with a particular political leaning p are exposed to a news (activated), any new activated individual with political leaning p cannot further improve the utility. This will give individuals with different political leanings a higher chance of being exposed to the news.

We note that the utility function, i.e., $f_p : 2^p \to \mathbb{Z}^+$, is a non-negative, monotone, submodular set function [29]. The submodularity is an intuitive notion of diminishing returns, stating that for any sets $A \subseteq A' \subseteq V$ and any node $a \in V \setminus A'$, it holds that:

$$f(A \cup \{a\}) - f(A) \geq f(A' \cup \{a\}) - f(A').$$

Although problem (3) is NP-hard in general [52], for maximizing a submodular function the following greedy algorithm provides a logarithmic approximation guarantee. The greedy algorithm starts from an empty set, add a new node to the set which provides the maximal marginal gain in terms of utility, and stops whenever the desired Q_p fraction of individuals with political leaning p are activated.

5.3 Spreading through Neutrals to Break the Chambers

To break the echo chambers we wish individuals with different political leanings to get a similar exposure to various news. In other works, we assume similar values for Q_l, Q_c, Q_n. Moreover, the news posted by individuals with neutral leanings have a higher chance of spreading among individuals with liberal and conservative political leanings. Therefore, to break the echo chambers we aim at finding the smallest subset $S \subseteq_n$ that when post a news, at least a fraction Q_p of individuals with political leaning p get exposed to the news. Formally, we have

$$\min_{S\subseteq_n} |S| \quad \text{subject to} \qquad (4)$$
$$\sum_{p\in\{l,c,n\}} \min\Big(f_p(S),\ Q_p \cdot |_p|\Big) \geq \sum_{p\in\{l,c,n\}} Q_p.$$

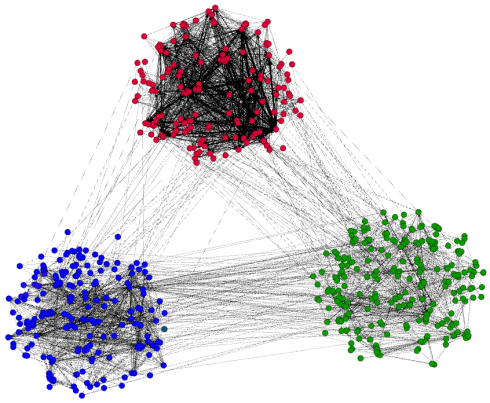


Figure 4: The sample graph from the real Twitter data set collected in 2009. Blue, red, and green nodes indicate users with liberal, conservative, and neutral political leanings.

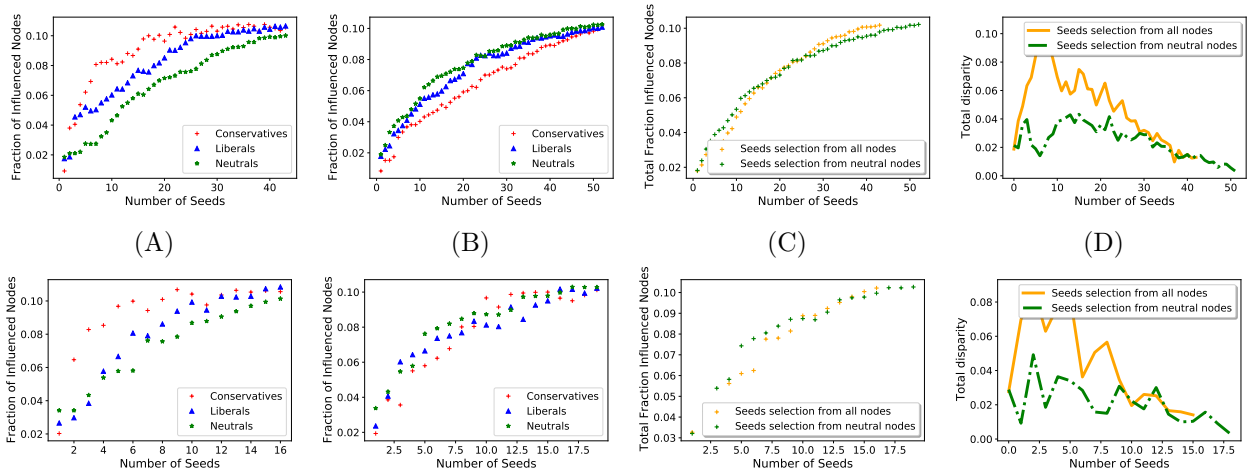


Figure 5: Fraction of individuals with liberal, conservative, and neutral political leanings who are exposed to a high consensus news. Top row shows the result on our Twitter network, and the bottom row shows the result on the smaller sampled Twitter network. (A), (E) show the fraction of exposed individuals when the seeds are selected from the entire network by solving Problem (3). (B), (F) show the fraction of exposed individuals when the seeds are selected from neutral users by solving Problem (4). (C), (G) compare the fraction of exposed individuals when the initial seed is selected from the entire network vs. neutrals. (D), (H) compare the disparity of diffusion when the initial seed set is selected from the entire network vs. neutrals.

6 Experimental Results

In this section we investigate the effect of spreading high consensus news posted by neutral users among individuals with conservative, liberal, and neutral leanings in Twitter. In particular, we show that our proposed strategy is very effective in spreading information among individuals with various political leanings and lowering societal polarization for news consumption. We first describe our instance of Twitter network. We then explain our experimental setup, and present our findings.

Twitter Network. Our network is collected from Twitter in September 2009 [3, 13], and includes: 52 million user profiles, 1.9 billion directed follow links among the users, and 1.7

billion public tweets posted by the users. In order to obtain a static network, we consider the tweets published on July 1, 2009, and filter out users that did not tweet before July 1. After this filtering, we have 70,000 active users. We then extract the strongest connected community, including 69,687 users and 2,907,026 link between them, yielding 19162, 3449, 47076 nodes with liberal, conservative, and neutral leanings, respectively. The average degree of network is 41.5. Figure 4 shows an induced random sample from our final Twitter network.

Sampled Twitter Network. We also created a smaller network by sampling 10% of nodes uniformly at random from our original Twitter network, and connecting the users if they have a connection in the original network. The strongest connected community includes 3,753 users and 6,993 connections with average degree of 1.83. Our sampled Twitter network includes 812 liberals, 186 conservatives, and 2,755 neutrals. Note that the structure of the original Twitter network is very different than the sampled Twitter network. In particular, the sampled Twitter network is significantly sparser than the original Twitter network.

Experimental Setup. For a pair of users $u \in_i$ and $w \in_j$, we calculate the success probability of activation p_{uw} as the expected fraction of users with political leaning j who retweeted news posted by users with political leaning i. The retweeting probabilities are listed in Table 3.

We apply the greedy algorithm to find a near optimal subset of users that can spread a news over a certain fraction $Q_l = Q_c = Q_n = 0.1$ of liberals $(_l)$, conservatives $(_c)$, and neutrals $(_n)$ in the Twitter network. To evaluate the utility function $f_p(.)$ in Problem 3 and Problem 4, we estimate it by using Monte Carlo sampling [26]. We used 200 samples for this estimation, which yielded a stable estimation of the utility function.

Note that using equal values for Q_l, Q_c, Q_n in Problem (3), we retrieve the fair influence maximization formulation proposed by [1]. In our experiments, we compare our proposed strategy to fair influence maximization.

6.1 Neutrals Can Break Echo Chambers

In our first set of experiments, we apply the greedy algorithm to Problem (3) and Problem (4) to find the initial set of users to spread news in Twitter. Figure 5 compares the fraction of individuals with liberal, conservative, and neutral political leanings who got exposed to a high consensus news spread through an initial seed set obtained by solving Problem (3) vs. Problem (4). The goal is to expose $Q_p = 10\%$ of individuals with liberal, conservative, and neutral leanings to the news. The top row shows the result on our original Twitter network, and the bottom row shows the result on the smaller sampled Twitter network. Note that the sampled network is much sparser than the original Twitter network.

Figures 5(A), 5(E) show the fraction of exposed individuals when the seeds are selected from the entire network by solving Problem (3). Figures 5(B), 5(F) show the fraction of exposed individuals when the seeds are selected from the users with neutral leanings by solving Problem (4). We note that as more individuals are added to the initial seed set by the greedy algorithm, the disparity in the number of exposed users with different political leanings is much smaller in Figures 5(B), 5(F) compared to Figures 5(A), 5(E). This clearly confirms the effectiveness of our proposed strategy in breaking the echo chambers.

We note that if we do not take into account the different pattern of diffusion among users of various political leanings, the neutral users may not be the ones that can maximize the spread of information. 0

		Retweeters			
		Liberal	**Conservative**	**Neutral**	**Sum**
Publishers	**Liberal**	H: 76 L: 104	H: 9 L: 5	H: 32 L: 15	H:117 L:124
	Conservative	H: 9 L: 10	H: 58 L: 94	H: 18 L: 9	H:87 L:113
	Neutral	H: 45 L: 49	H: 43 L: 47	H: 43 L: 32	H:131 L:128

Table 4: Average number of retweeting a high and low consensus news post (indicated H and L in the table) by users from various political leanings. Rows and columns correspond to publishers and retweeters. For instance, in the first cell (first row and column), liberal users on average retweets high/low consensus news posts published by liberals publishers 76 times. We note that news posted by neutrals is retweeted by an even larger number of users, compared to news posted by liberal or conservative publishers.

6.2 Neutrals Can Spread News Widely

Figures 5(C), 5(G) compare the fraction of exposed individuals when the initial seed is selected from the entire network vs. neutrals. There are two interesting observations: the initial seeds selected from neutral users can spread the news even more than users selected from the entire network. Moreover, as we continue the selection process, selected neutral seeds can spread the news as well as the seed set selected from the entire network. This interesting observation confirms the power of neutral users in spreading news in Twitter. As Table 4 depicts, the average number of retweeting of a tweet posted by liberal, conservative, and neutrals are almost equal. There is a interesting observation. High consensus tweets posted by neutrals are retweeted with many democrats and republicans in addition to neutrals. Interestingly, news posted by neutrals is retweeted by an even larger number of users compared to news posted by liberal or conservative publishers.

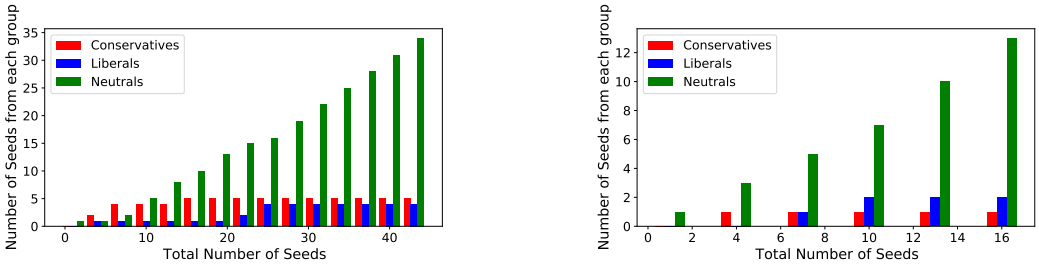
Figure 6(A) shows the number of users selected with liberal, conservative, and neutral leanings for varying number of seeds selected greedily to solve Problem (3). Figure 6(A) shows the result on the Twitter network, and Figure 6(B) shows the result on the sampled Twitter network. We see that in the set of seeds greedily selected from the entire network, the majority of the users have neutral leanings. This further shows that neutral users are highly effective in spreading information in Twitter. This is consistent with our initial observation, that the news posted by neutrals has a higher probability of spreading among users with different political leanings.

6.3 Neutrals Spread News with Low Disparity

Figures 5(D), 5(H) compare the total disparity of diffusion when the initial seed set is selected from the entire network vs. neutrals. We define the total disparity as the sum of all disparity (differences) between exposure for each pair of political leanings. Formally, we have: Total disparity=

$$\left|\frac{f_l(S)}{l} - \frac{f_c(S)}{c}\right| + \left|\frac{f_l(S)}{l} - \frac{f_n(S)}{n}\right| + \left|\frac{f_c(S)}{c} - \frac{f_n(S)}{n}\right|.$$

We observe that the total disparity is much smaller when the initial seed set is selected from users with neutral political leanings (Problem (3)) compared to the case when the initial seed



(A) Twitter network (B) Sampled Twitter network

Figure 6: Number of users selected with liberal, conservative, and neutral leanings for varying number of seeds selected greedily to solve Problem (3). Figure shows the results on (A) the Twitter network, and (B) the smaller sampled Twitter network.

set is selected from the entire network (Problem (4)). The difference is larger when the size of the initial seed set is smaller.

6.4 Echo Chambers Grow Larger over Time

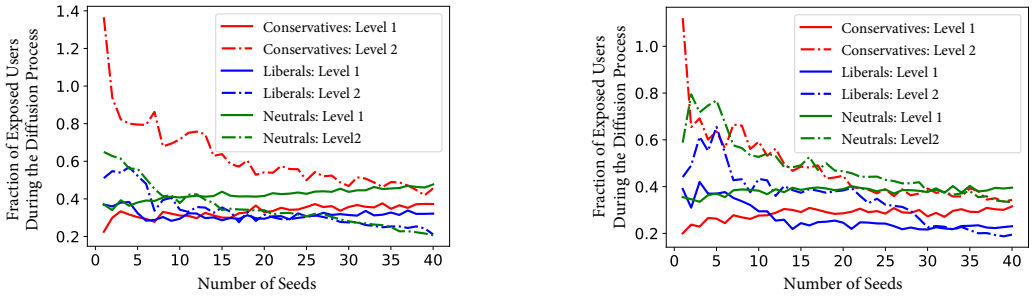
Figure 7 shows the fraction of individuals with various political leanings who got exposed to a high consensus news during the diffusion process (IC), for varying number of seeds. More precisely, for a given seed set information diffusion proceeds in discrete time steps $t = \{0, 1, 2, \ldots, \}$. Figure 7 compares the fraction of users with various political leanings who received the news in the first time-step, $t = 1$, and second time-step $t = 2$ in our original Twitter network. Figure 7(A) shows the result when the seeds are selected from the entire network, by solving Problem (3). Figure 7(B) shows the result when the seeds are selected from the users with neutral political leanings, by solving Problem (4). It can be seen that when seeds are selected from the entire network, the disparity becomes larger as the diffusion process continues. On the other hand, the disparity is much smaller when seeds are selected from neutral users.

The above result confirms our observation that the echo chambers grow larger as the diffusion continues over time. In other words, when the seeds are selected from the entire network, as the we get farther away from the initial set of seeds, the disparity in the number of users with different political leanings who are exposed to the news becomes larger. On the other hand, when diffusion is originated from neutral seeds, users with different political leanings get exposed to the information at the same time. This is crucial while spreading time-critical information, such as health-related information or emergency warnings, in the network.

7 Discussion

Since we propose increasing exposure to high consensus news, we would like to check that such stories carry important information for public discourse. Babaei et al. [4] compared low and high consensus posts on social media by empirically analyzing their properties. They showed that both types of posts are equally popular and cover similar topics. We checked this by analyzing 400 randomly selected posts including examples of high and low consensus news, along with their sources, number of retweets, replies, and likes. We highlight the following observations:

I. For both types of news posts, a variety of news sources exists across the ideological spectrum. Figure 1 also shows several publishers with different political leaning that



(A) (B)

Figure 7: Fraction of users who are exposed to the news from different groups in first and second time step of propagation process in the Twitter network. (A) Shows the result when seed are selected from the entire network (Problem 3), (B) shows the result when seed are selected from neutral users (Problem 4).

posts both types of high and low consensus news.

II. On average, high and low consensus tweets are retweeted 158 and 177 times respectively. On average high consensus tweets are liked 532 times, whereas, low consensus news are liked 488 times. Thus, high and low consensus news stories have similar popularity.

III. Galtung and Ruge [20] introduce newsworthiness theory in which they propose several news factors such as frequency, meaningfulness, continuity, etc. Eilders [18] showed that these factors impact news' worthiness. Weber [51] proposes the following hypothesis: "The news factors of a news item influence the level of participation in commenting in an article's comments section". Weber also noted several other factors, such as having a high social impact or being controversial, that may attract more comments as participation [50]. Weber emphasized that if a news story attracts more comments, then it has higher worthiness. Here we can consider the number of replies as participating comments. On average, high consensus and low consensus news stories received 100 and 114 replies (comments), respectively, suggesting that both types of news have similar worthiness.

In summary, we observe that high and low consensus news are similar along multiple dimensions, including variety of news source, popularity, topic covering, and worthiness.

8 Conclusion

In this work, we studied the diffusion of news in Twitter. We investigated how users with various political leanings (liberals, conservatives and neutrals) get exposed to low and high consensus news posted by different publishers (e.g. CNN, FoxNews, etc.). We found that (1) while low consensus news stories are more likely to proliferate amongst users with a particular political leaning, high consensus news has a much higher chance of spreading among users with different political leanings and can reach a greater total number of users; (2) high consensus news posted by neutral publishers has the lowest disparity for spreading among liberal and conservative users and reaches the greatest total number of users; and (3) as users get farther away from the publishers, they get a more biased exposure to the news.

Based on the above observations, we studied the effect of spreading high consensus news through neutral users on decreasing the disparity in users' exposure. Our extensive simulation experiments on Twitter showed that our proposed strategy can be highly effective in spreading information while decreasing the disparity of information across users with differing views.

While some of our results may appear unsurprising, we believe we are the first to empirically verify, and theoretically justify, that seeding high consensus news stories with neutral publishers is superior to other strategies (such as seeding with several publishers from each political leaning, particularly for a low number of initial seeds, see Fig 5 C) in reaching a large number of users, who are also diverse across their views. We hope our findings may be helpful for spreading newsworthy stories broadly and equitably, breaking echo chambers, reducing fragmentation in online social media, lowering segregation in consumption of important news, and help lead to a healthier society.

Acknowledgements

AW acknowledges support from a Turing AI Fellowship under grant EP/V025279/1, The Alan Turing Institute, and the Leverhulme Trust via CFI.

References

[1] J. Ali, M. Babaei, A. Chakraborty, B. Mirzasoleiman, K. P. Gummadi, and A. Singla. On the fairness of time-critical influence maximization in social networks. *ArXiv*, abs/1905.06618, 2019.

[2] J. An, D. Quercia, M. Cha, K. Gummadi, and J. Crowcroft. Sharing political news: the balancing act of intimacy and socialization in selective exposure. *EPJ Data Science*, 3(1): 12, 2014.

[3] M. Babaei, P. Grabowicz, I. Valera, K. P. Gummadi, and M. Gomez-Rodriguez. On the efficiency of the information networks in social media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 83–92, 2016.

[4] M. Babaei, J. Kulshrestha, A. Chakraborty, F. Benevenuto, K. P. Gummadi, and A. Weller. Purple feed: Identifying high consensus news posts on social media. In *Proceedings of the AAAI/ACM Conference on Artifical Intelligence, Ethics & Society (AIES)*, 2018.

[5] C. A. Bail. *Terrified: How anti-Muslim fringe organizations became mainstream.* Princeton University Press, 2014.

[6] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37): 9216–9221, 2018.

[7] E. Bakshy, S. Messing, and L. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 2015.

[8] E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[9] D. Bhattacharya and S. Ram. Sharing news articles using 140 characters: A diffusion analysis on twitter. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 966–971. IEEE, 2012.

[10] A. W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis.* Clarendon Press, 2004.

[11] E. Bozdag. Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, 15(3):209–227, Sep 2013. ISSN 1572-8439. doi: 10.1007/s10676-013-9321-6. URL `https://doi.org/10.1007/s10676-013-9321-6`.

[12] C. Budak, S. Goel, and J. M. Rao. Fair and Balanced? Quantifying Media Bias Through Crowdsourced Content Analysis. *http://dx.doi.org/10.2139/ssrn.2526461*, 2014.

[13] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in Twitter: the million follower fallacy. In *Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM'2010)*, May 2010.

[14] A. Chakraborty, S. Ghosh, N. Ganguly, and K. P. Gummadi. Dissemination biases of social media channels: On the topical coverage of socially shared news. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[15] A. Chakraborty, M. Ali, S. Ghosh, N. Ganguly, and K. P. Gummadi. On quantifying knowledge segregation in society. *arXiv preprint arXiv:1708.00670*, 2017.

[16] C.-F. Chiang and B. Knight. Media bias and influence: Evidence from newspaper endorsements. *The Review of Economic Studies*, 78(3):795–820, 2011.

[17] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Proc. AAAI ICWSM*, 2011.

[18] C. Eilders. News factors and news decisions. theoretical and methodological advances in germany. *Communications*, 31(1):5–24, 2006.

[19] S. Flaxman, S. Goel, and J. M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, 2016. doi: 10.1093/poq/nfw006. URL `+http://dx.doi.org/10.1093/poq/nfw006`.

[20] J. Galtung and M. H. Ruge. The structure of foreign news: The presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of peace research*, 2(1): 64–90, 1965.

[21] M. Gentzkow and J. M. Shapiro. What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, 78:35–71, 2010.

[22] A. Goyal, F. Bonchi, L. V. Lakshmanan, and S. Venkatasubramanian. On minimizing budget and time in influence propagation over social networks. *Social network analysis and mining*, 3(2):179–192, 2013.

[23] T. Groseclose and J. Milyo. A measure of media bias. *The Quarterly Journal of Economics*, 120:1191–1237, 2005.

[24] A. Gruzd and P. Mai. Going viral: How a single tweet spawned a covid-19 conspiracy theory on twitter. *Big Data & Society*, 7(2):2053951720938405, 2020.

[25] J. Hartline, V. Mirrokni, and M. Sundararajan. Optimal marketing strategies over social networks. In *Proceedings of the 17th international conference on World Wide Web*, pages 189–198, 2008.

[26] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970. ISSN 0006-3444. doi: 10.1093/biomet/57.1.97. URL `https://doi.org/10.1093/biomet/57.1.97`.

[27] I. Himelboim, S. McCreery, and M. Smith. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. *Journal of Computer-Mediated Communication*, 18(2):40–60, 2013.

[28] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma. Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2751–2754. ACM, 2012.

[29] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[30] J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, K. P. Gummadi, and K. Karahalios. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 417–432, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4335-0. doi: 10.1145/2998181.2998321. URL `http://doi.acm.org/10.1145/2998181.2998321`.

[31] J. Lichterman. New Pew data: More Americans are getting news on Facebook and Twitter, 2010. *http://tinyurl.com/News-on-Social-Media*.

[32] Z. Liu and I. Weber. Is twitter a public sphere for online conflicts? a cross-ideological and cross-hierarchical look. In *International Conference on Social Informatics*, pages 336–347. Springer, 2014.

[33] C. G. Lord, L. Ross, and M. R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.

[34] S. Munson, S. Lee, and P. Resnick. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, pages 419–428, Boston, USA, 2013. AAAI press.

[35] B. Nyhan and J. Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

[36] E. Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group, 2011. ISBN 1594203008, 9781594203008.

[37] S. Park, S. Kang, S. Chung, and J. Song. Newscube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 443–452. ACM, 2009.

[38] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[39] F. N. Ribeiro, K. Saha, M. Babaei, L. Henrique, J. Messias, F. Benevenuto, O. Goga, K. P. Gummadi, and E. M. Redmiles. On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 140–149. ACM, 2019.

[40] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, 2002.

[41] M.-A. Rizoiu, T. Graham, R. Zhang, Y. Zhang, R. Ackland, and L. Xie. # debatenight: The role and influence of socialbots on twitter during the 1st 2016 us presidential debate. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[42] D. Schkade, C. R. Sunstein, and R. Hastie. What happened on deliberation day? *California Law Review*, 95(3):915–940, 2007.

[43] P. J. Shoemaker and T. Vos. *Gatekeeping theory*. Routledge, 2009.

[44] K. Starbird, A. Arif, T. Wilson, K. Van Koevering, K. Yefimova, and D. Scarnecchia. Ecosystem or echo-system? exploring content sharing across alternative media domains. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[45] C. R. Sunstein. The law of group polarization. *Journal of Political Philosophy*, 10(2): 175–195, 2002. ISSN 1467-9760. doi: 10.1111/1467-9760.00148. URL `http://dx.doi.org/10.1111/1467-9760.00148`.

[46] C. S. Taber and M. Lodge. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769, 2006.

[47] J. Teevan, D. Ramage, and M. R. Morris. # twittersearch: a comparison of microblog search and web search. In *WSDM '11 Proceedings of the fourth ACM international conference on Web search and data mining*, pages 35–44, Hong Kong, China, 2011. ACM.

[48] A. Tsang, B. Wilder, E. Rice, M. Tambe, and Y. Zick. Group-Fairness in Influence Maximization. *arXiv preprint arXiv:1903.00967*, 2019.

[49] E. K. Vraga and L. Bode. Using expert sources to correct health misinformation in social media. *Science Communication*, 39(5):621–645, 2017.

[50] P. Weber. The virtual get-together. determinants interpersonal - "o public communication on news websites. *Social Media and Web Science. Frankfurt am Main: DGI*, pages 457–459,

2012.

[51] P. Weber. Discussions in the comments section: Factors influencing participation and interactivity in online newspapers' reader comments. *New media & society*, 16(6):941–957, 2014.

[52] L. A. Wolsey. An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, 2(4):385–393, 1982.

[53] T. Wood and E. Porter. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 41(1):135–163, 2019.

Engineering systems

FRCCS

Approximate information for efficient exploration-exploitation strategies

Alex Barbier–Chebbah[1✓], Christian L. Vestergaard[1], Etienne Boursier[2] and Jean-Baptiste Masson[1]

[1] *Institut Pasteur, Université Paris Cité, CNRS UMR 3571, Paris France; Épimethée, Inria, Paris, France; alex.barbier-chebbah@pasteur.fr, christian.vestergaard@pasteur.fr, jbmasson@pasteur.fr.*
[2] *INRIA, Université Paris Saclay, LMO, Orsay, France; etienne.boursier1@gmail.com.*
[✓] *Presenting author*

Abstract. The exploration-exploitation dilemma is a fundamental challenge in decision-making, inherently present in multi-armed bandit problems. These problems involve an agent deciding whether to exploit current knowledge for immediate gains or to explore new avenues for potential long-term benefits. Leveraging the information maximization principle, we developed a novel and efficient class of bandit algorithms that employs a carefully chosen analytical approximation of entropy to forecast the information gain of each action and greedily chooses the one with the highest information gain at each point in time.

Keywords. *Learning Theory; Multi-armed bandits; Exploitation and exploration; Information maximization*

1 Contextes

The exploration-exploitation dilemma is a fundamental challenge in decision-making and is ubiquitous in various fields, from anomaly detection [2] to the modeling of biological search strategies [14] and human decision-making [1, 3]. It occurs when an agent faces the decision of using its existing knowledge to maximize short-term gains or seeking new information that could result in higher long-term benefits.

The multi-armed bandit (MAB) problem is a paradigmatic example that embodies the explore-exploit tradeoff. The classical MAB model is a simple slot machine game where the goal is to maximize the payout. As such, the agent is presented with a set of possible actions, or "arms" to pull, each associated with a probabilistic reward. Since pulling sub-optimal solutions is costly, MAB algorithms must carefully quantify their exploration time and have to be robust to noisy inputs. As a direct consequence, this abstract framework finds application across a wide spectrum of domains, comprising neuroscience [7], clinical trials [15], epidemic control [6], and reinforcement learning [11], among others. In addition, there is a rich mathematical literature on algorithms that provide optimal results [12, 8, 13, 4]. Even if these approaches efficiently utilize currently available information, they do not aim directly to acquire more information.

Information-maximization approaches provide a decision-making strategy where the agent tries

to maximize their information about one or a set of relevant stochastic variables. The information-maximizing principle has shown to be efficient in a broad range of domains [14] where decisions must be made in fluctuating or unknown environments, and its application to classical bandit settings, coined Infomax [9], has shown promising empirical performance.

As a consequence, we have developed a novel class of algorithms based on this information maximization principle, focusing on analytical tractability, computational efficiency, and extensibility, while demonstrating their robustness across a range of priors.

2 Results

Our main contribution is the introduction of a new class of asymptotically optimal algorithms, denoted AIM, that rely on approximations of a functional representing the current information of interest about the entire bandit system. More precisely, we choose the entropy of the posterior distribution of the value of the maximal mean reward. However, because it often cannot be computed in closed form, we have developed a second, simplified, and analytic functional that mirrors the entropy.

This analytic result strengthens the information maximization principle by providing novel algorithms that are analytical, tractable, and computationally efficient, while also preserving the main advantages of Infomax, a crucial challenge for information methods, as stressed in [10].

Our resulting algorithms, denoted as AIM, have shown strong empirical performance compared to state-of-the-art baselines for classic rewards settings, while providing outstanding effectiveness when facing multiple arms.

Furthermore, for the 2-armed Gaussian rewards case, we demonstrate that AIM attains the Lai and Robbins bound [5], thereby ensuring optimal guarantees in the infinite horizon setting.

Our work has already led to a first article published in Physical Review E at `https://journals.aps.org/pre/accepted/bd076R7eT24E3e1c71565c369af71607c217fce3d`. More recently, a second study, currently under review, which proves the asymptotic optimality of AIM and addressing its potential generalization to more correlated information structures can be found at `https://arxiv.org/abs/2310.12563`.

References

[1] Jonathan D Cohen, Samuel M McClure, and Angela J Yu. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942, March 2007.

[2] Kaize Ding, Jundong Li, and Huan Liu. Interactive Anomaly Detection on Attributed Networks. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pages 357–365, New York, NY, USA, January 2019. Association for Computing Machinery.

[3] Thomas T. Hills, Peter M. Todd, David Lazer, A. David Redish, and Iain D. Couzin. Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, 19(1):46–54, January 2015.

[4] Junya Honda and Akimichi Takemura. An Asymptotically Optimal Bandit Algorithm for Bounded Support Models. In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT*

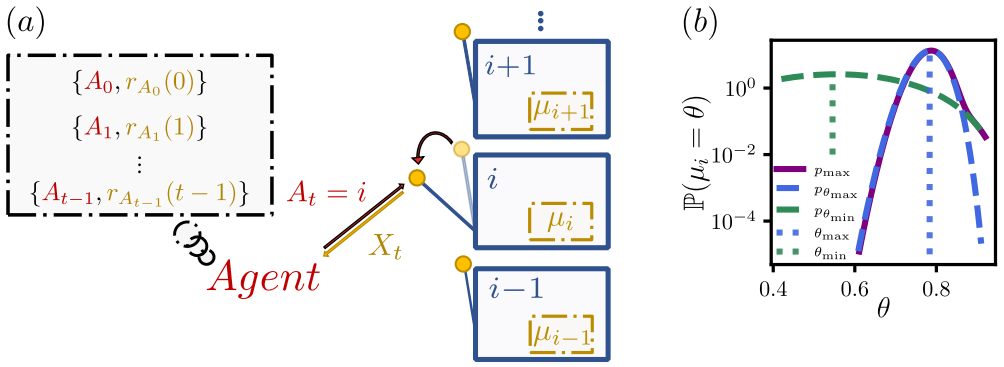


Figure 1: **(a)** Illustration of the multi-armed bandit problem. At each time step t the agent chooses an action $i = A_t$ that returns a reward drawn from a distribution of unknown mean μ_i. The agent's goal is to minimize the cumulative regret. **(b)** Posterior distributions of bandit values after playing the 2-armed Bernoulli game where $r_i(t)$, $n_i(t)$ are respectively the cumulative reward and number of draws of arm i. In blue, the posterior distribution, $p_{\theta_{\max}}$, of the reward of the current best arm. Vertical green and blue lines are the posterior mean rewards of the suboptimal (denoted $\theta_{\min}$) and better empirical arm ($\theta_{\max}$). In green, the posterior distribution, $p_{\theta_{\min}}$, of the current suboptimal arm. Infomax algorithm will choose the arm minimizing the entropy associated to the posterior distribution, $p_{\max}$, of the maximum reward of all arms.

2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010, pages 67–79. Omnipress, 2010.

[5] T. L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, March 1985.

[6] Baihan Lin and Djallel Bouneffouf. Optimal Epidemic Control as a Contextual Combinatorial Bandit with Budget. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8, July 2022.

[7] Dimitrije Marković, Hrvoje Stojić, Sarah Schwöbel, and Stefan J. Kiebel. An empirical evaluation of active inference in multi-armed bandits. *Neural Networks*, 144:229–246, December 2021.

[8] Sebastian Pilarski, Slawomir Pilarski, and Dániel Varró. Optimal Policy for Bernoulli Bandits: Computation and Algorithm Gauge. *IEEE Transactions on Artificial Intelligence*, 2(1):2–17, February 2021.

[9] Gautam Reddy, Antonio Celani, and Massimo Vergassola. Infomax Strategies for an Optimal Balance Between Exploration and Exploitation. *Journal of Statistical Physics*, 163(6):1454–1476, April 2016.

[10] Daniel Russo and Benjamin Van Roy. Learning to Optimize via Posterior Sampling. *Mathematics of OR*, 39(4):1221–1243, November 2014.

[11] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016.

[12] Aleksandrs Slivkins. Introduction to Multi-Armed Bandits. *MAL*, 12(1-2):1–286, November 2019.

[13] William R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another

in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285–294, 1933.

[14] Massimo Vergassola, Emmanuel Villermaux, and Boris I. Shraiman. 'Infotaxis' as a strategy for searching without gradients. *Nature*, 445(7126):406–409, January 2007.

[15] Sofía S. Villar. Bandits strategies evaluated in the context of clinical trials in rare life-threatening diseases. *Probability in the Engineering and Informational Sciences*, 32(2):229–245, April 2018.

FRCCS

Fine-Tuning LLMs Or Zero/Few-Shot Prompting for Knowledge Graph Construction?

Hussam Ghanem[1✓] and Christophe Cruz[2]

[1] *ICB, UMR 6306, CNRS, Université de Bourgogne, 21000 Dijon, France*
DAVI The Humanizers, Puteaux, France
[2]*ICB, UMR 6306, CNRS, Université de Bourgogne, 21000 Dijon, France*

✓ *Presenting author*

Abstract. This paper explores Text-to-Knowledge Graph (T2KG) construction,, assessing Zero-Shot Prompting (ZSP), Few-Shot Prompting (FSP), and Fine-Tuning (FT) methods with Large Language Models (LLMs). Through comprehensive experimentation with Llama2, Mistral, and Starling, we highlight the strengths of FT, emphasize dataset size's role, and introduce nuanced evaluation metrics. Promising perspectives include synonym-aware metric refinement, and data augmentation with LLMs. The study contributes valuable insights to KG construction methodologies, setting the stage for further advancements.[a]

Keywords. *Text-to-Knowledge Graph; Large Language Models; Zero-Shot Prompting; Few-Shot Prompting; Fine-Tuning*

[a]Our code at `https://github.com/ChristopheCruz/LLM4KGC/`

1 Introduction

The term "knowledge graph" has been around since 1972, but its current definition can be traced back to Google's 2012. This was followed by similar announcements from companies such as Airbnb, Amazon, eBay, Facebook, IBM, LinkedIn, Microsoft, and Uber, among others, leading to an increase in the adoption of Knowledge graphs(KGs) by various industries. As a result, academic research in this field has seen a surge in recent years, with an increasing number of scientific publications on KGs [1]. These graphs utilize a graph-based data model to effectively manage, integrate, and extract valuable insights from large and diverse datasets [2].

KGs serve as repositories for structured knowledge, organized into a collection of triples, denoted as $KG = (h, r, t) \subseteq E \times R \times E$, where E represents the set of entities, and R represents the set of relations [1]. Within a graph, nodes represent various levels, entities, or concepts. These nodes encompass diverse types, including person, book, or city, and are interconnected by relationships such as located in, lives in, or works with. The essence of a KG emerges when it incorporates multiple types of relationships rather than being confined to a single type. The overarching structure of a KG constitutes a network of entities, featuring their semantic types, properties, and interconnections. Thus, constructing a KG necessitates information

about entities (along with their types and properties) and the semantic relationships that bind them. For the extraction of entities and relationships, practitioners often turn to NLP tasks like Named Entity Recognition (NER), Coreference Resolution (CR), and Relation Extraction (RE).

KGs play a crucial role in organizing complex information across diverse domains, such as question answering, recommendations, semantic search, etc. However, the ongoing challenge persists in constructing them, particularly as the primary sources of knowledge are embedded in unstructured textual data such as press articles, emails, and scientific journals. This challenge can be addressed by adopting an information extraction approach, sometimes implemented as a pipeline. It involves taking textual inputs, processing them using Natural Language Processing (NLP) techniques, and leveraging the acquired knowledge to construct or enhance the KG.

If we envision the Text-to-Knowledge Graph (T2KG) construction task as a black box, the input is textual data, and the output is a knowledge graph. Achieving this can be approached through methods that directly convert text into a graph or by implementing NLP tasks in two ways: 1) through an information extraction pipeline incorporating the mentioned tasks independently, or 2) by adopting an end-to-end approach, also known as joint prediction, using Large Language Models (LLMs) for example. In the realm of LLMs and KGs, their mutual enhancement is evident. LLMs can assist in the construction of KGs. Conversely, KGs can be employed to validate outputs from LLMs or provide explanations for them [3]. LLMs can be adapted to KG construction task (T2KG) through various approaches, such as fine-tuning [4] (FT), zero-shot prompting [5] (ZSP), or few-shot prompting (FSP) [6] with a limited number of examples. Each of these approaches has their pros and cons with respect to the performance, computation resources, training time, domain adaption and training data required.

In-context learning, as discussed by [7], coupled with prompt design, involves telling a model to execute a new task by presenting it with only a few demonstrations of input-output pairs during inference. Instruction fine-tuning methods, exemplified by InstructGPT [8] and Reinforcement Learning from Human Feedback (RLHF) [9], markedly enhance the model's ability to comprehend and follow a diverse range of written instructions. Numerous LLMs have been introduced in the last year, as highlighted by [3], particularly within the ChatGPT [42] like models, which includes GPT-3 [10], LLaMA [11], BLOOM [12], PaLM [13], Mistral [14], Starling [18] and Zephyr [16]. These models can be readily repurposed for KG construction from text by employing a prompt design that incorporates instructions and contextual information.

This study does not entail a comparison with traditional methods of constructing KGs; rather, it delves into the developments and challenges associated with KG construction methodologies, and aiming at providing formal evaluation of T2KG task. Specifically, we focus on the utilization of LLMs, and explore the three approaches mentioned before, Zero-shot, Few-shot and Fine-tuning (Fig. 1). Each of these approaches addresses specific challenges, contributing significantly to the evolution of T2KG construction techniques.

The present study is organized as follows, Section 2 presents a comprehensive overview of the current state-of-the-art approaches for Text to KG (T2KG) Construction. In the Section 3, we present the general architecture of our proposed implementation (method), with datasets, metrics, and experiments. Section 4 then encapsulates the findings and discussions, presenting the culmination of results. Finally, Section 5 critically examines the strengths and limitations of these techniques.

2 Background

The current state of research on knowledge graph construction using LLMs is discussed. Three main approaches are identified: Zero-Shot, Few-Shot, and Fine-Tuning. Each approach has its own challenges, such as maintaining accuracy without specific training data or ensuring the robustness of models in diverse real-world scenarios. Evaluation metrics used to assess the quality of constructed KGs are also discussed, including semantic consistency and linguistic coherence. This section highlight methods and metrics to construct KGs and evaluate the result.

The figure 1 illustrates the black box joint prediction of the T2KG construction process using LLMs. It demonstrates how two French examples on the left are converted into an expected result (KG) on the right using ZSP, FSP or FT approaches with LLMs.

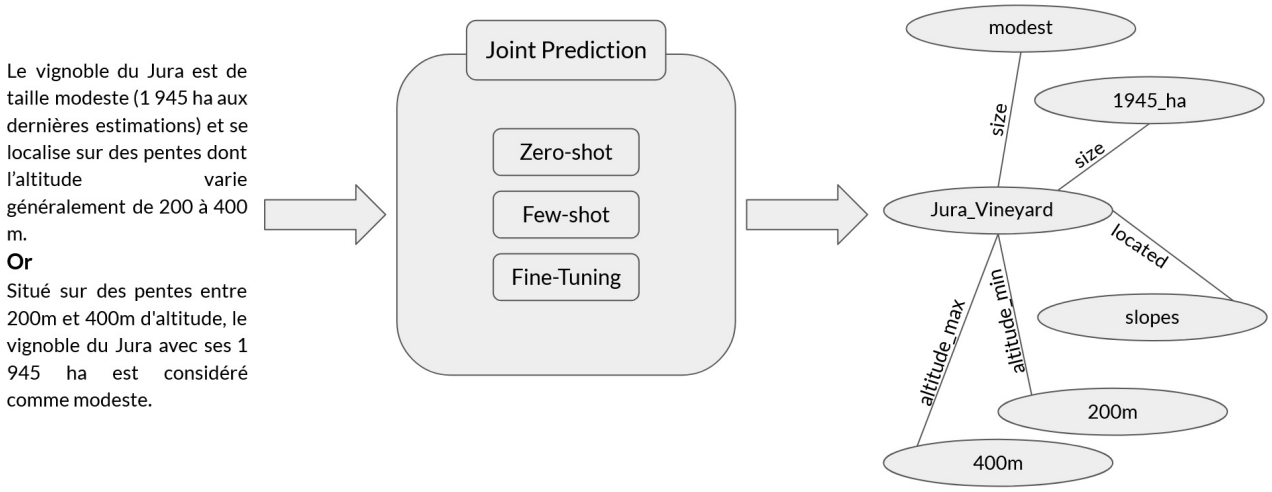


Figure 1: T2KG Task

2.1 Zero Shot

Zero Shot methods enable KG construction without task-specific training data, leveraging the inherent capabilities of large language models. [17] introduce an innovative approach using large language models (LLMs) for knowledge graph construction, employing iterative zero-shot prompting for scalable and flexible KG construction. [18] evaluate the performance of LLMs, specifically GPT-4 and ChatGPT, in KG construction and reasoning tasks, introducing the Virtual Knowledge Extraction task and the VINE dataset, but they do not take into account open sourced LLMs as LLaMA [11]. [19] assess ChatGPT's abilities in information extraction tasks, identifying overconfidence as an issue and releasing annotated datasets. [20] tackle zero-shot information extraction using ChatGPT, achieving impressive results in entity relation triple extraction. [21] propose a method for Knowledge Graph Construction (KGC) using an analogy-based approach, demonstrating superior performance on Wikidata. [22] address the limitations of existing generative knowledge graph construction methods by leveraging large generative language models trained on structured data. The most of these approaches having the same limitation, which is the use of closed and huge LLMs as ChatGPT or GPT4 for this task. Challenges in this area include maintaining accuracy without specific training data and addressing nuanced relationships between entities in untrained domains.

2.2 Few Shot

Few Shot methods focus on constructing KGs with limited training examples, aiming to achieve accurate knowledge representation with minimal data. [6] introduce PiVe, a framework enhancing the graph-based generative capabilities of LLMs, and the authors create a verifier which is responsable to verifie the results of LLMs with multi-iteration type. [23] explore the potential of LLMs for knowledge graph completion, treating triples as text sequences and utilizing LLM responses for predictions. [24] automate the process of generating structured knowledge graphs from natural language text using foundation models. [25] present OpenBG, an open business knowledge graph derived from Alibaba Group, containing 2.6 billion triples with over 88 million entities. [26] explore the integration of LLMs with semantic technologies for reasoning and inference. [27] investigate LLMs' application in relation labeling for e-commerce Knowledge Graphs (KGs). As ZSP approaches, FSP approaches use closed and huge LLMs as ChatGPT or GPT4 [42] for this task. Challenges in this area include achieving high accuracy with minimal training data and ensuring the robustness of models in diverse real-world scenarios.

2.3 Fine-Tuning

Fine-Tuning methods involve adapting pre-trained language models to specific knowledge domains, enhancing their capabilities for constructing KGs tailored to particular contexts. [4] present a case study automating KG construction for compliance using BERT-based models. This study emphasizes the importance of machine learning models in interpreting rules for compliance automation. [28] propose an approach for knowledge extraction and analysis from biomedical clinical notes, utilizing the BERT model and a Conditional Random Field layer, showcasing the effectiveness of leveraging BERT models for structured biomedical knowledge graphs. [29] propose Knowledge Graph-Enhanced Large Language Models (KGLLMs), enhancing LLMs with KGs for improved factual reasoning capabilities. These approaches that applied FT, they do not use new generations of LLMs, specially, decoder only LLMs as Llama, and Mistral. Challenges in this domain include ensuring the scalability, interpretability, and robustness of fine-tuned models across diverse knowledge domains.

2.4 Evaluation metrics

As we employ LLMs to construct KGs, and given that LLMs function as Natural Language Generation (NLG) models, it becomes imperative to discuss NLG criteria. In NLG, two criteria [30] are used to assess the quality of the produced answers (triples in our context).

The first criterion is semantic consistency or Semantic Fidelity which quantifies the fidelity of the data produced against the input data. The most common indicators are :

- **Hallucination**: It is manifested by the Presence of information (facts) in the generated text that is absent in the input data. In our scenario, hallucination exists if the generated triples (GT) contain triples not present in the ground truth triples (ET) (T in GT and not in ET);
- **Omission**: It is manifested by the omission of one of the pieces of information (facts) in the generated text. In our case, omission occurs if a triple is present in ET but not in GT;
- **Redundancy**: This is manifested by the repetition of information in the generated text. In our case, the redundancy exists if a triple appears more than once in GT;

- **Accuracy**: The lack of accuracy is manifested by the modification of information such as the inversion of the subject and the direct object complement in the generated text. Accuracy increases if there is an exact match between ET and GT. ;

- **Ordering**: It occurs when the sequence of information is different from the input data. In our case, the ordering of GT is not consider.

The second criterion is linguistic coherence or Output Fluency to evaluate the fluidity of the text and the linguistic constructions of the generated text, the segmentation of the text into different sentences, the use of anaphoric pronouns to reference entities and to have linguistically correct sentences. However, in our evaluation, we do not take into account the second criterion.

In their experiments, [3] calculated three hallucination metrics - subject hallucination, relation hallucination, and object hallucination - using certain preprocessing steps such as stemming. They used the ground truth ontology alongside the ground truth test sentence to determine if an entity or relation is present in the text. However, a limitation could arise when there is a disparity in entities or relations between the ground truth ontology and the ground truth test sentence. If the generated triples contain entities or relations not present in the ground truth text, even if they exist in the ground truth ontology, it will be considered a hallucination.

The authors of [6] evaluate their experiments using several evaluation metrics, including Triple Match F1 (T-F1), Graph Match F1 (G-F1), G-BERTScore (G-BS) from [36] and Graph Edit Distance (GED) from [37]. The GED metric measures the distance between the predicted graph and the ground-truth graph, which is equivalent to computing the number of edit operations (addition, deletion, or replacement of nodes and edges) needed to transform the predicted graph into a graph that is identical to the ground-truth graph, but it does not provide a specific path for these operations to calculate the exact number of operations. To adhere with semantic consistency criterion, we use the terms "omission" and "hallucination" in place of "addition" and "deletion," respectively.

3 Propositions

This section describes our approach to evaluate the quality of generated KGs. We explain how we use evaluation metrics such as T-F1, G-F1, G-BS, GED, Bleu-F1 [38] and ROUGE-F1 [39] to assess the quality of the generated KGs in comparison to ground-truth KGs. Additionally, we discuss the use of Optimal Edit Paths (OEP) metric [1] to determine the precise number of operations required to transform the predicted graph into an identical representation of the ground-truth graph. This metric serves as a basis for calculating omissions and hallucinations in the generated graphs. We employ examples from the WebNLG+2020 dataset [40] for testing with ZSP and FSP techniques. Additionally, we utilize the training dataset of WebNLG+2020 to train LLMs using the FT technique. Subsequent subsections delve into a detailed discussion of each phase.

[1] NetworkX - optimal edit paths : `https://networkx.org/documentation/stable/index.html`

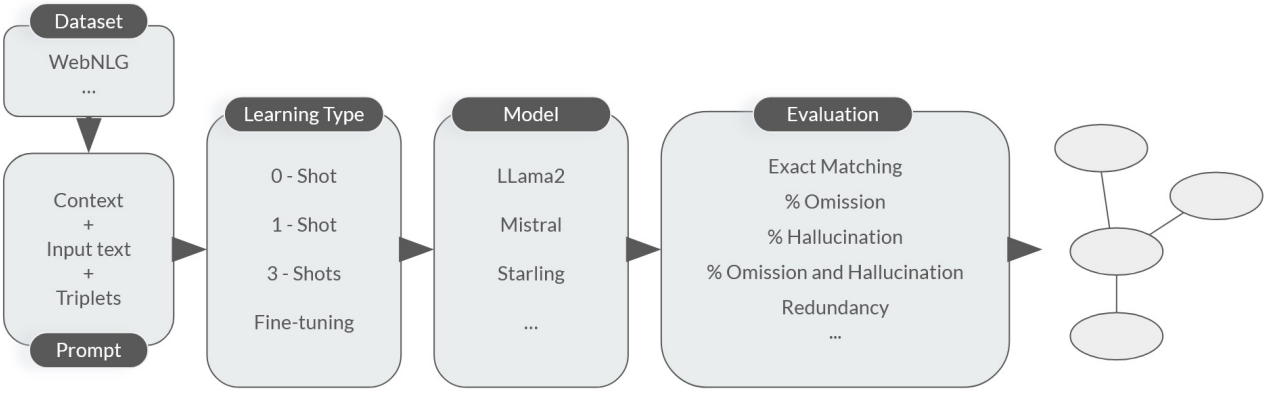


Figure 2: Overall experimentation's process

3.1 Overall experimentation's process

We leverage the WebNLG+2020 dataset, specifically the version curated by [6]. Their preparation of graphs in lists of triples proves beneficial for evaluation purposes. We utilize these lists and employ NetworkX [41] to transform them back into graphs, facilitating evaluations on the resultant graphs. This step is instrumental in performing ZSP, FSP, and FT LLMs on this dataset.

The figure 2 illustrates the different stages of our experimentation process, including data preparation, model selection, training, validation, and evaluation. The process begins with data preparation, where the WEBNLG dataset is preprocessed and split into training, validation, and test sets. Next, the learning type is selected, and different models are trained using the training set. The trained models are then evaluated on the validation set to evaluate their performance. Finally, the best-performing model is selected and validated on the test set to estimate its generalization ability.

3.2 Prompting learning

During this phase, we employ the ZSP and FSP techniques on LLMs to evaluate their proficiency in extracting triples (e.g. construction of the KG). The application of these techniques involves merging examples from the test dataset of WebNLG+2020 with our adapted prompt. Our prompt is strategically modified to provide contextual guidance to the LLMs, facilitating the effective extraction of triples, without the inclusion of a support ontology description, as demonstrated in [3]. The specific prompts used for ZSP and FSP are illustrated in Fig 3(a) and Fig 3(b),

In our approach for ZSP, we began with the methodology outlined in [6], initiating our prompt with the directive "Transform the text into a semantic graph." However, we enhanced this prompt by incorporating additional sentences tailored for our LLMs, as illustrated in Fig 3.(a).

For FSP, we executed 7-shots learning. The rationale behind employing 7-shots learning lies in the fact that the maximum KG size in WebNLG+2020 is 7 triples. Consequently, we fed our prompt with 7 examples of varying sizes; example 1 with size 1, example 2 with size 2, example 3 with size 3, and so forth. In Figure 3-b, we depict a prompt containing two examples.

To demonstrate the efficacy of our refined prompt (including additional sentences), we conducted zero-shot experiments on ChatGPT [42], comparing the outcomes with those of [6]. Our results consistently reveal that our prompt yields more coherent answers in terms of struc-

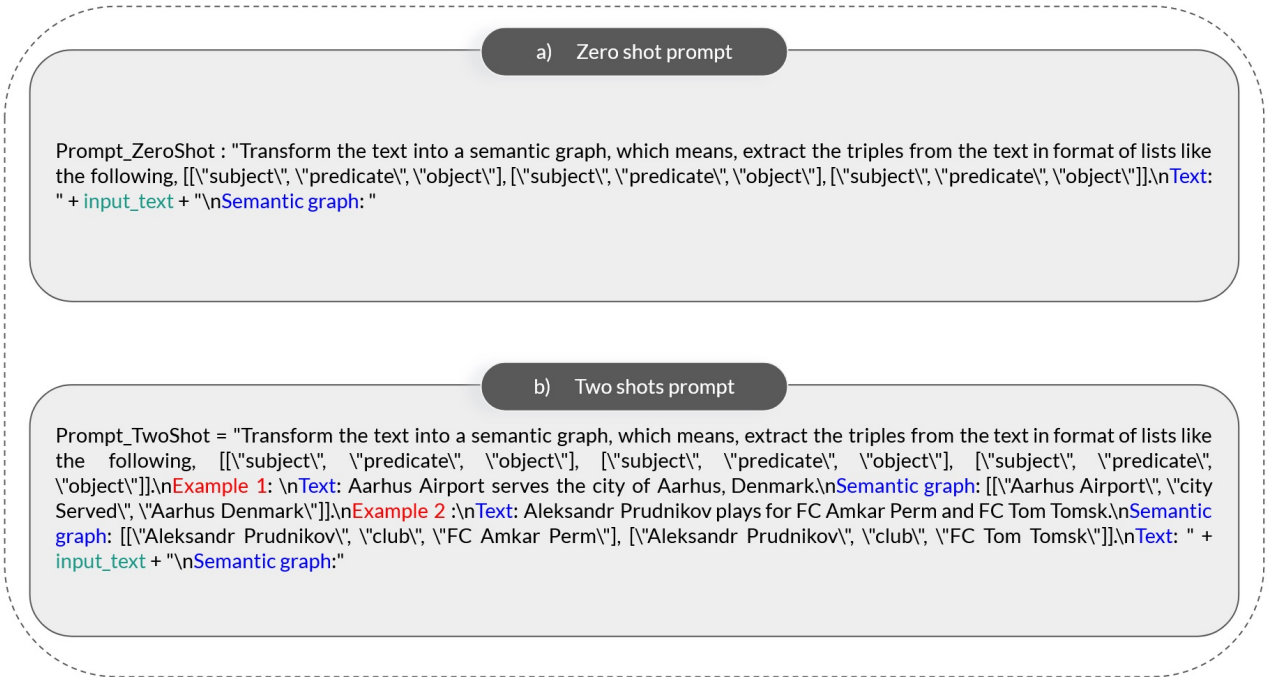


Figure 3: Prompting examples

ture.

3.3 Finetuning

If the initial results from the ZSP and FSP on LLMs prove reasonable, we proceed to the FT phase. This phase aims to provide the LLMs with a more specific context and knowledge related to the task of extracting triples within the domains covered by the WebNLG+2020 dataset. Using the example "a)" illustrated in Fig 3, we passe in the FT prompt, at once for each line of the training dataset, the input text and the corresponding KG (the list of triples). To do this phase (FT), we employ QLoRA [31], a methodology that integrates quantization [32] and Low-Rank Adapters (LoRA) [33]. The LLM is loaded with 4-bit precision using bitsandbytes [34], and the training process incorporates LoRA through the PEFT library (Parameter-Efficient Fine-Tuning) [35] provided by Hugging Face.

3.4 Postprocessing

Given our focus on KG construction, our evaluation process involves assessing the generated KGs against ground-truth KGs. To facilitate this evaluation, we take a cleaning process for the LLMs output. This involves transforming the graphs generated by LLMs into organized lists of triples, subsequently transferred to textual documents.

The transformation is executed through rule-based processing. This step is applied to remove corrupted text (outside the lists of triples) from the whole text generated by LLMs in the preceding step. The output is then presented in a list of lists of triples format, optimizing our evaluation process. This approach proves especially effective when calculating metrics such as G-F1, GED, and OEP, as we will see in more detail in 3.5

A potential problem arises when instructing LLMs to produce lists of triples (KGs), as there may be instances where the generated text lacks the desired structure. In such cases, we

address this issue by substituting the generated text with an empty list of triples, represented as '[["","",""]]', allowing us to effectively evaluate omissions. However, this approach tends to underestimate hallucinations compared to the actual occurrences.

3.5 Experiment's evaluation

For assessing the quality of the generated graphs in comparison to ground-truth graphs, we adopt evaluation metrics as employed in [6]. These metrics encompass T-F1, G-F1, G-BS [36], and GED [37]. Additionally, we incorporate the Optimal Edit Paths (OEP) metric, a tool aiding in the calculation of omissions and hallucinations within the generated graphs.

Our evaluation procedure aligns with the methodology outlined in [6], particularly in the computation of GED and G-F1. This involves constructing directed graphs from lists of triples, referred to as linearized graphs, utilizing NetworkX [41].

In contrast to [3], our methodology diverges by not relying on the ground truth test sentence of an ontology. As previously mentioned, we opt for a distinct approach wherein we assess omissions and hallucinations in the generated graphs using the OEP metric. Unlike the global edit distance provided by GED, OEP gives the precise path of the edit, enabling the exact quantification of omissions and hallucinations, either in absolute terms or as a percentage across the entire test dataset.

For example, in the illustrated nodes path labeled 'a)' in Fig 4-(b), we observe 2 omissions, while the edges path in Fig 4-(a) exhibits 1 hallucination. In our evaluation, the criterion for incrementing the global hallucination metric for all graphs is set at finding >=1 hallucinations or 1 omission in a generated graph. This approach ensures a comprehensive assessment of the presence of omissions and hallucinations across the entirety of the generated graphs.

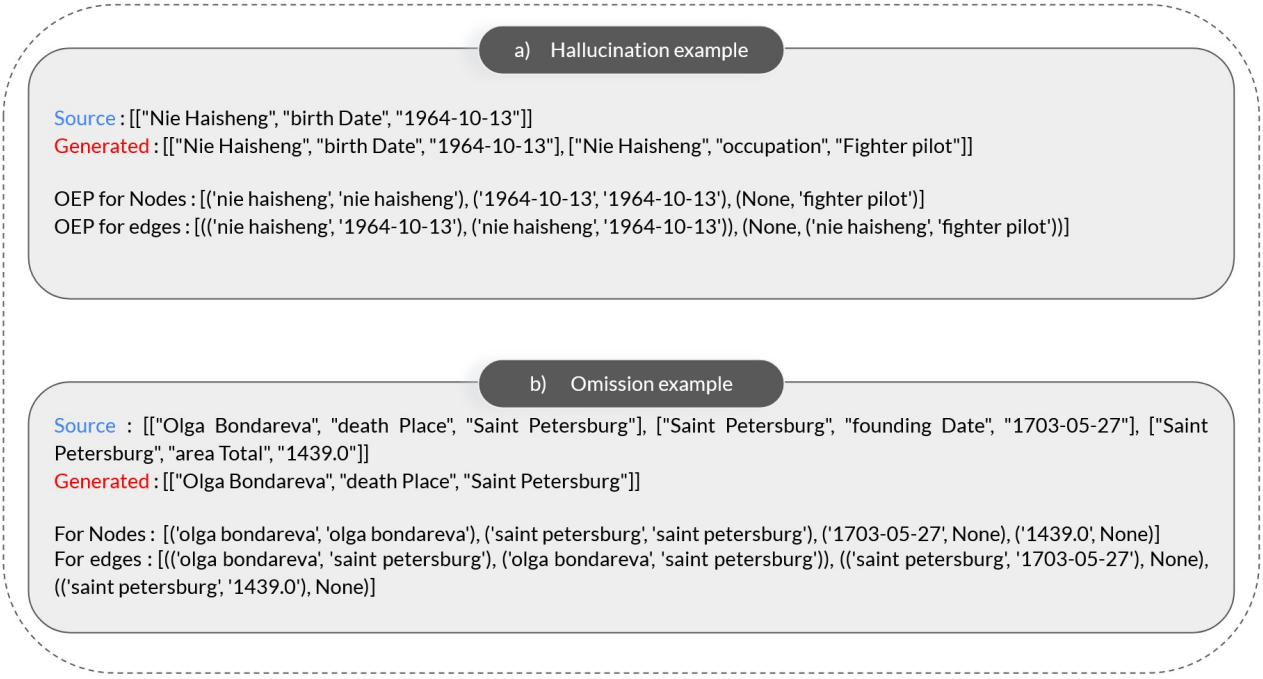


Figure 4: Results examples

As mentioned earlier, the evaluation of the three methods is conducted using examples sourced from the test dataset of WebNLG+2020. The primary goal is to enhance G-F1, T-F1, G-BS, Bleu-F1, and ROUGE-F1 metrics, while reducing GED, Hallucination, and Omission.

4 Experiments

This section provides insights into the LLMs utilized in our study for ZSP, FSP, or FT, followed by the presentation of our experimental results.

In this section, we provide a brief overview of the LLMs utilized in our experiments. Our selection criteria focused on employing small, open-source, and easily accessible LLMs. All models were sourced from the HuggingFace platform[2]

- **Llama 2** [43] is a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. In our experiments, we deploy the 7B and 13B pretrained models, which have been converted to the Hugging Face Transformers format.
- Introduced by [14], **Mistral-7B-v0.1** is a pretrained generative text model featuring 7 billion parameters. Notably, Mistral-7B-v0.1 exhibits superior performance to Llama 2 13B across all benchmark tests in their experiments.
- In the work presented by [18], **Starling-7B** is introduced as an open LLM trained through Reinforcement Learning from AI Feedback (RLAIF). This model leverages the GPT-4 labeled ranking dataset, berkeley-nest/Nectar, and employs a novel reward training and policy tuning pipeline.

In our review of the state-of-the-art, we observed that, apart from [3], which incorporates hallucination evaluation in their experiments, other studies primarily focus on metrics such as precision, recall, F1 score, triple matching, or graph matching. In our approach to evaluating experiments, we consider also hallucination and omission through a linguistic lens.

Upon examining Table 1, we observe the superior performance of the FT method compared to ZSP and FSP for the T2KG construction task. Of particular interest is the finding that, with the exception of Llama2-7b, applying ZSP to the fine-tuned Llama2-7b results in worse performance than FSP on the original Llama2-7b. Overall, this table provides a clear visualization of the relative performance of each method, highlighting the strengths and limitations of each approach for T2KG construction.

Furthermore, it is evident that better results are achieved by providing more examples (more shots) to the same model, whether original or fine-tuned. The results underscore the positive correlation between the quantity of examples and the model's performance. Comparing the fine-tuned Mistral and fine-tuned Starling, they exhibit similar performance when given 7 shots, surpassing the two Llama2 models by a significant margin. The standout performer with ZSP on the fine-tuned LLM is Mistral, showcasing a considerable lead over other LLMs, including Starling. To corroborate these findings, future versions of our study plan to assess our fine-tuned models using an alternative dataset with diverse domains.

As depicted in Figure 2, Hall. represents Hallucinations, while Omis. denotes Omissions.

Taking into account our strategy of introducing an empty graph when LLMs fail to produce triples, we note that even with LLama2-13b with ZSP exhibiting the least favorable results across all metrics, it displays minimal hallucinations. Nonetheless, it's crucial to recognize that the model with the fewest hallucinations may not necessarily be the most suitable choice. To overcome this limitation in our evaluation metric, we aim to improve it by considering the

[2]Hugging Face: `https://huggingface.co/`

Table 1: Comparison of performance metrics and models

Model | Metric	G-F1	T-F1	G-BS	GED	F1-Bleu	F1-Rouge	Hall.	Omis.
PiVE	14.00	18.57	89.82	11.22	-	-	-	-
Mistral-0	2.30	0.00	77.87	15.93	54.97	55.15	20.63	31.48
Mistral-7	18.72	28.44	87.54	10.13	55.09	63.94	17.88	21.14
Mistral-FT-0	31.93	44.08	86.89	8.25	63.88	69.08	13.55	18.27
Mistral-FT-7	34.68	**49.11**	**91.99**	6.69	**71.78**	**77.43**	15.01	**14.45**
Starling-0	5.23	7.83	86.29	13.35	34.64	14.61	17.48	33.24
Starling-7	21.30	33.77	90.41	8.96	60.47	69.34	17.31	14.61
Starling-FT-0	21.47	28.29	72.86	11.87	44.07	47.69	10.17	42.78
Starling-FT-7	**35.69**	48.49	91.95	**6.60**	71.51	76.67	11.35	18.27
Llama2-7b-0	0.00	0.46	54.20	18.29	20.23	17.98	4.83	81.53
Llama2-7b-7	11.80	20.88	82.78	12.66	45.48	54.29	20.74	30.02
Llama2-7b-FT-0	3.82	15.41	59.19	15.78	16.82	17.95	6.07	79.20
Llama2-7b-FT-7	18.77	32.63	87.19	10.16	58.48	66.35	25.24	18.66
Llama2-13b-0	0.00	0.79	57.42	17.79	20.50	18.23	**4.78**	81.23
Llama2-13b-7	13.49	23.99	84.89	11.59	50.18	58.71	26.36	19.06
Llama2-13b-FT-0	20.52	32.18	75.88	11.38	46.53	50.78	11.64	39.63
Llama2-13b-FT-7	23.55	37.29	88.77	8.94	63.26	70.12	23.55	16.19

prevalence of empty graphs in the generated results before assessing them against ground truth graphs.

The G-BS consistently remains high, indicating that LLMs frequently generate text with words (entities or relations) very similar to those in the ground truth graphs. Among the models, the finetuned Starling with 7 shots achieves the highest G-F1, which focuses on the entirety of the graph and evaluates how many graphs are exactly produced the same, suggesting that it accurately generates approximately 36% of graphs identical to the ground truth. For various metrics, the finetuned Mistral with 7 shots performs exceptionally well, particularly in T-F1, where F1 scores are computed for all test samples and averaged for the final Triple Match F1 score. Additionally, it excels in metrics such as "Omis.," F1-Bleu, and F1-Rouge. F1-Bleu and F1-Rouge represent n-gram-based metrics encompassing precision (Bleu), recall (Rouge), and F-score (Bleu and Rouge). These metric could potentially yield even better results if synonyms of entities or relations are considered as exact matches.

The authors in [6] conduct evaluations using WebNLG+2020. Consequently, we adopt their approach (PiVE) as a baseline for comparison with our experiments. Upon analyzing the results, it becomes evident that nearly all fine-tuned LLMs outperform PiVE, which is applied on both ChatGPT and GPT-4 as mentioned before.

5 Conclusion and perspectives

This study delves into the Text-to-Knowledge Graph (T2KG) construction task, exploring the efficacy of three distinct approaches: Zero-Shot Prompting (ZSP), Few-Shot Prompting (FSP), and Fine-Tuning (FT) of Large Language Models (LLMs). Our comprehensive experimentation, employing models such as Llama2, Mistral, and Starling, sheds light on the strengths and

limitations of each approach. The results demonstrate the remarkable performance of the FT method, particularly when compared to ZSP and FSP across various models. Notably, the fine-tuned Llama2-7b with ZSP gaved worst results than FSP with the original Llama2. Additionally, the positive correlation between the quantity of examples and model performance underscores the significance of dataset size in training. An essential part of our study involves the evaluation metrics employed to assess the generated graphs. Particularly, we introduced nuanced considerations for refining these metrics to measuring hallucination and omission in the generated graphs, offering valuable insights into the fidelity of the constructed knowledge graphs.

Looking forward, there are promising perspectives for further enhancement. One is to involve refining evaluation metrics to accommodate synonyms of entities or relations in generated graphs, employing advanced methods or tools for synonym detection. Furthermore, leveraging LLMs for data augmentation in the T2KG construction task shows promise. Notably, during experimentation, LLMs, particularly Starling, exhibited the ability to provide continuity in generated results for T2KG, proposing texts alongside corresponding KGs (triples).

References

[1] Hogan A, Blomqvist E, Cochez M, d'Amato C, Melo GD, Gutierrez C, Kirrane S, Gayo JE, Navigli R, Neumaier S, Ngomo AC. Knowledge graphs. ACM Computing Surveys (CSUR). 2021 Jul 2;54(4):1-37. Author, Article title, Journal, Volume, page numbers (year)

[2] N. F. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor. 2019. Industry-scale knowledge graphs: Lessons and challenges. ACM Queue 17, 2 (2019).

[3] Mihindukulasooriya, Nandana, et al. "Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text." International Semantic Web Conference. Cham: Springer Nature Switzerland, 2023.

[4] Ershov, Vladimir. "A Case Study for Compliance as Code with Graphs and Language Models: Public release of the Regulatory Knowledge Graph." arXiv preprint arXiv:2302.01842 (2023).

[5] Caufield, J. Harry, et al. "Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning." arXiv preprint arXiv:2304.02711 (2023).

[6] Han, Jiuzhou, et al. "PiVe: Prompting with Iterative Verification Improving Graph-based Generative Capability of LLMs." arXiv preprint arXiv:2305.12392 (2023).

[7] Min, Sewon, et al. "Rethinking the role of demonstrations: What makes in-context learning work?." arXiv preprint arXiv:2202.12837 (2022).

[8] Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in Neural Information Processing Systems 35 (2022): 27730-27744.

[9] Stiennon, Nisan, et al. "Learning to summarize with human feedback." Advances in Neural Information Processing Systems 33 (2020): 3008-3021.

[10] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

[11] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).

[12] Workshop, BigScience, Scao, Teven Le, et al. "Bloom: A 176b-parameter open-access multilingual language model." arXiv preprint arXiv:2211.05100 (2022).

[13] Chowdhery, Aakanksha, et al. "Palm: Scaling language modeling with pathways." Journal of Machine Learning Research 24.240 (2023): 1-113.

[14] Jiang, Albert Q., et al. "Mistral 7B." arXiv preprint arXiv:2310.06825 (2023).
[15] Zhu, Banghua, et al. "Starling-7b: Improving llm helpfulness & harmlessness with RLAIF." (2023).
[16] Tunstall, Lewis, et al. "Zephyr: Direct distillation of lm alignment." arXiv preprint arXiv:2310.16944 (2023).
[17] Carta, Salvatore, et al. "Iterative zero-shot llm prompting for knowledge graph construction." arXiv preprint arXiv:2307.01128 (2023).
[18] Zhu, Yuqi, et al. "LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities." arXiv preprint arXiv:2305.13168 (2023).
[19] Li, Bo, et al. "Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness." arXiv preprint arXiv:2304.11633 (2023).
[20] Wei, Xiang, et al. "Zero-shot information extraction via chatting with chatgpt." arXiv preprint arXiv:2302.10205 (2023).
[21] Jarnac, Lucas, Miguel Couceiro, and Pierre Monnin. "Relevant Entity Selection: Knowledge Graph Bootstrapping via Zero-Shot Analogical Pruning." Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 2023.
[22] Bi, Zhen, et al. "Codekgc: Code language model for generative knowledge graph construction." arXiv preprint arXiv:2304.09048 (2023).
[23] Yao, Liang, et al. "Exploring large language models for knowledge graph completion." arXiv preprint arXiv:2308.13916 (2023).
[24] Khorashadizadeh, Hanieh, et al. "Exploring In-Context Learning Capabilities of Foundation Models for Generating Knowledge Graphs from Text." arXiv preprint arXiv:2305.08804 (2023).
[25] Deng, Shumin, et al. "Construction and applications of billion-scale pre-trained multimodal business knowledge graph." 2023 IEEE 39th International Conference on Data Engineering (ICDE). IEEE, 2023.
[26] Trajanoska, Milena, Riste Stojanov, and Dimitar Trajanov. "Enhancing Knowledge Graph Construction Using Large Language Models." arXiv preprint arXiv:2305.04676 (2023).
[27] Chen, Jiao, et al. "Knowledge Graph Completion Models are Few-shot Learners: An Empirical Study of Relation Labeling in E-commerce with LLMs." arXiv preprint arXiv:2305.09858 (2023).
[28] Harnoune, Ayoub, et al. "BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis." Computer Methods and Programs in Biomedicine Update 1 (2021): 100042.
[29] Yang, Linyao, et al. "ChatGPT is not Enough: Enhancing Large Language Models with Knowledge Graphs for Fact-aware Language Modeling." arXiv preprint arXiv:2306.11489 (2023).
[30] Ferreira, Thiago Castro, et al. "Neural data-to-text generation: A comparison between pipeline and end-to-end architectures." arXiv preprint arXiv:1908.09022 (2019).
[31] Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." Advances in Neural Information Processing Systems 36 (2024).
[32] Zhang, Xishan, et al. "Adaptive precision training: Quantify back propagation in neural networks with fixed-point numbers." arXiv preprint arXiv:1911.00361 (2019).
[33] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).
[34] Belkada, Y., et al. "Making llms even more accessible with bitsandbytes, 4-bit quantization and qlora." (2023).
[35] Mangrulkar, Sourab, et al. "PEFT: State-of-the-art Parameter-Efficient Fine-Tuning meth-

ods." 2022. Hugging Face, https://github.com/huggingface/peft.
[36] Saha, Swarnadeep, et al. "ExplaGraphs: An explanation graph generation task for structured commonsense reasoning." arXiv preprint arXiv:2104.07644 (2021).
[37] Abu-Aisheh, Zeina, et al. "An exact graph edit distance algorithm for solving pattern recognition problems." 4th International Conference on Pattern Recognition Applications and Methods 2015. 2015.
[38] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.
[39] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. 2004.
[40] Gardent, Claire, et al. "The WebNLG challenge: Generating text from RDF data." Proceedings of the 10th International Conference on Natural Language Generation. 2017.
[41] Hagberg, Aric, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using NetworkX. No. LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
[42] OpenAI. GPT-4 technical report. CoRR, abs/2303.08774 (2023).
[43] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).

FRCCS

Interpretable Control of Modular Soft Robots

Giorgia Nadizar[1✓] and Eric Medvet[1]

[1] *University of Trieste; giorgia.nadizar@phd.units.it, emedvet@units.it.*
[✓] *Presenting author*

Abstract. Modular Soft Robots (MSRs) are ensembles of elastic modules which achieve movement through their synergy of contractions and expansions. The inherent complexity of the dynamics of MSRs poses challenges in hand-crafting effective controllers for task execution. While Artificial Neural Networks (ANNs) have often been employed to this end, they lack transparency, preventing understanding the agent's inner functionality and problem-solving strategy. To tackle this issue, we propose to adopt interpretable controllers, in the form of graphs, to be optimized with Graph-based Genetic Programming (GGP). This methodology enables the optimization of effective controllers, which can also facilitate gaining insights into information flow and decision-making processes within MSRs. From preliminary experiments, we find our approach feasible and promising.

Keywords. *Interpretability; Modular Soft Robots; Collective Intelligence.*

1 Introduction

Modular Soft Robots (MSRs) have garnered significant attention in recent years due to their remarkable versatility and potential applications. Their softness confers them compliance and adaptability, allowing them to navigate challenging environments and undertake tasks that conventional rigid robots struggle with, e.g., exploration in confined spaces [2]. However, unlike their rigid counterparts, the complex dynamics of MSRs makes them more challenging to study and optimize.

Traditional approaches to MSRs control include open-loop controllers [5], which lack adaptability to dynamic environments, and various types of Artificial Neural Networks (ANNs), which can achieve more sophisticated behaviors [13]. Recent advancements have also explored distributed control strategies leveraging the collective intelligence of multiple ANNs within the robot [14], even removing the communication requirements among modules [15]. However, while ANN-based control offers greater flexibility, it lacks transparency. Namely, the combination of the complex body dynamics of MSRs with an opaque controller makes it nearly impossible to understand and *interpret* the processes occurring within the agent.

Therefore, achieving *interpretability* in the control of MSRs presents a significant endeavor. The notion of interpretability per se is often highly subjective and elusive [7, 12], especially when it involves control systems [16, 11], due to the intrinsic dynamics of the agent-environment system. Nevertheless, employing symbolic expressions or graph representations in controllers, can undoubtedly facilitate the understanding of causal relationships among variables and elucidate

information flow within the system [15, 8], ultimately enabling insights into the dynamics and the decision-making processes of MSRs.

Thus, we propose a novel approach for achieving interpretable control of MSRs, based on graphs. Building upon insights from our previous work [11], we aim to leverage Graph-based Genetic Programming (GGP) for optimizing controllers in the form of graphs. Graph-based controllers promise to be not only effective at guiding the MSR towards the achievement of a task, but also directly understandable with simple analyses. Ultimately, with this proposition we seek to elucidate the dynamics governing MSRs when accomplishing various tasks.

2 Background

Modular Soft Robots. MSRs represent a paradigm of robots comprising an agglomerate of modules interconnected to form a flexible structure [5]. These robots achieve motion through the coordinated actuation of their constituent modules. To facilitate research and development, numerous MSR simulators have been developed [6, 9, 1].

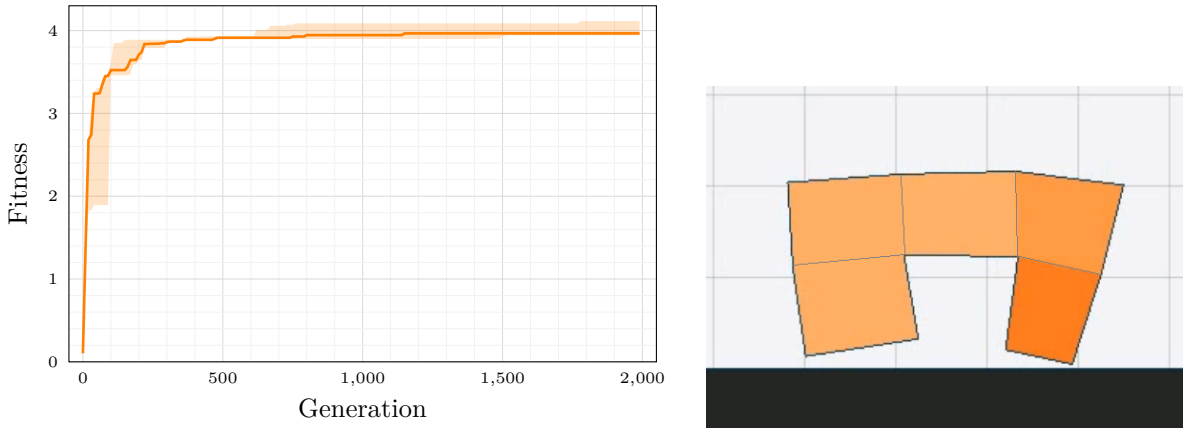
We consider EvoGym [1], a 2D discrete-time simulator, which models MSRs as composed by four main types of modules, each modeling distinct material properties: rigid inactive, soft inactive, horizontal actuator, and vertical actuator. The arrangement of these modules within the robot's body, represented as a spatial 2D grid, defines its *morphology*. In addition to their physical structure, MSRs are equipped with sensors that vary according to the task at hand and are necessary to provide information about the robot itself and its surrounding environment. The control of MSR behavior is dictated by a *controller*, which relies on sensor information to compute values for all actuators within the robot's body.

MSRs can be designed and optimized to perform a wide range of *tasks*, from simple locomotion to object manipulation. These tasks are defined by specific objectives, with rewards provided to promote behaviors that lead to their successful completion.

Graph-based Genetic Programming. GGP represents an evolutionary optimization technique for functional/program graphs inspired by natural evolution [4]. In evolutionary optimization, a population of candidate solutions undergoes iterative refinement through processes of selection, crossover, and mutation to produce increasingly fit individuals [3]. Each candidate solution is represented by a *genotype*, which encodes its genetic information, and a *phenotype*, which represents its observable characteristics. The *fitness* of each individual is evaluated based on its performance in solving a specific problem, with the goal of finding the most suitable solution.

One popular variant of GGP is Cartesian Genetic Programming (CGP), which encodes graphs as grids of nodes in a Cartesian plane [10]. Every node represents a function, which can use either the outputs of the nodes of the previous layers or the n_{in} inputs as arguments. The final outputs are collected from the outputs of n_{out} selected nodes.

The genotype of a CGP graph is a sequence of integers indicating, for each node, the inputs positions and the function to be applied to them. Moreover, the genotype also indicates the nodes to be used as output nodes. Thanks to the indirect encoding, optimization can be performed with standard Evolutionary Algorithms (EAs) suitable for evolving integer strings.



(a) Fitness progression. (b) Example of MSR.

Figure 1: Fitness during the optimization (left) and the simulated MSR (right). In 1b, the color of each of the five modules represents the actuation (*i.e.*, contraction): the darker, the more contracted.

3 Preliminary evaluation

To test the feasibility of our proposition, we conducted a preliminary experimental analysis. We considered the MSR depicted in Figure 1b, and optimized a graph controller with CGP for the task of locomotion on a flat terrain. As inputs to the graph we employed the sensors of the robot, while we fed the graph output to the 5 actuation modules (the chosen MSR only has horizontal actuation modules). We conducted optimization with a standard Genetic Algorithm (GA) with a population of 100 individuals, for 2 000 generations. As fitness we considered the total displacement along the x-axis, to promote the locomotion behavior.

We report our preliminary results in Figure 1a, where we show the progression of fitness over generations, in terms of median and inter-quartile range (across 5 independent runs). From the plot we can confirm the effectiveness of our method, as the fitness increases over evolutionary time, *i.e.*, optimization is occurring. However, from a brief comparison with the ANN results from the literature our results fall behind. This descends from two main points: (1) graphs are naturally less expressive than ANNs due to their fewer parameters, and (2) the discrete representation of CGP is more prone to converge to local optima because of the complex fitness landscape induced by the MSRs dynamics—a similar observation has been made also in [8] for a different discrete representation.

Final remarks. These results constitute a proof-of-concept, yet call for more in-depth research on the topic. Moreover, we have not yet analyzed systematically the computation graph within the MSR, which shall be the core part of our future studies on the subject to investigate the processes underlying the behavior of MSRs.

References

[1] Jagdeep Bhatia et al. “Evolution gym: A large-scale benchmark for evolving soft robots”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 2201–2214.

[2] Nick Cheney, Josh Bongard, and Hod Lipson. "Evolving soft robots in tight spaces". In: *Proceedings of the 2015 annual conference on Genetic and Evolutionary Computation*. 2015, pp. 935–942.

[3] Agoston E Eiben and James E Smith. *Introduction to evolutionary computing*. Springer, 2015.

[4] Léo Françoso Dal Piccol Sotto et al. "Graph representations in genetic programming". In: *Genetic Programming and Evolvable Machines* 22.4 (2021), pp. 607–636.

[5] Jonathan Hiller and Hod Lipson. "Automatic design and manufacture of soft robots". In: *IEEE Transactions on Robotics* 28.2 (2011), pp. 457–466.

[6] Sam Kriegman et al. "Simulating the evolution of soft and rigid-body robots". In: *Proceedings of the genetic and evolutionary computation conference companion*. 2017, pp. 1117–1120.

[7] Zachary C Lipton. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57.

[8] Eric Medvet and Giorgia Nadizar. "GP for Continuous Control: Teacher or Learner? The Case of Simulated Modular Soft Robots". In: *Genetic Programming Theory and Practice XX*. Springer, 2024, pp. 203–224.

[9] Eric Medvet et al. "2D-VSR-Sim: A simulation tool for the optimization of 2-D voxel-based soft robots". In: *SoftwareX* 12 (2020), p. 100573.

[10] Julian F. Miller and Peter Thomson. "Cartesian Genetic Programming". In: *Genetic Programming*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 121–132. ISBN: 978-3-540-46239-2.

[11] Giorgia Nadizar, Eric Medvet, and Dennis G Wilson. "Naturally Interpretable Control Policies via Graph-based Genetic Programming". In: *European Conference on Genetic Programming (Part of EvoStar)*. Springer. 2024.

[12] Giorgia Nadizar et al. "An analysis of the ingredients for learning interpretable symbolic regression models with human-in-the-loop and genetic programming". In: *ACM Transactions on Evolutionary Learning* (2024).

[13] Giorgia Nadizar et al. "An experimental comparison of evolved neural network models for controlling simulated modular soft robots". In: *Applied Soft Computing* (2023), p. 110610.

[14] Giorgia Nadizar et al. "Collective control of modular soft robots via embodied Spiking Neural Cellular Automata". In: *arXiv preprint arXiv:2204.02099* (2022).

[15] Federico Pigozzi et al. "Evolving modular soft robots without explicit inter-module communication using local self-attention". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. 2022, pp. 148–157.

[16] Mathurin Videau et al. "Multi-objective genetic programming for explainable reinforcement learning". In: *European Conference on Genetic Programming (Part of EvoStar)*. Springer. 2022, pp. 278–293.

FRCCS

Nash Equilibrium Analysis of Attack and Defense Strategies in the Air Transportation Network

Renaud Horacio Gaffan[1], Issa Moussa Diop[2✓], Ndeye Khady Aidara[1], Cherif Diallo[1] and Hocine Cherifi[3]

[1] *LACCA, Gaston Berger University, Saint-Louis, Senegal ; renaudgaffan@gmail.com,aidara.ndeye-khady@ugb.edu.sn, cherif.diallo@ugb.edu.sn*
[2] *I3S, Cote d'Azur University, Nice, France ; issa-moussa.diop@univ-cotedazur.fr*
3 *ICB Lab, UMR 6303 CNRS, University of Burgundy, Dijon, France; hocine.cherifi@u-bourgogne.fr*
✓ *Presenting author*

Abstract. This study explores the strategic dynamics within interconnected systems by integrating game theory with complex networks. It presents a static zero-sum game model to analyze attack and defense strategies in such networks. Investigating three strategies for attackers and defenders—random, degree centrality, and betweenness centrality—the study examines Nash equilibrium under equal resource assumptions. Analyzing the payoff matrix and players' responses identifies the dominant strategy as combining random attacks and betweenness-based defenses.

Keywords. *Game theory; Complex networks; Attacker-defender game; Attack and defense strategies; Nash equilibrium*

1 Introduction

The emergence of attack-defense games provides a strategic perspective for evaluating complex network security. Previous studies have explored these games extensively, focusing on achieving Nash Equilibrium, where players employ optimal strategies. For example, one study introduced a zero-sum game model to understand network robustness during attacker-defender confrontations[8]. Other studies proposed a game model considering network topology and system performance, analyzing interactions with limited budgets and targeted strategies [6][7]. This study extends this research by introducing novel defense strategies based on alternative centrality measures [11, 12, 13]. It shows that prioritizing nodes with high betweenness centrality offers an effective defense strategy against random attacks.

The game model assumes the presence of an attacker and a defender. It is a one-shot game, meaning players make their decisions simultaneously without knowing each other's choices. The model is based on two assumptions. The first is decision-maker rationality, which implies that players decide based on their interests and seek to maximize their payoffs. The second is their knowledge of each other's strategies, which means that players have perfect knowledge of the network and other players' strategies.

For the **attacker**, V^A is the set of attacked nodes, with $V^A \subseteq V$. θ_A is the attack range

parameter, with $\theta_A = \frac{|V^A|}{N}$. X is an attack strategy, with $X \in S^A = [x_1, x_2, ..., x_N]$, where S^A is the set of attack strategies. x_i is a binary variable for each node in the network. $x_i = 1$ if the corresponding node v_i is selected as the target of an attack ($v_i \in V^A$) and $x_i = 0$ otherwise. We obtain $\theta_A = \frac{1}{N}\sum_{i=1}^{N} x_i$. N is the number of nodes of the graph.

For the **defender**, one replaces A by D and X par Y. For example, Y is an defense strategy, with $Y \in S^D = [y_1, y_2, ..., y_N]$.

The payoff is defined as the reduction in network performance caused by the attack. In the attacker-defender game, it is important to note that both players move simultaneously without knowing each other's decisions. A node v_i is removed when attacked and not defended ($x_i = 1$ and $y_i = 0$). The removed nodes are denoted $\hat{V}$, where $\hat{V} \subseteq V$. The network after removing vulnerable nodes is denoted $\hat{G}$, with $\hat{G} = (V - \hat{V}, \hat{E})$, and we have $\hat{V} = V^A - V^A \cap V^D$. The performance measure function $\Gamma(G)$ is used in the study to evaluate network performance under different attack and defense strategies. It is defined as the size of the largest connected component of the network G after removing attacked but undefended nodes. The attacker's payoff is defined as follows:

$$U^A(X, Y) = \frac{\Gamma(G) - \Gamma(\hat{G})}{\Gamma(G)} \in [0, 1] \tag{1}$$

$\Gamma(G)$ is the network's performance before the attack, and $\Gamma(\hat{G})$ is the network's performance after the attack. The sum of the attacker's and defender's payoffs is always equal to zero, as the game is a zero-sum game. $U^D(X, Y)$ represents the defender's payoff.

$$U^A(X, Y) + U^D(X, Y) = 0 \tag{2}$$

2 Experimental Results

We examine the air transport network described in [2]. We assume that both players have an equal capacity to attack and defend nodes ($\theta_A = \theta_D$). Therefore, if both players choose the same targeted strategy, all attacked nodes are defended, resulting in zero payoffs. The strategies for attack and defense can either be random or targeted. Targeted strategies are centrality-based, with nodes prioritized in descending centrality order.

Figure 1 (left) depicts the attacker's payoffs as a function of the attack (defense) scope parameter, θ, when different profiles of targeted strategies compete. Neither player has a dominant strategy, so only one mixed-strategy Nash equilibrium exists. When $\theta \leq 0.5$, the gain of the Betweenness attack against the Degree defense is greater than that of the Degree attack against the Betweenness defense. Thus, attack and defense strategies based on the Betweenness are better than those based on Degree. On the other hand, when $\theta > 0.5$, the gain of the Betweenness attack against the Degree defense is equal to that of the attack-based Degree against the Betweenness defense. This means that attack and defense strategies based on the Betweenness give equivalent gain to Degree attack. In other words, when the proportion of airports to be attacked is high, the hubs are similar in the two strategies. However, with a limited budget for nodes and given the centrality anomaly [4] in the global air transport network, the hubs resulting from the two strategies differ. In addition, nodes with a very high Betweenness tend to connect distinct groups of nodes. Thus, neglecting to protect these critical nodes could potentially compromise the integrity of the network's giant component.

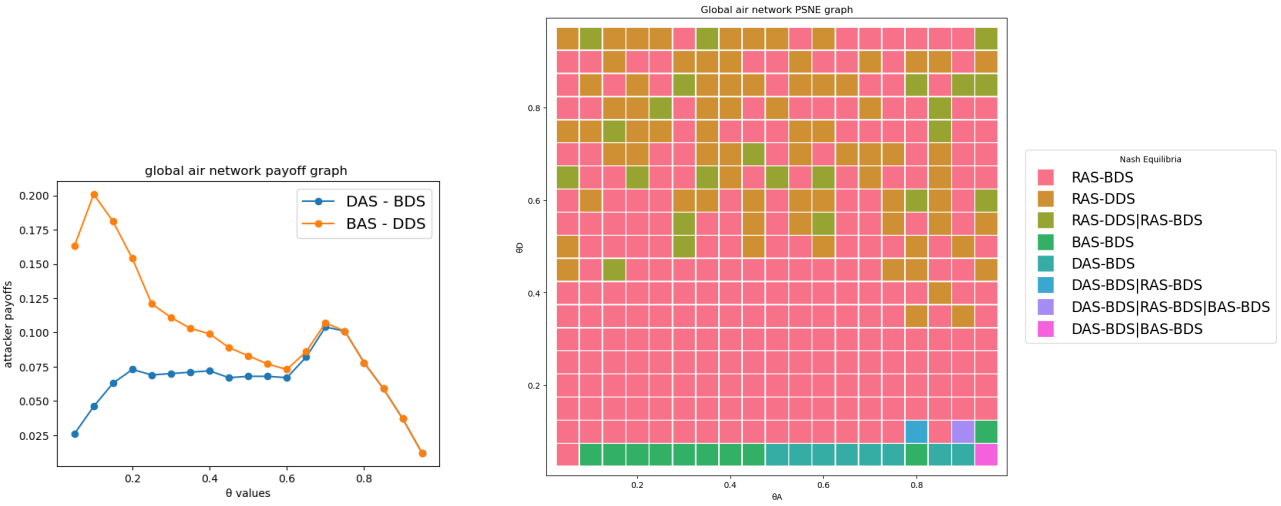


Figure 1: Left): The payoffs of the attacker as a function of the attack (defense) range parameter θ when Degree and Betweenness strategies clash in the global air network. On the x-axis, we represent the values of the attack (defense) range parameter θ, and on the y-axis the attacker's payoffs. DAS means Degree Attack Strategy, and BDS means Betweenness Defense Strategy. Right): Graph of pure strategy Nash equilibrium in the global airport network. On the x-axis, we represent attack resources θ_A, and on the y-axis, defense resources θ_D. Each tile represents an equilibrium strategy. The legend on the right shows the different game balances and the corresponding colors. RAS and RDS mean, respectively, Random Attack Strategy and Random Defense Strategy.

When analyzing conflicts between the random and targeted strategies (Degree or Betweenness). The attacker gains as a function of the attack (defense) range parameter θ when different targeted strategy profiles confront the random strategy in the global air network shows that the Degree or Betweenness attack against the random defense strategy gives a better gain than the random attack strategy against the random defense strategy which itself provides a better gain than the random attack strategy against the Degree or the Betweenness defense. So there is a Nash equilibrium in pure strategy (random attack, Degree defense) or (random attack,Betweenness defense).

In the following, we consider the game's three typical strategies. Figures 1 (right), show the pure strategy Nash equilibrium of the game in the different networks when attack resource θ_A and defense resource θ_D are different. In real-world attack scenarios, attack and defense resources are unlikely to be equal. The primary Nash equilibrium occurs when the attacker employs a random attack strategy while the defender safeguards nodes with high betweenness. Following this equilibrium, the subsequent dominant Nash equilibrium emerges when the attacker selects airports arbitrarily, and the defender prioritizes the protection of airports with high degrees. In more detail, whether $\theta_A \geq 0.1$ and $\theta_D = 0.05$, the attacker chooses the attack-based Betweenness, and the defender also chooses the defense Strategy Based on Betweenness. On the other hand, when $\theta_A \geq 0.5$, and $\theta_D = 0.05$, the attacker chooses the Degree-based attack, and the defender maintains the nodes with the high Betweenness. There are also situations where we can have several Nash equilibria in pure strategy. Indeed, when $\theta_A > \theta_D$, the defender chooses the airports with high Betweenness and the attacker chooses the nodes randomly. When $\theta_A < \theta_D$, the defender chooses the defense-based Degree or the defense-based Betweenness, and the attacker chooses the random attack strategy.

3 Conclusion

This work analyzes attack and defense strategies in complex networks where attackers and defenders have equal resources. It shows that targeted attacks focusing on betweenness centrality are more effective than those with degree centrality, primarily when targeting only a few nodes. The analysis identifies critical strategies for both attackers and defenders, highlighting how differences in their available resources affect their equilibrium strategies. Future research will evaluate additional strategies based on different centrality measures [1, 5, 10]and networks[9, 3, 9]. Furthermore, we plan examining Nash equilibrium in mixed strategies.

References

[1] Debayan Chakraborty, Anurag Singh, and Hocine Cherifi. Immunization strategies based on the overlapping nodes in networks with community structure. In *Computational Social Networks: 5th International Conference, CSoNet 2016, Ho Chi Minh City, Vietnam, August 2-4, 2016, Proceedings 5*, pages 62–73. Springer International Publishing, 2016.

[2] Issa Moussa Diop, Chantal Cherifi, Cherif Diallo, and Hocine Cherifi. Revealing the component structure of the world air transportation network. *Applied Network Science*, 6:1–50, 2021.

[3] Zakariya Ghalmane, Chantal Cherifi, Hocine Cherifi, and Mohammed El Hassouni. Extracting backbones in weighted modular complex networks. *Scientific Reports*, 10(1):15539, 2020.

[4] Roger Guimera, Stefano Mossa, Adrian Turtschi, and LA Nunes Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *PNAS*, 102(22):7794–7799, 2005.

[5] Manish Kumar, Anurag Singh, and Hocine Cherifi. An efficient immunization strategy using overlapping nodes and its neighborhoods. In *Companion Proceedings of the The Web Conference 2018*, pages 1269–1275, 2018.

[6] Ya-Peng Li, Suo-Yi Tan, Ye Deng, and Jun Wu. Attacker-defender game from a network science perspective. *Chaos*, 28(5), 2018.

[7] Yapeng Li, Ye Deng, Yu Xiao, and Jun Wu. Attack and defense strategies in complex networks based on game theory. *Journal of Systems Science and Complexity*, 32(6):1630–1640, 2019.

[8] Yapeng Li and Jun Wu. Modeling confrontations in complex networks based on game theory. In *2018 Int. Conf. on Computer Science, Electronics and Communication Engineering (CSECE 2018)*, pages 109–112. Atlantis Press, 2018.

[9] Khubaib Ahmed Qureshi, Rauf Ahmed Shams Malick, Muhammad Sabih, and Hocine Cherifi. Deception detection on social media: A source-based perspective. *Knowledge-Based Systems*, 256:109649, 2022.

[10] Stephany Rajeh and Hocine Cherifi. Ranking influential nodes in complex networks with community structure. *Plos one*, 17(8):e0273610, 2022.

[11] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Interplay between hierarchy and centrality in complex networks. *IEEE Access*, 8:129717–129742, 2020.

[12] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Characterizing the interactions between classical and community-aware centrality measures in complex networks. *Scientific reports*, 11(1):10088, 2021.

[13] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Comparative evaluation of community-aware centrality measures. *Quality & Quantity*, 57(2), 2023.

FRCCS

Predicting the impact of communication outages in swarm collective perception

Dari Trendafilov[1,2✓], Ahmed Almansoori[1], Nicolas Bredeche[2], Timoteo Carletti[1] and Elio Tuci[1]

[1] *naXys, University of Namur, Namur, Belgium ; { dari-borisov.trendafilov, ahmed.almansoori, timoteo.carletti, elio.tuci } @unamur.be*
[2] *ISIR, Sorbonne University, Paris, France ; nicolas.bredeche@sorbonne-universite.fr*
[✓] *Presenting author*

Abstract. We present an application of a recently introduced information-theoretic complexity measure for predicting the impact which obstacles to swarm communication have on swarm performance in the collective perceptual discrimination task. Our formalism is built on the notion of Empowerment – a task-independent, universal and generic utility function, which characterizes the level of perceivable control an embodied agent has over its environment. We conducted series of simulations with an empowerment model of the collective perception scenario, including simple communication obstacles of the same size and shape placed in varying positions and/or orientations in one particular environmental pattern used previously for assessing collective decision-making. The results indicate the potential detrimental impact communication disruptions in particular locations of the arena could have on swarm performance, while suggesting no effect when the same obstacles are placed elsewhere. Such analysis could provide a characterization of critical spots in the arena for a given environmental pattern.

Keywords. *Information theory; Complexity measures; Swarm robotics; Collective perception; Empowerment*

1 Introduction

Swarm robotics studies multi-robot systems in which each robot has its own controller, perception is local and communication is based on spatial proximity [11]. The group-level response emerges from a self-organisation process [5], based on the interaction between the robots and their physical environment. However, the autonomous nature of this process poses a challenge for designers, since it is notoriously difficult to infer which set of individual actions leads to the emergence of a desired collective response. Moreover, traditional design methods lack the ability to tackle problems and swarms of increasing complexity in uncertain and unpredictable environments. Real-world contexts may include unexpected obstacles of different size, shape, position and orientation that affect the unrestricted movement and/or the effective communication of the swarm. This further intensifies the need for fundamental and generic automated methodologies for modulating collective behaviour, with the potential to circumvent tedious trial-and-error model tuning. Generic theoretic measures of behavioural diversity could facilitate the assessment of the interactions and trade-offs between individual robots, swarm and

environment, and could help predict swarm performance without resorting to costly empirical studies. In this regard, information-theoretic utilities have been proposed as potential generic measures, since they can abstract from implementation details and focus on the interactions and dynamics related to information processing only [23].

In this paper, we apply a recently introduced information-theoretic measure for the characterization of task difficulty in the collective perception paradigm. In this task, a swarm of robots aims to find a consensus on the most salient perceptual cue among those available in the environment, following a particular decision-making mechanism. We explore the potential of the empowerment measure [14] to capture and predict the effect of different communication obstacles in one particular environmental pattern in collective perception. The aim of the study is to demonstrate the ability of this measure to assess the impact of obstacles with a purely theoretical treatment. To our knowledge, this is the first study to consider the effect of communication occlusions on the behavioural dynamics and swarm performance in this task.

2 Background

For designing large groups of robots, which coordinate and cooperatively perform a task, swarm robotics takes inspiration from natural self-organizing systems and attempts to recreate the emergence of collective behaviour from simple local interaction rules [see 15, 36]. Through the design of individual robot behaviour, swarm robotics aims to achieve locally coordinated interaction that results in a self-organized collective behaviour [10, 12]. Information theory has previously been applied to formalise guided self-organization [22, 21] in which complex global patterns emerge from relatively simple local interactions [see 20, 9]. Shannon entropy-based measures, used to characterise self-organized emergent robot behaviour, range from mutual information [25, 27] and transfer entropy [26], to predictive and integrated information [7, 2]. Information-theoretic methods allow for a quantitative study of robot-environment systems [29], and are fundamental in embodied systems research [19]. Generic information-theoretic complexity measures have been used to study system dynamics [16, 4], to characterise information flows in the sensorimotor loop [17], and to analyse robot behaviour [23], due to their ability to capture salient features of robot behaviour based on generic information processing principles, while abstracting from system-specific details [see 23]. The information-theoretic concept of empowerment [14] has been applied to problems in various domains, such as, dynamical control systems [13], robotics [24], and human-computer interaction [30, 31], and more recently in swarm robotics for providing a complexity (i.e., task difficulty) measure in the perceptual discrimination task [33]. The potential of the empowerment measure, demonstrated in initial investigations, provides motivation for its further exploration for facilitating the automatic design of robot swarms and the analysis of their behavioural dynamics.

The collective perceptual discrimination task for swarm of robots has been originally introduced by [18], who used a binary version of this task to design and evaluate individual mechanisms underpinning the collective decision-making process. In this task, the swarm explored a close arena patched with tiles, randomly painted in black and white, with the aim to collectively decide which colour is dominant. Various individual mechanisms for opinion selection have been developed since, from the classical hand-crafted solutions, based on the voter model, the majority rule, and their variants [see 34], to more recent ones, based on the synthesis of artificial neural networks [1]. The performance of decision-making strategies has been investigated for varying options quality by [35], while multi-options scenarios have been studied by [8]. Some studies explored the presence of byzantine robots, i.e., robots that communicate decep-

tive messages with the intent to entice the swarm to converge on a consensus to a non-optimal choice [28]. Research in this domain generally considers situations where the robot movement is disturbed only by collisions with close neighbours and arena walls, and communication is unobstructed and effective in a range of up to 50cm based on the e-puck2 platform [32]. Typically, the unpredictable disturbances are modelled uniformly as a random additive noise perturbing both actuation (i.e., movement) and perception (i.e., floor sensing and communication). However, such uncertainty accounts only for sensor and actuator imprecision, while ignoring the possibility of encountering obstructive foreign objects in the field. Earlier research focused on the environmental feature ratio for modulating task difficulty, whereas more recently, [3] proposed that the key determinant of the difficulty in this task is the features' distribution and introduced a set of variations in the environmental topology. Building on their work, [33] introduced an empowerment-based universal and generic measure of task difficulty, which takes into account not only the environmental complexity (i.e., the features distribution), but also the agent's capabilities – arguably a key factor influencing swarm performance. Further extending this work, we demonstrate the ability of the empowerment measure to quantify salient features (i.e., obstacles) in the environment, independent from the task or goal of the swarm, which makes this approach directly applicable to various scenarios in this domain. This initial study provides important insights regarding the effect of communication obstacles in collective perception and sheds light on the interaction between obstacle placement and the predicted impact on swarm performance.

3 Collective Perception

This study is based on the collective perceptual discrimination task as described in [3, 1], which takes place in a square arena whose floor is covered by black and white tiles and where the dominant colour (black or white) covers 55% of the arena floor, while the other colour covers the remaining 45%. The goal of the swarm of robots is to reach a consensus on the dominant colour by randomly exploring the arena and by communicating their opinions on what is the dominant colour to spatially proximal robots. The most frequently used features' distribution in this task is the random distribution of colour patches (see Figure 1/left), which, however, has its limitations with respect to generalization of swarm behaviour; that is, decision-making strategies designed for randomly distributed patches are not equally successful in environments where features are distributed in a different way (for one such example see Figure 1/right). To study these limitations, [3] proposed a set of nine structurally different patterns, which revealed that swarm performance tends to deteriorate when the perceptual evidence is spatially arranged in distinctive clusters, regardless of the nature of control mechanisms (hand-coded [3] or neural network-based [1]). Overall, the less clustered the distribution of perceptual evidence, the higher the swarm accuracy in the collective decision-making [see 3, 1].

Figure 1: The Random environmental pattern, typically investigated in swarm collective perception research (left). The Stripe pattern, which was explored in our study (right).

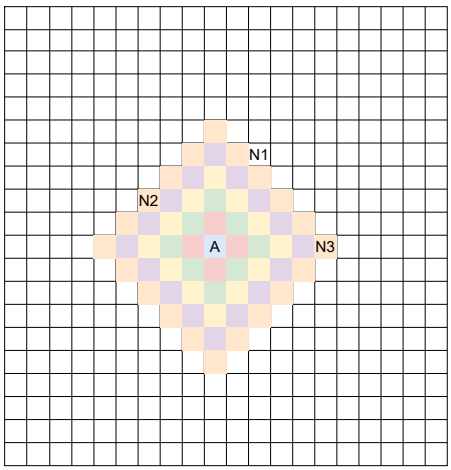


Figure 2: The experimental 2-D grid (20×20 cells) used in our study. The perceivable range of agent A is denoted with a colour map: range0 (blue), range1 (red), range2 (green), range3 (yellow), range4 (purple), and range5 (orange). E.g., neighbours $N2$ and $N3$ are in range5, and $N1$ is out of range.

Drawing from this work, [33] proposed a generic way of measuring task complexity with respect to the distribution of features. This approach places a robot with a particular morphology into a specific environmental condition and attempts to quantify the complexity of the environment as perceived by the agent, which essentially depends on the agent's perception–action loop. For facilitating the analysis, the following simplifications with respect to the original robot-based scenario, as illustrated in [3, 1], were made. A single agent is placed in a discretized square grid of size of 20×20 cells in which each cell corresponds to a tile, that can be either black or white. The agent can perceive the colour of the cell in which it is located and the colours of neighbouring cells. The number of perceivable cells can vary from 5 (range 1) to 61 (range 5). The neighbourhood ranges of an agent A are illustrated in Figure 2. The access to the colour of neighbouring cells intends to simulate the information generated by social influence. Within this metaphor, different ranges correspond to different levels of the maximal robot–robot communication distance, which maps directly this model to studies based on the e-puck2 robotic platform with a communication range of 50 cm and an arena of $2m \times 2m$, patched with tiles $10cm \times 10cm$ each [see, for example 1].

To compute empowerment for each neighbourhood size (i.e., range), the agent is located in every cell of the grid. Thus, empowerment provides a measure of perceivable features with respect to the current position and range. By computing this measure for all possible positions of the agent in the arena, a task complexity estimate integrating both the environmental structure and the agent's sensory capabilities is obtained. This measure of task difficulty has shown its ability to predict swarm performance in collective perception using a number of different decision-making mechanisms for opinion selection and various environmental patterns (see [33]). We extend this work in the current study, by focusing on one particular environmental pattern – Stripe (see Figure 1/right) – while introducing communication obstacles in the arena at different locations. All obstacles have the shape of a straight thin line, three tiles long, placed between tiles and act as impenetrable walls, further restricting the communication range of 50 cm. To explore the effect of such communication obstacles, we varied the location and/or orientation according to the six layouts presented in Figure 3 and computed the empowerment levels for each one using the above approach.

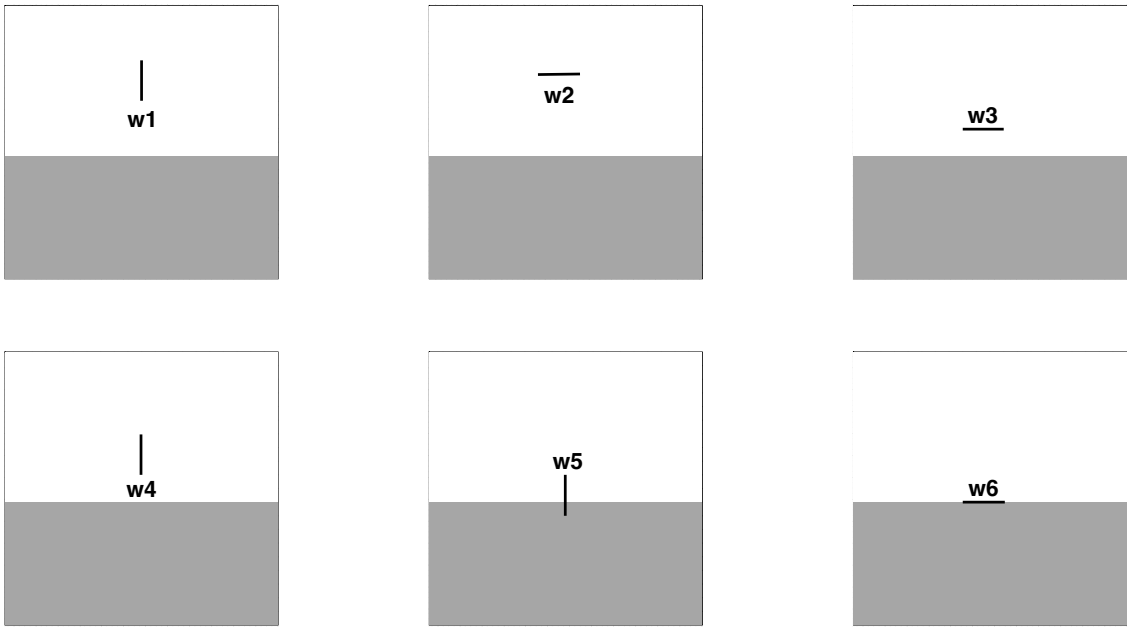


Figure 3: The six conditions investigated in this study, corresponding to six different placements (i.e., w1, w2, w3, w4, w5, and w6) of the communication obstacle (locations and/or orientations) in the discrete 2-D grid (20×20 cells) reflecting the Stripe environmental pattern.

4 Empowerment Model

The information-theoretic model of the Collective Perception paradigm introduced in [33] is based on the empowerment formulation [14] of the perception–action loop of an embodied agent and its environment, represented as a communication channel. Using the causal Bayesian network representation of the perception-action loop (see Figure 4), empowerment is defined as the Shannon channel capacity from the sequence of actions $U_t, U_{t+1}, ..., U_{t+n-1}$ to the perception Y_{t+n} through the environment $X_{t+1}, X_{t+2}, ..., X_{t+n}$ after an arbitrary number of (n) time steps,

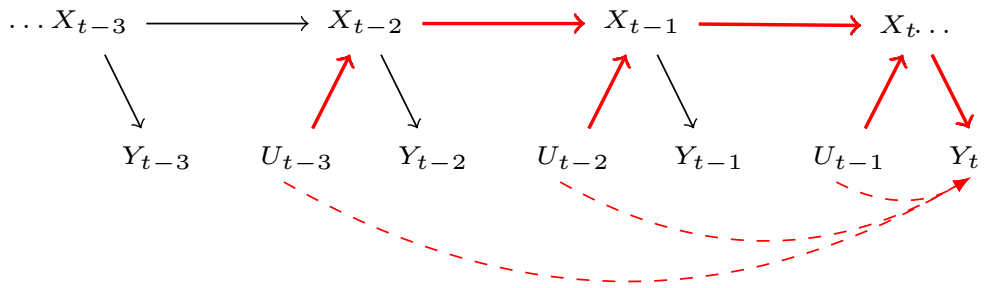


Figure 4: Perception–action loop as a causal Bayesian network – an agent performs an action U and injects information into the environment X, and subsequently reacquires part of this information via its sensors Y. Empowerment is the channel capacity from the action sequence (e.g., $U_{t-3}, U_{t-2}, U_{t-1}$) to the resulting observation (e.g., Y_t) after n (e.g., 3) time-steps.

using the following formulation

$$C(U_t, ..., U_{t+n-1} \to Y_{t+n}) = \sup_{p(\vec{u})} I(U_t, ..., U_{t+n-1}; Y_{t+n})$$

where $\vec{u} = (u_t, ..., u_{t+n-1})$ and the mutual information between two discrete random variables U and Y is defined by

$$I(U;Y) = \sum_u p(u) \sum_y p(y|u) \log \frac{p(y|u)}{p(y)}.$$

Empowerment is a task and representation independent utility function, fully specified by the dynamics of the perception–action loop of the agent–environment coupling unrolled over time. It reflects the capacity of an agent to control or influence its environment as perceived by its sensors. Empowerment depends on the agent's embodiment, i.e., its sensory apparatus and motor abilities, and on the degree of interaction between agents, i.e., agents need freedom to act and at the same time they need certain constraints imposed by other agents [6].

The decision-making mechanisms for collective perception are based on the agent's own perception and the opinions of its neighbours, which contain information about the environment at various remote positions and are transmitted from a distance within a specific communication range. This enables the agent to extend its sensing abilities and to acquire information about (perceive) the environment at distant locations. The collective perception scenario can be transformed into the empowerment formalism by re-framing the task into a communication problem, using swarm communication as an action space and representing the action horizon with the communication range. In this model, the state space consists of the position of a single agent in the grid. For simplicity, only the main four orthogonal directions are used from which neighbourhoods of a particular size are constructed with an action space $\mathcal{U}$ of the following five primitive actions

$$\mathcal{U} = \{north, south, east, west, idle\}.$$

The first four actions correspond to communicating with (i.e., polling the opinions of) the immediate neighbours in the four respective directions, while the last (*idle*) action reflects the agent's own sensor reading. N-step action sequences represent communication with agents in a neighbourhood of a particular range. The borders of the environment are hard and constrain the actions. Following this representation, Figure 2 depicts the perceivable range of agent A, in a blank 2-D grid, defined by a colour map – range0 (blue), range1 (red), range2 (green), range3 (yellow), range4 (purple), and range5 (orange).

We evaluate empowerment in all positions across the grid, using the environmental features as sensor readings. For any state $x \in \mathcal{X}$ in the grid empowerment is computed by

$$\mathfrak{E}(x) = \max_{p(\vec{u})} I(U_t, ..., U_{t+n-1}; Y_{t+n}|x),$$

where the action space $\mathcal{U}$ consists of the above five actions and the perception space $\mathcal{Y}$ is defined by a binary random variable

$$\mathcal{Y} = \{0, 1\},$$

representing the environmental feature (black or white) in state $y \in \mathcal{X}$, where y is the resulting state after applying the action sequence $U_t, ..., U_{t+n-1}$ starting from x. Note that x is a starting position on the 2-D grid, while the perception $Y_{t+n} \in \mathcal{Y}$ is a binary value representing the feature in the final position.

5 Results

Employing the above model, we computed the empowerment levels for every starting position in the 2-D grid for all six conditions presented in Figure 3, with a range of empowerment horizons from one to five, which corresponds to a discrete communication radius of one to five cells and is in line with previous swarm robotics studies in this scenario [see 1]. Furthermore, to facilitate the assessment of the experimental conditions, we computed the empowerment for the baseline case consisting of the Stripe environmental pattern excluding obstacles.

For brevity, in this baseline condition, the evolution of the empowerment levels as the communication horizon increases is presented only for the minimal (1-step) and the maximal (5-step) horizon (see Figure 5). The results reveal an empowerment increase to its maximal level (in this case 1 bit) around the borderline between the black and the white patches, and is zero elsewhere. The larger the horizon, the wider the area of high empowerment is, as expected.

We found no difference in the empowerment levels between the baseline and the experimental conditions with two exceptions, namely conditions w3 and w6, which are presented in full details in Figures 6 and 7. For condition w3 (see Figure 6), the empowerment levels follow the trends of the baseline condition for 1-step and 2-step horizons, however, with the gradual increase of the horizon a drop from 1 bit to 0 bit in empowerment appears in the vicinity of the communication obstacle. For condition w6 (see Figure 7), this drop in empowerment is symmetrical, as expected, since the communication obstacle is placed exactly between the black and the white patches, and furthermore, affects all horizons from 1-step to 5-step. The impact for horizon steps 1 and 2 leads to zero empowerment on the borderline. We present the overall empowerment levels averaged across the grid for the baseline, w3 and w6 conditions in Figure 8. It reveals that the closer the obstacle is to the borderline, the stronger the overall effect on empowerment is. The impact may appear negligible, however, one should keep in mind the rather modestly sized obstacle used in this study. These results indicate that both the position and the orientation of the communication obstacles are crucial for maintaining optimal empowerment levels across the arena and suggests that swarm performance may be affected to a different degree by such obstacles depending on their particular placement. This insight highlights the intricate relationship between communication obstacles and specific environmental patterns, and calls for more thorough future investigations.

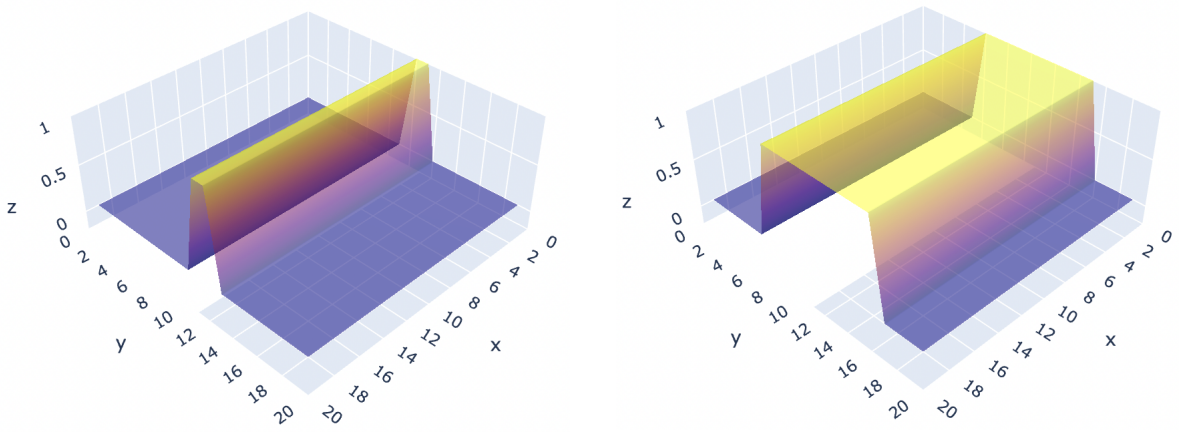


Figure 5: 1-step (left) and 5-step (right) empowerment levels computed across the 2-D grid in the Stripe condition without communication obstacles. Empowerment is at its maximum of 1 bit around the borderline between black and white patches, and zero elsewhere.

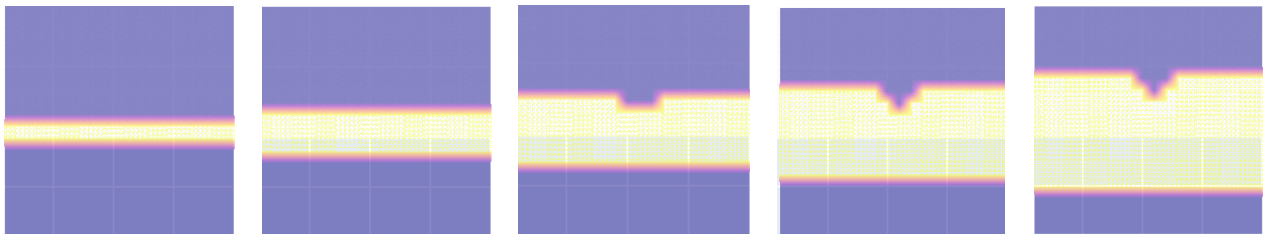


Figure 6: 1, 2, 3, 4, and 5-step (from left to right) empowerment levels, computed across the 2-D grid in condition $W3$ and projected on the X-Y plane. The close proximity of the obstacle to the borderline between black and white regions in this particular orientation leads to its detrimental effect on empowerment levels for 3, 4, and 5-step horizons.

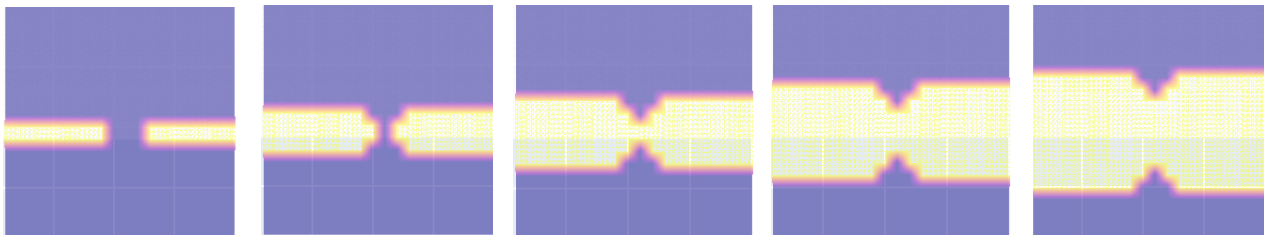


Figure 7: 1, 2, 3, 4, and 5-step (from left to right) empowerment levels, computed across the 2-D grid in condition $W6$ and projected on the X-Y plane. The placement of the obstacle exactly on the borderline between black and white regions explains the symmetry of the profiles and its detrimental effect on empowerment levels for all horizons.

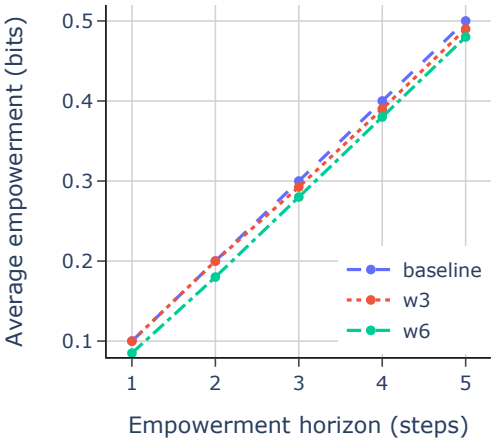


Figure 8: Average empowerment levels aggregated over the 2-D grid for five horizon spans in the baseline, w3 and w6 conditions. The drop in empowerment is minor for condition w3 with respect to the baseline and appears at larger horizons, whereas for w6 the drop is pronounced and affects all horizons. Note that the maximal empowerment level in this case is 1 bit.

6 Discussion

We have explored the potential impact communication obstacles might have on swarm performance with respect to environmental topology, building on a recently introduced [33] generic information-theoretic approach for characterizing task difficulty in the collective perception paradigm. This approach does not characterise the topological structure of the environment based on number, size, shape and inter-connectivity of clusters, but instead, it explores the environment with the given agent morphology, which is critically relevant in determining task difficulty. Our earlier study [33] revealed a significant correlation between the empowerment levels and the accuracy of state-of-the-art decision-making strategies, which suggests the potential of the empowerment measure to predict swarm performance based solely on properties of the environment and independent of the particular task. Building on this insight, we investigated the effect communication obstacles have on empowerment levels across various horizons in one particular environmental pattern. We applied the empowerment formalism to characterise the effect various locations and orientations of linear obstacles imply on task difficulty. Two key parameters influencing swarm performance in this scenario are the environmental pattern type and the swarm communication abilities. For this – first of its kind study – we have selected the simplest environment reported previously in this field, which is composed of two coherent colour patches. For obstacle, we have chosen a straight line placed in six different locations and/or orientations with respect to the borderline between the two colour patches, which appears to be a performance-critical spot. The rational behind the simplistic configurations used in our initial study was grounded in the search for unambiguous and clear interpretations of the interactions between communication range, environmental pattern and communication obstacles. Different and more complex environments with multiple obstacles of various sizes and shapes would have had a less predictable effect on empowerment which is more difficult to interpret, however, is an important direction for future research.

The results demonstrate that obstacles located sufficiently far – with respect to the empowerment horizon – from the borderline have no effect on empowerment, as expected, since the communication exchanges in homogeneous areas carry no new information for the swarm. Therefore, such communication obstacles have no detrimental impact. However, when placed closer to the borderline, the same obstacles inflict a drop in empowerment levels for sufficiently large horizons with respect to the distance between an obstacle and the borderline. Obstacles on the borderline have the highest impact on empowerment for all horizons, as expected, and result in a symmetric empowerment profile with respect to the borderline. An interesting finding of this study is the fact that only obstacles parallel to the borderline have influence on empowerment levels, which opens up new questions with regard to obstacle orientation for future research.

The key benefits of the applied information-theoretic treatment are that it is universal, general and could enable the analytical comparison of scenarios with different computational models. The proposed approach elucidates the trade-off between task difficulty (i.e., swarm performance) and the cost of enabling particular agent capacities, and provides information-theoretic bounds, which are fundamental properties of agent–environment systems. The empowerment levels could reveal critical points in the environment (e.g., obstacles), which might inflict significant drops in swarm performance, and thus raise designer's attention for a more careful consideration. Empowerment captures in a uniform measure salient features of the agent–environment perception–action loop, such as topology, morphology, noise in the sensing, actuation and communication channels, with a generic information-theoretic model. We believe that theories and tools from complex systems and information theory can successfully be applied for facilitating the automated design of robot collectives and for the analysis of their dynamics.

7 Conclusion

This paper presents an initial study on the effects communication obstacles might have on swarm performance in the collective perceptual discrimination task. The analysis, based on an application of the information-theoretic capacity of empowerment to the field of swarm robotics, highlights the benefits of utilising such a generic utility measure. Our approach is task-independent and the same model could be applied to further scenarios in this domain, e.g., shortest-path or site-selection. Leveraging Shannon's information theory by way of creating generative mathematical models and artificial simulations, empowerment offers a novel perspective for swarm robotics, building on objective quantitative measures and analytical tools, which could support the automated design of robotic swarms. Our study opens up new directions for research into how environmental factors and communication obstacles influence swarm behaviour and decision-making. Future work will focus on developing more robust, empirically validated models of swarm intelligence that have direct implications for the design and optimization of swarm-based technologies in various domains. To bridge the gap between theoretical insights provided here and their practical applications, related domain-specific methodologies will be explored for contextualizing and enhancing the findings of this study in more realistic settings and scenarios that reflect the complexity of real-world applications.

8 Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101034383 and the CERUNA doctoral fellowship by the University of Namur. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region. The authors would like to thank the anonymous reviewers for their constructive and insightful feedback.

References

[1] A. Almansoori, M. Alkilabi, and E. Tuci. Further investigations on the characteristics of neural network based opinion selection mechanisms for robotics swarms. *Proceedings of EvoStar – 26th European Conference on Applications of Evolutionary Computation*, pages 737–750, 2023.

[2] D. Balduzzi and G. Tononi. Integrated information in discrete dynamical systems: Motivation and theoretical framework. *PLOS Computational Biology*, 4(6):1–18, 06 2008.

[3] P. Bartashevich and S. Mostaghim. Benchmarking collective perception: New task difficulty metrics for collective decision-making. In *Progress in Artificial Intelligence: 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Proceedings, Part I*, page 699–711, 2019.

[4] R. D. Beer and P. L. Williams. Information processing and dynamics in minimally cognitive agents. *Cognitive Science*, 39(1):1–38, 2015.

[5] S. Camazine, J.L. Deneubourg, N.R. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau. *Self-Organization in Biological Systems*. Princeton University Press, United States, 2001.

[6] P. Capdepuy, C.L. Nehaniv, and D. Polani. Maximization of potential information flow as a universal utility for collective behaviour. In *Proceedings of the First IEEE Symposium on Artificial Life*, pages 207–213, 2007.

[7] R. Der, F. Güttler, and N. Ay. Predictive information and emergent cooperativity in a chain of mobile robots. *Artificial life XI, 166-172*, 01 2008.

[8] J.T. Ebert, M. Gauci, and R. Nagpal. Multi-feature collective decision making in robot swarms. In *17th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2018*, pages 1711–1719, 2018.

[9] N. Fernández, C. Maldonado, and C. Gershenson. Information measures of complexity, emergence, self-organization, homeostasis, and autopoiesis. In M. Prokopenko, editor, *Guided Self-Organization: Inception*, pages 19–51. Springer Berlin Heidelberg, 2014.

[10] E. C. Ferrer, T. Hardjono, A. Pentland, and M. Dorigo. Secure and secret cooperation in robot swarms. *Science Robotics*, 6, 2021.

[11] H. Hamann. *Swarm Robotics: A Formal Approach*. Springer, 2018.

[12] K. Hasselmann, A. Ligot, J. Ruddick, and M. Birattari. Empirical assessment and comparison of neuro-evolutionary methods for the automatic off-line design of robot swarms. *Nat Commun*, 12(4345):1–11, 2021.

[13] T. Jung, D. Polani, and P. Stone. Empowerment for continuous agent-environment systems. *Adaptive Behavior*, 19(1):16–39, 2011.

[14] A. Klyubin, D. Polani, and C. Nehaniv. Keep your options open: An information-based driving principle for sensorimotor systems. *PLOS ONE*, 3(12):1–14, 12 2008.

[15] C. Ronald Kube and H. Zhang. Collective robotics: From social insects to robots. *Adaptive Behavior*, 2(2):189–218, 1993.

[16] J. Lizier. *The Local Information Dynamics of Distributed Computation in Complex Systems*. 2013.

[17] M. Lungarella and O. Sporns. Mapping information flow in sensorimotor networks. *PLOS Computational Biology*, 2(10):1–12, 10 2006.

[18] G. Morlino, V. Trianni, and E. Tuci. Collective perception in a swarm of autonomous robots. *In: Proceedings of the International Joint Conference on Computational Intelligence*, 1:51–59, 2010.

[19] R. Pfeifer and C. Scheier. *Understanding Intelligence*. Cambridge, MA: The MIT Press, 2001.

[20] D. Polani. Foundations and formalizations of self-organization. In M. Prokopenko, editor, *Advances in Applied Self-organizing Systems*, pages 19–37. Springer London, 2008.

[21] D. Polani, M. Prokopenko, and L. S. Yaeger. Information and self-organization of behavior. *Advances in Complex Systems*, 16(2&3):1–12, 2013.

[22] M. Prokopenko. *Guided self-organization: Inception*. New York: Springer, 2014.

[23] A. Roli, A. Ligot, and M. Birattari. Complexity measures: Open questions and novel opportunities in the automatic design and analysis of robot swarms. *Frontiers in Robotics and AI*, 6, 2019.

[24] C. Salge, C. Glackin, D. Ristić-Durrant, M. Greaves, and D. Polani. Information-theoretic measures as a generic approach to human-robot interaction: Application in corbys project. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, page 282–283, 2014.

[25] C. Salge and D. Polani. Local information maximisation creates emergent flocking behaviour. *ECAL*, pages 688–696, 2011.

[26] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, (85):461–464, 2000.

[27] V. Sperati, V. Trianni, and S Nolfi. Mutual information as a task-independent utility function for evolutionary robotics. *Guided Self-Organization: Inception*, pages 389–414, 2014.

[28] V. Strobel, E. Ferrer, and M. Dorigo. Managing byzantine robots via blockchain technology in a swarm robotics collective decision making scenario. *In: Proceedings of the 17th*

International Conference on Autonomous Agents and Multi-Agent Systems, page 541–549, 2018.

[29] D. Tarapore, M. Lungarella, and G. Gómez. Quantifying patterns of agent–environment interaction. *Robotics and Autonomous Systems*, 54(2):150–158, 2006.

[30] D. Trendafilov and R. Murray-Smith. Information-theoretic characterization of uncertainty in manual control. In *Proc. IEEE SMC*, 2013.

[31] D. Trendafilov and D. Polani. Information-theoretic sensorimotor foundations of fitts' law. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–6, 2019.

[32] Dari Trendafilov, Ahmed Almansoori, Timoteo Carletti, and Elio Tuci. Generalizations of evolved decision-making mechanisms in swarm collective perception. volume WIVACE 2023: XVII International Workshop on Artificial Life and Evolutionary Computation, Venice, Italy, 09 2023.

[33] Dari Trendafilov, Ahmed Almansoori, Timoteo Carletti, and Elio Tuci. The role of the environment in collective perception: A generic complexity measure. volume ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference, page 13, 07 2023.

[34] G. Valentini. Achieving consensus in robot swarms: Design and analysis of strategies for the best-of-n problem. *SCI*, 706, 2017.

[35] G. Valentini, D. Brambilla, H. Hamann, and M. Dorigo. Collective perception of environmental features in a robot swarm. In *Swarm Intelligence*, pages 65–76, 2016.

[36] J. Werfel, K. Petersen, and R. Nagpal. Designing collective behavior in a termite-inspired robot construction team. *Science*, 343(6172):754–758, 2014.

FRCCS

Natural Language Processing for Requirements Model Extraction in Systems Engineering

Stella Zevio[1✓]

[1] *Cénotélie, Paris, France ; stella.zevio@cenotelie.fr*
[✓] *Presenting author*

Abstract. Requirements describe the needs and expectations of the customers. Consequently, they have a central role in the design of a complex system in systems engineering. They also constitute the raw material for a large number of applications among which requirements elicitation, analysis, modeling, verification and validation or management. Requirements are mostly described in specification documents, in natural language, which is inherently ambiguous. To encourage the reuse of requirements within further applications, we need to extract requirements from specification documents in a machine-readable and unambiguous format. For this purpose, we propose an automated tool that takes advantage of a combination of deep learning and natural language methods to transform documentation into a set of semantic triples. These triples are used to generate a knowledge graph-based representation of requirements. We designed our tool to integrate seamlessly into a model-based systems engineering framework and foster collaboration around complex systems engineering.

Keywords. *Knowledge extraction; Natural language processing; Systems engineering; Requirements engineering; Model-based systems engineering*

Designing a complex system is a very challenging task. It implies a perfect coordination between the stakeholders during the development phase (1). Indeed, these latter are in charge of building several heterogeneous systems and components that need to comply with strong requirements to meet the needs and expectations of the customers (1). As requirements express the costumers' needs, they constitute the central element on which the stakeholders rely to design the complex system. Requirements are often found in specification documents, written in natural language, which is inherently ambiguous (2). Therefore, it is essential to capture and

represent them in a unambiguous way to encourage collaboration between the stakeholders. These problematics are the issues of concern of requirements engineering (RE), which is the discipline that consists in defining, evaluating, validating and managing requirements.

Natural language processing (NLP) methods have been applied to RE since the early 1980s (3). Since then, natural language processing supported requirements engineering (NLP4RE) has emerged as a research field. It has recently known a rapid growth of interest due to the the technological advances in NLP (3). Indeed, NLP has drastically changed since the introduction of large pre-trained transformer-based language models such as GPT-3, GPT-4 or BERT (4). Most recent studies propose large language models (LLM) based methods for requirements engineering. For example, an annotated aerospace corpus has been used and the BERT language model has been fine-tuned to create a model for identifying named entities in aerospace requirements (5). However, a recent review of the field has stated that there is a huge gap between the state of the art and the state of the practice in the field (3). The authors suggest that this gap is mainly induced by the lack of industrial validation of NLP4RE research and the lack of adoption of proposed tools.

We propose a tool that integrates into our model-based system engineering (MBSE) framework deployed on an industrial scale. MBSE involves replacing documentation with a visual representation of the system. This representation constitutes a common frame of reference that promotes information sharing and considerably facilitates collaboration between the different stakeholders in the development of a complex system. However, MBSE requires standardized, machine-readable requirement and its adoption has been slowed down by the lack of adapted methods and tools to support its induced change of paradigm in industries. Our tool aims at supporting the adoption of MBSE by extracting requirements from specification documents in the form of a set of triples that are further used to automatically supply the visual representation of a complex system. The tool takes advantage of a combination of deep learning and NLP methods. We use an optical character recognition (OCR) algorithm to extract text from heterogeneous documents. Then, we apply a NLP workflow on the text. This workflow is composed of a requirements extraction component based on a sentence segmentation component, along with a named entity recognition model and a relation extraction component.

Requirements are extracted thanks to a rule-based sentence segmentation method and are weighted. For example, if the requirement contains the word « shall », it is a low-level requirement, while if it contains the word « must », it is a high-level requirement. To identify named entities in the text, we use SpaCy, which is a tool based on a deep learning method and supervised learning. The method takes advantage of a word embedding strategy that uses subword features and « Bloom »

embeddings, a deep convolutional neural network with residual connections, and a transition-based approach to named entity parsing. To generate a custom SpaCy NER model for requirements extraction in systems engineering, we annotated a large specification document in aeronautics. We identified named entities labels to annotate, among which ORG for organizations, ROLE for entities such as the *Supplier*, the *Provider* or the *equipment manufacturer*, COMPONENT for hardware parts, DOCUMENT used to describe documents such as technical specification or reports, or CRITERIA for criteria such as maintainability or safety. To identify relations existing between named entities, we use a rule-based method that takes advantage of semantic features and text patterns or regular expressions. We provide a generic workflow that helps extract any relation between source and target named entities labels sets, whether a triggering text is needed to identify the relation or not and whether the matching method is regular (the source named entities being found before the target named entities in the text) or specific. For a requirements engineering use case, we defined relations thanks to our relation extraction component, among which COLLABORATION, which exists between two ROLES that collaborate together, COMPLY_WITH, which exists between a named entity and a standard, DEFINED_BY, which exists between a standard or a process and a document, or COMPOSED_BY, which exists between two components, a system and its component or a hardware and its component.

For illustration purposes, we consider the following requirement : « The supplier must comply with the requirements defined in the specification documents. The purchaser shall provide the engine control system of the Airbus A380. ». With our method, we identify the following named entities in the text : *supplier* (ROLE), *requirements* (REQUIREMENT), *specification documents* (DOCUMENT), *engine control system* (COMPONENT), *Airbus* (ORG) and *A380* (HARDWARE). We also identify the following relations : COMPLY_WITH(supplier, requirements), DEFINED_BY(requirements, specification documents), PROVIDE(purchaser, engine control system) and COMPOSED_BY(A380, engine control system).

We obtained promising preliminary results with our method, supported by a deployed workflow at industrial scale. This workflow integrates with a MBSE framework that is already adopted by the industry and enhance its features by enabling to automatically supply a visual representation of a complex system from specification documents. The visual representation of the complex system is automatically constructed from specification documents thanks to the extracted semantic triples (named entities, relations) that are used to generate a knowledge graph-based representation of requirements. Evaluation of this work is still ongoing. In the future, we want to fine-tune the BERT language model to create a model for identifying named entities and compare its performances with our actual named entity recognition component.

References

[1] Henderson, Kaitlin and Salado, Alejandro (2021) *Value and Benefits of Model-based Systems Engineering (MBSE): Evidence from the Literature*, Systems Engineering, 24(1):51—66

[2] Dalpiaz, Fabiano and Ferrari, Alessio and Franch, Xavier and Palomares, Cristina (2018) *Natural Language Processing for Requirements Engineering: The Best is Yet to Come*, IEEE Software, 35(5):115-119

[3] Zhao, Liping and Alhoshan, Waad and Ferrari, Alessio and Letsholo, Keletso J and Ajagbe, Muideen A and Chioasca, Erol-Valeriu and Batista-Navarro, Riza T (2021) *Natural Language Processing for Requirements Engineering: A Systematic Mapping Study*, ACM Computing Surveys (CSUR), ACM New York, 54(3):1–41

[4] Min, Bonan and Ross, Hayley and Sulem, Elior and Veyseh, Amir Pouran Ben and Nguyen, Thien Huu and Sainz, Oscar and Agirre, Eneko and Heintz, Ilana and Roth, Dan (2023) *Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey*, ACM Computing Surveys, ACM New York, 56(2):1–40

[5] Tikayat Ray, Archana and Pinon-Fischer, Olivia J and Mavris, Dimitri N and White, Ryan T and Cole, Bjorn F (2023) *aeroBERT-NER: Named-Entity Recognition for Aerospace Requirements Engineering using BERT*, AIAA SCITECH 2023 Forum, 2583

Infrastructure, planning, and environment

FRCCS

Benchmarking algorithms for matching geospatial vector data

Paul Guardiola[1✓], Juste Raimbault[1], Ana-Maria Olteanu-Raimond[1] and Julien Perret[1]

[1] *Univ Gustave Eiffel, ENSG, IGN, LASTIG, F-94160 Saint-Mandé, France; first-name.lastname@ign.fr*
[✓] *Presenting author*

Abstract. Matching algorithms for geospatial vector data are useful tools for change detection in spatial systems. Different algorithms have been proposed but not systematically compared, and they furthermore require a parametrisation. This contribution benchmarks two such algorithms, and proceeds to their sensitivity analysis, providing Pareto optimal instantiations regarding accuracy and runtime.

Keywords. *Urban dynamics; Matching geospatial vector data; Sensitivity analysis*

1 Introduction

Cities and territories are at the core of pressing sustainability issues, and an understanding of processes driving the evolution of urban systems is necessary to design and implement sustainable planning policies [5]. In that context, change detection in GIS data using data matching algorithms is a well established method to quantify urban dynamics [8], along other methods such as machine learning [7]. Geospatial vector data matching provides links between spatial features based on proximity; it can be used to enrich or control the quality of a dataset [1], or to detect changes by matching the two datasets of the same spatial area at two different dates. It can be declined into several modalities, including point data matching, line data matching for road network evolution, or polygon data matching for building change detection, for example. Several methods were developed for matching polygons, such as the Geometric Matching of Areas algorithm [2] (GMA) or multi-criteria algorithms [4]. They are developped in different contexts and perform accordingly on different cases: as it is usually, the case with real-world isues and geographical data, there is no "one-size-fits-all" algorithm, and furthermore never out-of-the-shelf as they require a detailed parametrisation. For this reason, a benchmark and sensitivity analysis of these methods is necessary for an accurate application to real world case studies, as well as to facilitate the use of the existing algorithms.

This contribution thus proposes a sensitivity analysis [6] and benchmark of the two aforementioned matching algorithms for geospatial vector data. We use two complementary measures of performance, namely the accuracy given by the F-score obtained on a ground truth dataset, and the algorithm run time, as different parametrisations induces different algorithmic complexities.

2 Data and methods

In general, sensitivity analysis is used to quantify the uncertainty in the output of a model depending on the uncertainty in inputs [6]. In our case, algorithm parameters (detailed below) are used as inputs. Two outputs are considered: F-score which gives the accuracy taking into account all cases of True/False-negative/positive , and execution time. We carry out a sensitivity analysis with a one-factor sampling on algorithm parameters. We benchmark two data matching algorithms. The first is Geometric Matching of Areas algorithm (GMA) which was developped specifically for polygons [2]. For each reference object (i.e. an object to be matched), the algorithm will select all candidates which intersect with it. Then, according some parameters and conditions on geometrical properties, the algorithm reduces the number of candidates to retain the minimum possible (see [2] for a full description of this filtering procedure, and Table 1 for the names of parameters we study - these have all a normalised range in $[0; 1]$). The second is a multi-criteria algorithm based on Dempster–Shafer theory [4], which combines matching beliefs of several criteria. The criteria evaluate the similarity between the reference object and candidates. Criteria we include in our implementation are euclidian, surface, radial and Hausdorff distances [3]. This algorithm is run twice, as it is not commutative regarding the reference dataset: this allows to switch from having 1-to-n matching links to n-to-m matching links, and be comparable with the GMA.

We benchmark the algorithms in a context of building change detection: our layers are buildings from BDTOPO (open data from IGN-France National Mapping Agency), respectively from 2013 for the reference and 2023 for the comparison dataset. The ground truth dataset was established manually by comparing aerial photographs for a medium-sized suburban area (around 2000 buildings) located in the outskirts of Rennes, France. We implement the two algorithms in python, and source code and data are available on an open git repository[1].

3 Results

Sensitivity analysis results reveal the importance of parameters and their influence on the model, as shown in Fig. 1 and Table 1 for the GMA. Some parameters have little influence on outputs, such as MIS or PIS while MCCF is strongly correlated with F-score. A second analysis revealed an anti-correlation between MIP and MSD. Indeed, a low value of MSD combined with a high value of MIP gets a higher F-score. These exploration of various parametrisation furthermore provide an optimised algorithm performance by selecting best performing parameter values.

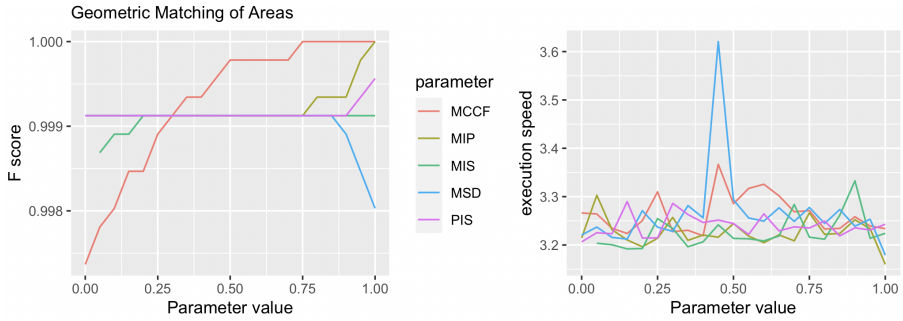


Figure 1: Variations of speed (ms) and accuracy of the GMA algorithm as a function of parameters.

[1] `https://github.com/paulguardiola/projetAppariement`

Table 1: Sensitivity analysis results for the GMA algorithm.

Parameter	Default	F-score Average	F-score variance	F-score correlation	Average speed	Speed variance	Speed correlation
MIS	1	0.99908	$1.32e-08$	0.59	3.23	0.0012	0.46
MIP	0	0.99923	$5.56e-08$	0.66	3.23	0.0008	-0.19
PIS	0.8	0.99915	$1.10e-08$	0.49	3.24	0.0005	0.02
MSD	0.6	0.99903	$7.47e-08$	-0.55	3.27	0.0074	0.01
$MCCF$	0.3	0.99930	$6.66e-07$	0.91	3.26	0.0016	-0.01

Algorithm parameters: MIS: Minimum Intersection Surface, MIP: Minimun Intersection Percentage, PIS: Percentage Intersection Surface, MSD: Maximum Surface Distance, $MCCF$: Minimum Completness Correctness Final.

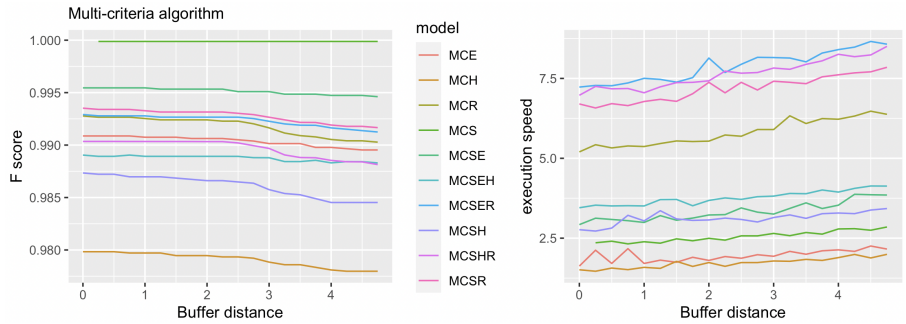


Figure 2: Variations of speed and accuracy of the multi-criteria algorithm (MC) as a function of buffer distance and for different combination of criteria (S: surface, E: euclidean, R: radial, H: Hausdorff).

Table 2: Sensitivity analysis results for the multi-criteria algorithm.

Parameter	Criteria	F-score Average	F-score variance	F-score correlation	Average speed (ms)	Speed variance	Speed correlation
d_b	S	0.999	0	$--$	2.55	0.027	0.95
d_b	E	0.990	$2.32e-07$	-0.96	1.95	0.032	0.65
d_b	H	0.979	$4.84e-07$	-0.96	1.71	0.025	0.93
d_b	R	0.991	$8.28e-07$	-0.95	5.80	0.170	0.96
d_b	S, E	0.995	$8.99e-08$	-0.96	3.33	0.084	0.92
d_b	S, R	0.992	$4.00e-07$	-0.96	7.17	0.161	0.96
d_b	S, H	0.986	$1.09e-06$	-0.96	3.12	0.037	0.75
d_b	S, H, E	0.989	$6.974e-08$	-0.89	3.76	0.048	0.96
d_b	S, H, R	0.989	$7.13e-07$	-0.90	7.64	0.200	0.97
d_b	S, E, R	0.992	$2.93e-07$	-0.93	7.88	0.228	0.95

Parameter: d_b buffer distance (m); **Criteria:** S: surface, E: euclidean, R: radial, H: Hausdorff.

For the multi-criteria algorithm, the sensitity analysis is made on buffer distance for selecting candidates and by varying the criteria included (enumeration of all combinations up to three criteria). Results, as shown in Fig. 2 and Table 2, highlight a model strongly dependant on criteria used. Models using radial distance criteria tend to be the slowest with no better F-score. The best F-score is obtained when using surface criteria. The buffer distance parameter is anti-correlated with the F-score, what means that considering only objects within a short distance of the reference is better for algorithm performance.

We finally show in Fig. 3 a comparison of algorithm instances regarding the two performance

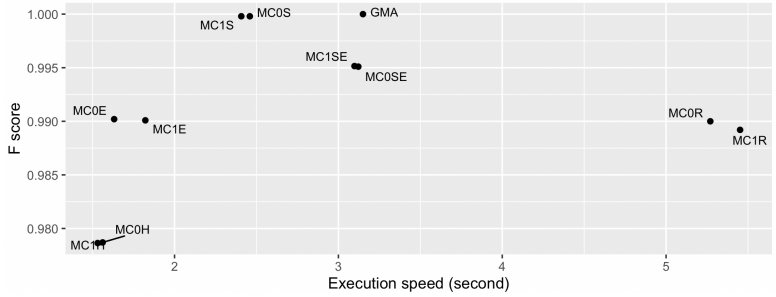


Figure 3: Pareto comparison of optimal instanciations of the matching algorithms. Optimal GMA parameters: $MIS = 1$, $MIP = 1$, $PIS = 0.8$, $MSD = 0.25$, $MCCF = 0.8$; MCOSE: Multi-Criteria ($d_b = 0$) with Surface and Euclidien criteria.

measures, suggesting a Pareto optimisation between accuracy and speed. In our case, only GMA obtains the highest F-score but at a cost of twice the execution speed of algorithm MC0S. This suggests that in real world settings, the choice of the algorithm will depend on compromises to be made between conflicting performance objectives.

Future work includes a more thorough sensitivity analysis using Global Sensitivity Analysis [6], the inclusion of other algorithms in the benchmark, the test on other types of polygon GIS data such as parcel data, and the test on several ground truth datasets to ensure results robustness.

References

[1] Atef Bel Hadj Ali. *Qualité géométrique des entités géographiques surfaciques. Application à l'appariement et définition d'une typologie des écarts geométriques.* PhD thesis, Université de Marne La vallée, 2001.

[2] Francis Harvey, F Vauglin, and A Bel Hadj Ali. Geometric matching of areas, comparison measures and association links. In *Proceedings of the 8th International Symposium On Spatial Data Handling*, pages 557–568, 1998.

[3] I Maidaneh Abdi, Arnaud Le Guilcher, and A-M Olteanu-Raimond. a classification model for the inference of spatial precision of openstreetmap buildings with intrinsic indicators. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10:167–174, 2023.

[4] Ana-Maria Olteanu. *Fusion de connaissances imparfaites pour l'appariement de données géographiques: proposition d'une approche s' appuyant sur la théorie des fonctions de croyance.* PhD thesis, Université Paris-Est, 2008.

[5] Celine Rozenblat and Denise Pumain. Conclusion: Toward a methodology for multi-scalar urban system policies. *International and Transnational Perspectives on Urban Systems*, 385, 2018.

[6] Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global sensitivity analysis: the primer.* John Wiley & Sons, 2008.

[7] Junye Wang, Michael Bretz, M Ali Akber Dewan, and Mojtaba Aghajani Delavar. Machine learning in modelling land-use and land cover-change (lulcc): Current status, challenges and prospects. *Science of the Total Environment*, 822:153559, 2022.

[8] Emerson MA Xavier, Francisco J Ariza-López, and Manuel A Urena-Camara. A survey of measures and methods for matching geospatial vector datasets. *ACM Computing Surveys (CSUR)*, 49(2):1–34, 2016.

FRCCS

Driver Deviation: A Measure of Traffic Changes in Low Traffic Neighbourhoods in London, UK

Shazia'Ayn Babul[1✓], Nicola Pedreschi[1] and Renaud Lambiotte[1]

[1] *Mathematical Institute, University of Oxford, UK ; shazia.babul@maths.ox.ac.uk, nicola.pedreschi@maths.ox.ac.uk*

✓ *Presenting author*

Abstract. Traffic barriers are used in London, UK, to encourage people to prioritise sustainable transport, and decrease motor traffic through residential neighbourhoods. These barriers block off certain roads, preventing drivers from taking their normal routes. In this work, we develop two network centrality measures to estimate the effectiveness of traffic barriers placed on roads around London. We present a node-based and edge-based measure, which we will compare with empirical data of how traffic density changed on roads after the implementation of the barriers.

Keywords. *Spatial Networks;Urban Planning;Network Centrality*

1 Introduction

Spatial networks are an extremely important tool to be used for predicting and understanding human mobility. They are often motivated by applications to urban planning, for example, predicting traffic flows in and around cities [3, 9, 5]. Such questions are becoming increasingly important for designing environmental policies that motivate a transition to green transport, and decrease motor vehicle traffic inside cities.This work is a collaboration with Sustrans Charity, a UK based organisation working to promote green transport. Here, we are particularly interested in the effects of "Low Traffic Neighbourhoods" (LTNs) inside the city of London, United Kingdom. LTNs are part of a scheme introduced in 2020, to reduce motor vehicle traffic in residential neighbourhoods through the addition of physical barriers that block traffic on certain inner roads of the neighbourhood [6]. Figure 1.a shows the LTN covering the London borough of Islington. The Islington LTN is split into sub-LTN areas, including the highlighted Canonbury East sub-LTN (Figure 1.b). The LTN is shown on top of the driving network of London retrieved from [2].

Car traffic in cities has been studied in various models that take into account different levels of area-specific information including population density, access to public transport, and employment [8, 4, 1]. In this work, however, we develop a network based approach using the changes to the city's driving road network to estimate the effectiveness of each LTN at reducing motor traffic. To this end, we introduce two measures (1) a node-centric measure of the average driver inconvenience for residents of the neighbourhood seeking to exit the sub-LTN in a motor vehicle, and (2) an edge-centric measure of the change in traffic flow density along different

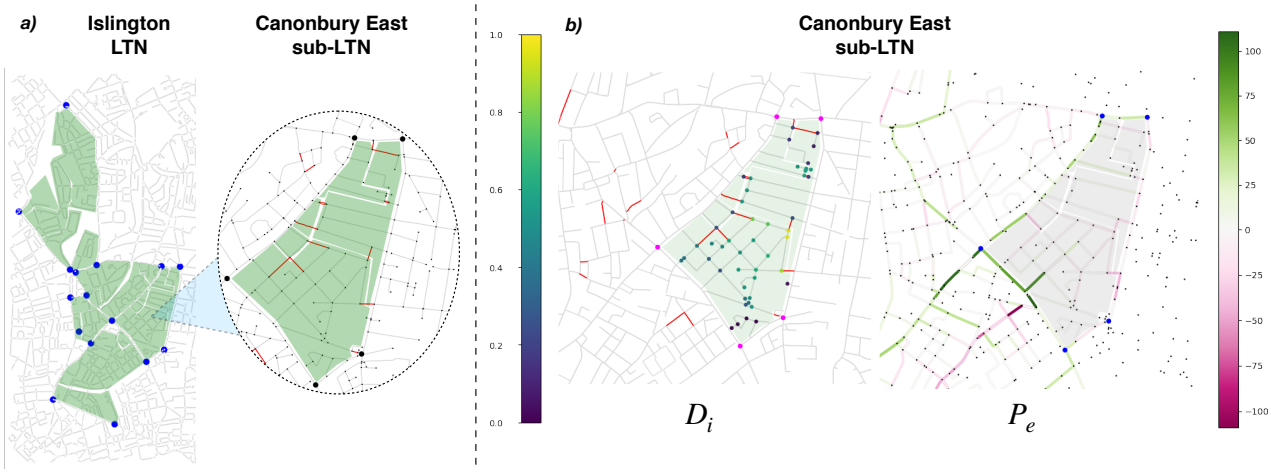


Figure 1: **a)** Islington LTN (left) and Canonbury East sub-LTN (right), highlighted in green on top of the road network of London. Nodes are junctions, and edges are driving roads. Blocked roads are shown in red. **b)** Canonbury East sub-LTN. Left: Nodes coloured by normalised driver deviation score. Right: Edge coloured by path density change score.

roads both inside and around the sub-LTN area. We then compare these measures with the observed change in motor vehicle traffic after the LTN implementations.

2 LTN Measures

We define $\mathcal{V}$ as the set of nodes contained inside the sub-LTN polygon, and $\mathcal{W}$ as the set of important junctions used to exit the sub-LTN, shown in bold black for Canonbury East in Figure 1.a (right). These junctions were identified by experts at the Sustrans organisation. We use $\mathcal{E}$ for the set of edges contained inside the sub-LTN polygon, on the road network without traffic filters. For each node $i \in \mathcal{V}$, we define a node-centric measure of driver deviation, denoted D_i, which measures the level of inconvenience for a driver starting from node i and trying to exit the sub-LTN through one of the important junctions. We define this by:

$$D_i = \frac{1}{|\mathcal{W}|} \sum_{j \in \mathcal{W}} (d'_{ij} - d_{ij}) \quad , \tag{1}$$

where d_{ij} is the weighted shortest path distance (in meters) between the nodes i and j on the road network graph without the LTN filters, and d'_{ij} is the same on the graph with the LTN filters in place. For each edge $e \in \mathcal{E}$, we define an edge-centric measure of path density change, to estimate the traffic density change on that road, after the placement of the traffic filters, denoted ΔP_e. We first define σ^e_{ij} to be 1 if the shortest path between node i and j passes through edge e, and 0 if not. The expected number of paths passing through edge e is:

$$P_e = \sum_{i \in \mathcal{V}, j \in \mathcal{W}} \sigma^e_{ij} \quad . \tag{2}$$

We look at this change before and after the LTN filters are put in place, defining $\Delta P_e = P'_e - P_e$, where P'_e is the same quantity on the road network with LTN filters in place.

3 Results: Canonbury East sub-LTN

Figure 1.b (left) shows the interior nodes coloured by their normalised driver deviation centrality (i.e., $D_i = 1$ means maximum deviation, $D_i = 0$ corresponds to minimum deviation instead)

given in Equation 1. The nodes most effected by the filters, in yellow, are those for which the paths to the important junctions (magenta) have been blocked off most effectively by the filters. Journeys beginning at these nodes now have to take longer routes to exit the sub-LTN, likely through the boundary roads. Each sub-LTN is assigned a driver deviation score by averaging over the driver deviation scores of the nodes contained inside, results are given for all the sub-LTNs that comprise the Islington LTN in Table 1. Figure 1.b (right) shows the Canonbury East sub-LTN, with the edges coloured according to the edge centric path density change measure. The path density increases on certain boundary roads, while decreasing on many inner roads, an expected result of the filters. However, we also observe that the path density can increases on certain inner roads, as certain journeys are now routed through them, and also density can decrease on other boundary roads, as the filters can actually block access to the boundary. After computing these measures for each sub-LTN, we will compare with actual traffic density changes on a sample of roads, using data obtained from [7]. Traffic density data is available only for certain roads, as the location of the sensors are up to the discretion of the local authorities.

Table 1: Average Driver Deviation Per Sub-LTN

sub-LTN	Avg. Driver Deviation (m)
Canonbury East	290.52
Canonbury West	124.31
St. Peter's LTN	335.85
Highbury West	186.68
St. Mary's Church	339.48

References

[1] Ana LC Bazzan and Franziska Klügl. A review on agent-based technology for traffic and transportation. *The Knowledge Engineering Review*, 29(3):375–403, 2014.

[2] Geoff Boeing. Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65:126–139, 2017.

[3] David E Boyce and Huw CWL Williams. *Forecasting urban travel: Past, present and future*. Edward Elgar Publishing, 2015.

[4] David A Hensher and Kenneth J Button. *Handbook of transport modelling*. Emerald Group Publishing Limited, 2007.

[5] Ryuichi Kitamura, Cynthia Chen, Ram M Pendyala, and Ravi Narayanan. Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation*, 27:25–51, 2000.

[6] TfL Community Team. Low traffic neighbourhoods: What, why and where?, Aug 2021.

[7] Asa Thomas and Rachel Aldred. Changes in motor traffic in london's low traffic neighbourhoods and boundary roads. *Case Studies on Transport Policy*, page 101124, 2023.

[8] Vincent Verbavatz and Marc Barthelemy. Critical factors for mitigating car traffic in cities. *PLoS one*, 14(7):e0219559, 2019.

[9] Pu Wang, Timothy Hunter, Alexandre M Bayen, Katja Schechtner, and Marta C González. Understanding road usage patterns in urban areas. *Scientific reports*, 2(1):1001, 2012.

FRCCS

Internal Migration in Rhineland-Palatinate - The evolution of the migration network

Christian Wolff[1✓], Markus Schaffert[1], Christophe Cruz[2] and Hocine Cherifi[2]

[1] *Mainz University of Applied Sciences, i3mainz–Institute for Spatial Information and Surveying Technology, Germany ; Christian.Wolff@HS-Mainz.de, Markus.Schaffert@HS-Mainz.de.*
[2] *Université de Bourgogne, Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB UMR CNRS 6303), France ; Christophe.Cruz@U-Bourgogne.fr, Hocine.Cherifi@U-Bourgogne.fr*
[✓] *Presenting author*

Abstract. The understanding of migration flows is an important research area for various disciplines, such as infrastructure and transportation planning. Instead of statistical analyses of in- and out migrations of individual communities, this work uses complex network analyses to analyze the evolution of the internal migration network in 2012-2021 between association communities in the federal state of Rhineland-Palatinate. The results show peculiarities in some years, which might be related to a high level of international migration.

Keywords. *Internal Migration; Complex Network Analysis; Spatial Data*

1 Introduction

Migration is one of the most important forms of human mobility. Understanding migration connections is important for various areas, especially for spatial planning. The understanding of migration, as well as its spatial patterns and causes, makes an important contribution to understanding urbanization, re-urbanization, sub-urbanization, or counter-urbanization since migration is often a driver of these effects. The work presented herein provides an overview of the development of internal migration between the various association municipalities in the federal state of Rhineland-Palatinate (RLP) in Germany from 2012 to 2021. Internal migration is analyzed as a network consisting of the municipalities as nodes and the migration flows between them as edges. This representation generates an additional perspective that expands the existing analyses. While these approaches mainly rely on statistical analyses of in- and out-migration flows of the individual municipalities, the network-based approach facilitates the study of the properties of the migration network. Statistical analyses of migration movements in RLP by the Rhineland-Palatinate State Statistical Office (RLPSSO) show a strong population growth since 2011. However, this is not due to a high birth rate but rather to high numbers of immigrants, especially from abroad. In addition to this international migration, internal migration in RLP shows different migration patterns depending on age. It is primarily the urban communes that benefit from the immigration of young people between 20 and 35, while middle-aged and older adults leave the urban communes for the rural districts. However, this migration does not lead to rejuvenation but rather to the aging of the local population, although

it often includes the migration of the children of middle-aged adults. Since 2012, more people have migrated from urban communes every year than the other way around. It should be noted that in the case of international migration, the strong increases in asylum seekers, especially in 2015, were often initially recorded in Trier and then distributed among the districts. This procedure led to a distortion of external and internal migration. However, the pattern of increasing internal migration from the urban communes to the districts remains unaffected.[11] The data basis for the analysis presented herein is the migration flows between the association municipalities in the federal state of RLP from 2012 to 2021. The data was provided by the RLPSSO ((c) Rhineland-Palatinate State Statistical Office; Reproduction and Dissemination, also excerpts, are permitted provided the source is cited). The data set Administrative Areas 1: 250,000 (VG250-EW) (© GeoBasis-DE / BKG 2022) from the Federal Agency for Cartography and Geodesy (BKG) is used to represent the data spatially.

2 Results

Using the python libraries NetworkX and CDlib [12], the analyses of the migration network are carried out on macroscopic (topology of the whole network)[5, 6], mesoscopic (community and core structure)[2, 3, 1], and microscopic levels (centralities)[7, 9, 8, 10] to enable the most complete representation of the changes in the network over time.

Macroscopic Analysis: At the macroscopic level, abnormalities are particularly evident around 2015. This period has been affected by strong international immigration. Figure 1 shows the migration rates over the district borders for the state RLP divided into Germans (red '>'), foreigners (blue '<'), and total (green 'x'). Statistical analyses by the RLPSSO also showed that the years 2014 and 2015 were significant for the average values of internal migration in the 2011 - 2020 period. In 2012, more than half of the urban communes and districts still had deficits in internal migration. In 2015, apart from the independent city of Trier, all administrative districts showed internal migration surpluses due to the peculiarity mentioned at the beginning. In 2020, only 11 urban communes or districts had deficits in their internal migration balances.[11] In line with these analyses, the highest number of edges in the undirected networks examined herein is in 2016. In that year the number of edges increases by around 10% from 2012 (6546 edges) to 2016 (7251 edges) and falls continuously in the following years until 2020. In 2021, there is another increase in the number of edges. Although the networks consist of 170 nodes, they show a low diameter of two, which could indicate the existence of hub municipalities that share migration connections with all other cities. A similar pattern to the number of edges can also be seen in the number of degrees in most association municipalities. The highest value for the median degree is also in 2016 (approx. 80). We calculate the global parameters transitivity, density, and degree correlation, with and without considering the total migration flows as weight. All parameters show a similar trend in the undirected networks (Figure 1). The internal migration network in the RLP shows a high complexity. The highest density value is 0.505 in 2016. So, more than half of all possible migration links exist between any two municipalities in the network. The course of transitivity suggests that migration flows are not concentrated on single hubs, e.g., because of the distribution of international migrants from Trier to the surrounding rural areas, but also between the surrounding areas. Interestingly, this development reverses after 2016 and only increases again in 2021. We also calculate the degree correlation to analyze the assortativity in the networks (Figure 1 bottom left chart). Overall, the degree correlation is at a low level. The results show that in 2016, the network becomes more disassortative. Indeed, the highest value for a negative correlation is -0.119. The effect could be due to increased urban-rural connections where residents from

urban communes with higher degrees migrate to rural municipalities with lower degrees. We also include the total migration flows as weights for the analysis of the assortativity (Figure 1 middle left chart). The results show that considering the weight, correlation values are even smaller. The year 2015 is conspicuous but in the opposite direction, making the network more non-assortative (-0.04).

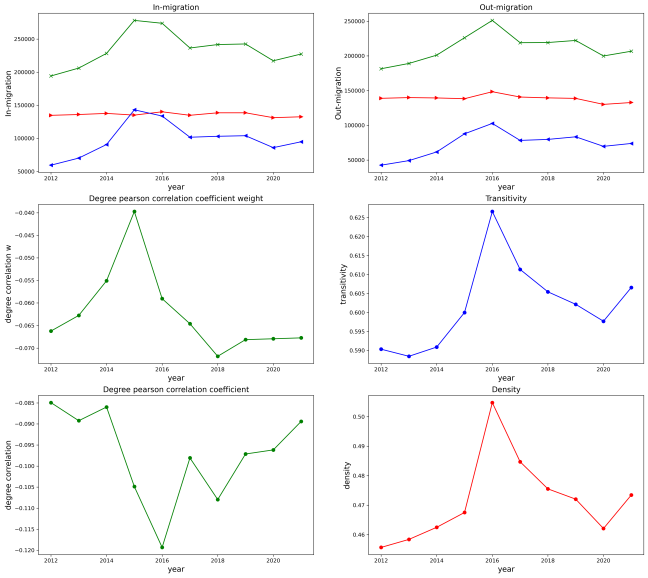


Figure 1: Migration rates (over district borders) for the state of RLP (© Federal Statistical Office of Germany, 2024.) and global parameters for the same period

Mesoscopic Analysis: Mesoscopic Analysis: An analysis of the community structure using the well-known Louvain algorithm and the total migration flows as a weight for the undirected network reveals a community structure with five to seven communities. Each of the five permanent communities contains one of the urban communes with the highest population, Mainz, Koblenz, Trier, Ludwigshafen, or Kaiserslautern, and surrounding municipalities which fits the urban-rural migration pattern mentioned in the first section. Although the modularity drops in 2015 and 2016 from about 0.6 to 0.5, there is no major change in the community structure in 2015 or 2016. **Microscopic Analysis:**Further analyses using degree-based centralities of the directed networks also show insightful results. Figure 2 shows the out-degree centrality for the directed networks. The urban communes with the highest population are also the nodes with the highest out-degree centrality. In addition to the trend of migrations from the urban communes to the districts, the destinations of the urban leavers are also dispersed. What is also interesting is the increase in the out-degree centrality of individual association municipalities in individual years. The municipality of Ingelheim, west of Mainz has a particularly high out-degree centrality in 2014/2015. This also applies to the municipality of Hermeskeil near Trier from 2015 to 2020.

Acknowledgements This work is part of the project Spatial Intelligence for the Integrated Care of Senior Citizens in Rural Neighbourhoods (Raumintelligenz für die integrierte Versorgung von Seniorinnen und Senioren in ländlichen Quartieren (RAFVINIERT)). It is funded by the Carl Zeiss Foundation in the program Society Transfer - Intelligent Solutions for an Ageing Society ("Transfer - Intelligente Lösungen für eine älter werdende Gesellschaft").

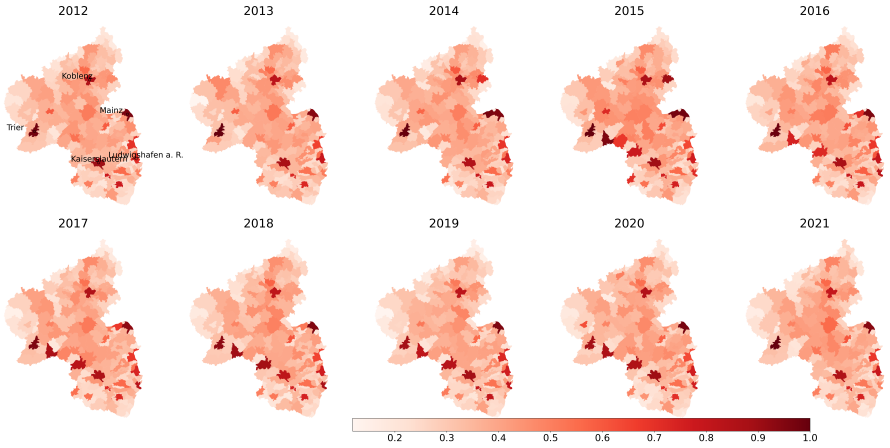


Figure 2: Out-Degree-Centrality

References

[1] Issa M. Diop, Chantal Cherifi, Cherif Diallo, and Hocine Cherifi. Revealing the component structure of the world air transportation network. *Applied Network Science*, 6:1–50, 2021.

[2] Zakariya Ghalmane, Chantal Cherifi, Hocine Cherifi, and Mohammed El Hassouni. Extracting backbones in weighted modular complex networks. *Scientific Reports*, 10(1):15539, 2020.

[3] Zakariya Ghalmane, Chantal Cherifi, Hocine Cherifi, and Mohammed El Hassouni. Extracting modular-based backbones in weighted networks. *Information Sciences*, 576:454–474, 2021.

[4] Aric Hagberg, Pieter Swart, and Daniel Chult. Exploring Network Structure, Dynamics, and Function Using NetworkX. *Proceedings of the 7th Python in Science Conference*, 2008.

[5] Khubaib Ahmed Qureshi, Rauf Ahmed Shams Malick, Muhammad Sabih, and Hocine Cherifi. Complex network and source inspired covid-19 fake news classification on twitter. *IEEE Access*, 9:139636–139656, 2021.

[6] Khubaib Ahmed Qureshi, Rauf Ahmed Shams Malick, Muhammad Sabih, and Hocine Cherifi. Deception detection on social media: A source-based perspective. *Knowledge-Based Systems*, 256:109649, 2022.

[7] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Interplay between hierarchy and centrality in complex networks. *IEEE Access*, 8:129717–129742, 2020.

[8] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Characterizing the interactions between classical and community-aware centrality measures in complex networks. *Scientific reports*, 11(1):10088, 2021.

[9] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Investigating centrality measures in social networks with community structure. In *Complex Networks & Their Applications IX: Volume 1*, pages 211–222. Springer, 2021.

[10] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Comparative evaluation of community-aware centrality measures. *Quality & Quantity*, 57(2):1273–1302, 2023.

[11] Statistisches Landesamt Rheinland-Pfalz. Demografischer Wandel in Rheinland-Pfalz, 2022.

[12] Giulio Rossetti, Letizia Milli, and Rémy Cazabet. CDLIB: A python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, 2019.

Social complexity

FRCCS

Geographic Distance and Equity Within a Collaboration Network

Andrew Estrada[1✓], Colin Chun[1], Mitashi Parikh[1], Zoë J. Wood[1], Theresa Migler[1]

[1] *Computer Science and Software Engineering Department, California Polytechnic State University San Luis Obispo, CA, USA ; aestra46@calpoly.edu, cjchun2001@gmail.com, mdparikh@calpoly.edu, zwood@calpoly.edu, tmigler@calpoly.edu,*
[✓] *Presenting author*

Abstract. We analyze publication data within the realm of computing to explore geographic distance between authors who publish together. We compare trends between female and male authors with a sample of schools from the U.S. public university system, specifically within the state of California. We find noticeable differences in the reach of female and male authors over their career. These differences were consistent over time despite reach having increased for both genders.

Keywords. *Collaboration Network; Geographic Distance; Gender; Equity; Computing*

1 Introduction

Research frequently requires more than a team of one and collaborations can take many forms, from two people who share an office to tens or hundreds of people who are spread across the globe. International collaboration, from a city-to-city perspective, has increased over the last three decades along with the average distance to the strongest collaboration partners [8]. The geographic distance of research collaborations has mostly been investigated on a larger scale, either by city [8] or by university [16]. We are interested in investigating whether differences in collaboration distance at the scale of individual authors exist with regard to gender identity.

While it has been shown that diversity of perspectives [19] and, by extension, attributes such as gender [22] is beneficial to problem-solving and research, research and academia in the U.S. remains predominantly male [24]. The tendency for researchers to collaborate with others of the same gender, known as gender homophily, is still prevalent in academia [12]. Researchers of underrepresented race and gender are more likely to produce novel or innovative connections between ideas, yet their work is less likely to be impactful and they are less likely to sustain careers in research [11].

We define 'collaboration distance' as a measure of the geographic distance between an author and everyone they have published with. In this paper we contribute findings that, within the realm of computing, reflect a difference in collaboration distance between female and male authors.

2 Related Work

There has been much research detailing gender related differences in research and collaboration. Spoon et al., looking at employment census data and survey responses, found that not only do women leave academia at higher rates than men, they do so (or are considering doing so) for gendered reasons across the entire U.S. university system [21]. With more focus on collaboration, Whittington used global life-science patent information to analyze the collaboration network of 216,000 inventors. The analysis showed that women are less likely to be in positions of strategic advantage within the network – they are less likely to have "brokerage" ties where they connect two otherwise unconnected inventors [25].

Not much research exists comparing geographic distance along gender although, Abramo et al., proved that among Italian researchers, female researchers have a lower tendency to collaborate internationally than males, despite overall having a greater propensity to collaborate [5]. More generally in regards to collaboration distance, Csomos et al. looked at Web of Science (WoS) data for three 2-year spans from 1994-2016 and found that international collaboration has increased along with the average distance to the strongest collaboration partners. While it is the case across all three decades that the farther the distance, the weaker the collaboration ties, high-impact collaborations tend to span large distances, suggesting impact can be gained from a larger geographic reach [8].

Previously, this project has focused on different aspects of collaboration for specific schools or subsets of schools within our sample as the database has grown. For example, Nakamichi et al. and McNichols et al. analyzed a specific master's-level public California university at the department scale and university-wide respectively; their findings suggest that previous trends seen at the university level are not always consistent among different sectors of a school (colleges and departments) [18, 14]. McNichols et al. studied inferred male and female subnetworks of a public California university and found that the male network had a larger span of collaborators while the female network had stronger collaboration ties [15]. Carroll et al. analyzed collaboration data for PhD-level public California universities and showed evidence for a strong tendency to collaborate with European institutions; the paper also found that including publications with larger collaborator counts dramatically skews the network away from California to Europe [7]. The project now centers around a master's level and a majority of PhD-level public California universities.

3 Methods

3.1 Data Collection

We use data that was gathered from Scopus, an extensive database of authors and publications from Elsevier. Scopus was used to retrieve author and publication data from publications surrounding the topic of computing with contributions from a sample of California public university authors primarily from 1970 to 2020. See Figure 1 for the distribution of publications by year. We refer to the universities that we specifically retrieved data for as *base-schools*.

From Scopus, the name, organization, document count, citation count, h-index, and field of each author was collected. For publications, the title, year, venue, language, cited by count, DOI, and contributors list were collected. Latitude and longitude coordinates were obtained from data fields available in Scopus using the open source library GeoPy [15].

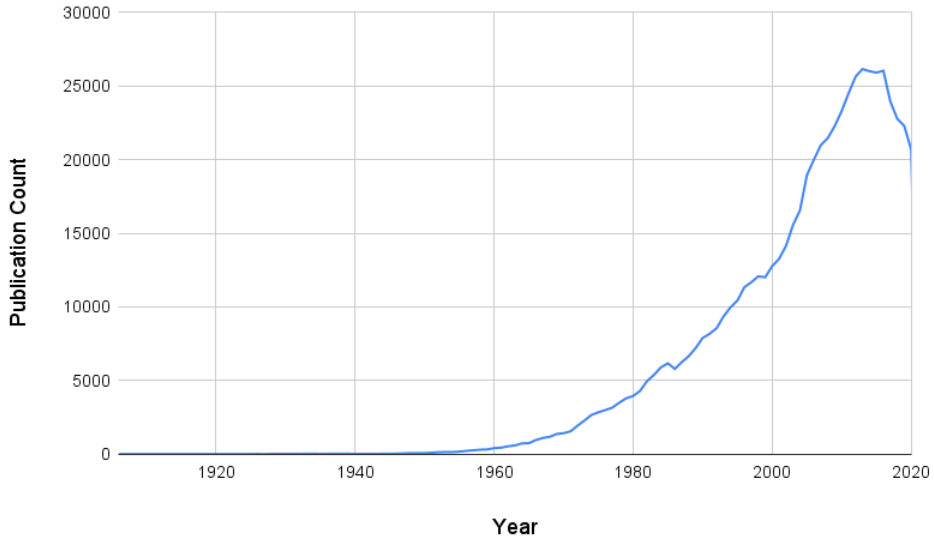


Figure 1: Count All Publication by Year Published

The resulting database has collaboration information on 555,827 publications spanning 851,077 authors. NamSor [20] was used to infer both the binary gender of authors as well as their race and ethnicity based on their name. With NamSor's inferred gender, 30.9% of the authors are female and 69.1% are male. The data was validated to appropriately reflect the true state of the real network by crosschecking with publicly available data such as the gender distributions of research within California public universities.

3.2 Network Building and Analysis

For the creation of the network, vertices represent authors and 2 authors are connected with an edge if they have collaborated on a publication. For each publication, we create a clique connecting all pairs of contributing authors.

Publications with more than 20 authors were ignored when creating the network (and consequently when conducting analysis) to avoid skewing of the data. The thought behind a threshold of 20 is that it is unlikely that an author would be able to meaningfully engage with each collaborator in a larger group and that is what we want our findings to capture. The restriction of at most 20 authors brings the number of analyzed publications to 535,178 spanning 619,804 authors.

From this data, the resulting collaboration network has 619,804 vertices, 4,207,930 unique edges, and 7,099,714 total connections between authors – counting duplicate edges from different publications.

3.3 Distance Calculation

For this paper, distance calculation is a measure of the distance between an author and everyone they have worked with using their institution as their site of work or 'location'. There are two pieces of location data associated with authors within our database – the current organization at the time data was retrieved (which we'll call the *current*), and the organization they were affiliated with at the time of each publication (which we'll call *publication-specific*). With this information, there were two ways that we calculated distance:

1. ***Average Distance of Collaborations -*** A collaboration is two authors working together for a specific publication. The distance of a collaboration is the distance between the authors' publication-specific organizations. This calculation looks specifically at publications for which an author was at a base school.
   For example, let us suppose we have two authors, A and B, who worked together on three publications $P1$, $P2$, and $P3$. Further suppose author A was at base school $S1$ for $P1$, base school $S2$ for $P2$ and was not a base school for $P3$ while author B was at organization O for all three. Author A's average distance of collaborations, assuming no others exist, is

$$\frac{(\text{distance from } S1 \text{ to } O) + (\text{distance from } S2 \text{ to } O)}{2} \tag{1}$$

   Notice that P3 is ignored since author A was not affiliated with a base school when contributing.
   Collaboration data from 340,431 publications, all of which have at least one author who was affiliated with a base school at the time of publication was analyzed in this manner.
2. ***Average Distance to Collaborators -*** This is a bit more straightforward using the current organization to calculate distance. For each author whose current organization is a base school, we average the distance to all unique collaborators.
   Using the same example above, let's suppose that author A's current organization is a base school $S3$, their average distance to collaborators is:

$$\frac{(\text{distance from } S3 \text{ to O})}{1} \tag{2}$$

   Collaboration data from 485,114 publications, all of which have at least one author who is currently affiliated with a base school, was analyzed in this manner.

The Average Distance of Collaborations gives information to the effect of "at any given time, how far are an author's collaborators?", while the Average Distance to Collaborators is more a measure of an author's reach over time.

Currently, distances to organizations which we do not already have longitude and latitude data for (NULL values within the database) are ignored. Of all organizations in the database 6.03% did not have location data. In both versions of the distance calculations, there exist authors who have no distance data, either because they did not collaborate with anyone or because they only collaborated with authors from organizations which we do not have location data for.

"As the crow flies" distance is calculated between two latitude and longitude coordinates converted to radians (ϕ_1, λ_1) and (ϕ_2, λ_2) using the Haversine formula:

$$dist = 2r \arcsin \sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos\phi_1 * \cos\phi_2 * \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)} \tag{3}$$

4 Findings

Section 4.1 details the results of analyzing individual collaborations – average distance of collaborations, Section 4.2 details the results of analyzing reach – average distance to collaborators, and Section 4.3 details reach over time.

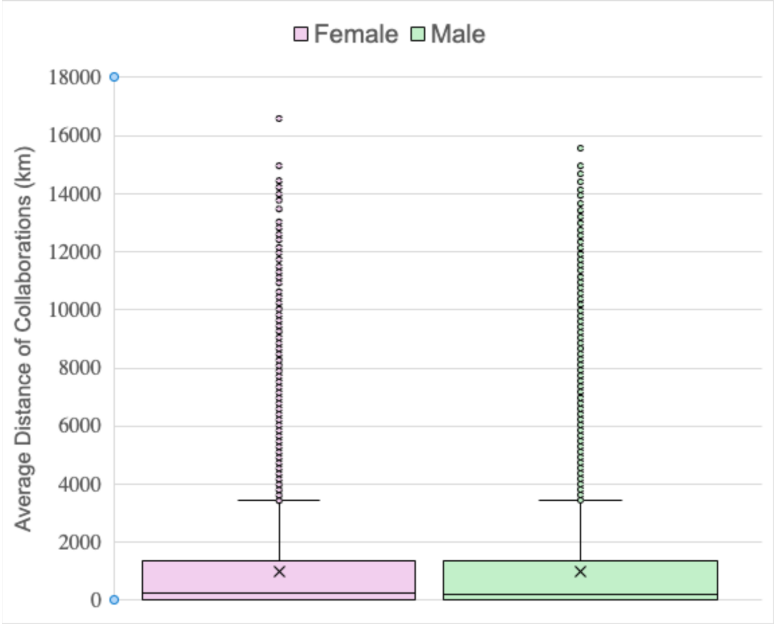


Figure 2: Average Distance of Collaborations Across All Authors Who Were Affiliated with Base Schools

4.1 Average Distance of Collaborations

From analyzing the collaboration data of all publications which have at least one author who was affiliated with a base school at the time of publication, there was little difference between genders. See Figure 2 for the distributions and Table 1 for aggregate data. This suggests that for any given collaboration, the distance will be the same regardless of the author's gender.

Table 1: Statistics for Average Distance of Collaborations

Data Point	Female Authors	Male Authors
Average (km)	994.9	993.9
Max (km)	16590.3	15567.0
Median (km)	243.9	186.2
Mode	0 km (32.9%)	0 km (36.5%)
Total Author Count	51,822	111,840
Authors Without Collaborators	264	647

4.2 Average Distance to Collaborators

From analyzing the collaboration data of all publications which have at least one author who is currently affiliated with a base school, we saw that male authors tend to have a sightly further reach by approximately 144 km further than female authors. Note this shows female author's reach being ∼92% that of male authors. See Figure 3 for the distributions and Table 2 for aggregate data.

Table 2: Statistics for Average Distance to Collaborators

Data Point	Female Authors	Male Authors
Average (km)	1556.7	1700.7
Max (km)	14979.6	15569.3
Median (km)	965.5	1098.8
Mode	0 km (14.3%)	0 km (15.7%)
Total Author Count	43,715	87,527
Authors Without Collaborators	198	500

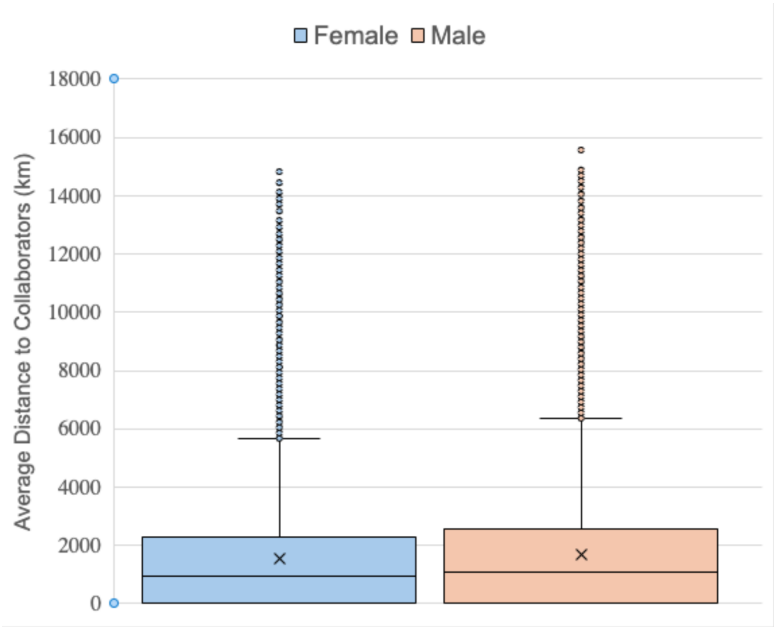


Figure 3: Distribution of Average Distance to Collaborators Across All Authors Currently at Base Schools

4.3 Reach Over the Decades

We then analyzed the publication data of three separate decades (1970–1980, 1990–2000, and 2010–2020) in the same fashion. We saw that this trend (i.e. male authors having a slightly further reach than female authors) was consistent over the decades even though the average distance (reach) increased for both genders over time. See Figure 4 and Table 3. Among the three decades, the reach of female authors was on average ∼85% that of male authors.

We hypothesize that this may be a reflection of previous findings that may compound. Since females are less likely to be rewarded for their position in a collaboration network [25], they are less likely to be pulled towards better opportunities and more likely to leave academia [21]. Then, since female authors are less likely to relocate, are more likely to leave academia, [21] and are less likely to collaborate internationally during their academic careers [5], they are less likely to accumulate as large a reach as male authors. They may also be less likely to take on remote international opportunities which is measured in reach as well.

Table 3: Statistics for Average Distance to Collaborators Across Different Time Periods

	1970 - 1980		1990 - 2000		2010 - 2020	
Data Point	Female	Male	Female	Male	Female	Male
Average (km)	1042.1	1302.1	1508.7	1684.5	1633.7	1908.4
Max (km)	12741.3	15569.3	14404.0	17832.7	14979.6	15136.0
Median (km)	174.0	358.6	844.2	1026.1	1030.7	1309.8
Mode	0 km (30.8%)	0 km (27.3%)	0 km (18.4%)	0 km (17.7%)	0 km (12.2%)	0 km (12.2%)
Total Author Count	909	4,994	6,436	19,013	29,694	52,272
Authors Without Collaborators	44	211	83	257	117	256

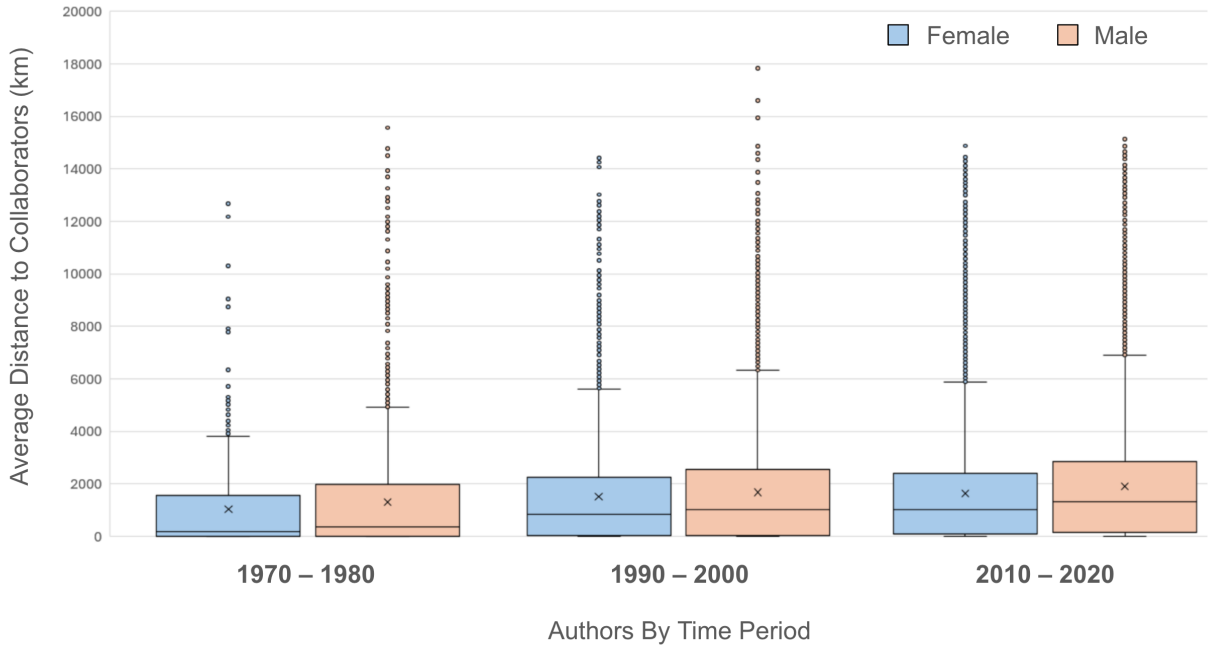


Figure 4: Distribution of Authors' Average Distance to Collaborators Across Different Time Periods

5 Conclusions and Future Work

Collegiate research has long been an important source of contribution to a wide variety of fields; it has impacted technology, industry, policy, and so much more. It is interesting, particularly in such an emerging field as computing, to analyze patterns in the work being done. In this study we concentrate on the geographic distance from individual authors and those that they have worked with. Looking at collaboration data from California public universities, we analyze two separate measures of distance and find that in regard to gender, a difference lies in the geographic reach of an author over all their publications. We acknowledge that since

collaborations with organizations for which we do not have location data is ignored, it may be the case that this missing data impacts some groups more than others skewing the results.

These finding are relevant to discussions surrounding university-level research objectives. Refinement of this work may include analyzing the two distance types for the same author population (currently the populations overlap but are not the same) as well as excluding graduate students who are likely to have a collaboration distance of 0 which skews the average. This can be seen in the histograms of Figure 5 and 6 which both show a large number of authors with a collaboration distance of 0 or close to 0. Further work could explore distance patterns of international collaboration, distance patterns within specific subject areas, and distance patterns between schools vs. industry.

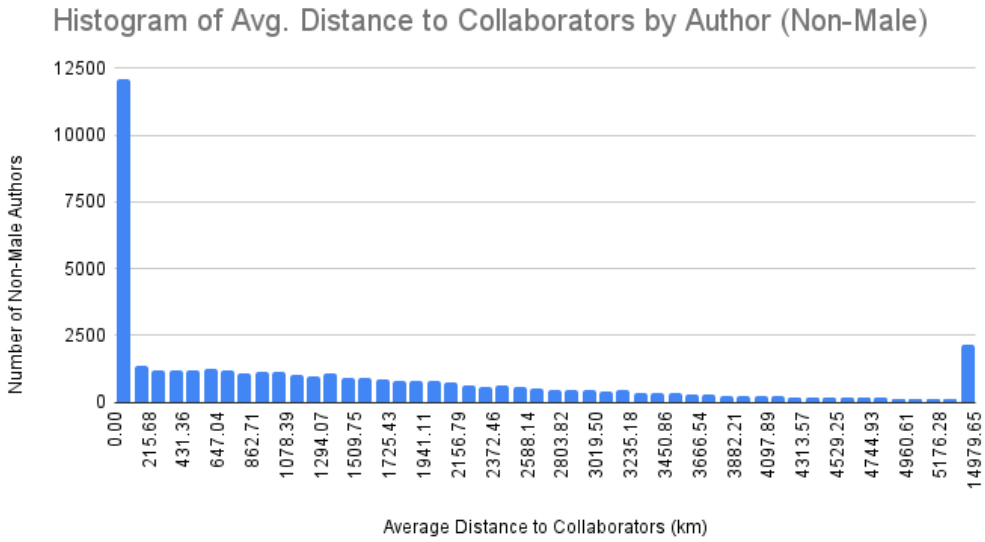


Figure 5: Distribution of Average Distance to Collaborators for Female Authors Currently at Base Schools

Acknowledgements

We would like to thank the BEACoN Program for funding and support. We would like to express our sincere gratitude to all previous contributors to this project: Mugizi Rwebangira, Logan McNichols, Steven Pineda, Emma Sauerborn, Brandon Tat, Kevin Yoo, Lauren Nakamichi, Conor Carroll, Nupur Garg, Gabriel Medina-Kim, Viet Lien Nguyen, Christian Rapp, Leticia Siqueira, Polina Volnuhina, Meghan Tran, Riley Mae Badnin, Justin Brunings, Bill Chan, Alex Liang, Eeshan Mishra, Drew Soderquist, and Alex Johnson.

References

[1] About BEACoN Research Scholars. `https://diversity.calpoly.edu/BEACoN`.
[2] Cal Poly Github. `http://www.github.com/CalPoly`.
[3] Strategic Research Initiatives Program: Areas for Strategic Growth, Leadership and Innovation. `https://research.calpoly.edu/strategic-research-initiatives`.
[4] Teacher-Scholar Model. `https://academicaffairs.calpoly.edu/teacher-scholar-model`.

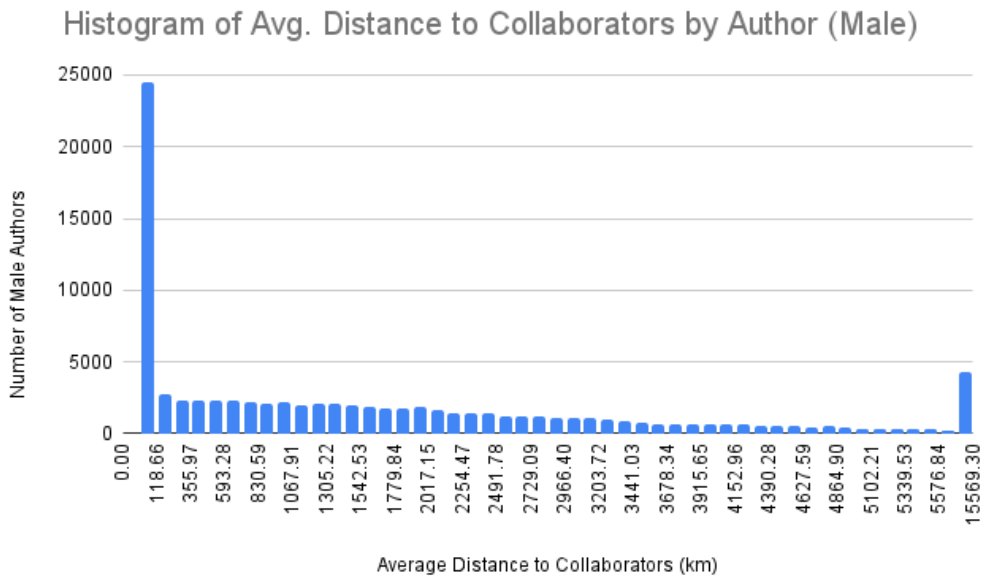


Figure 6: Distribution of Average Distance to Collaborators for Male Authors Currently at Base Schools

[5] Giovanni Abramo, Ciriaco Andrea D'Angelo, and Gianluca Murgia. Gender differences in research collaboration. *Journal of Informetrics*, 7(4):811–822, 2013.

[6] Peter S. Bearman and James Moody. Suicide and friendships among american adolescents. *American journal of public health*, 94(1):89–95, 2004.

[7] Conor Carroll, Nupur Garg, Theresa Migler, Barbara Walker, and Zoë J Wood. Mapping and visualization of publication networks of public university faculty in computer science and electrical engineering. In *CATA*, volume 2, pages 1–12, 2020.

[8] Gyorgy Csomos, Zsofia Viktoria Vida, and Balazs Lengyel. Exploring the changing geographical pattern of international scientific collaborations through the prism of cities. *PLoS ONE*, 15(11), 2020.

[9] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8:128–140, 1741.

[10] WASC/Academic Senate Teacher-Scholar Model Task Force. AS-725-11 Resolution on Defining and Adopting the Teacher-Scholar Model. *Academic Senate Resolutions*, March 2011.

[11] Bas Hofstra, Vivek V. Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A. McFarland. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291, 2020.

[12] Luke Holman and Claire Morandin. Researchers collaborate with same-gendered colleagues more often than expected across the life sciences. *PLoS ONE*, 14(4):e0216128, 2019.

[13] Hagar Mahmoud and Nadine Akkari. Shortest path calculation: A comparative study for location-based recommender system. In *2016 World Symposium on Computer Applications & Research (WSCAR)*, pages 1–5, 2016.

[14] Logan McNichols, Gabriel Medina-Kim, Viet Lien Nguyen, Christian Rapp, and Theresa Migler. Gender's influence on academic collaboration in a university-wide network. In *International Conference on Complex Networks and Their Applications*, pages 94–104. Springer, Cham, 2019.

[15] Logan McNichols, Steven Pineda, Emma Sauerborn, Brandon Tat, Kevin Yoo, Jane Lehr,

Zoë Wood, and Theresa Migler. Mavac: Mapping and visualization of academic collaborations with a focus on diversity. In *Complex Networks XII: Proceedings of the 12th Conference on Complex Networks CompleNet 2021*, pages 86–97. Springer International Publishing, 2021.

[16] Alessandro Muscio. University-industry linkages: What are the determinants of distance in collaborations?*. *Papers in Regional Science*, 92(4):715–740, 2013.

[17] Lauren Nakamichi. Analyzing gender in the cal poly collaboration network, 2020.

[18] Lauren Nakamichi, Theresa Migler, and Zoë Wood. An analysis of four academic department collaboration networks with respect to gender. In *Complex Networks & Their Applications IX: Volume 1, Proceedings of the Ninth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2020*, pages 262–272. Springer International Publishing, 2021.

[19] Stephanie M Reich and Jennifer A Reich. Cultural competence in interdisciplinary collaborations: a method for respecting diversity in research partnerships. *American journal of community psychology*, 38(1–2):51–62, 2006.

[20] Paul Sebo. Performance of gender detection tools: a comparative study of name-to-gender inference services. *Journal of the Medical Library Association*, 109(3):414+, 2021.

[21] Katie Spoon, Nicholas LaBerge, K. Hunter Wapman, Sam Zhang, Allison C. Morgan, Mirta Galesic, Bailey K. Fosdick, Daniel B. Larremore, and Aaron Clauset. Gender and retention patterns among u.s. faculty. *Science Advances*, 9(42):eadi2205, 2023.

[22] Talia H Swartz, Ann-Gel S Palermo, Sandra K Masur, and Judith A Aberg. The science and value of diversity: Closing the gaps in our understanding of inclusion and diversity. *The Journal of Infectious Diseases*, 220(220 Suppl 2):S33–S41, 2019.

[23] Jay Thompson. Cal poly ranked best overall master's-level university in the west, 2023. `https://www.calpoly.edu/news/cal-poly-ranked-best-overall-masters-level-university-west`.

[24] K. Hunter Wapman, Sam Zhang, Aaron Clauset, and Daniel B. Larremore. Quantifying hierarchy and dynamics in us faculty hiring and retention. *Nature*, 610:120 – 127, 2022.

[25] Kjersten Whittington. A tie is a tie? gender and network positioning in life science inventor collaboration. *Research Policy*, 47(2):511–526, 2018.

FRCCS

Geographical variations of social mobility In France and its determinants

Andrea Russo[1✓], Floriana Gargiulo[1], Maxime Lenormand [2] and Cyril Jayet[1]

[1] *Sorbonne Université - CNRS (GEMASS) Paris (France) ; Andrea.russophd@gmail.com, cyril.jayet@sorbonne-universite.fr.*
[2] *INRAE - Institut national de recherche pour l'agriculture, l'alimentation et l'environnement. Montpellier (France) ; maxime.lenormand@inrae.fr .*
[✓] *Presenting author*

Abstract. The term "Inter–generational Social mobility (ISM)" refers to the change of the socio-economic situation of an individual with respect to his/her parents.
In this study we aim to study this particular phenomenon adding to the traditional statistical methods used in Sociology new analytical tools coming from the analysis of human dynamics in complex systems.
In particular, our study develops in two parallel directions. We first analyze how the geographical specificity of the territories, and furthermore the dynamics of the spatial migration networks among the territories, impacts social mobility patterns. Second, introducing a semantic space to better characterize the available jobs in the job market, we embed the social-mobility network in a metric space. This operation allow us to study ISM with scores derived from the job distance between father and son, generational variances, and a time network weighted by distance, reaching a resolution scale of the social phenomenon that could not be observed through the traditional statistical methods.

Keywords. *Social mobility; Spatial Inequalities; Geography; Computational method.*

Inter-generational social mobility (ISM), which explores the correlation between an individual's social status and that of their parents, is a longstanding matter of study in Sociology. This phenomenon is indeed the basic ingredient for the persistence of social inequality in our society. ISM has been a subject of extensive international comparisons. These comparisons aim to shed light on the extent to which changes in social mobility can be attributed to the expansion of education and the democratization of schooling [1]. More recently, significant attention has been devoted to investigating spatial variations in social mobility within the borders of a single

country [2, 3]. This body of work underscores the crucial role of spatial inequalities and enables the identification of the impacts of territorial characteristics on social mobility.

Previous research on France [4, 5] has predominantly depicted geographical variations in social mobility in a static manner, lacking an exploration of temporal shifts in the geography of social mobility. Consequently, these studies do not provide insights into the dynamic processes that shape territorial inequalities and the impact of technological transformations and globalization. While the repercussions of globalization and technological transformations have been extensively scrutinized in economic studies [6] and within the realm of sociology focusing on inequalities [7], the existing body of work on technological innovation has primarily concentrated on understanding social inequalities rather than delving into the intricate dynamics of social mobility.
Remarkably, in the France's context, there is a notable gap in the research on ISM concerning the impact of fundamental societal phenomena, like the technological transformations and the globalization processes, that have deeply transformed the geography of the territories and the physical mobility of society.
Filling this gap requires, first of all, to go beyond the traditional methods to analyze ISM, based on classical statistical tools and introduce new analytical frameworks, inspired by human mobility studies and, eventually new data sources, to integrate the labour surveys provided by the national statistical offices. Our study is therefore grounded in this interdisciplinary challenge to develop a novel analytical approach, based on Complex Systems methods, to investigate the finer structure of ISM and its temporal evolution.
In the context of this ambitious research program we first address two parallel research objectives:

- To identify the social and economic characteristics of the territories than influence social mobility, and to explore the co–evolution of territorial transformations and ISM based on sociological attributes.
- To analyze ISM at a finer scale of the job-market, investigating the secondary factors that could affect individual career choices (i.e. continue working in a rural environment, being grown up in a geographical mobility situation, etc.)

To develop this project, we use different types of data. The central dataset comes from the French Labour Force survey conducted from 1982 to 2022. This survey covers millions of individuals, with detailed information on birthplace, residence, level of study, occupation and employ of the parents.
From this survey we can build the basic objects that we manipulate in our study: the geographical "origin-destination" tables of the individual trajectories ("social mobil-

ity tables", SMT), $P_{ij}^{M/F}(x,t)$. Namely we count for each time-step, t, the number of individuals living in x, that having a parent with a job i exert a job j (we build two different matrices for the fathers' and the mothers' jobs). For each of those mobility tables, we compute a large set of indicators, like the social mobility score (matrix trace divided by total sum minus one), Phi score (categorical association measure), the 'perfect immobility' (ideal immobility) and the dissimilarity index. The list of these indicators,$\kappa^{M/F}(x,t)$, that we name "mobility vector", constitute the mobility fingerprint of each matrix, allowing to identify patterns of mobility characteristics of each region and each generation.
We identify the proximity between regions and generation by the distance between their mobility vector, $D_{x,y}(t) = d(\kappa^{M/F}(x,t), \kappa^{M/F}(y,t))$. To measure this distance, we use various techniques of dimensionality reductions and from complex network analysis. This distance in patterns of mobility among geographical regions is afterward explained by their territorial characteristics such as economic development of the region, degree of urbanization and the economic innovations produced in the regions.
The dynamics of the geographical network between the regions, based on their distance according to the mobility vector, and the reorganization of its meso-scale structure, gives a better understanding of the association between the geographical determinants and the socio-economic drivers (associated to the transformations of the regions) of ISM.

The second difficulty in analyzing ISM is that this type mobility is not associated to a proper metric space: occupation categories have a hierarchical structure based on the French classification, the PCS. This classification comes with a precises definitions for each type of job (each type of job has a 4-digit code, for a total of 486 types of jobs).
This classification doesn't allow to properly identify the real distance between the position of parents ant that of their descendants : parents can for example work as farmers for a wine producer and their children as directors of the market for the same company. This socio-economic progression however spams a smaller distance respect to a case where the children were employed as directors of the market for a car producer. Since getting this data isn't possible (even from the INSEE site), we first enrich the description of the PCS description with some homogeneous pieces of information extrated with the ChatGPT4 API, for each of the 486 jobs (working places, possible studies, level of study, rural job, etc.), considering that ChatGPT 4 has the advantage to be trained on an enormous volume of information, giving access to ordinary information and representations of occupations in webpages and books. Merging the definitions given in the PCS manual and the information from ChatGPT, it is possible to create a Job's Map based on the processing of textual information by word-embedding with BERT. This map, represented in figure 1, allows to embed the occupation landscape in a metric space and to perform finer

measures of ISM.In the Maps each nodes is a PCS code where peoples are associated with. By doing so, the maps allows to generate a list of parent(s) and offspring for each specific job by the numerical distance, providing a numerical value and variance between the distance of jobs parent and they child's for each generation gaps. This information contributes the scientific community by allow to study ISM with a totally new level, as there is no parameter for comparison between different types of jobs and between generations and people via the cosine distance in a two-dimensional space, giving the possibility of weighting (through distance) the link in a temporal network (between generations).

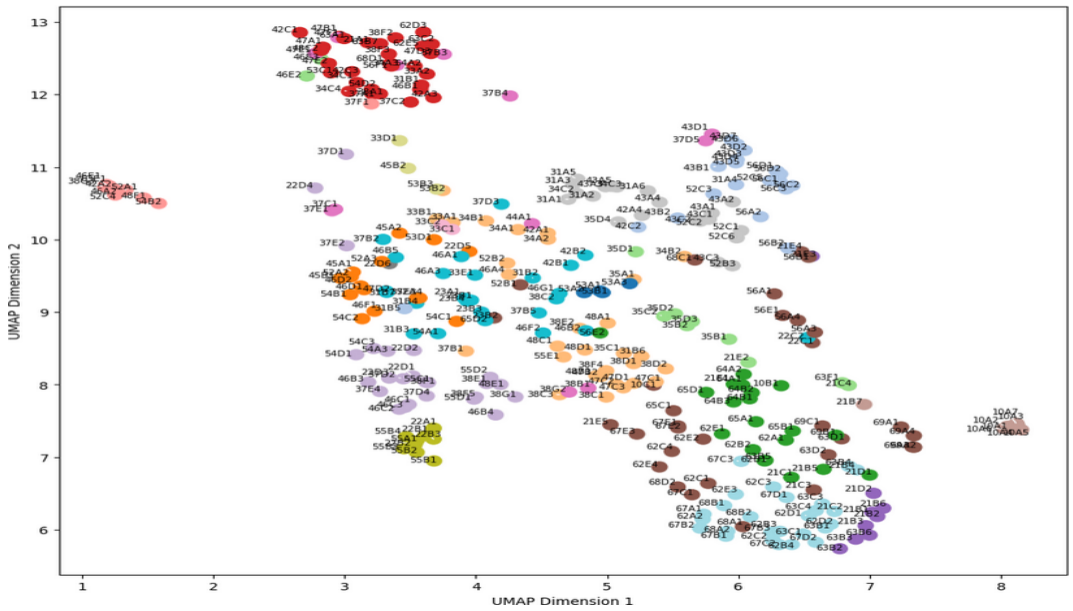


Figure 1: Word-Embedding with BERT about PCS category definition

References

[1] Richard Breen (Ed.). *Social mobility in Europe*. Oxford University Press, Oxford, 2004.

[2] Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. *Where is the land of opportunity? The geography of intergenerational mobility in the United States. The Quarterly Journal of Economics*, 129(4):1553-1623, 2014.

[3] Gabriele Ballarino and Nazareno Panichella. *Social origins, geographical mobility and occupational attainment in contemporary Italy. Genus*, 77(3), 2021.

[4] Cécile Dherbécourt. *La géographie de l'ascenseur social français. Document de travail France Stratégie*, Vol. 6, 2015.

[5] Gábor Kenedi and Laurent Sirugue. *The Anatomy of Intergenerational Income Mobility in France and its Spatial Variations*. Working paper, 2021.

[6] Daron Acemoglu and Pascual Restrepo. *Robots and jobs: Evidence from US labor markets. Journal of Political Economy*, 128(6):2188-2244, 2020.

[7] Melinda Mills. *Globalization and Inequality. European Sociological Review*, 25(1):1-8, 2009.

FRCCS

Investigation of Social Networks In University upon Belongingness and Mental Health

Rachel Izenson[1✓], Lauren Allen[1], Deric Alvarez[1], Zoe Chen[1], Cameron Hardy[1], Tony Li[1], Julissa Romero[1], Julia Ye[1], Zoë J. Wood[1], and Theresa Migler[1]

[1] *Department of Computer Science and Software Engineering, California Polytechnic State University, San Luis Obispo, CA, USA ; rizenson@calpoly.edu, lallen15@calpoly.edu, dkalvare@calpoly.edu, schen184@calpoly.edu, chardy02@calpoly.edu, tli30@calpoly.edu, jhern430@calpoly.edu, jye13@calpoly.edu, zwood@calpoly.edu, tmigler@calpoly.edu*

[✓] *Presenting author*

Abstract. It has been shown that computing students have a statistically significantly lower overall sense of *belongingness* compared to other science students. A sense of community is important for many reasons. For example, there are studies that show that a student's sense of belonging correlates with improved academic performance. Our research aims to analyze students at a predominantly undergraduate public university in the United States for their sense of belonging through a network science lens. We surveyed for their sense of belonging, as well as their social network, to understand how friendships impact one's sense of belonging. When student responses were split by gender, males reported having a higher sense of belonging than females, and females reported higher belongingness than transgender, non-binary, or gender non-conforming individuals. The four nodes with the highest out-degree on the social network that was constructed were all professors, indicating the importance of student-faculty relationships.

Keywords. *Graph Theory; Imposter Syndrome; Social Networks*

1 Background

The imposter phenomenon (i.e. imposter syndrome) was first coined by Clance & Imes in 1978 and defined as "internal experience of intellectual phoniness" among high achieving individuals [4]. In a study in 2018, 58% of tech employees reported they experience imposter phenomenon [5] and a similar percentage (57%) was reported among Computer Science graduate and undergraduate students when asked if experiencing imposter phenomenon in college [13].

Prior research has found that a student's sense of belongingness has a direct correlation with their academic success, more specifically that "positive [student] interpersonal experiences at university increase student belongingness, and belongingness leads to student success" [15]. Related work using the Framingham Heart Study social network also shows that "people's happiness depends on the happiness of others with whom they are connected" [7]. This study was able to determine that happiness spreads within a social network, with each person unknowingly

impacting the happiness of others.

Mental health and belongingness have previously been studied, and researchers were able to identify patterns between national belongingness statistics in Computer Science and students at public universities. Prior research found that women felt more connected than men, and 86.4% of female students did not feel like typical computer scientists [16]. At a different university, researchers found the same result: women, gender minorities, and students with underrepresented ethnic/racial identities had a lower rate of considering themselves a typical computer scientist [10].

This study was conducted at a public university located in California that is mainly undergraduate, with the average student-to-faculty ratio being 19:1. The department that this survey was conducted in has around 1,200 students total with 67 faculty members. The department contains Computer Science, Computer Engineering and Software Engineering majors and throughout the study we refer to students in this department as "computing students." 75.2% of computing students are male, 24.8% are female, and <0.1% are non-binary. The gender ratio is similarly reflected within the computing faculty population: 76.2% of the departments faculty members are male, and 23.8% are female.

This paper explores the social network of a computing department at a university with respect to belongingness scores and demographic factors. The network is split into subnetworks by gender identity and race and ethnicity to see how these individual factors effect a students sense of belongingness.

2 Methodology

2.1 Survey Distribution

We conducted a survey containing questions relating to mental health, social networks, and demographic information. The survey was sent out to all computing majors and across a variety of platforms (described next). 40% of students responded. The survey was distributed through a weekly department email, flyers around campus (specifically in the computing buildings), sent in the department Discord, word of mouth (with students in this research group encouraging peers to respond), and via asking professors to email their students. Some instructors gave students time in class to fill out the survey, and further buy-in was garnered via sending an email from the department chair encouraging students to fill out the survey. To give students time to fill out the survey, we attended 35% (39 / 111) of computing classes offered in the fall term. For the remaining 65% of sections, the professors teaching those sections were emailed asking if they could send an announcement to their students asking them to fill out the survey. As an incentive to fill out the survey we offered the chance to win a gift card. There were nine raffle winners, each of whom was awarded a $20 gift card.

Social Network Questions: We constructed our social network questions to build an accurate social network without causing survey burnout. The social network questions allowed students to list up to five people they are closest to within the computing department. After they listed an individual, they were asked questions to determine more about the relationship. Questions included: how close they feel to the person they listed, if they usually hang out for fun or to do school work, and if, to their knowledge, the person they listed is friends with the other people they previously listed. After entering up to five individuals, the survey included

an open text question for respondents to enter any other friends or acquaintances within the department. We were cautious of the total time it would take a student to fill out the survey and chose to limit social network questions to keep the survey under five minutes.

Belongingness Questions: The questions which focused on understanding respondents sense of belonging were taken from prior related surveys to enable comparison. These questions were motivated by Stewart [16] and Metcalf et al [10]. See appendix for all survey questions.

2.2 Anonymization

In order to analyze data while protecting student privacy, the data was anonym-ized to remove student names. We use the phrase "anonymized data" to reference the data after all names have been replaced with a unique identifier. This data is still not completely anonymized as there is demographic data that can be traced back to individual students.

Although professors were not invited to participate in the study, students could still list them in their social network.

Anonymization involved three steps.

1. **Step 1: Professor Name Conversions** We asked students to list up to five people they were closest to within the department, but also gave them an option to list any extra people they feel are important in their department social network. We were surprised to find that many responses included professors, but the professors were listed in a variety of versions (example: Dr. vs Professor vs Prof etc.). In order to make sure all of the variations of a professors name referred to the same person, the first step in the anonymization process was to create a mapping between the spellings of a professors name and their full name. Variations within the data were then replaced with a single full name.
2. **Step 2: Find and Correct Name Misspellings** We knew it was extremely unlikely that everyone would be able to list their friends' names without any misspellings, and we did not want our data to reflect the same person as two different nodes. In order to account for this, our anonymization script used fuzzy matching for all names.

   Fuzzy matching uses Levenshtein Distance to compare two sequences and returns a number indicating their similarity [6]. In our case, a high fuzzy matching score might indicate a misspelling of a name. Any other name that had a fuzzy matching score of over 70 was added to a list and reviewed. For example, "John Doe" and "Jonathan Doe" produce a fuzzy matching score of 80, while "John Doe" and "Jonathan Smith" produce a score of 45.

   A human data reviewer then determined if there were any duplicated names in the list containing misspellings. If there was a duplicated name, they were able to specify which misspellings should be grouped together to point to one node. For fuzzy matching, we used the `ratio()` function from the "fuzzywuzzy" package and used a threshold of 70 after a variety of experimentation. There was concern that setting the threshold higher could cause the algorithm to miss out on potentially matches. Note that this step did not change any of the data, but rather created a mapping from all names to their correct spellings that was then written to a file.
3. **Step 3: Create Unique Identifier Mapping** Using the map we created from the previous step, we took all of the values of the map to be valid names in our data. Then we created a map from a valid name to a unique identifier using the `uuid4()` function

from the uuid package. The algorithm then iterated through all the names in the data, mapping the name to a valid name, and then to the unique identifier, which replaced the name. The new data was written to a new file.

2.3 Network Analysis

In the network, nodes represent people and edges represent friendships. Two types of edges exist on the graph:

1. An edge connecting a respondent and a friend which they listed
2. An edge where a respondent indicated two people they listed were friends with each other

For example, if person A listed person B and person C as two of their friends, and also indicated that person B and person C were friends, the graph would look like Figure 1. If person B also took the survey and listed person A and person C as friends, but did not indicate person A and person C were friends, the graph would look like Figure 2. Only a new edge going from person B to person A was created because the edge from person B to person C already exists.

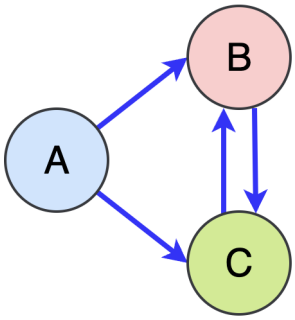


Figure 1: Graph Snippet including Person A's response

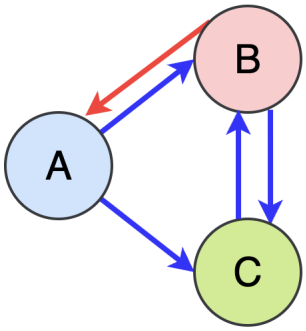


Figure 2: Graph Snippet including Person A's and Person B's response

To create subgraphs of the full social network, we created a graph of all respondents who identified as the specific characteristic we want to include in the subgraph and included all friendships that respondent listed.

We studied the social network with respect to degree distribution and clustering coefficient. In networks, a degree distribution is defined as the probability distribution of the degree of a specific node over the whole network [17]. The clustering coefficient of a network indicates the measure in which nodes in a graph are likely to cluster together. Clustering coefficients can vary between 0 and 1, where 0 indicates highly disconnected network and 1 indicates a highly connected network.

3 Results

3.1 Overall Results

In total, we received 494 survey responses. We cleaned the data for accuracy. For example, we found that eight students took the survey twice, two professors took the survey, and five people did not give consent, so we omitted all invalid submissions. We also found that 12 students were not in the department, thus their data was also omitted. In total, there were 467

valid responses, representing 38.9% of the student body. Out of these remaining responses, we found that 342 (73.2%) respondents were male, 112 (24.0%) were female, and 13 (2.8%) were non-binary, transgender, or gender-nonconforming.

3.2 Graph Analysis

When taking the survey, students could list up to five people who make up their social network as well as a question allowing them to list "any other friends or acquaintances" who are a part of their social network. For the first five friends students listed, they could indicate if any of their friends were friends with each other. To construct the social network from the data we collected, we used the NetworkX package. On average, students reported 4.359 friends total, and the mode being two friends reported. The distribution of number of friends reported can be seen in Fig 3.

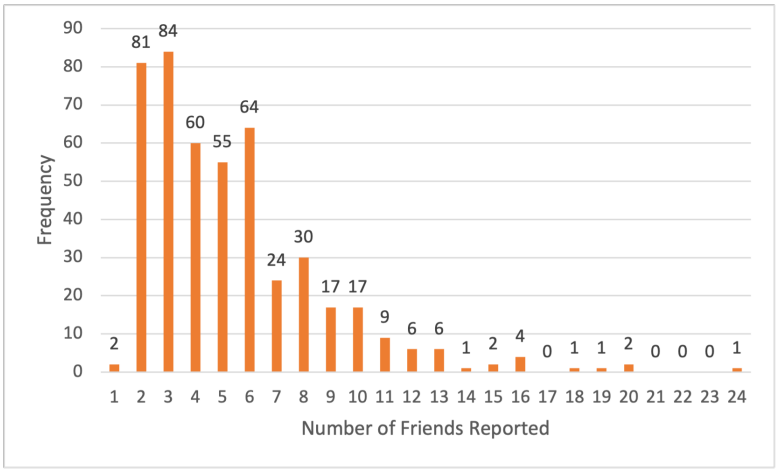


Figure 3: Histogram showing the number of friends reported vs frequency
The mean of the distribution was 4.359 and the mode was 2.

The social network contained 1325 nodes and 2054 edges. The degree distribution of the social network can be seen in Fig 4.

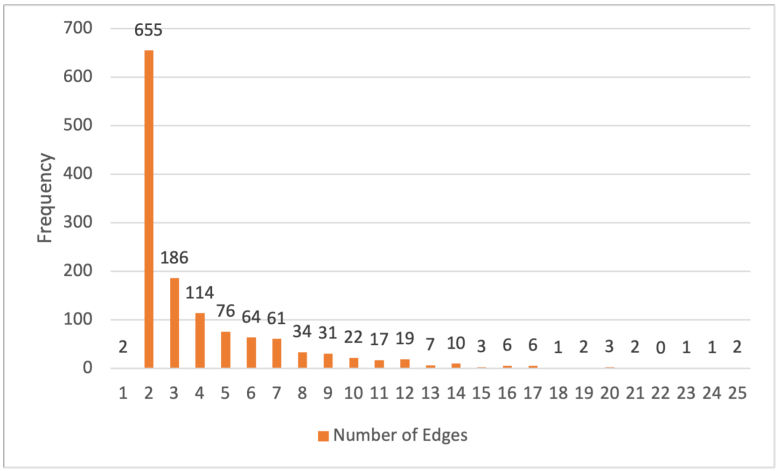


Figure 4: Degree Distribution of Social Network
There were 1325 nodes and 2054 edges total

The node with the highest degree is not particularly interesting because there was an option to enter as many friends as the student wanted. This would skew the data towards someone who

listed the most friends. Looking at a node's out degree tells us the degree of others who listed them as a friend. The four nodes with the highest out-degree (11, 15, 16, 17 respectively) were all professors. While this survey was directed towards students to understand students social network, many professors were listed for the question "List any other friends or acquaintances in your majors department." In total 101 different professors were listed from 110 different students representing 23.6% of students listing at least one professor in their social network.

3.3 Gender Identity and Social Networks

When split by gender, the degree distributions for the subgraphs can be seen in figures 5, 6 and 7. The number of nodes, edges, and clustering coefficients for each graph can be found on Table 1. A higher clustering coefficient in female identified data allows us to conclude that females in general are more closely connected than males. Although females reported lower scores on the mental health and success questions, it is not because of their lack of community. One reason for this might be because of active student clubs which support females in computing in this institution. This impact can also be seen in that 64% of female respondents reported that they were part of one of these clubs.

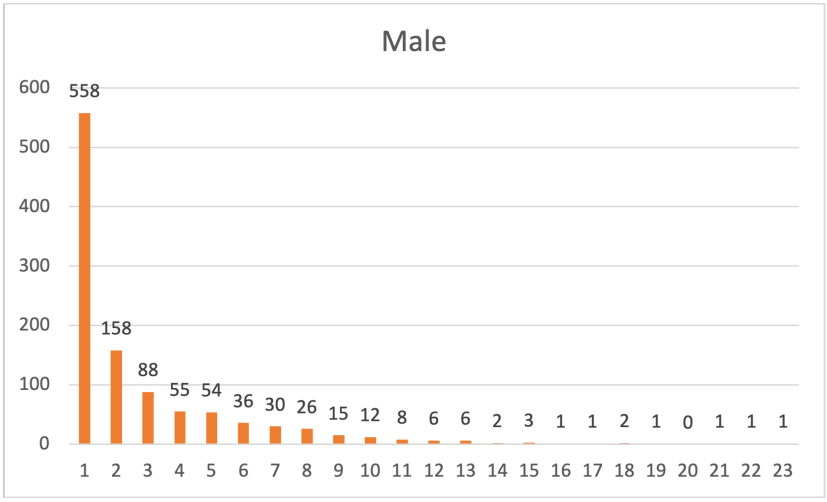


Figure 5: Degree Distribution for male subgraph

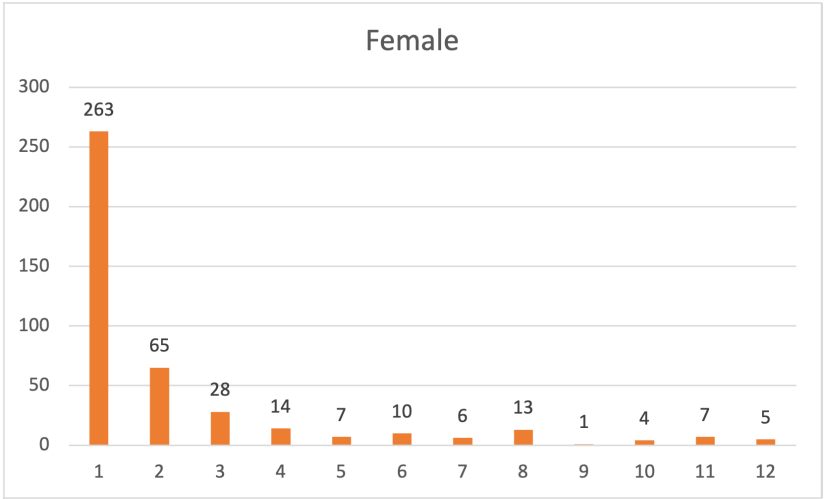


Figure 6: Degree Distribution for female subgraph

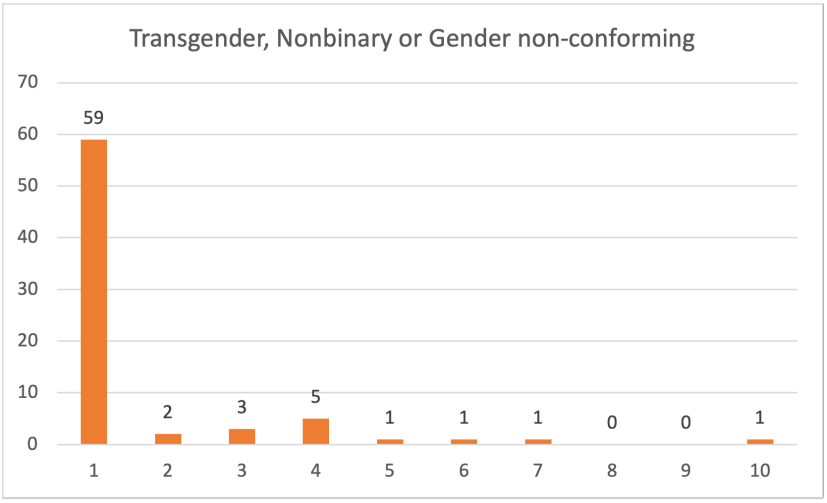


Figure 7: Degree Distribution for transgender, non-binary or gender non-conforming subgraph

Table 1: Graph characteristics on networks split by gender

Gender	Number of Nodes	Number of Edges	Clustering Coefficient
Female	427	512	0.0593
Male	1066	1054	0.0679
Transgender, Non-binary or Gender Non-conforming	74	66	0.0679

The subgraph containing only transgender, non-binary and gender non-conforming students has a low clustering coefficient indicating a disconnected network. There is no specific club at the university with the sole goal to connect and provide support to non-binary computing students. Many clubs work to be inclusive but another option could be an initiative to create a club with the mission to support non-binary computing majors.

3.4 Typical Computer Scientist

A study in 2018 found that there are four key factors that contribute to a STEM students sense of belongingness, one of which is science identity. Science identity can be defined with relation to "one's personal connection to their field". Students with a high science identity were found a more strong sense of belongingness [12].

When asked the survey question: "I consider myself a typical computer scientist", 144 (30.8%) respondents answered "Agree" or 'Highly Agree", while in contrast 201 (43.0%) respondents answered "Disagree" or "Highly Disagree" and 122 (26.1%) answered "Neutral" as seen in Figure 10. In a 2019 survey conducted at the same institution when asked the same question, 67.8% and in 2020 57.9% of students did not feel like typical computer scientists, which can be seen in Figures 8 and 9 [16].

In the survey conducted in 2019 and 2020, respondents could only choose between "Yes" and "No" when responding to if they felt like a typical computer scientist, while the survey in 2023 respondents could select 'Strongly Disagree', 'Disagree', 'Neutral', 'Agree' or 'Strongly Agree'. When "Neutral" is considered an agreement to the statement, our results show an improvement in the students sense of belonging over time, notably, a decrease of 14% since 2020. When "Neutral" is considered a disagreement to the statement, the results show that 69.1% of students disagree. This statistic falls between the average taken in 2019 and 2020.

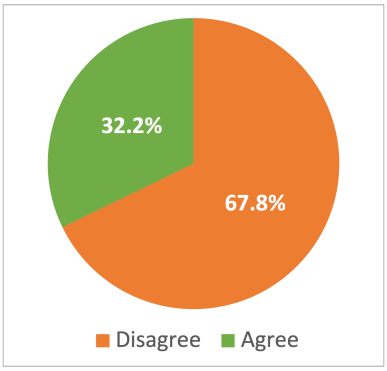


Figure 8: Responses to "Typical Computer Scientist" in 2019

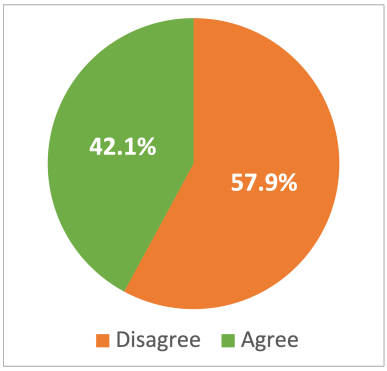


Figure 9: Responses to "Typical Computer Scientist" in 2020

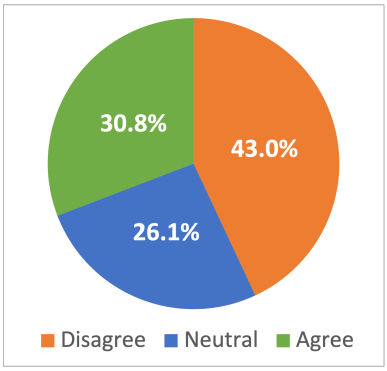


Figure 10: Responses to "Typical Computer Scientist" in 2024

If only 'Strongly Agree', 'Agree', 'Strongly Disagree' and 'Disagree' responses are considered, 41.7% of students agree with the statement while 58.3% of students disagree.

3.5 Gender Identity and Belongingness

When comparing male and female respondents we found a difference of 8.3%, where 40.0% of males and 48.3% of females responding that they did not feel like a typical computer scientist. When compared to those who identified as transgender, non-binary, or gender non-conforming, 77.2% responded they do not feel like typical computer scientists. We acknowledge this is a small group of students. However, this is significantly higher rate of students questioning their computing identity than the male and female students and suggests there is room for improvement within the department for non-binary students.

For questions pertaining to mental health and success, on average males reported higher fit and happiness. Questions pertaining to mental health and success include:

- I can be myself and still fit in in the department
- I feel happy more often than I am sad
- I consider myself a typical computer scientist
- Do you feel successful at this university?
- I have a healthy balance between my academic and personal responsibilities

On every question listed above, males agreed more with these statements on average than females, and females agreed more with these statements on average than those who identified as transgender, non-binary or gender non-conforming. These questions were asked on a scale from strongly disagree to strongly agree, and when converted to a 1-5 scale (1-strongly disagree, 5-strongly agree), males ranked the question: "I can be myself and still fit in in the department" with an average of 3.79 indicating they agree with the statement. Female identified students ranked this question 3.43 on average, while those who identified as transgender, non-binary or gender non-conforming on average ranked this question 2.54. It is apparent that those who do not identify as male or female are much more likely to feel they do not fit into the department. The distribution of responses to this question, split by gender can be seen in Figure 11.

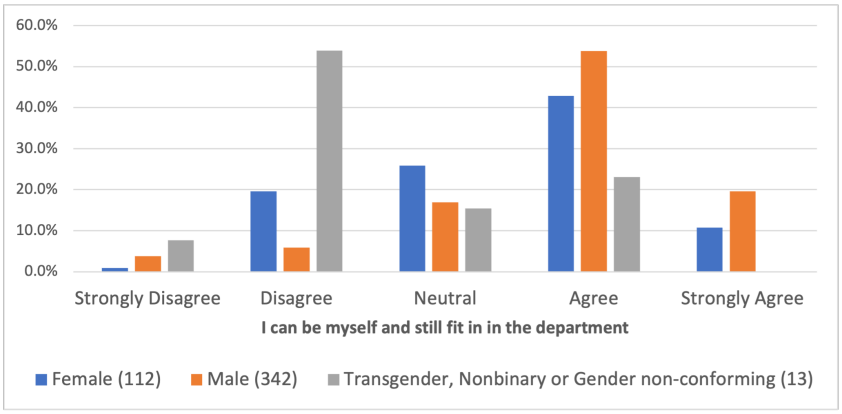


Figure 11: Student responses to "I can be myself and still fit in in the department" split by gender

3.6 Race and Ethnicity and Belongingness

Figures 12, 13, 14, 15 and 16 display the same five mental health and success questions from above when split by race/ethnicity. Race and ethnicity was split by the three most common identities reported (White, Asian/Pacific islander, Hispanic/Latino), mixed race, and historically marginalized races and ethnicities. On average students with a mixed race identity reported the most positive mental health scores closely followed by White students. Students with an identity from historically marginalized race and ethnicities scored the lowest overall on every mental health and success question compared to other race and ethnicities.

The question: "I consider myself a typical computer scientist" had the lowest score compared to the other questions overall for historically underserved students.

All race/ethnicities besides Asian/Pacific Islander, had more than 50% of their population disagree to the statement regarding feeling like a typical computer scientist. For students who are members of a historically underserved race/ethnicity group, 57% of students listed that they disagreed with the question, matching the national imposter syndrome average within the Computer Science industry.

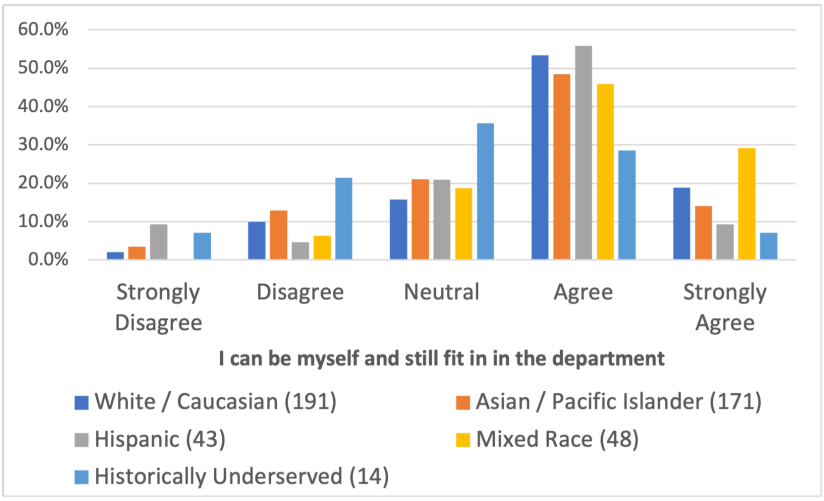


Figure 12: Student responses to "I can be myself and still fit in in the department" split by race/ethnicity

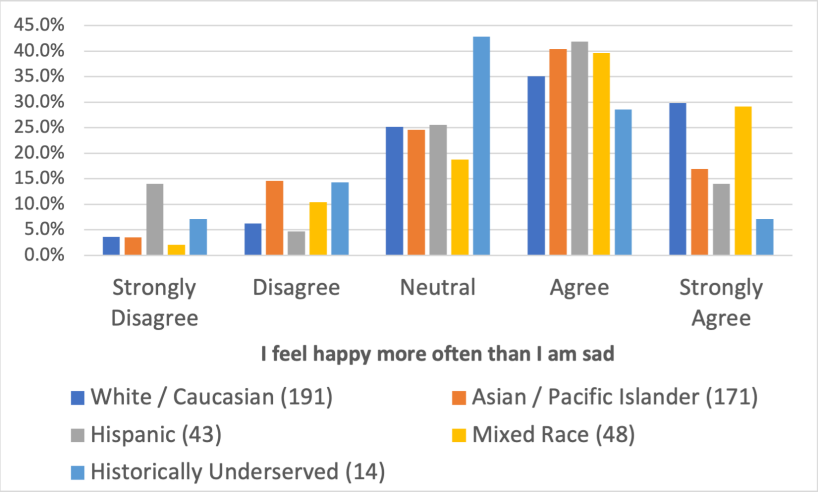


Figure 13: Student responses to "I feel happy more often than I am sad" split by race/ethnicity

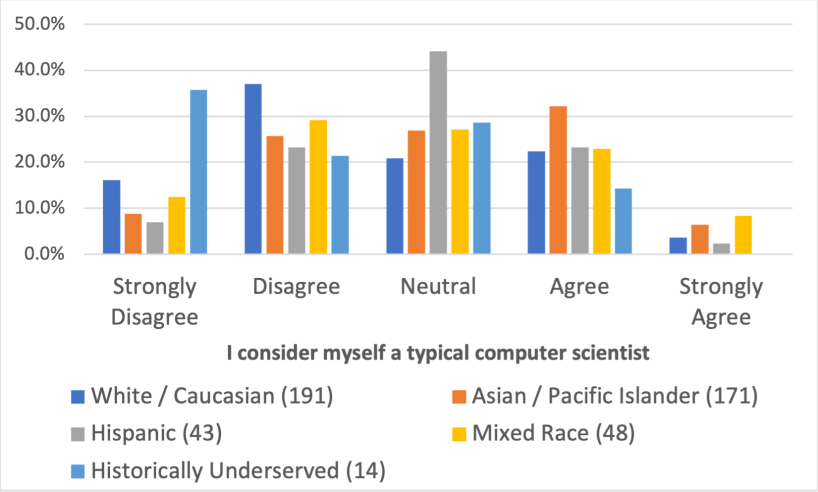


Figure 14: Student responses to "I consider myself a typical computer scientist" split by race/ethnicity

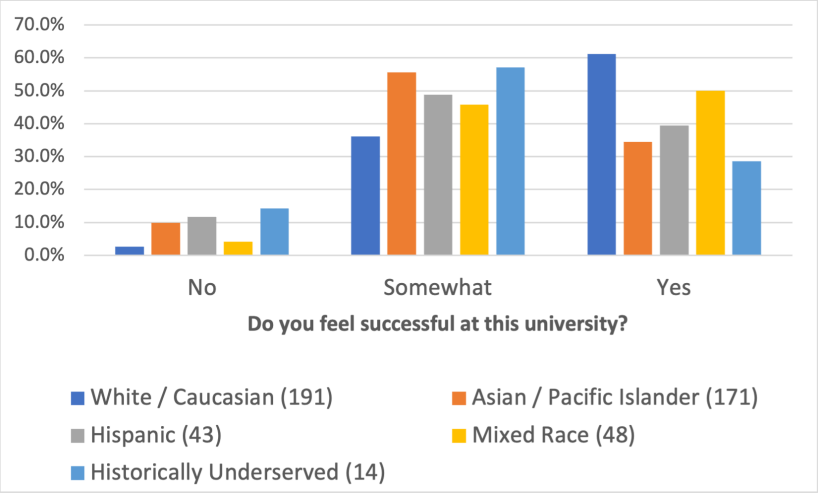


Figure 15: Student responses to "Do you feel successful at this university?" split by race/ethnicity

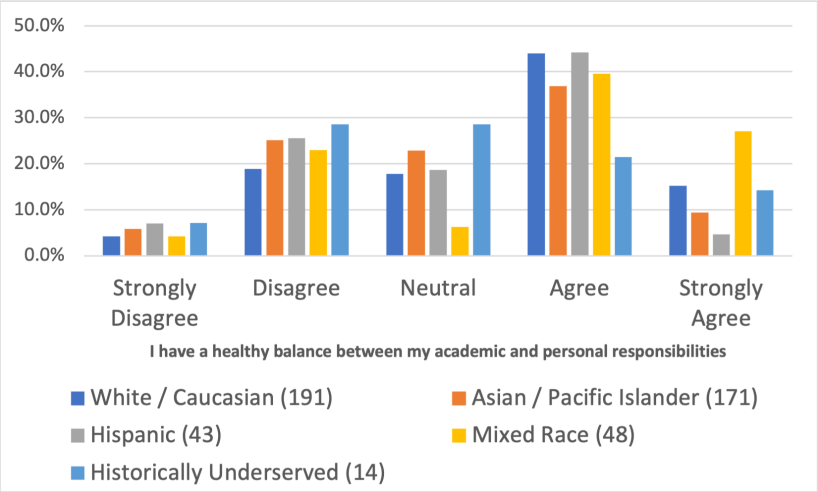


Figure 16: Student responses to "I have a healthy balance between my academic and personal responsibilities" split by race/ethnicity

The degree distribution of the network split by race/ethnicity can be seen in Figures 17, 18, 19, 20 and 21. The number of nodes, edges and clustering coefficients can be seen in Table 2. Both White students and students who are members of a historically underserved race/ethnicity have high clustering coefficients, indicating their network connectedness.

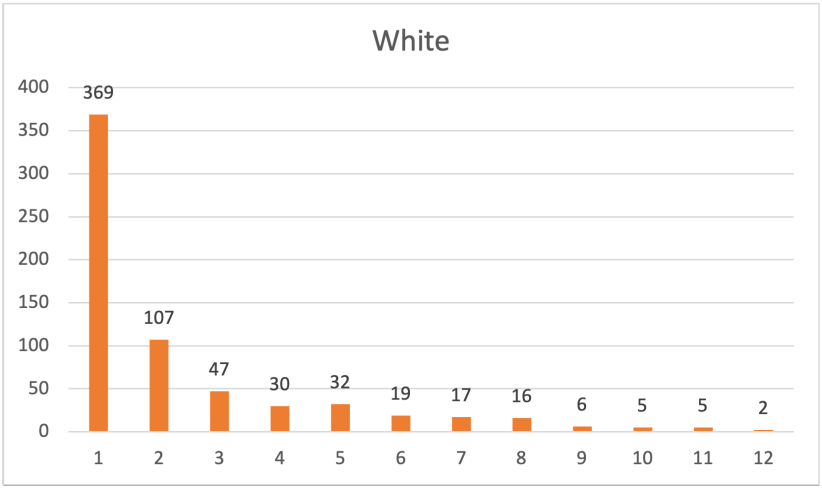


Figure 17: Degree Distribution for White students subgraph

Table 2: Graph characteristics on networks split by race/ethnicity

Race/Ethnicity	Number of Nodes	Number of Edges	Clustering Coefficient
White	661	818	0.063771162
Asian / Pacific Islander	666	805	0.044124318
Hispanic	189	177	0.026455026
Mixed Race/Ethnicity	263	264	0.037467286
Historically Underserved	67	67	0.064676617

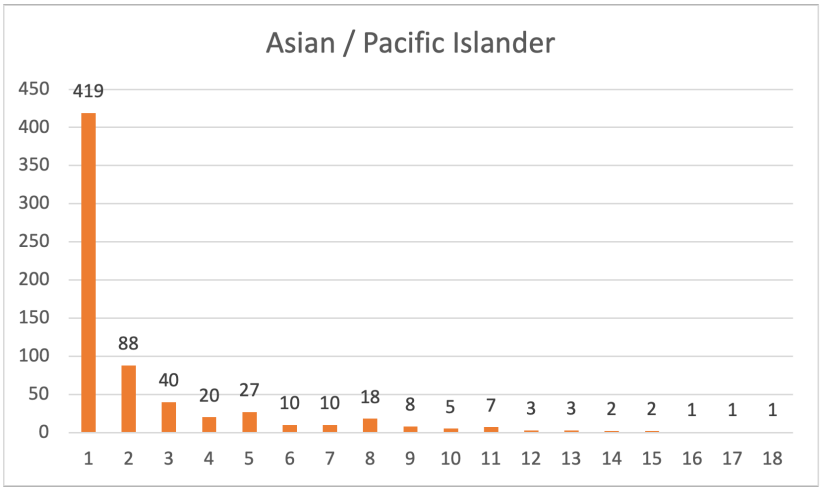


Figure 18: Degree Distribution for Asian/Pacific Islander students subgraph

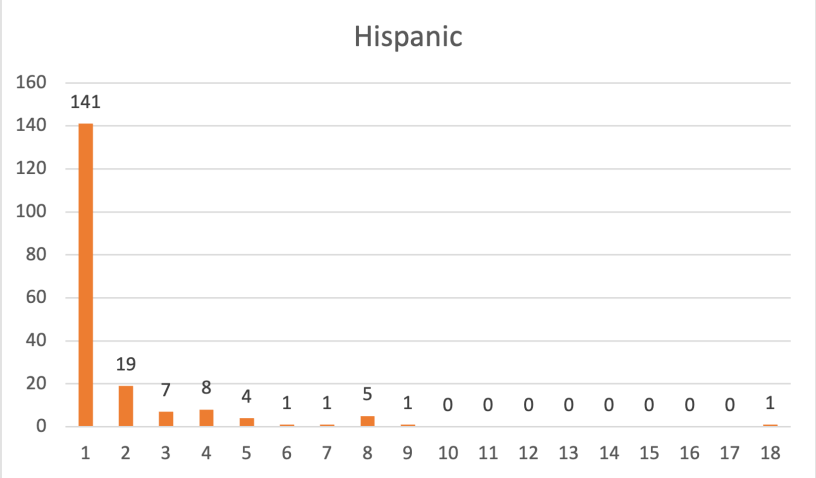


Figure 19: Degree Distribution for Hispanic students subgraph

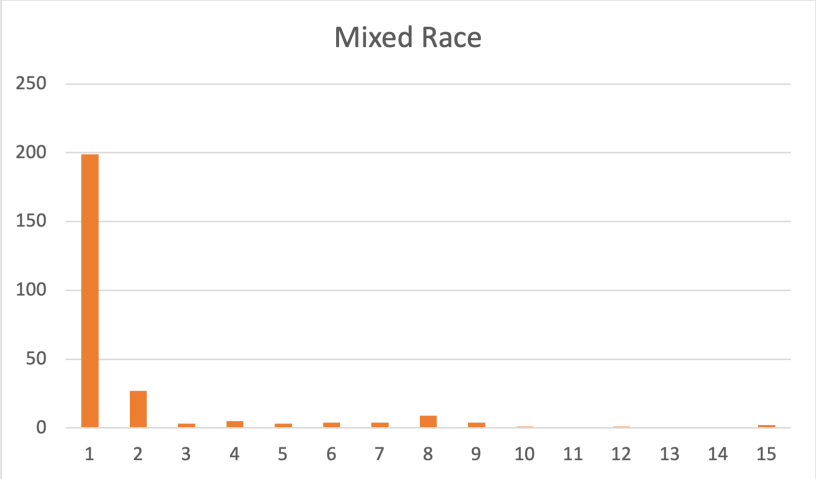


Figure 20: Degree Distribution for students of mixed race/ethnicity subgraph

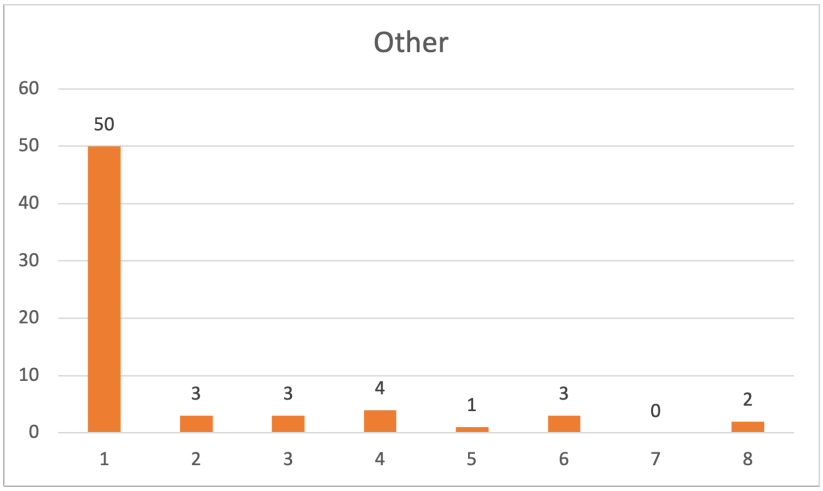


Figure 21: Degree Distribution for students of a historically underserved race subgraph

Note that we also analyzed the various subgraphs by year the student has been attending the university, but no notable differences were noticed.

3.7 Threats to Validity

While our results indicate future avenues of research and notable results about various demographic subgroups social networks, the potential threats to validity include selection bias. Specifically since our methodology used as gift card incentives, and not a global requirement for all students to fill out the survey, our 38.9% coverage of the department's student body could include selection bias.

4 Future Work

Since we have historical data from 2019 and 2020 about belongingness within the department, we can continue to administer the survey to students to create a longitudinal study of the sense of belonging within the department. Based on the results, we can implement interventions to help increase students sense of belongingness and continue to survey students to see the impacts and which intervention proved the best results.

This survey could easily be scaled to include bigger populations. Staying within the department, the survey could be given out to all professors to understand how professor-student relationships impact student success. Since many students listed professors as a part of their social network, it would be interesting to analyze if professors list the same, or see how big of a role students play in a professors network. A professor's network could give us insight into how their social network impacts their promotion history, job retention and overall well-being.

Another area of future work could be surveying other adjacent departments at the same university (such as the Mathematics, Electrical Engineering or Statistics) for their departments belongingness. We can use these results to find similarities and differences within the student body. We can discover what methods other departments are using to facilitate student belongingness, and if that is working, implement it within the department.

The survey only targeted current computing students. This population does not include students who were one of these majors and switched out of computing or the university. If the department has a goal of maintaining student success and belongingness, it is important to

include the population of students who are no longer in the program.

Other network analysis measures can be used to study the built social network. This includes studying keystone vertices to understand which types of students are essential in the network, betweenness centrality to understand how mutual friends can be important in the spread of belongingness, average path length to see how connected or disconnected certain groups of students are, among other graph theory measures. We also plan to study the network in relation to academic performance to understand which factors are most influential.

5 Acknowledgements

We would like to thank The Cal Poly Institutional Review Board for their input and advice regarding prioritizing and maintaining student privacy. We would also like to thank the the support of the CSSE department and the BEACoN research scholars program. We would also like to thank everyone who took time to take the survey and contribute their responses to this project.

References

[1] Degree programs. `https://csc.calpoly.edu/degree-programs/`, Last accessed on 2024-02-10.

[2] Joseph A. Anistranski and B. Bradford Brown. A little help from their friends? how social factors relate to students' sense of belonging at a large public university. *Journal of College Student Retention: Research, Theory & Practice*, 25(2):305–325, 2023.

[3] Joseph A. Anistranski and B. Bradford Brown. A little help from their friends? how social factors relate to students' sense of belonging at a large public university. *Journal of College Student Retention: Research, Theory & Practice*, 25(2):305–325, 2023.

[4] Pauline Rose Clance and Suzanne Ament Imes. The imposter phenomenon in high achieving women: Dynamics and therapeutic intervention. *APA*, 1978.

[5] Pauline Rose Clance and Suzanne Ament Imes. 58 percent of tech workers feel like impostors. 2018.

[6] Adam Cohen. fuzzywuzzy: Fuzzy string matching in python.

[7] James H Fowler and Nicholas A Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *BMJ*, 337, 2008.

[8] Teaching + Learning Lab. Students' sense of belonging matters: Evidence from three studies, 2023. `{https://tll.mit.edu/sense-of-belonging-matters/}`.

[9] ACM Digital Library. Student sense of belonging: The role of gender identity and minoritisation in computing and other sciences, 2023. `{https://dl.acm.org/doi/fullHtml/10.1145/3576123.3576133}`.

[10] Heather E. Metcalf, Tanya L. Crenshaw, Erin Wolf Chambers, and Cinda Heeren. Diversity across a decade: A case study on undergraduate computing culture at the university of illinois. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, SIGCSE '18, page 610–615, New York, NY, USA, 2018. Association for Computing Machinery.

[11] CITI Program. Social-behavioral-educational (sbe) refresher 1, 2023. `{https://about.citiprogram.org/course/human-subjects-research-social-behavioral-educational-sbe-refresher-1/}`.

[12] Katherine Rainey, Melissa Dancy, Roslyn Mickelson, Elizabeth Stearns, and Stephanie

Moller. Race and gender differences in how sense of belonging influences decisions to major in stem. *International Journal of STEM Education*, 5(1):10, Apr 2018.

[13] Adam Rosenstein, Aishma Raghu, and Leo Porter. Identifying the prevalence of the impostor phenomenon among computer science students. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, SIGCSE '20, page 30–36, New York, NY, USA, 2020. Association for Computing Machinery.

[14] Adam Rosenstein, Aishma Raghu, and Leo Porter. Identifying the prevalence of the impostor phenomenon among computer science students. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, SIGCSE '20, page 30–36, New York, NY, USA, 2020. Association for Computing Machinery.

[15] Valerie A. Sotardi. On institutional belongingness and academic performance: mediating effects of social self-efficacy and metacognitive strategies. *Studies in Higher Education*, 47(12):2444–2459, 2022.

[16] Kylan Stewart. An equity-minded assessment of belonging among computing students at cal poly. 2021.

[17] Wikipedia contributors. Degree distribution — Wikipedia, the free encyclopedia, 2024. [Online; accessed 28-March-2024].

A Survey Questions

Hosted on https://bit.ly/CSSESurvey using TypeForm Platform

Question 1: I agree to participate in this survey
(a) Yes (b) No

Question 2: What is your first and last name?

Question 3: What is your Cal Poly username?

Question 4: List the first and last name of the first person you are closest to in the CSSE or CPE department: friends, study buddies, classmates.

Question 5: On a scale from 1-5, how close do you feel to *<person listed in question 4>*? 1 - we've had a conversation, 5 - we are close friends
(a) 1 (b) 2 (c) 3 (d) 4 (e) 5

Question 6: Are the majority of the times you hang out with *<person listed in question 4>* to study / in class or for fun?
(a) Study or in class (b) For fun

Question 7: Would you like to list another person? (You can list up to 5)
(a) Yes (b) No

Question 8: List the first and last name of the second person you are closest to in the CSSE or CPE department: friends, study buddies, classmates.

Question 9: On a scale from 1-5, how close do you feel to *<person listed in question 8>*? 1 - we've had a conversation, 5 - we are close friends
(a) 1 (b) 2 (c) 3 (d) 4 (e) 5

Question 10: Are the majority of the times you hang out with *<person listed in question 8>* to study / in class or for fun?
(a) Study or in class (b) For fun

Question 11: To your knowledge, is *<person listed in question 8>* friends with *<person listed in question 4>*?
(a) Yes (b) No

Question 12: Would you like to list another person? (You can list up to 5)
(a) Yes (b) No

Question 13: List the first and last name of the third person you are closest to in the CSSE or CPE department: friends, study buddies, classmates.

Question 14: On a scale from 1-5, how close do you feel to *<person listed in question 13>*? 1 - we've had a conversation, 5 - we are close friends
(a) 1 (b) 2 (c) 3 (d) 4 (e) 5

Question 15: Are the majority of the times you hang out with *<person listed in question 13>* to study / in class or for fun?

(a) Study or in class (b) For fun

Question 16: To your knowledge, is *<person listed in question 13>* friends with *<person listed in question 4>*?
(a) Yes (b) No

Question 17: To your knowledge, is *<person listed in question 13>* friends with *<person listed in question 8>*?
(a) Yes (b) No

Question 18: Would you like to list another person? (You can list up to 5)
(a) Yes (b) No

Question 19: List the first and last name of the fourth person you are closest to in the CSSE or CPE department: friends, study buddies, classmates.

Question 20: On a scale from 1-5, how close do you feel to *<person listed in question 19>*? 1 - we've had a conversation, 5 - we are close friends
(a) 1 (b) 2 (c) 3 (d) 4 (e) 5

Question 21: Are the majority of the times you hang out with *<person listed in question 19>* to study / in class or for fun?
(a) Study or in class (b) For fun

Question 22: To your knowledge, is *<person listed in question 19>* friends with *<person listed in question 4>*?
(a) Yes (b) No

Question 23: To your knowledge, is *<person listed in question 19>* friends with *<person listed in question 8>*?
(a) Yes (b) No

Question 24: To your knowledge, is *<person listed in question 19>* friends with *<person listed in question 13>*?
(a) Yes (b) No

Question 25: Would you like to list another person? (You can list up to 5)
(a) Yes (b) No

Question 26: List the first and last name of the fifth person you are closest to in the CSSE or CPE department: friends, study buddies, classmates.

Question 27: On a scale from 1-5, how close do you feel to *<person listed in question 26>*? 1 - we've had a conversation, 5 - we are close friends
(a) 1 (b) 2 (c) 3 (d) 4 (e) 5

Question 28: Are the majority of the times you hang out with *<person listed in question 26>* to study / in class or for fun?
(a) Study or in class (b) For fun

Question 29: To your knowledge, is *<person listed in question 26>* friends with *<person listed in question 4>*?

(a) Yes (b) No

Question 30: To your knowledge, is *<person listed in question 26>* friends with *<person listed in question 8>*?

(a) Yes (b) No

Question 31: To your knowledge, is *<person listed in question 26>* friends with *<person listed in question 13>*?

(a) Yes (b) No

Question 32: To your knowledge, is *<person listed in question 26>* friends with *<person listed in question 19>*?

(a) Yes (b) No

Question 33: List any other friends or acquaintances in your majors department, separated by commas

Question 34: Out of these clubs, select the ones you've been a part of

(a) Women Involved in Software and Hardware (WISH)

(b) Hack4Impact Cal Poly

(c) Cal Poly Linux Users Group (CPLUG)

(d) Cal Poly Game Development Club (CPGD)

(e) Cal Poly App Dev Club

(f) CP Security Education Club (formerly White Hat)

(g) Cal Poly Computer Engineering Society

(h) Color Coded

(i) Cal Poly Robotics Club

(j) Cal Poly UX Club (CPUX)

(k) SLO Hacks

(l) Society of Women Engineers (SWE)

(m) STEM/Engineering Greek Life

(n) Computer Science and Artificial Intelligence (CS+AI)

Question 35: Would you say you get sick more often compared to others?

(a) Yes (b) No

Question 36: On average, how many hours of sleep do you get per night

(a) 4-5 hours (b) 6-7 hours (c) 8-9 hours (d) 10+ hours

Question 37: I have a healthy balance between my academic and personal responsibilities

Question 38: My academics cause me stress

(a)Highly Disagree (b)Disagree (c)Neutral

(d)Agree (e)Highly Agree

Question 39: I am more stressed than those around me

(a)Highly Disagree (b)Disagree (c)Neutral
(d)Agree (e)Highly Agree

Question 40: Balancing school and work is stressful

(a)Highly Disagree (b)Disagree (c)Neutral
(d)Agree (e)Highly Agree

Question 41: I am happy with where I'm currently at in life

(a)Highly Disagree (b)Disagree (c)Neutral
(d)Agree (e)Highly Agree

Question 42: I feel happy more often than I am sad

(a)Highly Disagree (b)Disagree (c)Neutral
(d)Agree (e)Highly Agree

Question 43: I consider myself a typical computer scientist

(a)Highly Disagree (b)Disagree (c)Neutral
(d)Agree (e)Highly Agree

Question 44: I can be myself and still fit in in the department

(a)Highly Disagree (b)Disagree (c)Neutral
(d)Agree (e)Highly Agree

Question 45: I frequently interact with instructors outside of class

(a)Highly Disagree (b)Disagree (c)Neutral
(d)Agree (e)Highly Agree

Question 46: How many years have you been attending Cal Poly?

(a) 1 (c) 3 (e) 5
(b) 2 (d) 4 (f) 6+

Question 47: Select your major

(a) Computer Science (c) Computer Engineering
(b) Software Engineering (d) Not Listed

Question 48: Are you a transfer student?

(a) Yes (b) No

Question 49: Have you changed your major at Cal Poly?

(a) Yes (b) No

Question 50: Have you had a CS/SE/CPE related internship before?
(a) Yes (b) No

Question 51: Do you work during the school year?
(a) Yes (b) No

Question 52: Are you an international student?
(a) Yes (b) No

Question 53: Are you in the 4+1 blended program?
(a) Yes (b) No

Question 54: Do you feel successful at Cal Poly?
(a) Yes (b) Somewhat (c) No

Question 55: What is your current cumulative GPA?
(a) 1.0 - 1.5 (c) 2.0 - 2.5 (e) 3.0 - 3.5
(b) 1.5 - 2.0 (d) 2.5 - 3.0 (f) 3.5 - 4.0

Question 56: What is your sexual orientation? Select all that apply, if unlisted please specify:
(a) Heterosexual (d) Pansexual (g) Queer
(b) Homosexual (e) Asexual (h) Not Listed
(c) Bisexual (f) Demisexual

Question 57: What is your gender identity? Select all that apply, if unlisted please specify:
(a) Male (c) Transgender (e) Gender-nonconforming
(b) Female (d) Non-binary (f) Not Listed

Question 58: Which race or ethnicity best describes you?
(a) American Indian or Alaskan Native
(b) Asian / Pacific Islander
(c) Black or African American
(d) Hispanic
(e) White / Caucasian
(f) Not Listed

B Consent Form

INFORMED CONSENT TO PARTICIPATE IN A RESEARCH PROJECT:
"Investigation of Social Networks upon Academic Performance and Mental Health"
INTRODUCTION
This form asks for your agreement to participate in a research project on building a social network in the Computer Science and Software Engineering (CSSE) and Computer Engineering (CPE) departments. Your participation involves filling out a survey. It is expected that your participation will take approximately 5 minutes. There are minimal risks anticipated with your participation. Others may benefit from your participation. If you are interested in participating,

please review the following information.
PURPOSE OF THE STUDY AND PROPOSED BENEFITS

- The purpose of the study is to create a recurring survey of the student body within the Computer Science and Software Engineering (CSSE) and Computer Engineering (CPE) departments that can be analyzed and used to make improvements towards the structure and care of both the education and well-being of students.
- Potential benefits associated with the study include helping the department gain insight on their student body and adapt to the needs of their students. Through the data gathered from the survey, the department will be able to adapt to encourage student success.

YOUR PARTICIPATION

- If you agree to participate, you will be asked to fill out an online survey.
- Your participation will take approximately 5 minutes.
- As an incentive, you will be offered donuts and cookies, and a chance to win a gift card if you opt in in the survey.
- At the end of the survey there is access to a raffle entry google form. There are ten $20 gift cards. The odds of winning are at least 1 in 1,000. Five winners will be drawn December 3rd, 2023 and five will be drawn January 19th, 2023. If you want to enter the raffle, but don't want to participate, you can select "No" to "I agree to participate in this survey" and fill out the Google Form linked.

PROTECTIONS AND POTENTIAL RISKS

- Please be aware that you are not required to participate in this research, refusal to participate will not involve any penalty or loss of benefits to which you are otherwise entitled, and you may discontinue your participation at any time. You may omit responses to certain questions you choose not to answer.
- There are minimal risks anticipated with your participation in this study. For example, disclosing your friends' names might expose your friends' social connections. This may place both you and your friends at risk of negative impacts to reputation.
- Your confidentiality will be protected by anonymizing the data collected by replacing names with unique IDs throughout the survey. Only aggregated results will be published and there will be no way back to figuring out who a participant is once the data is anonymized.
- Once we are ready to begin the analysis on the data we've collected, we plan to export the data into an excel spreadsheet and upload the spreadsheet to OneDrive. This file will then be shared with only the researchers of this project. Since data will be collected continuously, we will infrequently re-export the data and re-upload the excel file to OneDrive. Both the raw data and the anonymized data will be exported from TypeForm and uploaded into OneDrive and shared only with the researchers listed on this project.

RESOURCES AND CONTACT INFORMATION

- If you should experience any negative outcomes from this research, please be aware that you may contact Campus Health and Wellbeing at chw.calpoly.edu, for assistance.

- This research is being conducted by Rachel Izenson, Cameron Hardy, Lauren Allen, Tony Li, Ash Chen and Julia Ye (Researchers) and Theresa Migler (Advisor) in the Department of Computer Science and Software Engineering at Cal Poly, San Luis Obispo. If you have questions regarding this study or would like to be informed of the results when the study is completed, please contact the researcher(s) at rizenson@calpoly.edu and the student's faculty advisor at tmigler@calpoly.edu.
- If you have concerns regarding the manner in which the study is conducted, you may contact Dr. Michael Black, Chair of the Cal Poly Institutional Review Board, at (805) 756-2894, mblack@calpoly.edu, or Ms. Trish Brock, Director of Research Compliance, at (805) 756-1450, pbrock@calpoly.edu.

AGREEMENT TO PARTICIPATE
If you are 18 or older and agree to voluntarily participate in this research project as described, please indicate your agreement by checking the "I agree to participate in this survey" box in the survey used to collect data.

FRCCS

On the shape of illicit networks

Guy Melançon[1✓], Masarah Paquet-Clouston[2] and Martin Bouchard[3]

[1] *Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, France ; Guy.Melancon@u-bordeaux.fr*
[2] *School of Criminology/CICC, Université de Montréal, Canada ; m.paquet-clouston@umontreal.ca*
[3] *School of Criminology, Simon Fraser University, Canada ; mbouchard@sfu.ca*
[✓] *Presenting author*

Abstract. The inherent nature of illicit networks warrants their examination as a distinct category of networks. This analysis prompts us to examine the characteristics of the graphs used to model them. Authors have explored the structural properties of these networks, considering the sociological and criminological processes that shape them. For example, triadic closure is linked to trust-building, while preferential attachment indicates individuals with many connections are more likely to be apprehended. Based on a literature review, we suggest that most illicit networks embody a blend of three primary network models: small world, preferential attachment, and community-based. Furthermore, we argue that three specific random graph models effectively capture these diverse characteristics.

Keywords. *Illicit networks, Random models, Spectral distances*

1 Introduction

The inherent nature of illicit networks motivates their study as a distinct category of networks. Indeed, these networks show specific structural properties emerging from the sociological and criminological dynamics at play.

For instance, in [7] the authors claim that in covert settings, triadic closure (hence the small world effect) aligns with trust building, allowing individuals to protect themselves through the use of brokers to access the rest of the network. On the other hand, the degree distributions found in illicit networks often follows the preferential attachment principle [8, 17], with highly connected individuals being more at risks of being arrested [13]. At the mesolevel, large illicit networks have also displayed community structures comprised of well-defined and dense communities [5].

Hence, the literature supports the hypothesis that illicit networks consist of a mixture of three network archetypes: small world, preferential attachment, and community-based and we find that three random graph models genuinely capture these distinct characteristics: preferential attachment based on several sources [1, 14, 9], Wang's pseudo-fractal (small world) model [18] and a variation of Fortunato's relaxed caveman model [11].

We furthermore back this claim with a methodology that uncovers the inner-workings of a given illicit network by positioning it relative to instances of the three previously cited archetypes

using graph distances presented in [20]. A careful study of the structural difference between the archetypes (varying the parameters used in each model) allows us to get a clear picture of the parameter space they define. We moreover investigate how key statistics in illicit networks, such as local centralities (degree, clustering coefficient, betweenness centrality) and mesolevel measures (participation coefficient [12] and broker score [15]), behave when the structure of the illicit network studied changes or moves closer from one graph archetype to another. This helps to interpret these statistics and determine their conceptual relevance based on the graph's underlying structure.

2 Probing the shape of a criminal network

Taking [20] as the cornerstone of our work, and employing a methodology akin to that found in [6], we hypothesize that the structure of a graph is captured by its Laplacian or adjacency matrix, by leveraging various metrics for network comparison.

More precisely, given two graphs $G = (V, E), G' = (V', E')$, of size $N = |V| = |V'|$ one can define a distance $d_L(G, G') = \sqrt{\sum_{i=1}^{N}(\lambda_i - \lambda'_i)}$ where $\Lambda = (\lambda_1 \geq \lambda_2 \geq \ldots)$ (and similarly Λ') is the spectrum (ordered sequence of all eigenvalues) of the Laplacian matrix L of G (resp. L' of G').

We may also consider the *normalized Laplacian matrix* and define a distance $d_{L_{norm}}$, or opt for the adjacency matrix to define a distance d_A. In practice, it is often the case that only the first k eigenvalues are examined, where $k \ll N$, with the understanding that the adjacency spectral distance focuses on the *largest* k eigenvalues, while Laplacian spectral distances prioritize the *smallest* k eigenvalues. Note that spectral distances enable the comparison of graphs of different sizes, although it is most effective when the graphs are of similar dimensions.

To assess a graph $G = (V, E)$ that models a criminal network, we gauge its proximity to examples from three random graph models that encapsulate characteristics commonly observed in criminal networks. Specifically, we examine:

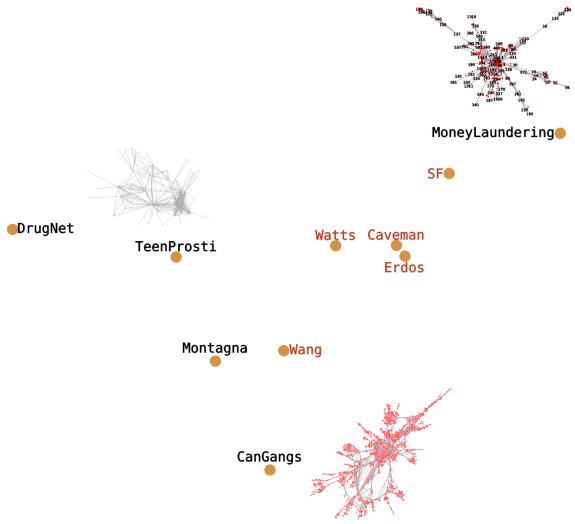


Figure 1: 2D mapping (MDS) of network models (red labels) and specific network instances (black labels) based on spectral distances. For sake of completeness, the map includes two additional models (Watts Strogatz [19], Erdos Renyi [10]).

- Scale-free graphs, which showcase preferential attachment, mimicking the degree distribution of G with an equivalent node count [1, 14, 9].
- Graphs of size $N = |V|$ created using Wang's pseudo-fractal model, generating graphs populated with triads [18].
- Community-structured graphs, mirroring the number of groups in G as identified by algorithms like those in [2, 14], generated via a variant of Fortunato's relaxed caveman

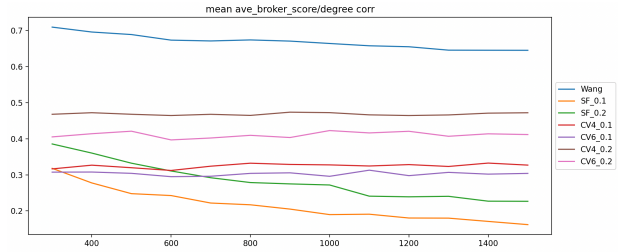
model [11].

Furthering our understanding of this conceptual space, we conducted simulations to generate numerous graphs from each family, measuring pairwise distances. This process enables a two-dimensional representation of the space, offering insights into the defining features of any given network.

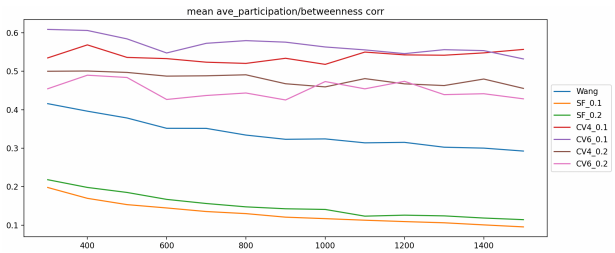
These three families of graphs delineate a conceptual landscape in which any criminal network can be positioned. Although this framework is also relevant to non-criminal networks, the attributes it emphasizes are particularly interesting for criminal networks. Figure 1 presents a mapping of various criminal networks in relation to the random models previously discussed, positioning the three models as benchmarks across a continuum that encapsulates the typical configurations found within criminal networks. Additionally, we incorporate two renowned models for comparison: the Watts-Strogatz model as an alternate representation of small-world networks, and the Erdős-Rényi model as an example of fully random graphs devoid of any specific structure. Notably, Wang's model excels, surpassing even the Watts & Strogatz model, in its ability to capture the quintessential small-world nature of these networks. One particular network, dedicated to Money Laundering, distinctly stands out, demonstrating a more pronounced scale-free structure.

3 Conclusion

Our study leads to a compelling although obvious conclusion: it is crucial to distinguish the network under study from the data describing it. This distinction is necessary because data often accentuates certain features of the network while neglecting others – sometimes induced from the investigation's imperatives, thereby challenging the accurate portrayal of the network's full complexity. For example, attempting to analyze the network's meso-level structure using data alone would be ineffective if field observations lead to building a scale-free graph. In such cases, a statistical indicator focused on meso-structure fails to provide insightful information in a network characterized by a scale-free pattern.



(a) Comparing the degree of nodes and their broker score [15].



(b) Comparing the participation coefficient of nodes [12] and their betweenness centrality.

Figure 2: Pearson's correlation coefficient between network statistics (according to network size) across different instances of random models help discern the distinct behaviors of the models and gain insights into the underlying mechanisms of specific criminal networks.

Our methodology serves as a diagnostic tool, evaluating the balance or the dominance of particular aspects within the data, illuminating the inherent characteristics of the network under analysis. By employing this approach, researchers and experts in criminal science can achieve a deeper, more nuanced understanding of the structure of illicit networks as revealed by their

data. Additionally, it aids in assessing the relevance and impact of the network statistics used in their analysis.

Further, we have conducted a detailed investigation of structural features to pinpoint the distinct behaviors of these models and enrich our comprehension of the forces that shape specific criminal networks, as depicted in Figure 2.

Building on this rationale, we plan to investigate the NetSimile distance [2], a metric that assesses networks based on their features for comparative analysis. By integrating both feature-based and spectral distances, along with comparisons of features at local, global, and meso-levels, we are in the process of developing a comprehensive framework. This framework aims to equip criminal science specialists with the ability to swiftly categorize the networks they examine, thus streamlining the development and verification of hypotheses or theoretical propositions.

References

[1] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[2] Michele Berlingerio, Danai Koutra, Tina Eliassi-Rad, and Christos Faloutsos. Network similarity via multiple social theories. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 1439–1440, 2013.

[3] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[4] Béla Bollobás, Christian Borgs, Jennifer T Chayes, and Oliver Riordan. Directed scale-free graphs. In *SODA*, volume 3, pages 132–139, 2003.

[5] Martin Bouchard. Collaboration and Boundaries in Organized Crime: A Network Perspective. *Crime and Justice*, 49:425–469, 2020.

[6] Ulrik Brandes, Jürgen Lerner, Uwe Nagel, and Bobo Nick. Structural trends in network ensembles. In *Complex Networks: Results of the 2009 International Workshop on Complex Networks (CompleNet 2009)*, pages 83–97, 2009.

[7] David Bright, Johan Koskinen, and Aili Malm. Illicit Network Dynamics: The Formation and Evolution of a Drug Trafficking Network. *Journal of Quantitative Criminology*, 35(2):237–258, 2019.

[8] Lucia Cavallaro, Annamaria Ficara, Francesco Curreri, Giacomo Fiumara, Pasquale De Meo, Ovidiu Bagdasar, and Antonio Liotta. Graph comparison and artificial models for simulating real criminal networks. In *Complex Networks & Their Applications IX: Volume 2, Proceedings of the Ninth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2020*, pages 286–297, 2021.

[9] Aaron Clauset, Cosma Rohilla Shalizi, and Mark E.J. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[10] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[11] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[12] Roger Guimera and Luís A. Nunes Amaral. Cartography of complex networks: modules and universal roles. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(02):P02001, 2005.

[13] Carlo Morselli. Assessing vulnerable and strategic positions in a criminal network. *Journal*

of Contemporary Criminal Justice, 26(4):382–392, 2010.

[14] Mark E.J. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[15] Masarah Paquet-Clouston and Martin Bouchard. A Robust Measure to Uncover Community Brokerage in Illicit Networks. *Journal of Quantitative Criminology*, 39(3):705–733, 2023.

[16] V. A. Traag. Faster unfolding of communities: Speeding up the louvain algorithm. *Phys. Rev. E*, 92:032801, Sep 2015.

[17] Federico Varese. The structure and the content of criminal connections: The russian mafia in italy. *European sociological review*, 29(5):899–909, 2013.

[18] L. Wang, F. Du, H. P. Dai, and Y. X. Sun. Random pseudofractal scale-free networks with small-world effect. *The European Physical Journal B*, 53(3):361–366, 2006.

[19] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[20] Peter Wills and François G. Meyer. Metrics for graph comparison: A practitioner's guide. *PLOS ONE*, 15(2):e0228728, 2020.

Workshop on Complex Network Sparsification

FRCCS

Generic Network Sparsification via Hybrid Edge Sampling

Zhen Su[1✓], Jürgen Kurths[2] and Henning Meyerhenke[1]

[1] *Humboldt-Universität zu Berlin, Berlin, Germany*
[✓] *Presenting author*

Abstract. Network (or graph) sparsification is a non-trivial task due to the need to preserve various (or at least representative) network properties. In this work, we propose a general hybrid edge sampling scheme, that is a combination of the **Lo**cal-filtering-based Random Edge sampling (LRE) [Hamann et al., SNAM 2016] and the **Ga**me-theoretic Sparsification with Tolerance (GST) framework [Su et al., ASONAM 2022]; we call the new method LOGA for brevity. LOGA fully utilizes the advantage of GST in preserving complex structural properties (especially the degree distribution) by preserving local node property, as well as the strength of LRE in preserving the connectivity of a given network. Technically, this is achieved by regarding GST as a technique to refine LRE – leading to LOGA, and by further including the preservation of the largest connected component and the weighted average clustering coefficient – leading to an algorithmic variant LOGA^{sc}. In this way, LOGA / LOGA^{sc} generalize the work on GST to graphs with weights and different densities, without increasing the asymptotic time complexity. Extensive experiments on 26 weighted and unweighted networks with different densities demonstrate that LOGA^{sc} performs best for all 26 instances, i.e., they preserve representative network properties better than using state-of-the-art sampling methods alone.

Keywords. *Graph sparsification; Edge sampling; Hybrid sampling*

1 Introduction

It is well known that one can speed up graph analyses by sparsification. Sparsification removes a large proportion of possibly redundant edges in a given network or graph $\mathcal{G} = (V, E, W)$ without the aggregation of nodes; $\mathcal{G}$ is therefore compressed into a sparser graph $\hat{G}$ by sparsification.

Sparsification requires the preservation of structural properties of $\mathcal{G}$ in $\hat{G}$ in a scaled manner. Preserving various structural properties of $\mathcal{G}$ is still a non-trivial problem. A pragmatic solution is to preserve *representative* ones (see Fig. 1). By doing so, we can expect other properties to be preserved to some extent, due to the correlations between different properties [3, 9, 10, 11].

When doing sparsification, time consumption is an important aspect to consider. For this, edge sampling without access to the entire network is often preferred [8, 7, 4, 6, 12, 5, 11]. Still, preserving a set of representative properties by edge sampling is also non-trivial, because it is hard to define an appropriate sampling objective characterizing well the selected representative properties. A practical option is to combine different edge sampling methods.

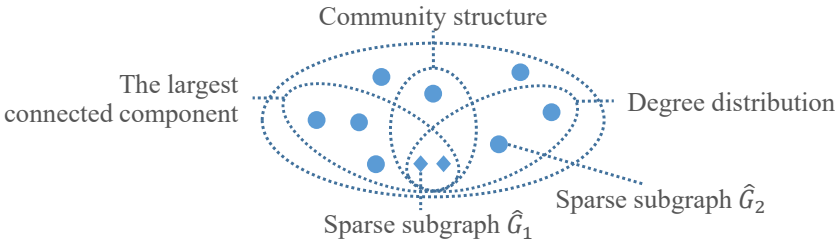


Figure 1: A schematic view of sparse subgraphs $\hat{G}$ preserving structural properties of a given graph $\mathcal{G}$. Shapes are sparse subgraphs of the same network density. Assuming the largest connected component, community structure, and degree distribution are selected *representative* properties, two sparse subgraphs in diamond are ideal choices for preserving these properties.

Therefore, we propose a hybrid sampling scheme LOGA, as a combination of two edge sampling methods: LRE [4] and GST [11]. Technically, we first improve GST by *initialization optimization*; that is, we provide GST with a good initialized sparse subgraph G^* using LRE, leading to a hybrid edge sampling LOGA. Then, we include *constrained update* in LOGA by the preservation of the largest connected component and the weighted average clustering coefficient based on G^*, leading to an algorithmic variant LOGA^{sc}. Thus, the contributions are as follows:

- We propose a hybrid edge sampling scheme LOGA and an algorithmic variant LOGA^{sc} for graph sparsification.
- LOGA^{sc} maintains representative properties better than the state of the art for functional climate, real-world, and synthetic networks (on average).
- We recommend using LOGA$^{sc}_{2,3,w}$ in practice, where subscripts '2', '3', and 'w' represent that LOGA^{sc} preserves the expected degrees of nodes, the expected number of triangles (i.e., closed wedges), and the expected number of non-closed wedges associated nodes.

Table 1: Summary of the performance of sampling methods, i.e., LOGA vs {GST, LD, LJS, RE, LRE, and CN}, out of 26 unweighted and weighted networks with different densities. The shooting score counts the number of data sets for which different methods perform better.

Method	GST$_{2,3}$	LD	LJS	RE	LRE	CN
Score	8	14	0	10	**25**	0
Method	GST$_{2,3,w}$	LD	LJS	RE	LRE	CN
Score	10	13	0	11	**23**	0
Method	LOGA$_{2,3}$	LD	LJS	RE	LRE	CN
Score	**21**	6	0	6	20	0
Method	LOGA$_{2,3,w}$	LD	LJS	RE	LRE	CN
Score	**22**	6	0	5	20	0
Method	LOGA$^{sc}_{2,3}$	LD	LJS	RE	LRE	CN
Score	**23**	4	0	5	18	0
Method	LOGA$^{sc}_{2,3,w}$	LD	LJS	RE	LRE	CN
Score	**26**	4	0	5	19	0

2 Hybrid Edge Sampling Scheme

Initialization optimization. The idea of optimizing initialization for GST stems from the existence of multiple equilibria in GST, which is not indicated in Ref. [11]. Due to this, it is natural to ask how to steer an algorithm to find a good optimum. One established strategy is to use a good starting solution, that in our context, already preserves representative properties

reasonably well. Therefore, we propose LOGA first by using LRE as an initializer optimizing GST due to the best performance of LRE in our preliminary study.

Constrained update. For GST, initialization optimization is not sufficient to ensure that, the representative properties preserved by the initialized sparse subgraph G^* can still be preserved after sparsification. For this, we preserve the largest connected component and the weighted average clustering coefficient during sparsification. These two structural properties are chosen, because they can be used for characterizing network resilience [2, 1].

The experimental evaluation is given in Table 1. The shooting score computes the number of data sets for which each method has relatively better performance. For example, $\text{LOGA}^{sc}_{2,3,w}$ achieves a shooting score of 26, indicating that it performs relatively better in all 26 networks.

References

[1] Oriol Artime, Marco Grassia, Manlio De Domenico, James P. Gleeson, Hernán A. Makse, Giuseppe Mangioni, Matjaž Perc, and Filippo Radicchi. Robustness and resilience of complex networks. *Nat Rev Phys*, 6(2):114–131, 2024.

[2] J. Ash and D. Newth. Optimizing complex networks for resilience against cascading failure. *Physica A: Statistical Mechanics and its Applications*, 380:673–683, 2007.

[3] Michele Benzi and Christine Klymko. On the Limiting Behavior of Parameter-Dependent Network Centrality Measures. *SIAM Journal on Matrix Analysis and Applications*, 36(2):686–706, 2015.

[4] Michael Hamann, Gerd Lindner, Henning Meyerhenke, Christian L. Staudt, and Dorothea Wagner. Structure-preserving sparsification methods for social networks. *Social Network Analysis and Mining*, 6(1):22, 2016.

[5] Can M. Le. Edge Sampling Using Local Network Information. *Journal of Machine Learning Research*, 22(88):1–29, 2021.

[6] Jianguo Lu and Hao Wang. Uniform random sampling not recommended for large graph size estimation. *Information Sciences*, 421:136–153, 2017.

[7] Veeru Sadhanala, Yu-Xiang Wang, and Ryan Tibshirani. Graph Sparsification Approaches for Laplacian Smoothing. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1250–1259. PMLR, 2016.

[8] Venu Satuluri, Srinivasan Parthasarathy, and Yiye Ruan. Local graph sparsification for scalable clustering. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 721–732, New York, NY, USA, 2011. Association for Computing Machinery.

[9] David Schoch, Thomas W. Valente, and Ulrik Brandes. Correlations among centrality indices and a class of uniquely ranked graphs. *Social Networks*, 50:46–54, 2017.

[10] Zhen Su, Chao Gao, Jiming Liu, Tao Jia, Zhen Wang, and Jürgen Kurths. Emergence of nonlinear crossover under epidemic dynamics in heterogeneous networks. *Phys. Rev. E*, 102(5):052311, 2020.

[11] Zhen Su, Jürgen Kurths, and Henning Meyerhenke. Network Sparsification via Degree- and Subgraph-based Edge Sampling. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9–16, 2022.

[12] Jianpeng Zhang, Kaijie Zhu, Yulong Pei, George Fletcher, and Mykola Pechenizkiy. Cluster-preserving sampling from fully-dynamic streaming graphs. *Information Sciences*, 482:279–300, 2019.

FRCCS

Hybrid method for graph reduction

Clément Aralou[1✓], Tobias Rupp[2], Samba Ndojh Ndiaye[1], Mohammed Haddad[1] and Hamida Seba[1]

[1] *UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France ; clement.aralou@univ-lyon1.fr, hamida.seba@univ-lyon1.fr, samba-ndojh.Ndiaye@univ-lyon1.fr, mohammed.haddad@univ-lyon1.fr.*
[2] *University of Stuttgart ; ruppts@fmi.uni-stuttgart.de*
✓ *Presenting author*

Abstract. This article is a summary of a paper accepted for presentation at the conference ECG2024 [Ara+24]. We focus on graph reduction using machine learning techniques to preserve distance in the graph. Either by grouping similar nodes or edges or sparsifying, removing less important edges for the downstream task. We propose a method that combines both approaches, first a sparsification step followed by an aggregation step.

Keywords. *Reinforcement Learning; Graphs; Reduction*

1 Introduction

Graphs or networks are used to model interaction between different types of entities. In many domains, complex data can be represented by graphs. Popular examples are social networks or protein-protein interactions. Most of these graphs are so large and complex that their study can be very challenging. To resolve this problem, one solution is to simplify the input graph by reducing its size. The objective is to get a smaller graph while keeping some of its properties. These properties depend on the downstream task at hand. There are two main types of methods for simplifying a graph, the first one is graph coarsening where the objective is to reduce the size of the graph by grouping similar nodes. For example, one can group nodes with the same neighborhood or nodes with the same features, etc. This creates a smaller graph where each node can be represented as an aggregation of several nodes. There are also sparsification methods that remove edges that are less important for the downstream task.

We will consider in this paper methods of reduction of graphs which preserve distance. In other words, two nodes at a distance d in the initial graph, are at the same distance with a multiplicative or additive factor in the simplified graph. The objective of preserving the distance is for example to be able to approximate path queries in the graph. The proposed algorithm will combine both sparsification and aggregation in order to build the smallest graph possible.

2 Preliminaries

Graph and path: A graph $G = (V, E)$ is composed of V a set of nodes (or vertices), and E a set of edges (or links), such that $E \subseteq V \times V$. The vertices u and v are neighbours if the edge $\{u, v\} \in E$. A leaf is a node linked to only another node, called its parent. A path is a sequence of nodes such that two subsequent nodes in the sequence are neighbours. The length of the path is its number of edges. The distance between two vertices u and v, denoted by $dist(u, v)$, is the length of the shortest path between the nodes. If no such path exists, the distance is considered to be $dist(u, v) = \infty$. A graph is connected if there exists a path between each pair of nodes. A cycle is a path where the first and last vertices are equal. A tree is a connected graph without any cycle. Let $G = (V, E)$ be a graph, $T = (V', E')$ is a k-dominating subgraph of G if T is a subgraph of G (i.e., $V' \subset V$ and $E' \subset E$) and $\forall v \in V, \exists u \in V^{'}$ such that $dist(u, v) \leq k$ in G.

Reinforcement learning: The principal objective of Reinforcement learning (RL) is to maximize a long-term reward [SB18]. In RL, the learner (the one who makes decisions) is called the agent. To maximize the reward, first it must explore the space of actions by making random decisions to discover which action gives the highest reward. Once he has explored enough, he will exploit his discovery by taking the best action. The problem will be represented as a Partially observable Markov decision process (POMDP) where only partial information about the current state is available to the agent. Solving POMDP problems is undecidable in general, but if we restrict ourselves to maximizing the cumulative reward, it can be solved using an approximation method. In our case, we will use Deep Learning because it has shown positive results in a similar context [WZL21].

3 Graph reduction using Reinforcement Learning

In this part, we will present our new distance-preserving graph reduction method inspired by SparRL [WZL21] and called HyRed. Let $G = (V, E)$ be a graph and $T = (V, E')$ the compressed graph of G such that $|E'| \subseteq |E|$. The distances are preserved if it exists a $k \in \mathbb{R}$ such that for every pair of nodes u and v in G, $d_T(u, v) \leq k \cdot d_G(u, v)$. This means that we can retrieve the distance in G from the distance in T up to the multiplicative factor k. Formally the distortion is:

$$\text{Distortion} = \frac{1}{\binom{|V|}{2}} \sum_{u,v \in V} \frac{dist_T(u, v)}{dist_G(u, v)}. \tag{1}$$

The sparsification algorithm: The goal is to choose a set of edges to remove from the graph while minimizing the distortion. The first step is to learn to select an edge to remove, taking into account a set of observations about the current subgraph. At the beginning of each training episode, we build a partial environment by removing random edges from the initial graph to obtain a partial environment. Then, we compute a minimum spanning tree (MST). Edges of this tree will not be removed ensuring the subgraph will remain connected. Then, we select a subgraph H_t induced by EH edges which do not belong to the MST, in order to reduce the computational cost. The edge to remove is selected within H_t using a set of observations d_{H_t} (degrees of nodes), η_t (ratio of degrees) and $\mathcal{N}_t$ (One-hop neighbourhood) which give a belief of the current state. From the past experiences, the agent will learn by trying to minimize the error between the true reward and the expected reward.

The k-domination algorithm: Intuitively, the algorithm recursively removes leaves until they are at a distance k from an already removed node. To achieve this, it constructs an associative array

which, for each parent node, maintains the distance at which the node was removed. The nodes of the resulting k-dominating tree are super-nodes that aggregate all the nodes attached to them and are not part of the k-dominating tree.

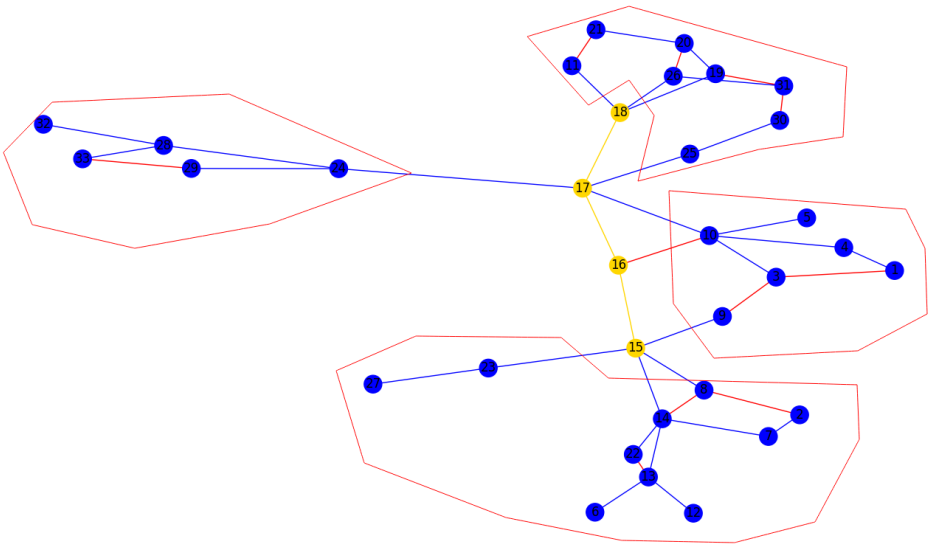


Figure 1: An example of HyRed applied to a graph which involves the following steps. Initially, red edges are removed because they are considered less important. From the resulting spanning tree, we then apply the 3-dominating algorithm. The resulting nodes are highlighted in gold, and the aggregate nodes are enclosed within the red clusters.

4 Experiments

A description of the datasets and algorithms used is given in the original paper [Ara+24]. The results are shown in Tables 1, 2 and 3.

Algorithm	50 edges †	150 edges †	225 edges	300 edges
BFS	1.930	1.919	**1.847**	1.997
Petal	1.883	1.910	1.949	1.940
Controlled Random	2.358	2.748	2.711	2.757
CKS	2.457	1.936	1.860	**1.929**
HyRed	**1.168**	**1.768**	1.944	2.332

Table 1: Average distortions over 10 executions for synthetic graphs using Watts-Stroggatz model. † indicates $|\text{EH}| = |E|$. We consider graphs with 20, 50, 75, and 100 nodes.

algorithm	184 edges	184 edges †	284 edges	284 edges †	384 edges	384 edges †
BFS	1.864	1.820	1.791	1.847	1.820	1.732
Petal	1.821	1.792	1.929	1.859	1.880	1.868
Controlled Random	2.735	2.569	2.668	2.591	2.811	2.708
CKS	**1.696**	1.696	**1.667**	**1.667**	**1.696**	1.696
HyRed	1.699	**1.637**	2.183	1.896	2.369	**1.603**

Table 2: Average distortions over 10 executions for synthetic graphs using **Barabási–Albert** model. † indicates $|\text{EH}| = |E|$. We consider graphs with 20, 50, 75, and 100 nodes.

Algorithm	51 edges	84 edges	120 edges	133 edges
BFS	1.597	1.382	1.660	1.703
Petal	1.443	1.297	1.596	1.643
Controlled Random	1.631	1.616	1.875	1.829
CKS	1.567	1.377	1.523	1.341
HyRed	**1.013**	**1.202**	**1.405**	**1.222**

Table 3: Average distortions over 10 executions for graphs of dataset **ENZYMES**. We consider graphs with approximately 100 nodes.

5 Conclusion and Discussion

HyRED is a hybrid method of graph reduction combining a step of sparsification followed by a step of aggregation. It obtains overall the best results on the considered graphs, and more specifically on **Enzymes** dataset. We notice that it becomes very challenging for the agent to act optimally when the graph size is large. In addition, even if the variance is high, we observe that there is often one run of HyRed that outperforms the results of all other methods. In the end, this work opened several perspectives: (1) The current task is non-markovian as the edge removed at timestep t will influence the future action and outcome. Thus it would be interesting to integrate the history of actions and observations in the embedding for better decisions at the current step using the previously discussed method. (2) To limit the training time, approximating the distortion could be interesting at the cost of a slight deterioration in performance.

Acknowledgement: For the research leading to these results, the authors received funding from Agence National de la Recherche (ANR) under Grant Agreement No ANR-20-CE23-0002.

References

[Ara+24] Clement Aralou et al. “Approche hybride basée sur l’apprentissage automatique pour la réduction de graphes”. In: *24ème conférence francophone sur l’Extraction et la Gestion des Connaissances EGC 2024*. Vol. RNTI-E-40. EGC 2024. Dijon, France, Jan. 2024, pp. 167–178. URL: https://hal.science/hal-04454820.

[SB18] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018. ISBN: 0262039249.

[WZL21] Ryan Wickman, Xiaofei Zhang, and Weizi Li. “SparRL: Graph Sparsification via Deep Reinforcement Learning”. In: *CoRR* abs/2112.01565 (2021). arXiv: 2112.01565. URL: https://arxiv.org/abs/2112.01565.

Workshop on Modeling Cities and Regions as Complex and Evolving Systems

FRCCS

Complex behaviour in day-to-day dynamics of Transportation systems

Jean-Patrick Lebacque[1✓] and Megan M. Khoshyaran[2]

[1] *UGE (University Gustave Eiffel) COSYS GRETTIA; jean-patrick.lebacque@univ-eiffel.fr*
[2] *Economics Traffic Clinic (ETC) Paris. France; etclinic@wanadoo.fr*
[✓] *Presenting author*

Abstract. The dynamics of large scale transportation systems result from: i) traffic flow in the system, a strongly nonlinear process, ii) traffic assignment, i.e. the route/departure time choice of travellers. Route and departure time choice occur on different time scales, real time and day-to-day. Travellers make their choices based on their assessment of their travel costs. Traffic assignment impacts the congestion patterns, thus retroacts with the travellers' choices. Available information and learning play a crucial role in this process. The aim of the paper is to analyse the day-to-day dynamics of the transportation system on a medium time scale. The questions of interest are: existence and unicity of equilibria, periodic orbits, possibly complex dynamics.

Keywords. *Nonlinear dynamical systems; infinite dimensional system; complex system; fixed point; equilibrium; dynamic traffic assignment; traffic flow model; GSOM model*

1 Introduction, setting the problem, aims.

In this contribution we consider the transportation system at a regional level, and focus on vehicular traffic, which contributes to the bulk of passenger transportation in most regions. The question addressed is: what is the evolution of the transportation system over a time range of a month to a year, and specifically does the system reach an equilibrium, and if yes, how is the equilibrium reached, is it unique, is it stable, do other dynamics eventually occur? The dynamics of networks are the result of the process of Dynamic Traffic Assignment (DTA), that is to say the choice by travellers of their route and their departure time. DTA modelling constitutes an essential tool for analysis, planning and management of the transportation system at the regional level. The reader is referred to [5] and [24]

Following Wardrop ([25] the general behavioral assumption for DTA is that travellers make their choices by minimizing their travel cost in order to achieve the object of their trip. In the case of DTA the travel cost includes mainly travel time and penalty for late/early arrival with respect to the desired arrival time, and possible financial costs (tolls). Travellers' choices have an obvious impact on the supply side of the transportation system. If many travellers chose a route because of its attraction (low travel cost) then this route becomes congested and its attraction diminishes. A similar process applies to departure time choice. Hence the fundamental question: does the system reach an equilibrium, and if the answer is positive, is

this equilibrium unique, is it stable, what are the day-to-day dynamics. All this questions have a direct impact on managing the system and on its planification.

In the case of static single mode traffic and monotonous diagonal travel costs (i.e. costs increasing with demand) the problem is well understood and the path flows can be obtained as the solution of a fixed point (or variational inequality) problem, which guarantees the existence of an equilibrium. The reader is referred for instance to [21, 22, 23]. Nevetheless it can be shown in some simple configurations that chaotic behaviour is liable to occur, depending on the assumptions made on the learning behaviour of travellers [6, 16, 4].

Analysis of the dynamic case is much more difficult. One reason is that models of very large transportation networks need to be both very fast and precise. Another reason lies with the innate complexity of the problem, the setting of which is a graph (the network) times a time interval (containing the possible departure times). Thus the natural functional setting is an infinite dimensional space. Many models have been considered for DTA: point queue models [12], simple and efficient, cellular automata (with a large-scale application in the NordRhein-Westphalen region [18]), hydrodynamic models [17], MFD (macroscopic fundamental diagram) based models [1], 2D (bidimensional) models [20, 10] which are emerging for addressing very large scale DTA problems, microsimulation/multiagent models [8, 2, 3] which constitute the core of simulation-based applications (MATSIM [7], commercial softwares such as AIMSUN, PTV, CALLIPER).

The present contribution uses a model of the GSOM family [15, 13, 14]. These models rely on the hydrodynamic paradigm for traffic modelling: they approximate the flow of traffic as a flow of liquid in a network and are well-suited for traffic on large networks. The approach outlined in the contribution also takes into account the evolution of technology which impacts directly DTA. Indeed, crowd sourcing, internet services and V2V (vehicle to vehicle) or V2I (vehicle to infrastructure) communicationprovide an increasing fraction of travellers with real-time information on the totality of the network. We will use the term ITT (instantaneous travel time) to designate the result of this real-time information. One important aspect of our approach is that travellers in DTA make choices at two different levels and time-scales. They base their route choice in real time on ITT, but base their departure time choice on a day-to-day basis on PTT (predictive travel times). Predictive travel times result from past experienced travel times (ETT) by a learning process. The learning process can also affect the route choice.

Three factors add to the complexity of the system. i) ITT is not a good motive for route choice, as traffic conditions change while the traveller moves along his route, which leads to suboptimal route choice and network dynamics [9, 11]. ii) The two different time-scales necessarily interact, especially since learning (a day-to-day process) can also impact route choice. iii) The learning behaviour of travellers is likely to retroact adversely on the dynamics of the system and may induce chaotic behaviour as shown in the static/quasi-static case [16, 4]. In [19] suggestion of possible chaotic behaviour has been shown with a simplified point queue model.

The contribution aims to extend these ideas with the more precise GSOM model, by considering various learning strategies in the iterative process of day-to-day dynamic assignment. We will analyze in this context the questions of multiplicity of equilibria, of the convergence and stability of the equilibrium reaching process, as well as possible bifurcation phenomena occurring in this process.

2 Outline of the traffic assignment process

The main data of the problem is $D_w^{t_a}$: the travel demand (number of trips per unit of time) of OD (origin-destination) couple $w \in \mathcal{W}$ of travellers with desired arrival time $t_a \in \mathcal{T}_a$. The main unknows are:

i) the distribution of departure times $\varphi_w^{t_a}(t)dt \quad \forall w \in \mathcal{W}, t_a \in \mathcal{T}_a$ with respect to departure time $t \in \mathcal{T}_d$. Thus

$$(\mathcal{K}) \left| \begin{array}{ll} \int_{t_d \in \mathcal{T}_d} \varphi_w^{t_a}(t_d) dt_d = 1 & \forall w \in \mathcal{W}, \forall t_a \in \mathcal{T}_a \\ \varphi_w^{t_a} \geq 0 & \forall w \in \mathcal{W}, \forall t_a \in \mathcal{T}_a \end{array} \right. \tag{1}$$

ii) the fraction $\varpi_p^{t_a}(t)$ of travellers departing at time $t \in \mathcal{T}_d$, with desired arrival time t_a, using path $p \in \mathcal{P}_w$ to complete their w OD trip. $\mathcal{P}_w$ denotes the set of plausible sets joining the OD $w \in \mathcal{W}$. Thus the fraction $\varpi_p^{t_a}$ are positive and satisfy:

$$\sum_{q \in \mathcal{W}} \varpi_q^{t_a}(t) = 1 \quad \forall t_a \in \mathcal{T}_a, t \in \mathcal{T}_d$$

Let us note by $f_p^{t_a}(t)$ the flow of travellers departing at time $t \in \mathcal{T}_d$, with desired arrival time $t_a \in \mathcal{T}_a$, using path $p \in \mathcal{P}_w$ to complete their trip joining the OD couple $w \in \mathcal{W}$. By the above definitions:

$$f_p^{t_a}(t) = \varphi_w^{t_a}(t) \varpi_p^{t_a}(t) D_w^{t_a} \quad \forall w \in \mathcal{W}, p \in \mathcal{P}_w, t_a \in \mathcal{T}_a, t \in \mathcal{T}_d$$

The $f_p^{t_a}(t)$ constitute the input of the GSOM traffic flow model. The output of this model includes: the instantaneous and experienced travel times and costs, and the predictive travel costs, after resolution of the Wardrop optimality conditions.

The late/early arrival time penalty takes the form $L(t_a - TA)$ where TA denotes the arrival time and L denotes a convex function which admits a minimum at $L(0) = 0$. Thus if the instantaneous travel time of path p at time t is $ITT_p(t)$ then the corresponding instantaneous travel cost $ITC_p^{t_a}(t)$ is obtained by $ITC_p^{t_a}(t) = ITT_p(t) + L\left(t_a - t - ITT_p(t)\right)$. Note that the instantaneous travel times are additive and express $ITT_p(t) \approx \int_p d\xi / V(\xi, t)$ where V denotes the velocity (an output of the GSOM model). The Wardrop principle applied to route choice, i.e. to the calculation of the fractions $\varpi_p^{t_a}(t)$ can be expressed as

$$\varpi_p^{t_a}(t).\left(ITC_p^{t_a}(t) - \min_{q \in \mathcal{P}_w} ITC_q^{t_a}(t)\right) = 0 \quad \forall p \in \mathcal{P}_w, t_a \in \mathcal{T}_a \tag{2}$$

which must be solved at any departure time $t \in \mathcal{T}_d$ and for all OD couples $w \in \mathcal{W}$.

The GSOM model also yields experienced travel times $ETT_p(t)$ for all paths $p \in \mathcal{P}_w, w \in \mathcal{W}$ and arrival times t. Note that the experienced travel times are not additive but satisfy a semi-group property. They are estimated on each path $p \in \mathcal{P}_w$ by keeping track of the departure time of travellers. Departure time is a traveller attribute which is advected by the traffic flow and thus is easily calculated in the GSOM model. The first step consists in inverting the experienced travel times in order to obtain the predictive travel times $PTT_p(t)$ with t being the departure time: $PTT_p(t) = ETT_p(t + PTT_p(t))$. Then the predictive path travel cost is obtained by $PTC_p^{t_a}(t) = PTT_p(t) + L\left(t_a - t - PTT_p(t)\right)$. Finally we obtain the predictive OD travel costs as the expectation of the predictive path travel costs:

$$PTC_w^{t_a}(t) = \sum_{q \in \mathcal{P}_w} \varpi_q^{t_a}(t).PTC_q^{t_a}(t) \quad \forall t_a \in \mathcal{T}_a, t \in \mathcal{T}_d$$

The Wardrop principle applied to the departure time choice can be expressed as

$$\int_{t \in \mathcal{T}_d} dt \left[\varphi_w^{t_a}(t).\left(PTC_w^{t_a}(t) - \min_{s \in \mathcal{T}_d} PTC_w^{t_a}(s)\right)\right] = 0 \quad \forall w \in \mathcal{W}, t_a \in \mathcal{T}_a, \varphi \in (\mathcal{K}) \tag{3}$$

where φ must be constrained by $(\mathcal{K})$.

3 Implementation, concluding remarks.

Let us consider the resolution of (3) in a day-to-day process. We denote by τ the day index. (3) can be viewed as a fixed point problem (a natural functional setting would be a L^2 space of square integrable functions with respect to $t \in \mathcal{T}_d$):

$$\varphi = P_{\mathcal{K}}\left[\varphi - PTC(\varphi)\right] \tag{4}$$

Here φ denotes the vector of OD costs, $P_{\mathcal{K}}$ the projector on $(\mathcal{K})$ and PTC the vector of OD costs, which can be considered as a function of φ because (2) must be solved with respect to ϖ given φ. Various schemes are conceivable in order to solve (4), for instance

$$\varphi^{\tau+1} = (1-\beta^\tau)\,\varphi^\tau + \beta^\tau P_{\mathcal{K}}\left[\varphi^\tau - \alpha^\tau PTC(\varphi^\tau)\right] \tag{5}$$

Here the coefficients α^τ and β^τ express learning behaviours of travellers. α^τ expresses the sensitivity to expected OD travel costs whereas β^τ would designate the fraction of travellers who actually react to OD travel costs. These coefficients can also depend on the day.

Preliminary results show various patterns: simple convergence, periodic orbits, lack of convergence suggestive of possible chaotic behaviour, depending on the learning strategies and on the demand level which appears also as a critical parameter. These results are consistent with previous results reported in the literature for the static/quasi-static case.

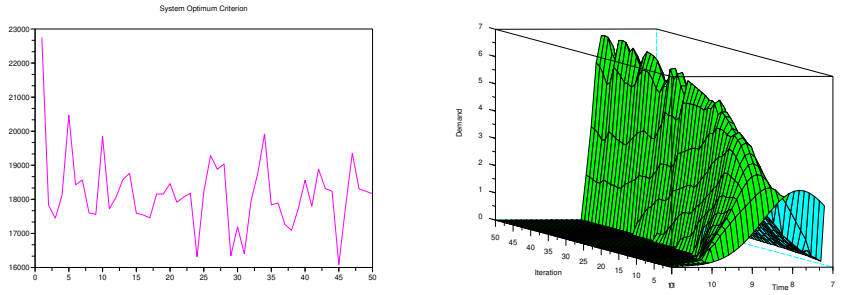


Figure 1: Non convergence. Left: Total network travel cost as a function of time, right: demand f for one path and desired arrival time, as a function of τ (iteration) and departure time

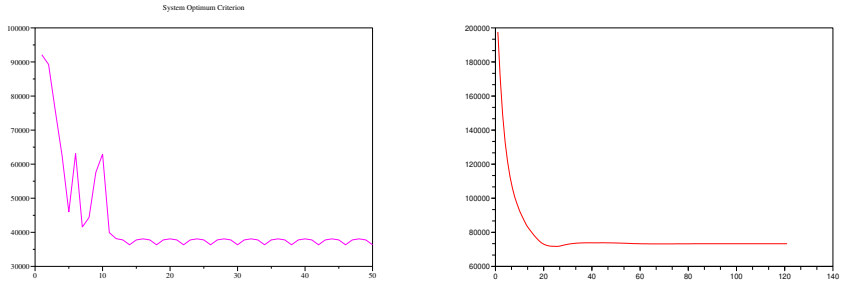


Figure 2: Left: the system converges towards an orbit of period 4, right the system converges to an equilibrium (index: total network travel cost as a function of iteration τ)

References

[1] Rafegh Aghamohammadi and Jorge A Laval. Dynamic traffic assignment using the macroscopic fundamental diagram: A review of vehicular and pedestrian flow models. *Transportation Research Part B: Methodological*, 137:99–118, 2020.

[2] Mostafa Ameli, Jean-Patrick Lebacque, and Ludovic Leclercq. Simulation-based dynamic traffic assignment: Meta-heuristic solution methods with parallel computing. *Computer-Aided Civil and Infrastructure Engineering*, 35(10):1047–1062, 2020.

[3] Mostafa Ameli, Jean-Patrick Lebacque, and Ludovic Leclercq. Computational methods for calculating multimodal multiclass traffic network equilibrium: Simulation benchmark on a large-scale test case. *Journal of Advanced Transportation*, 2021:1–17, 2021.

[4] Giulio Erberto Cantarella and Chiara Fiori. Multi-vehicle assignment with elastic vehicle choice behaviour: Fixed-point, deterministic process and stochastic process models. *Transportation Research Part C: Emerging Technologies*, 134:103429, 2022.

[5] Yi-Chang Chiu, Jon Bottom, Michael Mahut, Alexander Paz, Ramachandran Balakrishna, Steven Waller, and Jim Hicks. Dynamic traffic assignment: A primer (transportation research circular e-c153). 2011.

[6] Ren-Yong Guo and Hai-Jun Huang. Chaos and bifurcation in dynamical evolution process of traffic assignment with flow "mutation". *Chaos, Solitons & Fractals*, 41(3):1150–1157, 2009.

[7] Andreas Horni, Kai Nagel, and Kay W Axhausen. Introducing matsim. In *The multi-agent transport simulation MATSim*, pages 3–7. Ubiquity Press, 2016.

[8] Islam Kamel, Amer Shalaby, and Baher Abdulhai. Integrated simulation-based dynamic traffic and transit assignment model for large-scale network. *Canadian Journal of Civil Engineering*, 47(8):898–907, 2020.

[9] Megan M Khoshyaran and Jean-Patrick Lebacque. Gsom traffic flow models for networks with information. In *International Conference on Systems Science*, pages 210–220. Springer, 2016.

[10] Megan M Khoshyaran and Jean-Patrick Lebacque. Continuum traffic flow modelling: network approximation, flow approximation. In *Traffic and Granular Flow 2019*, pages 505–513. Springer, 2020.

[11] Megan M Khoshyaran and Jean-Patrick Lebacque. Reactive dynamic traffic assignment: impact of information. *Transportation Research Procedia*, 47:59–66, 2020.

[12] Masao Kuwahara and Takashi Akamatsu. Decomposition of the reactive dynamic assignments with queues for a many-to-many origin-destination pattern. *Transportation Research Part B: Methodological*, 31(1):1–10, 1997.

[13] Jean-Patrick Lebacque and Megan M Khoshyaran. A variational formulation for higher order macroscopic traffic flow models of the GSOM family. *Procedia-Social and Behavioral Sciences*, 80:370–394, 2013.

[14] Jean-Patrick Lebacque and Megan M Khoshyaran. Multimodal transportation network modeling based on the generic second order modeling approach. *Transportation Research Record*, 2672(48):93–103, 2018.

[15] Jean-Patrick Lebacque, Salim Mammar, and Habib Haj Salem. Generic second order traffic flow modelling. In *Transportation and Traffic Theory 2007. Papers Selected for Presentation at ISTTT17*, 2007.

[16] Shixu Liu, Lidan Guo, Said M Easa, Wensi Chen, Hao Yan, and Yingnuo Tang. Chaotic behavior of traffic-flow evolution with two departure intervals in two-link transportation network. *Discrete Dynamics in Nature and Society*, 2018:1–11, 2018.

[17] Hong Kam Lo and Wai Yuen Szeto. A cell-based dynamic traffic assignment model:

formulation and properties. *Mathematical and computer modelling*, 35(7-8):849–865, 2002.

[18] Siguraur F Marinósson, Roland Chrobok, Andreas Pottmeier, Joachim Wahle, and Michael Schreckenberg. Simulation des autobahnverkehrs in nrw. In *SCS/ASIM-16. Symposium in Rostock, Simulationstechnik, S*, pages 517–523, 2002.

[19] Khoshyaran M.M. and J.P Lebacque. Complex dynamics generated by simultaneous route and departure time choice in transportation networks. In Yiannis Dimotikalis Christos H Skiadas, editor, *Proceedings of the 16th Chaotic Modeling and Simulation International Conference*. Springer, 2024.

[20] Stéphane Mollier, Maria Laura Delle Monache, Carlos Canudas-de Wit, and Benjamin Seibold. Two-dimensional macroscopic model for large scale traffic networks. *Transportation Research Part B: Methodological*, 122:309–326, 2019.

[21] Anna Nagurney and Ding Zhang. Projected dynamical systems in the formulation, stability analysis, and computation of fixed-demand traffic network equilibria. *Transportation Science*, 31(2):147–158, 1997.

[22] Anna Nagurney and Ding Zhang. *Projected dynamical systems and variational inequalities with applications*, volume 2. Springer Science & Business Media, 2012.

[23] Michael Patriksson. *The traffic assignment problem: models and methods*. Courier Dover Publications, 2015.

[24] Yi Wang, Wai Y Szeto, Ke Han, and Terry L Friesz. Dynamic traffic assignment: A review of the methodological advances for environmentally sustainable road transportation applications. *Transportation Research Part B: Methodological*, 111:370–394, 2018.

[25] John Glen Wardrop. Road paper. some theoretical aspects of road traffic research. *Proceedings of the institution of civil engineers*, 1(3):325–362, 1952.

Detection of Anomalous Spatio-temporal Patterns of App Traffic in Response to Catastrophic Events

Sofia Medina[1✓], Nicola Pedreschi[1], Timothy LaRock[1], Shazia'Ayn Babul[1], Rohit Sahasrabuddhe[1], Renaud Lambiotte[1]

[1] *University of Oxford, UK; sofia.medina@maths.ox.ac.uk.*

[✓] *Presenting author*

Abstract. Catastrophic events can be captured through mobile phone data as people react in real time to virtual information. We investigate how mobile phone application usage patterns vary across apps in the aftermath of a catastrophic event. This is explored temporally by characterizing patterns on the day of and after the fire of the Notre-Dame cathedral. This is further explored spatially by examining the relationship between physical distance and virtual information spread. Our methods and findings give insight into how information spreads during a catastrophe in both time and space.

Keywords. *Mobility; Data Science*

1 Introduction

Understanding how information propagates during and after catastrophic events is an active field of investigation [4, 10, 6, 9]. Social media and online resources have been used to track the length and intensity of responses to breaking news stories [9, 5], or to categorize types of responses from the population [6]. Mobile phone data provides deep insight into the intricacies of human behavior, with high granularity on both temporal and spatial scales. These data-sets enable large-scale data driven analysis applied to a wide range of areas including social network analysis [3], population dynamics [2], and urban structure [7]. In this work, we analyze mobile phone data to understand how the temporal and spatial usage of different applications are perturbed in the aftermath of an unprecedented event. To this end, we use the NetMob2023 Data Challenge dataset [8] which provides mobile phone usage data for several cities in France for a range of applications over a 3 month period of time in 2019, at a spatial resolution of $100m^2$ and a time resolution of 15 minutes. We investigate dynamics caused by one of the most extreme events occurring during this period: the burning of the Notre-Dame cathedral in Paris and the collapse of its historic spire. We analyze the spread of information before, during, and after the fire, using volume of app traffic as a proxy for information transfer.

We first study the timeseries of app traffic per city for a select subset of the available apps. This allows us to characterize apps and look at application-function based differences, as well as user preferences across cities. Second, we analyze how traffic spikes were distributed spatially throughout the city of Paris. We consider if there is an association between information

spread, in both time and intensity, and the distance from the epicenter of the catastrophe. We discuss the dependency on radius from epicenter of catastrophe, quantitatively addressing the relaxation of an approximate radial spread of information over time.

Thus, we provide a thorough description of how the catastrophic event of the fire of the Notre-Dame cathedral perturbed the temporal and spatial patterns of app usage traffic. The methods utilized can be extended to other contexts to characterize mobile phone user response to unplanned catastrophic events, giving insight into how information spreads during a catastrophe in both time and space.

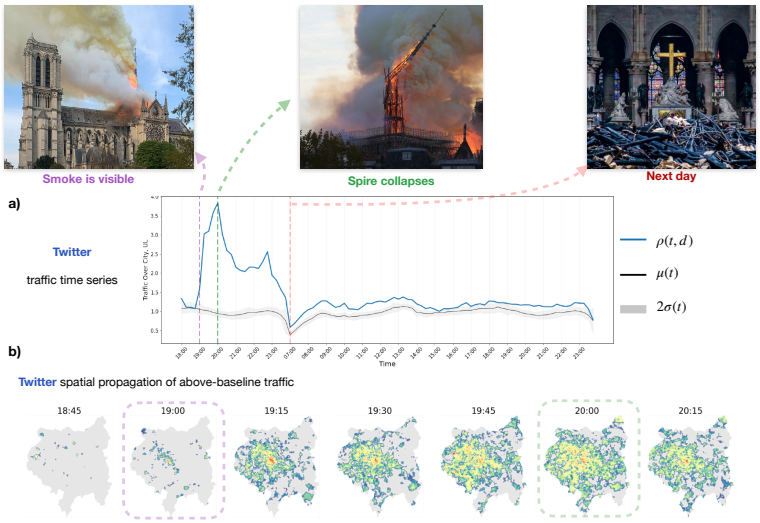


Figure 1: **Graphical introduction: a)** This panel shows the time series evolution of the application Twitter over a time period ranging from 18 : 00 to end of day on the day of the fire and 7 : 00 to end of day the day following the fire. The traffic on the day of the fire, ρ, is given in blue, with the mean traffic of previous weeks and two standard deviations from this mean also shown. **b)** The spatial evolution of application traffic of Twitter is shown in time on the outline of the city of Paris, with intensity of app traffic from an established baseline displayed.

2 Spatio-Temporal Methods

In order to indicate abnormal traffic spikes, we define the *distance from baseline* at time t as,

$$D(t) = \frac{\rho(t) - (\mu(t) + 2\sigma(t))}{\mu(t) + 2\sigma(t))}, \tag{1}$$

indicating the value of traffic volume in relation to two standard deviations from the baseline mean of traffic. We then represent the traffic time series of each app on the day of and after the fire as two features vectors of length 5, normalizing each vector by the maximum and minimum values for spiking apps within each feature for that day.

- (s_i^{start}, s_i^{end}): the starting and ending times of the ith spike
- $s_i^{duration}$: the duration of the spike
- s_i^{max}: the maximum value of $D(t)$ during the spike
- $s_i^{aggregate}$: The sum of all $D(t)$ values during the spike

Finally, we perform a k-means clustering of the feature vectors, clustering applications in Paris separately from all other cities in order to investigate changes in behavioral patterns between city the of incidence and other cites. We choose number of clusters $k = 5$ by computing the Sum of Squared Errors over values of k and manually inspecting the curves.

We now turn our attention to spatial considerations, defining the *tile-wise distance from baseline* at time t using the same form as Equation 1 for each $100m^2$ spatial tile in the city of Paris. We compute this distance on each tile for the app Twitter from t =18:45 on April 15th, the day of the fire, until t =24:00.

Taking inspiration from [1], we construct concentric squares of increasing radial size centered on the epicenter, with radius being the diagonal of each square, and measuring the distance from the epicenter in km's. We sum the total anomalous traffic, $\sum_{i \in r} D_i(t)$, within the square of radius r. For subsequent squares, we now subtract the total anomalous traffic of the previous square, resulting in a measure that gives the change in volume of application traffic as a function of the radius [1],

$$\Delta D(r) = \sum_{i \in r} D_i(t) - \sum_{j \in r-1} D_j(t). \tag{2}$$

We posit that a 'perfect' radial spread outward from the epicenter would be represented with Eq. 2 by an exponential decay curve. The authors in [1] generally consider change in anomalous traffic volume from the epicenter aggregated over some time period, giving the course-grained, general behavior of the response to catastrophe. Since we are interested in how the radial spread of information relaxes, we also consider the 'instantaneous' function of Eq. 2 at each measured time point. This more fine-grained approach to understanding radial spread shows how radial patterns evolve over time, relaxing from perturbation, and also ensures that the response of one measured time does not dominate the interval of measurement.

3 Results and Conclusions

We identify applications that present abnormal traffic in response to the fire, consistent with the assumption that human activity spreads across social-media as people react to catastrophic events. We also find that spiking patterns are observed for applications across all cities included in the analysis (Paris, Lyon, Marseille, Montpellier, Rennes, and Strasbourg). Interestingly, the applications spiking in cities other than Paris are always a subset of the applications that spike in the French capital, the city of incidence.

We find that applications with the same *'function'*, such a social media platforms, can exhibit very different response patterns. One might have better success in predicting app response patterns by a more fine grained type such as, messaging, broadcasting, image sharing, live video, etc. We also note that patterns of application traffic, and abnormal traffic in general, are prevalent the day of the fire but quickly relax and may be undetectable the day after the fire. The city of incidence experiences the most persistent abnormal application traffic, with traffic in other cities generally returning to normality.

While there are slight variations between the types of app behavior in Paris and behavior in other cities, there are notable similarities in user usage for applications in response to catastrophe. For example, in all cities, the live video streaming app Periscope is characterized by a very large, yet brief, spike in usage, taking place shortly after the start of the fire.

We also find that apps that spike in multiple cities tend to cluster together, for example, Facebook Messenger clusters together for all the cities. However, there is some slight variation of app usage based on within-city user behavior. This is evident in that there are apps that spike in some cities, and not others. Furthermore, there is some slight variation in which cluster an app can appear in for different cities. These results suggest the strong effect of user-preference on app traffic.

Finally, we investigate the association between information spread, in both time and intensity, and the distance from the epicenter of the catastrophe. We aim to identify and quantify spatial patterns of above baseline traffic volume and their evolution over time. We detect an approximately *radial* spreading pattern, emanating outward from the Notre-Dame Cathedral. Such a pattern appears to persist until the end of the day, relaxing over time in both intensity and *traveled* distance. This result is surprising in that the intensity of the spread appears to decay radially with the distance from the epicenter of the catastrophe, which is not immediately evident given that the means of communication being investigated, in principle, does not rely on physical distance.

References

[1] James P. Bagrow, Dashun Wang, and Albert-Laszlo Barabasi. Collective response of human populations to large-scale emergencies. *PLOS ONE*, 6(3), MAR 30 2011.

[2] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014.

[3] Nathan Eagle, Alex Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36):15274–15278, 2009.

[4] Ruth Garcia-Gavilanes, Milena Tsvetkova, and Taha Yasseri. Dynamics and biases of online attention: the case of aircraft crashes. *ROYAL SOCIETY OPEN SCIENCE*, 3(10), OCT 2016.

[5] Mihai Georgescu, Nattiya Kanhabua, Daniel Krause, Wolfgang Nejdl, and Stefan Siersdorfer. Extracting event-related information from article updates in wikipedia. In *European Conference on Information Retrieval*, 2013.

[6] Janette Lehmann, Bruno Gonçalves, José J. Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 251–260, New York, NY, USA, 2012. Association for Computing Machinery.

[7] Thomas Louail, Maxime Lenormand, Oliva G Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *Scientific reports*, 4(1):5276, 2014.

[8] Orlando E Martínez-Durive, Sachit Mishra, Cezary Ziemlicki, Stefania Rubrichi, Zbigniew Smoreda, and Marco Fiore. The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography, 2023.

[9] Miles Osborne, Saša Petrovic, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Bieber no more: First story detection using twitter and wikipedia. In *Sigir 2012 workshop on time-aware information access*, pages 16–76. Citeseer, 2012.

[10] Fang Wu and Bernardo A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, 2007.

FRCCS

The Effects of Climate Change on Internal Migration in South Africa

Valentina Antonaccio Guedes[1,2✓], Pete Barbrook-Johnson[1,2]

[1] *School of Geography and the Environment, University of Oxford, UK;* [2] *Institute of New Economic Thinking at the Oxford Martin School, University of Oxford, UK. Emails: valentina.antonacciogуedes@ouce.ox.ac.uk; peter.barbrook-johnson@ouce.ox.ac.uk*
[✓] *Presenting author*

Abstract. As climate impacts become more frequent and drastic, the importance of understanding how they will affect migration is growing. However, the complexity of migration dynamics means it is difficult to make conclusions about these effects. In this study, we focus on South Africa because there is a high prevalence of internal migrants, it is the most unequal country in terms of income, and it faces a high frequency of climate change impacts with long periods of droughts followed by intense floods, expected to get worse over time. These climate effects are affecting the country's infrastructure, productivity, human capital, and water resources, which in the absence of intervention, makes it one of the countries most likely to be negatively affected by unplanned climate migration. The challenge lies in understanding potential climate migration dynamics and designing effective policies to mitigate negative effects, from which other countries could learn from too. We will build a spatial agent-based model, exploring the decision making process behind migration and identify the *hot spots* of migration dynamics. That is, which are the sending places, where are people going to, what are the most salient patterns of people's movements within the country, and where do these overlap with climate risk. The goal is to use the agent-based model to study the differences between mitigation and adaptation policies for climate change migration, as well as economic growth and inequalities, especially those related to gender.

Keywords. *Climate Change; Migration; Agent Based Model; Public Policy; Spatial Networks*

1 Introduction

Climate change affects peoples' livelihoods both directly and indirectly through economic, political and social channels. One strategy that can be used to cope with its effects is to migrate [32, 15, 21]. As droughts and floods are becoming more frequent in time, internal migration to mitigate or adapt to these changes will become a central policy issue over the coming decades, especially so for developing countries. By internal migration we understand people's movement within a country's borders which is not necessarily from rural areas to urban ones, but also the reverse or even between cities [23, 32, 26, 28, 39].

If market forces are left to act, the impact of climate change on migration can lead to the unplanned growth of cities causing an increase in poverty, social divisions and general unrest

[23, 28, 40]. The Internal Displacement Migration Center [18] estimates that around 32.6 million people were displaced due to floods, windstorms, earthquakes or droughts in 2022. By 2050, it is expected that around 216 million people will be considered internal migrants due to climate change around the world. Of these, 86 million will be in Sub-Saharan Africa alone [31]. So, there is a clear need to better understand migration dynamics related to climate change and how these can be affected and managed by different socio-economic drivers and policy makers [40, 28].

Even though the number of publications on the effects of climate change on migration has increased exponentially in the past 10 years, there are many questions that remain to be answered [30]. The migrating process is considered a multi-causal complex phenomenon which makes isolating the environmental effect on it a challenge for researchers [4, 27, 2, 6, 28, 21, 30]. This is why many people have little confidence on current global estimates of climate-led migration [5, 34, 16, 17, 12, 21]. The effects go from climate having a small, indirect impact on migration to a direct large one [10, 25, 7]. What is more, each country is different, where similar shocks can result in different migration patterns, may be, due to the different frictions individuals face when it comes to moving within the country as climate impacts do not affect all people, households and communities in the same way [30, 32, 14, 26].

This is especially true for the African continent which is considered to be one of the most vulnerable regions to climate change, given its vast semiarid areas, rain-fed agricultural system and low adaptive capacity [7]. In fact, it is still not clear if migration should be seen as a symptom of climate change or an adaptive strategy to it, as it is only available to those who can afford it, tying the poor to a place where they are not able to support their livelihoods [30]. Finally, it is known that climate change has serious effects on women's physical and mental health [33, 9, 11]. However, it is not clear the implications that migration due to climate change may have for them, as they might see themselves tied to a place due to different social expectations, for example [32, 30].

1.1 An Agent Based Model on climate change and migration in South Africa

1.1.1 Model purpose

This study aims to explore the effects of on-set climate change on internal migration in South Africa using a spatial agent-based model (ABM). The main goal is to identify *hot spots* and characteristics of internal migrants. I will use the definition of *hotspot* as a *"geographical area where a strong climate event is combined with a large concentration of vulnerable and marginalised people"*, given the general characteristics of the area of study ([36]: p784.). That is, we will be looking at the *hotspots* of risk and vulnerability and given the occurrence of climatic events, forecast where people might decide to move to, as well as how often will we see these kinds of shifts. That is, at what point will people decide to leave a climatic affected area towards another, and what characterises a receiving place. Given that migration is a complex process, the model will require validation; based on the different climate scenarios that we will be basing our analysis upon, we will identify the particular surges of migration that may occur.

This research idea is based on the Groundswell reports by the World Bank [32, 31]. They build a gravity model, introducing environmental factors, in order to obtain four main results on migration: the number of climate change induced migrants, the share that these migrants constitute over the total number of people that migrate within and outside the country, maps of the hot spots of climate in-and out-migration. Then they estimate net in- and out-migration

for three areas: rural livelihood zones, coastal zones and urban areas. Gravity models aim to simulate aggregate human behaviour based on Push-Pull theories [24]. The relative attractiveness of a location relative to the un-attractiveness of the place of origin plus the distance is what drives the decision to migrate [13, 32, 24]. In this sense, they reduce human decision to depend on distance and the number of intervening opportunities without giving further insight on the mechanisms that push people to take the decision to move [13, 32, 22]. What is more, as [24] explains, factors in the destination area include not only jobs and social services available, but also social networks, or migration networks in the sending area, which are not possible to include in gravity models as well as the decision theory behind the decision making process.

ABMs are more flexible, as they allow us to include heterogeneous agents, who interact with each other, and make decisions based on the theories of choice, action, and social interactions, in order to have them react to the main drivers of migration flows: economic, political, demographic, social and environmental [24, 1, 32, 22]. As they also allow to incorporate networks, we can control for one of the strong determinants of where people decide to migrate [22]. In a nutshell, ABMs enable transitions to be determined by causal mechanisms, which are built from the bottom up. That is, macro level structures that stem from micro-level interactions. They also allow us to explore 'what-if' scenarios: you have a model that you can shock and provides the effect of if for different situations we create [22].

In summary, for this paper, we will build a spatial ABM that can identify the *hot spots* of where people are migrating to, and what the characteristics of those who are migrating to these places are, by exploring the different heuristics that people use to come to the decision to migrate, especially when affected by climate change.

1.1.2 Key inputs and data

As a first step, we will generate a realistic and spatially explicit synthetic population. For this, we will use NIDS data, as well as the population distribution and insights on the main predictors of migration and their effect on the time that it takes the individual to migrate, obtained from a previous data analysis [8, 35, 37].

The project will also rely on future climate scenarios - Representative Concentration Pathways (RCP) which is a greenhouse gas concentration trajectory adopted by the Intergovernmental Panel On Climate Change (IPCC) and the Shared Socioeconomic Pathways (RCP) presented by [20] and [29] which provides national-wide estimates of population, GDP and urbanization. We will test first the a situation in which we have "business as usual" and no policy is implemented. Then we will run the model including the SSP effects for population growth, GDP and urbanization. Finally, we will run it including the SSPs as well as the different scenarios of the RCP. The differences in results from the model with SSPs and the one that has both SSPS and RCP, we will attribute to the effects of climate change [32, 31].

1.1.3 Model design

When modelling human interaction and decision-making with an ABM, it is necessary to set up the rules that are going to determine how agents arrive to their outcome of choice; in this case, to migrate or not. As Arthur (1994) [3] argues, beyond a certain level of complexity, human reasoning is subject to bounded rationality. That is, humans tend to use heuristics - inductive rather than deductive methods - of thinking and reaching a conclusion when faced with a complex problem. When looking at the literature on the subject, the use of ABMs to study migration is a fairly unexplored topic.

We will adapt Jager, et al (2000) [19] cognitive processing models, adapting them to the case

of migration. These set of heuristics summarise in a parsimonious way how humans see the world and make decisions simplifying the complex situations in which they find themselves. Individuals will use heuristics to decide based on the levels of need satisfaction and uncertainty they are facing [19]. Following table 1, if the individual has a low degree of uncertainty and their needs are satisfied, then they will repeat their past behaviour. Only when their needs stop being satisfied (i.e. they lose their crop three seasons in a row) and they cannot sustain their livelihood anymore, they will become *deliberate*. That is, they have low uncertainty about their future, because they know what the best adaptive strategy at hand is, and they will do this in order to maximise their needs again. There are many adaptive strategies someone could take. In this particular case, some examples could be for the whole family to migrate, for a young adult to migrate to the city in order to diversify their income in case another crop fails, the government implements a loan system for farmers to get irrigation, etc. If the individual has high uncertainty but their needs are satisfied they will imitate another person who is similar to them, i.e. get irrigation for their corp because their neighbour did. But, if the individual is not able to afford this, then they will compare their own previous behaviour with that of their neighbours with similar characteristics, to then select the behaviour that would give them the highest need satisfaction: i.e. by choosing an alternative adaptive strategy like migrating.

Table 1: Dimensions of social processing and reasoned behaviour

	Needs Satisfied	**Needs Not Satisfied**
Low uncertainty	Repetition	Deliberation
High uncertainty	Imitation	Social comparison

We will also connect agents within a spatial social network, in which they will have stronger ties to those who live near them [22, 38]. Following, Valdano, at al (2021) [38] and Xiao, et al (2022) [40] we will use the additive and multiplicative effects model for networks (AMEN) with a Poisson distribution, because it allows to control for the spatial auto-correlation, already mentioned, and control for unobserved variables. Another important point to consider, and that can be tested with this methodology, is the fact that migration can happen at the household or the individual level. This is something that is presented in the New-Economics of Labour Migration Theory, which highlights that migration can be used as a strategy by the household to diversify its income sources and reduce its hardships and risks [32].

I will then use the model to explore migration outcomes under a range of different climate change and development scenarios presented in Jones and O'Neill (2016) and O'Neill et al (2014) respectively [20, 29] . These scenarios provide estimates of population, GDP and adaptation for each Shared Socioeconomic Pathway (SSP). A SSP is a different situation based on two indicators: challenges to adaptation to climate change, and challenges to mitigation of climate change. An adaptation measure is migration, for example, while a mitigation measure is reducing the amount of gas emissions into the atmosphere. For example, 'SSP1 - sustainability' is a shared socioeconomic pathway in which the challenges to both adaptation and mitigation are low, while 'SSP5 - conventional development' is one in which the challenges to adaptation are low but the ones for mitigation are high. These information is available through the SSP database [32].

References

[1] Guy J. Abel, Michael Brottrager, Jesus Crespo Cuaresma, and Raya Muttarak. Climate, conflict and forced migration. *Global environmental change*, 54:239–249, 2019. ISBN: 0959-3780 Publisher: Elsevier.

[2] Helen Adams and Susan Kay. Migration as a human affair: Integrating individual stress thresholds into quantitative models of climate migration. *Environmental Science & Policy*, 93:129–138, 2019. ISBN: 1462-9011 Publisher: Elsevier.

[3] W. Brian Arthur. Inductive reasoning and bounded rationality. *The American economic review*, 84(2):406–411, 1994. ISBN: 0002-8282 Publisher: JSTOR.

[4] Kelsea Best, Jonathan Gilligan, Hiba Baroud, Amanda Carrico, Katharine Donato, and Bishawjit Mallick. Applying machine learning to social datasets: a study of migration in southwestern Bangladesh using random forests. *Regional Environmental Change*, 22(2):52, March 2022.

[5] Frank Biermann and Ingrid Boas. Preparing for a warmer world: Towards a global governance system to protect climate refugees. *Global environmental politics*, 10(1):60–88, 2010. ISBN: 1526-3800 Publisher: MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ....

[6] Richard Black, W. Neil Adger, Nigel W. Arnell, Stefan Dercon, Andrew Geddes, and David Thomas. The effect of environmental change on human migration. *Global environmental change*, 21:S3–S11, 2011. ISBN: 0959-3780 Publisher: Elsevier.

[7] Marion Borderon, Patrick Sakdapolrak, Raya Muttarak, Endale Kebede, Raffaella Pagogna, and Eva Sporer. Migration influenced by environmental change in Africa. *Demographic Research*, 41:491–544, 2019. ISBN: 1435-9871 Publisher: JSTOR.

[8] Center for International Earth Science Information Network - CIESIN - Columbia University. *Gridded Population of the World, Version 4 (GPWv4): Basic Demographic Characteristics, Revision 11.* NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, New York, 2018.

[9] Matthew F. Chersich, Caradee Y. Wright, Francois Venter, Helen Rees, Fiona Scorgie, and Barend Erasmus. Impacts of climate change on health and wellbeing in South Africa. *International journal of environmental research and public health*, 15(9):1884, 2018. ISBN: 1660-4601 Publisher: MDPI.

[10] Hein De Haas. Mediterranean migration futures: Patterns, drivers and scenarios. *Global Environmental Change*, 21:S59–S69, 2011. ISBN: 0959-3780 Publisher: Elsevier.

[11] Zalak Desai and Ying Zhang. Climate change and women's health: A scoping review. *GeoHealth*, 5(9):e2021GH000386, 2021. ISBN: 2471-1403 Publisher: Wiley Online Library.

[12] Ottmar Edenhofer. *Climate change 2014: mitigation of climate change*, volume 3. Cambridge University Press, 2015.

[13] G. O. Ewing. Gravity and linear regression models of spatial interaction: a cautionary note. *Economic geography*, 50(1):83–88, 1974. ISBN: 0013-0095 Publisher: Taylor & Francis.

[14] Elizabeth Ferris. Research on climate change and migration where are we and where are we going? *Migration Studies*, 8(4):612–625, 2020. ISBN: 2049-5838 Publisher: Oxford University Press.

[15] U. K. Foresight. Migration and global environmental change: final project report. *The Government Office for Science, London. Available at: https://www. gov. uk/government/upl oads/system/uploads/attachment_d ata/file/287717/11-1116-migrationand-global-environmentalchange. pdf*, 2011.

[16] François Gemenne. How they became the human face of climate change. Research and policy interactions in the birth of the 'environmental migration'concept. *Cambridge Uni-*

versity Press, Cambridge, United Kingdom, 2011. Publisher: Cambridge University Press, Cambridge, United Kingdom.

[17] Betsy Hartmann. Rethinking climate refugees and climate conflict: Rhetoric, reality and the politics of policy discourse. *Journal of International Development: The Journal of the Development Studies Association*, 22(2):233–246, 2010. ISBN: 0954-1748 Publisher: Wiley Online Library.

[18] Internal Displacement Migration Center. IDMC | GRID 2023 | 2023 Global Report on Internal Displacement, 2023.

[19] Wander Jager, Marco A. Janssen, H. J. M. De Vries, J. De Greef, and C. A. J. Vlek. Behaviour in commons dilemmas: Homo economicus and Homo psychologicus in an ecological-economic model. *Ecological economics*, 35(3):357–379, 2000. ISBN: 0921-8009 Publisher: Elsevier.

[20] Bryan Jones and Brian C. O'Neill. Spatially explicit global population scenarios consistent with the Shared Socioeconomic Pathways. *Environmental Research Letters*, 11(8):084003, 2016. ISBN: 1748-9326 Publisher: IOP Publishing.

[21] David J. Kaczan and Jennifer Orgill-Meyer. The impact of climate change on migration: a synthesis of recent empirical insights. *Climatic Change*, 158(3):281–300, February 2020.

[22] Anna Klabunde and Frans Willekens. Decision-making in agent-based models of migration: state of the art and challenges. *European Journal of Population*, 32(1):73–97, 2016. Publisher: Springer.

[23] John Knight. The coming economic, social, and political apocalypse? *Centre for the Study of African Economies*, 2022. Publisher: Centre for the Study of African Economies.

[24] Everett S. Lee. A theory of migration. *Demography*, 3(1):47–57, 1966. ISBN: 1533-7790 Publisher: Springer.

[25] Luca Marchiori and Ingmar Schumacher. When nature rebels: international migration, climate change, and inequality. *Journal of Population Economics*, 24(2):569–600, 2011. ISBN: 1432-1475 Publisher: Springer.

[26] David McKenzie. Fears and Tears: Should More People Be Moving within and from Developing Countries, and What Stops this Movement? *The World Bank Research Observer*, 39(1):75–96, 2024. Publisher: Oxford University Press.

[27] Robert McLeman. Developments in modelling of climate change-related migration. *Climatic change*, 117(3):599–611, 2013. ISBN: 1573-1480 Publisher: Springer.

[28] Sara Mercandalli, Bruno Losch, Cristina Rapone, Robin Bourgeois, and Clara Aida Khalil. Rural migration and the new dynamics of structural transformation in sub-saharan Africa. *FAO*, 2017. Publisher: FAO.

[29] Brian C. O'Neill, Elmar Kriegler, Keywan Riahi, Kristie L. Ebi, Stephane Hallegatte, Timothy R. Carter, Ritu Mathur, and Detlef P. van Vuuren. A new scenario framework for climate change research: the concept of shared socioeconomic pathways. *Climatic change*, 122(3):387–400, 2014. ISBN: 1573-1480 Publisher: Springer.

[30] Etienne Piguet. Linking climate change, environmental degradation, and migration: An update after 10 years. *WIREs Climate Change*, 13(1):e746, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcc.746.

[31] Kanta Kumari Rigaud, Alex de Sherbinin, Bryan Jones, Susana Adamo, David Maleki, Anmol Arora, Anna Taeko Casals Fernandez, Tricia Chai-Onn, and Briar Mills. *Groundswell Africa: A Deeper Dive into Internal Climate Migration in the Lake Victoria Basin Countries*. Washington, DC: World Bank, 2021.

[32] Kanta Kumari Rigaud, Alex De Sherbinin, Bryan Jones, Jonas Bergmann, Viviane Clement, Kayly Ober, Jacob Schewe, Susana Adamo, Brent McCusker, and Silke Heuser. Groundswell: Preparing for internal climate migration. *World Bank, Washington, DC*,

2018. Publisher: World Bank, Washington, DC.

[33] Soheil Shayegh and Shouro Dasgupta. Climate change, labour availability and the future of gender inequality in South Africa. *Climate and Development*, pages 1–18, 2022. ISBN: 1756-5529 Publisher: Taylor & Francis.

[34] Sarah Opitz Stapleton, Rebecca Nadin, Charlene Watson, and Jan Kellett. *Climate change, migration and displacement: the need for a risk-informed and coherent approach*. Overseas Development Institute, 2017.

[35] Statistics South Africa. *General Household Survey*. Statistics South Africa, Pretoria, South Africa, 2021.

[36] Josephine Tucker, Mona Daoud, Naomi Oates, Roger Few, Declan Conway, Sobona Mtisi, and Shirley Matheson. Social vulnerability in three high-poverty climate change hot spots: What does the climate change literature tell us? *Regional Environmental Change*, 15(5):783–800, 2015. ISBN: 1436-378X Publisher: Springer.

[37] University of Cape Town's School of Economics. *National Income Dynamics Study*. University of Cape Town, 2017.

[38] Eugenio Valdano, Justin T. Okano, Vittoria Colizza, Honore K. Mitonga, and Sally Blower. Using mobile phone data to reveal risk flow networks underlying the HIV epidemic in Namibia. *Nature communications*, 12(1):1–10, 2021. ISBN: 2041-1723 Publisher: Nature Publishing Group.

[39] World Bank Group. South Africa Country Climate and Development Report. *CCDR Series. World Bank, Washington, DC*, 2022. Publisher: World Bank, Washington, DC.

[40] Tingyin Xiao, Michael Oppenheimer, Xiaogang He, and Marina Mastrorillo. Complex climate and network effects on internal migration in South Africa revealed by a network model. *Population and Environment*, 43(3):289–318, 2022. ISBN: 1573-7810 Publisher: Springer.

Beauty in Complexity

A Visual Representation of the Social Network of a Computer Science Department

Rachel Izenson [1✓], Julissa Hernandez [1] and Theresa Migler [1]

[1] *California Polytechnic State University, San Luis Obispo ; rizenson@calpoly.edu, jhern430@calpoly.edu, tmigler@calpoly.edu*
[✓] *Presenting author*

Abstract. This visual represents the biggest connected subgraph of a social network in the Computer Science department at a predominantly undergraduate public university in the United States. We conducted a survey for students within the department to answer questions pertaining to their mental health, social network, and demographic information. When looking at the diagram, edges represent a friendship between two nodes, colored nodes represent students who took the survey, while gray nodes represent other students or faculty who did not take the survey but were listed by someone who did. The colors are a representation of how long the student has been attending the university for and flow in rainbow order. Red represents the students who have been attending the university for 1 year all the way to purple representing the students who have been attending the university for 6 or more years.
The image was created through building a graph in Networkx and importing the graph into Cytoscape. Cytoscape was used to visualize the graph and color nodes based on the amount of time a student has been attending the university.
The social network is interesting when colored by year as the colors are not randomly distributed throughout the graph. We can see that 3rd year students are very central to the graph, while 1st years are closer to the outside of the graph. It's also interesting to see that while the similar years of students attending the university tend to group together, it is not a strict cluster as many students of different years have overlap and similar neighborhoods.

A Visual Representation of the Social Network of a Computer Science Department

Rachel Izenson, Julissa Romero and Theresa Migler

Department of Computer Science and Software Engineering, California Polytechnic State University, California, USA

FRCCS

All Roads Lead to Rome

Philippe Mathieu[1✓] and Jean-Paul Delahaye[1]

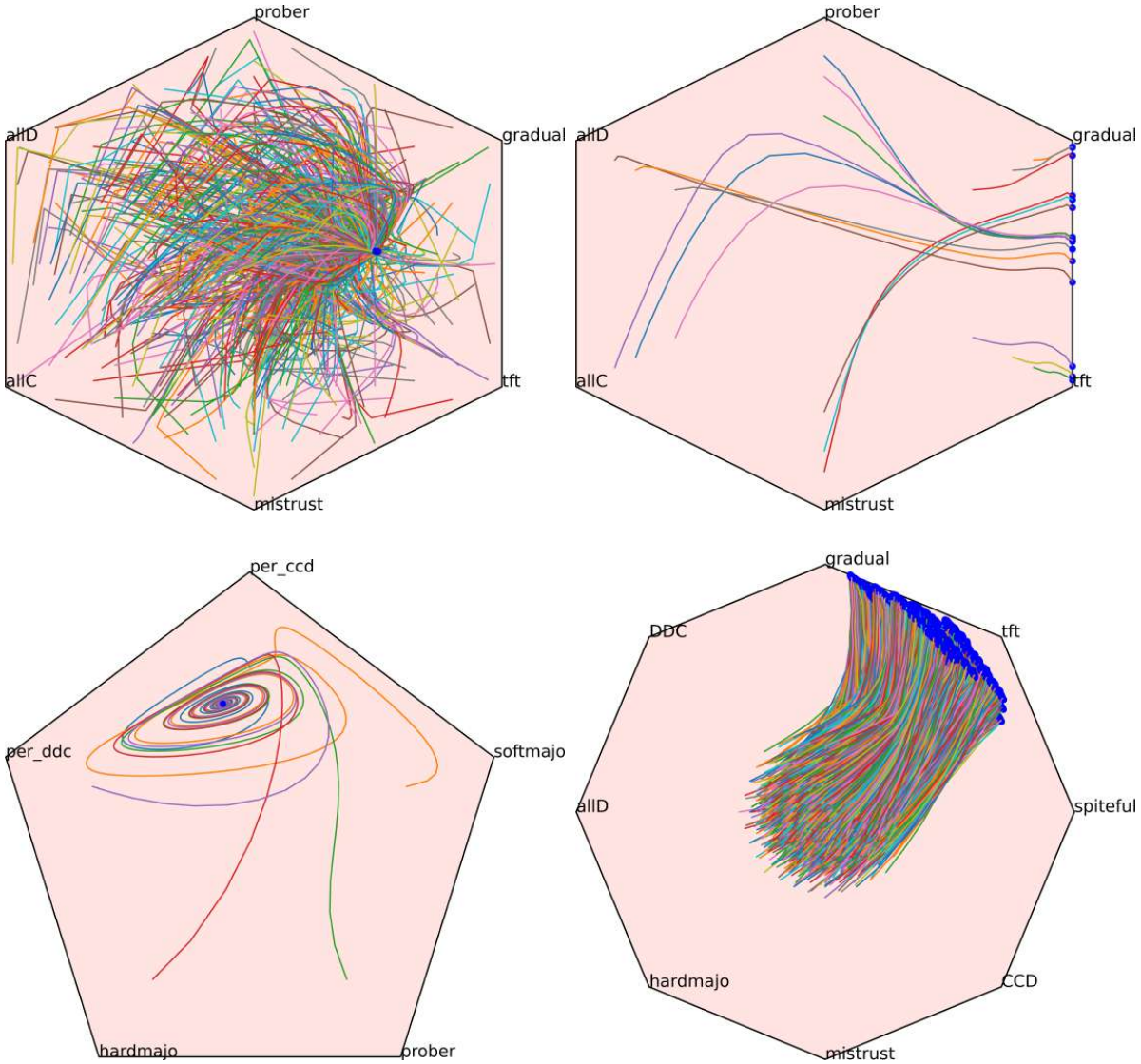
[1] *Univ. Lille, CNRS, Centrale Lille, UMR 9189, CRIStAL, F-59000 Lille, France ; philippe.mathieu@univ-lille.fr*
[✓] *Presenting author*

Abstract. There are a large number of methods for comparing sets of strategies. Among these, evolutionary models offer remarkable relevance and robustness. These images illustrate two remarkable models derived from evolutionary theory: the communitarian model in which an individual does not fight members of his own family (left), and the individualistic model in which an individual fights everyone. It is applied here to the famous iterated prisoner's dilemma with different well-known strategies, in four different situations. Each corner of a polygon corresponds to a proportion of the population. This population then evolves according to the rules of the chosen model, forming a complex system. Each of the two models shows an incredible phenomenon of convergence independent of the initial distribution, which occurs almost systematically, either towards a single attractor (on the left) or towards an axis of mutual cooperation (on the right). Each figure corresponds to a barycentric representation in which each point corresponds to a certain proportion of invividuals. The lines correspond to the evolution of this initial population. For example, in the first figure, we can see that starting with a lot of Probers and AllDs will lead to exactly the same point as starting with a lot of AllCs and Tfts, which is remarkable! This figure shows 720 evolutionary trajectories.

All Roads Lead to Rome

Population dynamics and convergence

Philippe MATHIEU & Jean-Paul DELAHAYE
Univ. Lille, CNRS, Centrale Lille, UMR 9189, CRIStAL
F-59000 Lille, France
philippe.mathieu@univ-lille.fr

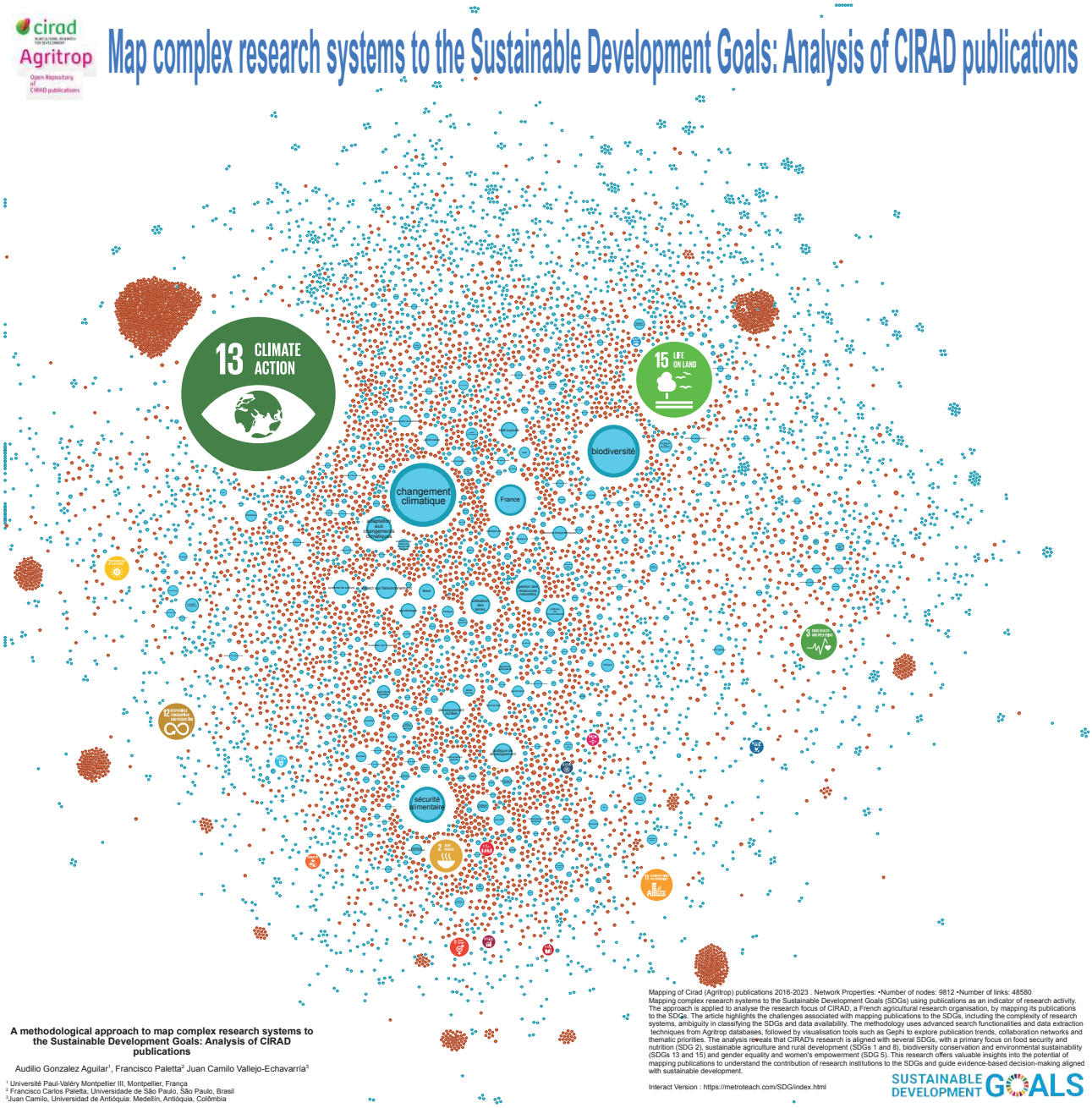


A methodological approach to map complex research systems to the Sustainable Development Goals: Analysis of CIRAD publications

Francisco Paletta [1✓], Audilio AGUILAR [2] and Juan Camilo Vallejo [3]

[1] *University of São Paulo ; fcpaletta@usp.br*
[2] *University Paul-Valéry Montpellier 3 ; audilio.gonzales@gmail.com*
[3] *Universidad de Antioquia ; juan.vallejo@udea.edu.co*
[✓] *Presenting author*

Abstract. Mapping of Cirad (Agritrop) publications 2016-2023. Network Properties: Number of nodes: 9812, Number of links: 48580. Mapping complex research systems to the Sustainable Development Goals (SDGs) using publications as an indicator of research activity. The approach is applied to analyse the research focus of CIRAD, a French agricultural research organisation, by mapping its publications to the SDGs. The article highlights the challenges associated with mapping publications to the SDGs, including the complexity of research systems, ambiguity in classifying the SDGs and data availability. The methodology uses advanced search functionalities and data extraction techniques from Agritrop databases, followed by visualization tools such as Gephi to explore publication trends, collaboration networks and thematic priorities. The analysis reveals that CIRAD's research is aligned with several SDGs, with a primary focus on food security and nutrition (SDG 2), sustainable agriculture and rural development (SDGs 1 and 8), biodiversity conservation and environmental sustainability (SDGs 13 and 15) and gender equality and women's empowerment (SDG 5). This research offers valuable insights into the potential of mapping publications to understand the contribution of research institutions to the SDGs and guide evidence-based decision-making aligned with sustainable development.
Interact Version : https://metroteach.com/SDG/index.html


cirad
Agritrop
Open Repository of CIRAD publications
Map complex research systems to the Sustainable Development Goals: Analysis of CIRAD publications
13 CLIMATE ACTION
15 LIFE ON LAND
changement climatique
biodiversité
France
sécurité alimentaire
A methodological approach to map complex research systems to the Sustainable Development Goals: Analysis of CIRAD publications
Audilio Gonzalez Aguilar[1], Francisco Paletta[2] Juan Camilo Vallejo-Echavarría[3]
[1] Université Paul-Valéry Montpellier III, Montpellier, França
[2] Francisco Carlos Paletta, Universidade de São Paulo, São Paulo, Brasil
[3]Juan Camilo, Universidad de Antióquia: Medellín, Antióquia, Colômbia
Mapping of Cirad (Agritrop) publications 2016-2023 . Network Properties: •Number of nodes: 9812 •Number of links: 48580
Mapping complex research systems to the Sustainable Development Goals (SDGs) using publications as an indicator of research activity. The approach is applied to analyse the research focus of CIRAD, a French agricultural research organisation, by mapping its publications to the SDGs. The article highlights the challenges associated with mapping publications to the SDGs, including the complexity of research systems, ambiguity in classifying the SDGs and data availability. The methodology uses advanced search functionalities and data extraction techniques from Agritrop databases, followed by visualisation tools such as Gephi to explore publication trends, collaboration networks and thematic priorities. The analysis reveals that CIRAD's research is aligned with several SDGs, with a primary focus on food security and nutrition (SDG 2), sustainable agriculture and rural development (SDGs 1 and 8), biodiversity conservation and environmental sustainability (SDGs 13 and 15) and gender equality and women's empowerment (SDG 5). This research offers valuable insights into the potential of mapping publications to understand the contribution of research institutions to the SDGs and guide evidence-based decision-making aligned with sustainable development.
Interact Version : https://metroteach.com/SDG/index.html
SUSTAINABLE DEVELOPMENT GOALS

Cat's Cradle of Pain: Exploring Connections in Chronic Pain

Iris Ho[1✓]

[1] *California Polytechnic State University ; iwho@calpoly.edu*
[✓] *Presenting author*

Abstract. This visual depicts a cat playing with a ball of yarn. The cat is a word cloud of chronic pain symptoms, treatments, and factors. The ball of yarn is a graph consisting of nodes (in a circle) that represent chronic pain patients where two patients are connected if they share a similar pain characteristic: pain duration (green), pain frequency (blue), and pain intensity (yellow). The densely connected graph is reflective of the complexity of chronic pain diagnosis and treatment. The yarn is representative of how these individuals are now connected by their shared experiences of chronic pain. The cat playing with the ball of yarn is symbolic of how chronic pain "toys" with people's lives. We hope that our research will make it easier to untangle the complexity of chronic pain, by first effectively predicting treatment / support groups.
The graph was implemented in NetworkX which was then exported into Cytoscape to visualize the graph. The word cloud was generated with an online word cloud generator.

Cat's Cradle of Pain: Exploring Connections in Chronic Pain

Iris Ho
Computer Science and Software Engineering Department, California Polytechnic State University
San Luis Obispo, CA, USA

FRCCS

Chemical communication in life and AI

Antoni Hernández-Fernández[1✓] and Iván González Torre[2]

[1] *Institut de Ciències de l'Educació, Universitat Politècnica de Catalunya ; antonio.hernandez@upc.edu*
[2] *Oracle, Universitat Oberta de Catalunya ; ivangonzaleztorre@gmail.com*
[✓] *Presenting author*

Abstract. This work presents a montage of images inspired by our previous article: "Compression Principle and Zipf's Law of Brevity in infochemical communication". After training with the article, the images are generated by Bing's Copilot generative AI, selected for their beauty and representation of complexity in chemical communication. The montage juxtaposes these AI-generated images with the figures from the original article, emphasizing the infinite generative capacity of AI and raising questions about the role of AI in scientific creativity.
The visual represents the complex interplay of chemical communication within ecosystems, highlighting the role of infochemicals in shaping ecological communities. Results in artistic creation still contain errors but can produce beautiful and plausible images. But doesn't genetic replication or the generation of chemicals in organisms also cause errors? However, this is not the case for chemical science, despite there are research groups working on the automatic generation of proteins, following language models and physical constraints.
Admire these invented figures, and don't be frightened by chemical elements that don't exist, if any doubt: could they exist? How would they change our universe?
The complexity of chemical communication impacts research by challenging scientists to understand the nuances of these interactions and their implications for ecological dynamics. The montage reflects this complexity through its combination of scientific figures and AI-generated images, sparking contemplation on the potential future applications of AI in scientific discovery. We want it to be understood as a metaphor: can the AI errors of today, the inventions of formulas and chemical elements, be a reality tomorrow, can they inspire researchers for future work?
We invite viewers to admire the invented figures, stimulating curiosity about how AI could reshape our understanding of our universe. Maybe AI can give you the idea for your next paper?


Fit to LND(5.24, 0.42)
allelochemicals
P(MW)
MW
(Z)-9-Tetradecenyl acetate
Formula: $C_{16}H_{30}O_2$
MW: 254.41
allelochemicals
$\lambda = 0.054$
Degree
Compound Size (MW)
Fit to LND(5.42, 0.47)
pheromones
Number of C
Ethanol
Formula: C_2H_6O
MW: 46.07
pheromones
Pheromones
(intraspecific communication)
Allelochemicals
(interspecific communication)
Attractants
Allomones
Kairomones
Synomones
Example
Balance
Effect of aggregation pheromones in Pentatomid bug nymphs.
Emitter Receiver
CHIMRESSION PPRINCESIPLE AND ZIPFS IW'S LAW IF ZVORT
IN ANIMAL COMMUNINCATION NETWORKS

Dandelion Distance Network

Andrew Estrada[1✓]

[1] *California Polytechnic State University ; aestra46@calpoly.edu*
[✓] *Presenting author*

Abstract. This graph represents a sample of the collaboration network for the authors of California Polytechnic State University - San Luis Obispo (Cal Poly), a primarily undergraduate serving, master's-level California public university. It was created using publication data available in Scopus and was used to analyze patterns in geographic collaboration distance. Nodes represent authors either associated with Cal Poly (green nodes), or authors who have collaborated directly with a Cal Poly author (yellow nodes). Edges connect a Cal Poly author and a collaborator – there are no edges connecting yellow nodes since distance information was only calculated for specific California public university authors. Edge lengths, however, are not representative of geographic distance. Cal Poly is just one of several schools whose publication data was used to investigate differences in collaboration distance between male and female authors. The program Cytoscape was used for the visualization of the network. Smaller components were omitted for aesthetics and simplicity.
The network resembles a bed of dandelions, which is fittingly symbolic – researchers are seeds of knowledge, and just as dandelion seeds are carried on the wind to spread, so too does knowledge as researchers collaborate and relocate across varied distances. We hope that our research provides some insight and encourages further analysis of patterns of geographic collaboration distance among academic researchers.

Geographic Distance and Equity Within a Collaboration Network

Andrew Estrada

Computer Science and Software Engineering Department, California Polytechnic State University
San Luis Obispo, CA, USA

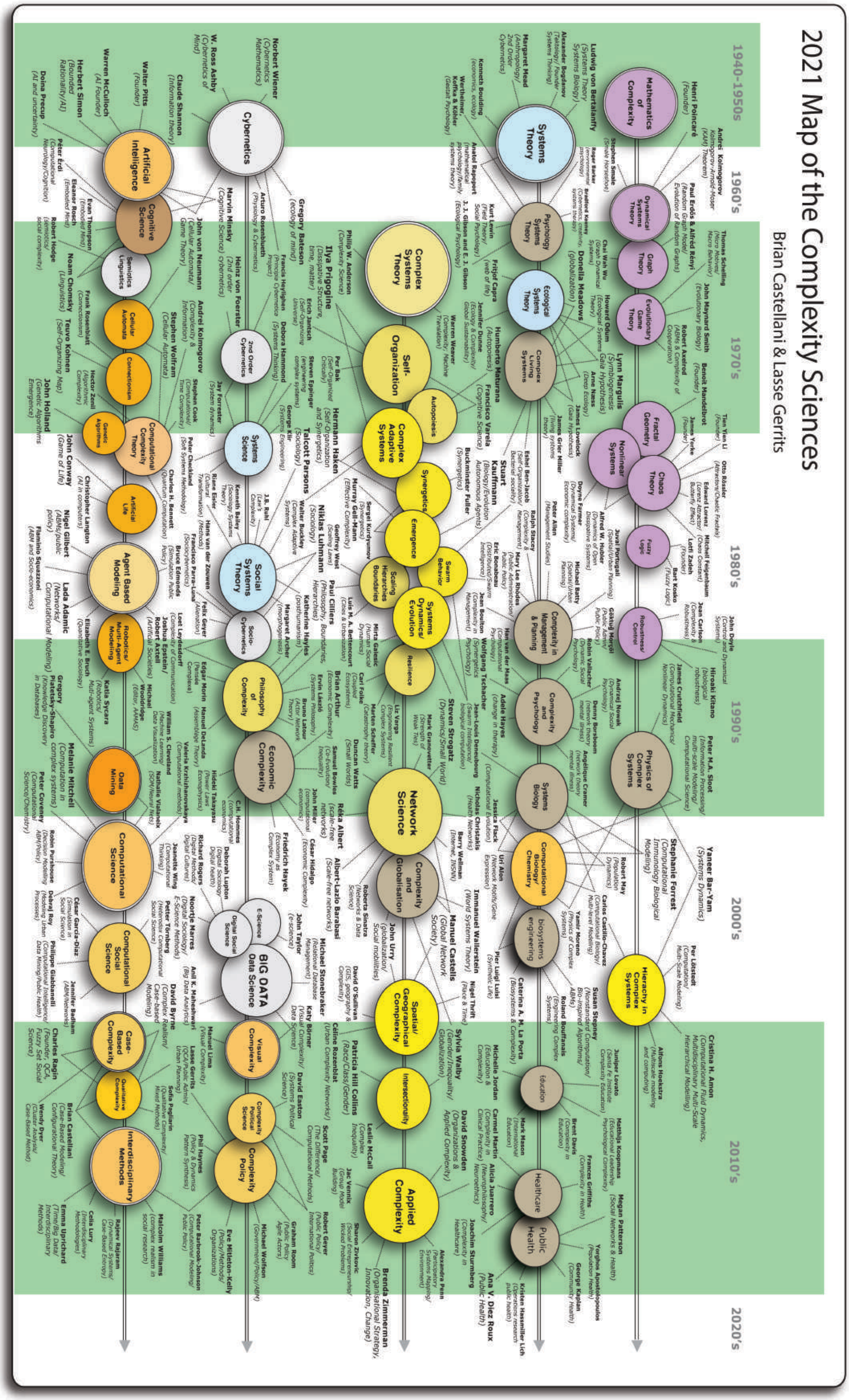
FRCCS

Map of the Complexity Sciences

Brian Castellani[1✓]

[1] *Durham University ; brian.c.castellani@durham.ac.uk*
[✓] *Presenting author*

Abstract. Moving from left to right, the map is read in a roughly historical fashion. The history of the complexity sciences is developed along five major intellectual traditions: dynamical systems theory (purple), systems science (BLUE), complex systems theory (YELLOW), cybernetics (GREY) and artificial intelligence (ORANGE). Placed along these traditions are the key scholarly themes and methods used across the complexity sciences. A theme's colour identifies the historical tradition with which it is "best" associated, even if a theme is placed on a different trajectory. Themes were placed roughly at the point they became a major area of study; recognizing that, from there forward, researchers have continued to work in that area, in one way or another. For example, while artificial intelligence (AI) gained significant momentum in the 1940s and therefore is placed near the start of the map, it remains a major field of study as of the 2020s. Themes in (BROWN) denote content/discipline specific topics, which illustrate how the complexity sciences are applied to different content. Finally, double-lined themes denote the intersection of a tradition with a new field of study, as in the case of visual complexity or agent-based modelling. It is important to point out that the positioning of scholars relative to an area of study does not mean they are from that time-period. It only means they are associated with that theme. Connected to themes are the scholars who "founded" or presently "exemplify" work in that area. In other instances, however, "up-and-coming scholars" are listed – mainly to draw attention to scholars early in their work. There was also an attempt to showcase research from around the world, rather than just the global north. Also, while some scholars have impacted multiple areas of study, given their position on the map only a few of their contributions can be visualized.


2021 Map of the Complexity Sciences
Brian Castellani & Lasse Gerrits
1940-1950s
1960's
1970's
1980's
1990's
2000's
2010's
2020's
Mathematics of Complexity
Systems Theory
Cybernetics
Artificial Intelligence
Complex Systems Theory
Self-Organization
Complex Adaptive Systems
Network Science
Computational Science
Computational Social Science
Economic Complexity
Philosophy of Complexity
Agent Based Modeling
Applied Complexity
Complexity Policy
Interdisciplinary Methods
Case-Based Complexity
BIG DATA Data Science
Physics of Complex Systems
Chaos Theory
Fractal Geometry
Social Systems Theory
Systems Science
Data Mining
Healthcare
Public Health

FRCCS

Pseudo-Fractals: Construction by stage-dependent rules

Andrew D. Irving[1] and Ebrahim L. Patel[2✓]

[1] *Freelance Researcher ; a_irving@btinternet.com*
[2] *University of Greenwich ; e.patel@greenwich.ac.uk*
[✓] *Presenting author*

Abstract. What makes a Fractal a Fractal? A rule of construction perhaps, one which determines each generation of structure. By convention, the same rule is applied to each generation. Our images adhere to this convention, each one guided by its own rule. Unconventionally though, each such rule is a function of g, the generation. Such an approach invites many lines of enquiry.

Inspired by von Koch's snowflake construction, we add a 'protrusion' to each line in our structure. Here, we use Asterisk-shaped seeds, some with just 3 prongs. As a first example, we add protrusions that are less symmetrical than von Koch's triangles. By adding a rhombus to the lines of our seeds, the result resembles a child's windmill toy. But what if the protrusion evolves from slanting 45 degrees 'to the left' to slanting 45 degrees 'to the right'?

Top left: We add a parallelogram protrusion to the central third of each line. Each protrusion makes an angle of $\pi/4 + (g-1)\,\pi/6$ radians with that line ($g = 1$ on the left, $g = 4$ on the right).

It is instructive to apply similar rules to simpler seeds.

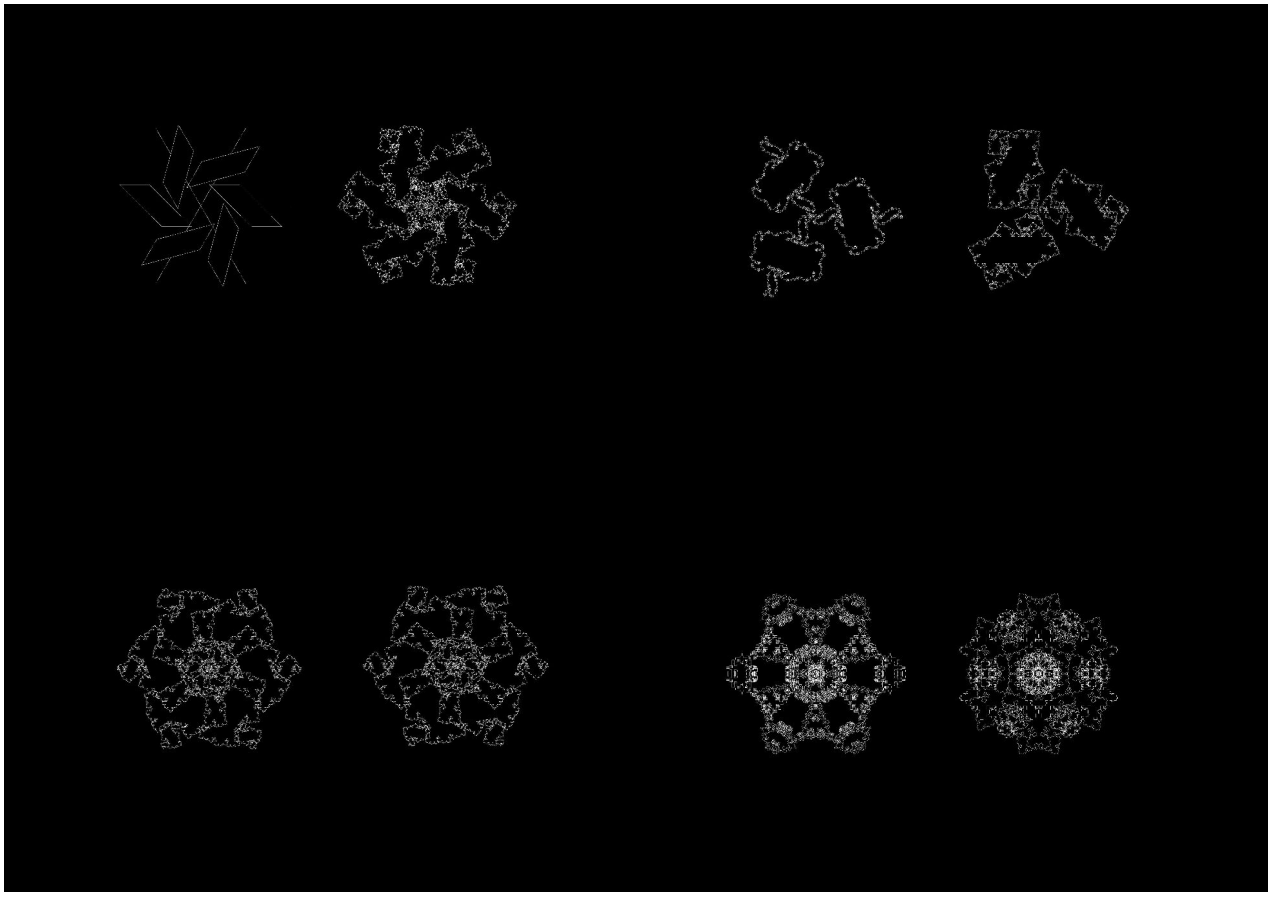
Top right: Each parallelogram protrusion makes an angle of $2\pi/18 + 6g\pi/18$ radians (left) and $5\pi/18 + g\pi/18$ radians with a line (right).

What if our rule of construction were cyclic?

Bottom left: We add a parallelogram protrusion to the central third of each line. Each protrusion makes an angle of $7\pi/18 + ((g-1) \bmod 2)\,4\pi/18$ radians (left) and $11\pi/18 - ((g-1) \bmod 2)\,4\pi/18$ radians with that line (right).

A rule of construction leads to a sequence of actions, one action at each generation. What would happen if such a sequence were reversed?

Bottom right: We add a rectangular protrusion to the central third of each line at each stage. This protrusion has length $(3\pi/12) + 2(g-1)\pi/12$ (left) and length $(9\pi/12) - 2(g-1)\pi/12$ (right). Going forwards, we wish to explore the dimensions of such structures.



FRCCS

Sarudango, selforganisation of extralarge huddling clusters in macaques

Cédric Sueur[1✓]

[1] *Université de Strasbourg ; cedric.sueur@iphc.cnrs.fr*
[✓] *Presenting author*

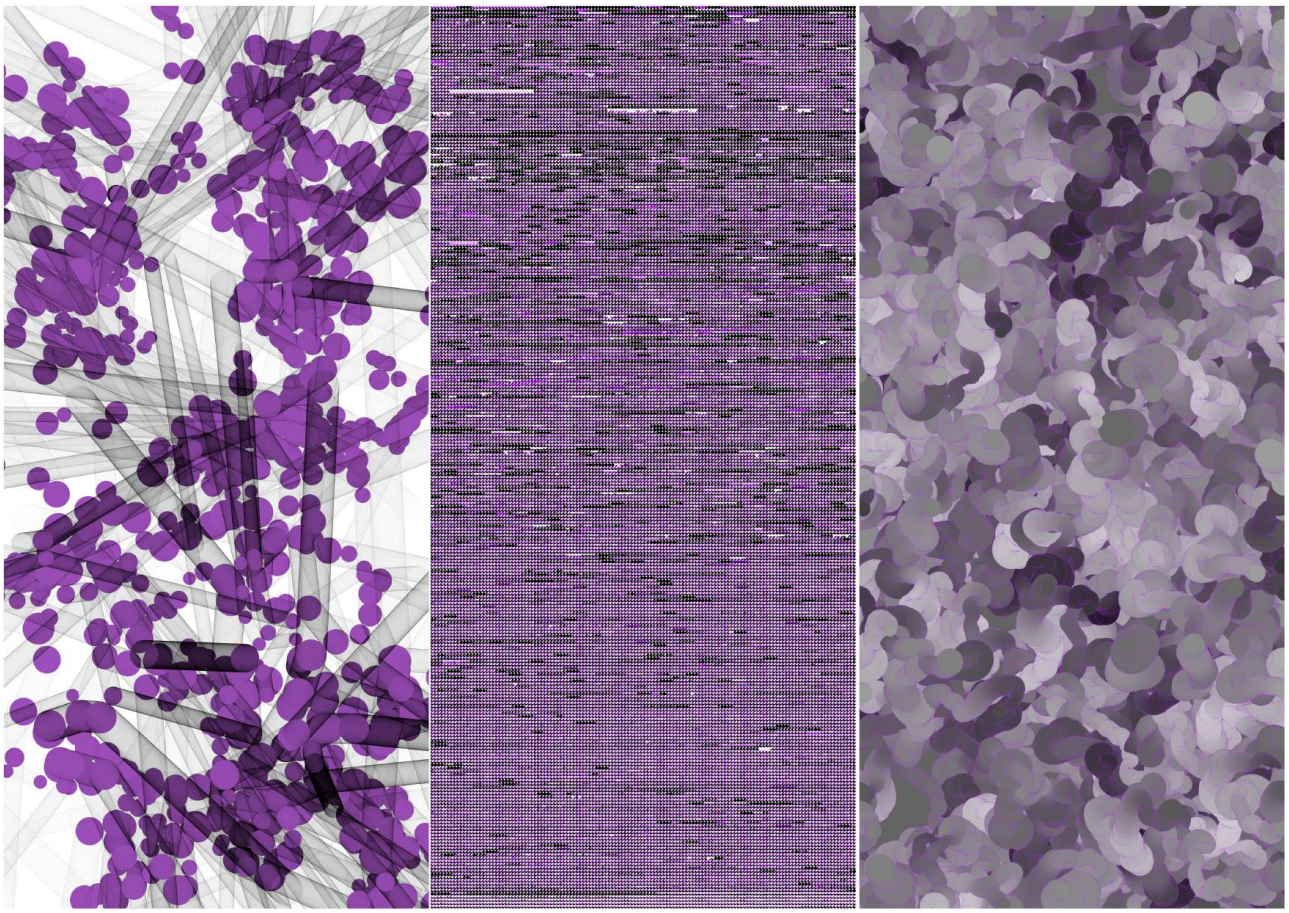
Abstract. Huddling behaviour is observed across various mammalian and avian species. Huddling, a behaviour wherein animals maintain close physical contact with conspecifics for warmth and social bonding, is widely documented among species in cold environments as a crucial thermoregulatory mechanism. Interestingly, on Shodoshima, Japanese macaques form exceptionally large huddling clusters, often exceeding 50 individuals, a significant deviation from the smaller groups observed in other populations (Arashiyama, Katsuyama, Taksakiyama) and climates. Our project aims to uncover the mechanisms behind the formation and size of these huddling clusters, proposing that such behaviours can be explained by simple probabilistic rules influenced by environmental conditions, the current cluster size, and individual decisions. Employing a computational model developed in Netlogo, we seek to demonstrate how emergent properties like the formation and dissolution of clusters arise from collective individual actions. We investigate whether the observed differences in huddling behaviour, particularly the larger cluster sizes on Shodoshima compared to those in colder habitats, reflect variations in social tolerance and cohesion. The model incorporates factors such as environmental temperature, cluster size, and individual decision-making, offering insights into the adaptability of social behaviours under environmental pressures. The findings suggest that temperature plays a crucial role in influencing huddling behaviour, with larger clusters forming in colder climates as individuals seek warmth. However, the study also highlights the importance of joining and leaving a cluster in terms of probability in the dynamics of huddling behavior. Social networks also play an important role. This study contributes to our understanding of complex social phenomena through the lens of self-organisation, illustrating how simple local interactions can give rise to intricate social structures and behaviours.

Science and complexity

Bruno Vianna[1✓]

[1] *CITM - Universitat Politècnica de Catalunya ; bruno.caldas@citm.upc.edu*
[✓] *Presenting author*

Abstract. This piece is inspired by the paper written in 1948 by Warren Weaver, named "Science and Complexity", which is one of the founding works of the complexity science.
This work is presented as a triptych where we have three panels reflecting the classification proposed by Weaver on his paper. On the left-hand side I depict what he calls "problems of simplicity", which is about classical mechanics and physics, like finding out the trajectory of objects and how they can be calculated using simple equations with few dimensions. The image shows circles that were moved randomly through the space. Whenever they bounced off each other, the color would change to purple. The trajectory of the circles is identified by the tracks that they leave.
The middle piece is on disorganized complexity, as Weaver says, which are problems that he proposed to be treated with statistical methods. I chose a simple and close subject to portray in these statistics, which is the memory of my computer. I took a snapshot of a random moment of the memory in my computer, made a dump of it, I analyzed the contents of 30,000 chunks approximately, The darker the circles, the more the chunks contain zero. Other integers distributions are represented by the intensity of the purple.
The last piece, the one on the right, reflects organized complexity. Again we have the circles running around and leaving tracks, but the difference is that these circles are not independent of each other anymore. They have behaviors, they are attracted to each other in some cases and repelled in some other cases, and that is what creates these intricate tracks which are represented in this image. They don't collide as in the first panel, because they can overlap. That would be the kind of science that was identified as organized because there are interdependent variables, and their relations will affect the final results.
I used the OpenProcessing platform for this work. The analysis of the memory dump was done in Python.



Index by Author

FRCCS 2024

French Regional Conference on Complex Systems

Montpellier, France
29-31 May

Applied Network Science

frontiers in Big Data

L'AVENIR DEPUIS 1220 UNIVERSITÉ DE MONTPELLIER

UNIVERSITÉ PAUL VALÉRY MONTPELLIER 3

tetis
TERRITOIRE ENVIRONNEMENT TÉLÉDÉTECTION INFORMATION SPATIALE

cirad

INRAE

LIRMM