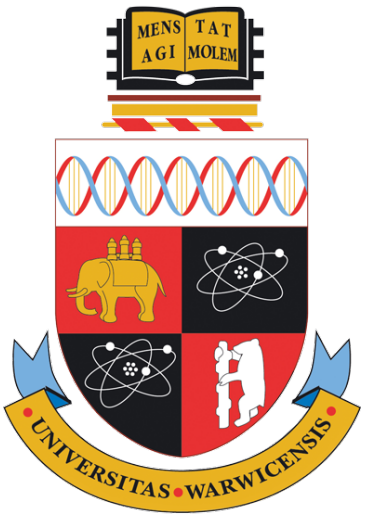
WARWICK
THE UNIVERSITY OF WARWICK

Network Science for Social and Technological Systems

by

Guillem Mosquera Doñate

Submitted to the University of Warwick
for admission to the degree of
Doctor of Philosophy



Centre for Complexity Science
September 2020

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank those who have helped me, consciously or inadvertently, during these last years. Going back to my first steps in research, I wish to thank Ricard Solé for inspiring me to become a physicist and study the fascinating world of complex systems. I am grateful to my supervisor, Samuel Johnson, for sharing his brilliant vision of complexity and his friendship. I would also like to acknowledge my second supervisor Weisi Guo, for his help and useful discussions.

I am thankful to my external collaborators. To Lluís Arola, for welcoming me in Tarragona and sharing interesting ideas and sleepless nights full of research. To Håvard Hegre and Mihai Croicu, two social scientist whom I admire, for treating me so well during my research visit in Uppsala.

Thanks to Ignasi Guillén and Ton Badal, whose refreshing non-academic points of view have helped me shaping ideas so many times. To both of you, and to the rest of my friends, thanks for being there. Knowing she will not be able to read this, I am also grateful to Maga, for taking me on walks when I needed them the most. For always helping me being myself and for their unconditional support, I am forever grateful to my parents.

To my beloved Helena, thank you for remaining patiently curious throughout the million scenes of our journey.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. This thesis has not been submitted for a degree at any other university. The work presented (including data generated and data analysis) was carried out by the author except in the cases outlined below, which contain work based on collaborative research:

- Chapter 4 was conducted with collaborators. The author contributed in research design, mathematical derivations, data analysis, numerical experimentation and writing of the associated paper.
- Chapter 5 was conducted with collaborators. The author contributed in research design, mathematical derivations, data analysis, numerical experimentation and proof-reading of the associated paper.

Parts of this thesis have been published by the author:

- Parts of Section 2.2 appear in:

  G. Mosquera-Doñate and M. Boguñá, "Follow the leader: Herding behavior in heterogeneous populations," *Physical Review E*, vol. 91, 2015.

- Chapter 3 is currently under peer-review for *Physical Review E*.
- Chapter 4 appears in:

  L. Arola-Fernández, G. Mosquera-Doñate, B. Steinegger, *et al.*, "Uncertainty propagation in complex networks: From noisy links to critical properties," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, 023 129, 2020.

- Chapter 5 appears in:

  A. Pagani, G. Mosquera, A. Alturki, *et al.*, "Resilience or robustness: Identifying topological vulnerabilities in rail networks," *Royal Society Open Science*, vol. 6, 181 301, 2019

Abstract

This thesis contains a collection of research outcomes from the field of complex networks. The results presented here have been divided in two parts, one devoted to theoretical methods and the other to data-driven applications. Although many of the results, especially in the first part, are general enough for describing many complex systems, a special focus on social systems has been used throughout the thesis.

The first part contains ideas that explore the interplay of topology and dynamics in complex systems, divided in three chapters dedicated to opinion dynamics, modular networks and weighted networks respectively. Regarding opinion dynamics, we study the emergence of self-organised leadership and herding behaviour in the voter model. Regarding modular networks, we present a generative model for networks with community structure and arbitrary bridgeness distribution. We also show how bridgeness interplays with functional behaviour in different dynamical systems. We use such interplay to define the concept of dynamical centrality, and show its applications to network dismantling under limited topological information. Finally, we demonstrate how topological uncertainty in link weights induces fluctuations on the critical threshold for multiple dynamical processes on networks. We also discuss the role of degree heterogeneity in this propagation, finding non-trivial dependencies for scale-free networks.

The second part contains two applications of network analysis to real-world systems. The first application is a data study on the rail network of London and its surrounding area. We show how topological resilience measures are strongly correlated to the performance of train operators in the network. The second application contains a network-based model of armed conflict prediction at city level of analysis. We use several centrality measures as features for machine learning models, showing how network information generates very significant improvements in out-of-sample prediction performance.

To my grandfather Joan

Chapter 1

Introduction

"... Thus you see, most noble Sir, how this type of solution bears little relationship to mathematics, and I do not understand why you expect a mathematician to produce it, rather than anyone else, for the solution is based on reason alone, and its discovery does not depend on any mathematical principle. Because of this, I do not know why even questions which bear so little relationship to mathematics are solved more quickly by mathematicians than by others."

— Leonhard Euler, letter to Carl Gottlieb Ehler 1736

"Psychohistory was the quintessence of sociology; it was the science of human behavior reduced to mathematical equations. The individual human being is unpredictable, but the reaction of human mobs, Seldon found, could be treated statistically. The larger the mob, the greater the accuracy that could be achieved"

— Isaac Asimov, *Second Foundation* (Prologue)

1.1 Scope and structure of this thesis

1.1.1 Scope

Society and its relation to the individual have been imagined, studied and reasoned about from all artistic and philosophical perspectives ever conceived since the dawn of society itself. Far from a peaceful topic, it seems that a conflict has persisted in all sociological endeavours up to modern days, a dialect between positivists and anti-positivists. Sociological positivism promotes the unity of the scientific method, directly applying methodological tenets from the natural sciences in the study of society. Challenging this view, anti-positivism considers a dualism between the natural and cultural world, regarding human behaviour as so irreducibly complex and special that its study requires methods of its own [4].

It is hard to judge who holds the higher ground, but nowadays we are undeniably used to the divide between so-called social and natural sciences — most likely a victory for team anti-positivist. Even today, young students are quickly segregated between those who will study the social world and those who will investigate the natural. For many, bringing the rules and ways of nature to the inquiries on society feels irreverent, almost unnatural.

But there have been numerous voices over the centuries who have claimed to derive the arithmetic or mechanics of society. Grounded on the believe that it could only be described to the extent that it could be quantified, entire disciplines such as economics or statistics have emerged from the positivist drive for measuring human behaviour. Seeking parallelisms with successful constructs such as Newtonian mechanics or Calculus, such voices have tried to build over-arching theories only based on a handful of fundamental principles of society.

The reader should not expect to find any of such "theories of society" in this thesis. Far from it, this is a dissertation about the relatively young field of complex networks, where some of its recent developments and techniques are reviewed and discussed. There is a strong research focus on the interplay between structural and dynamical features of networks, based on the assumption that complex systems can only be interpreted if the higher-order structure of their components' interactions are understood. The thesis has also been developed under a wide-angle lens in terms of the particular subfields that have been explored. From opinion dynamics, community detection, weighted networks, critical phenomena, transport networks or even armed conflict prediction, it is not easy to define a unitary narrative for all the work herein exposed.

In all truth, this variety is a reflection of how the author's curiosity has refused to focus on a single field of application. However, this also reflects the subjective perception that networks are a necessary part of virtually any phenomena we are exposed to in our daily lives. Going back to the initial discussion, the common hypothesis exploited throughout

all chapters below is that the theories and methods of complex networks can be used to model, describe and predict complex systems and, in particular, are a fundamental instrument to define the fabric of society.

1.1.2 Structure

Beyond the present introductory chapter, which provides below a very short context to complex systems and networks, the thesis is structured around two building blocks. The first part exposes theoretical methods developed around the concept of interplay between topology and dynamics in complex networks. By theoretical methods we mean mathematical models that do not contain direct empirical considerations using real-world data. These models clearly show how from relatively simple changes in the structure of complex networks unexpected or intricate dynamics may emerge. We believe our theoretical methods can illustrate how complex social phenomena is the result of how humans interact across different network scales. The second part of the thesis focuses on applications of complex networks as mathematical tools for the analysis of real-world social systems using data-driven methods. Below we provide detail on how these ideas are structured around specific chapters.

Theoretical Methods

Chapter 2 tackles the problem of opinion dynamics, a prototypic example of how methods from physics are used to study society. Opinions and beliefs are at the core of individual human behaviour, but at the same time are the result of complex dynamics of social influence. From a physical point of view, opinions can be regarded as the internal state of individuals interacting in social networks of influence. This is what the voter model does, a mathematical model that has gained popularity and success despite its numerous limitations and reductionism. One such limitation is the diffusive behaviour observed in the average opinion for the standard voter model. In this chapter, we provide a minimal set of heterogeneous influence structures that allow the voter model to exhibit herding behaviour, that is, rapid non-diffusive shifts on the average opinion promoted by emergent leaders in a population of voters. As a result of our mean-field approach, here we do not use an explicit network formalism, but our results can only be effectively understood in the context of a network of influences, e.g. in populations where popularity and activity are both related to connectivity.

Chapter 3 relates to the study of modular structures, which are pervasive in social, technological and biological networks. In particular, we study how cross-community link patterns affect the dynamical behaviour of a network. Such patterns are studied using bridgeness centrality, which measures to which extent each node participates in the different structural modules of a network. We present a generative network model of

community structure that controls the distribution of bridgeness in a network. Using two paradigmatic models of statistical physics, that of Potts spins and that of Kuramoto oscillators, we reveal an important interplay between the dynamical behaviour of individual nodes and their bridgeness centrality. In fact, we use such interplay to derive two novel measures of what we call dynamical centrality. Dynamical centralities are measures of local order parameters that allow us to differentiate the structural centrality of nodes just by observing their dynamical behaviour, without explicitly knowing the underlying connectivity. We show how they can be used, for example, to efficiently attack and dismantle networks even when we cannot observe their connections. Some of the concepts from this chapter are later used and tested using real-world data from Chapter 6.

Chapter 4 introduces a novel topic in complex networks, namely uncertainty propagation. Dynamical observations in real-world systems are usually noisy and fluctuating, driven in many cases by uncertainty on the structure of the underlying networks. Networks can fluctuate at two different scales: by addition or removal of links, or by uncertainty on the interaction weight of existent links. Here we focus on the latter, illustrating how weight uncertainty can be propagated towards dynamics, particularly on the critical threshold of physical models with phase transitions. We show how the critical range (i.e. the uncertainty in the critical threshold) of a network depends on the heterogeneity of its degree structure. Despite being very theoretically-oriented, the results of this chapter shed light into the important topic of network measurement-error and its consequences for real-world applications.

Applications

Chapter 5 shows a real-world application of the interplay between topology and dynamics in rail transport systems. We use several global network measures of resilience (vulnerability to cascade delays) and robustness (vulnerability to closure of stations) to the rail network in Greater London and surrounding commuter areas. We use public data on performance measures of several railway operators as proxies for resilience and robustness. We find that vulnerabilities to cascade delays are the most important topological factor related to the performance of train operators.

Finally, Chapter 6 presents our work on armed conflict research, a pioneering interdisciplinary field dealing with one of the most difficult to understand and catastrophic complex phenomena in social systems. Although initially a subfield of political sciences and international relations theory, peace and conflict research is nowadays in close contact with mathematics and statistics. We develop a novel framework for conflict prediction based on network models of geographical interactions of cities around the globe. Using several network centrality measures, including bridgeness (as studied in Chapter 3), we show how our network models have very significant out-of-sample performance in predicting armed conflict using conflict data from the last 30 years.

1.2 Perspectives on complex systems

1.2.1 Scientific landscape

The relatively young science of complex systems as we know it today brings together multiple disciplines in a pursue of laws and principles encompassing all imaginable scales, from buzzing molecular worlds to crowded societies and the echos of history. The study of complexity as an ultimate goal of science, however, is an old endeavour. Let us go back, for example, to the days of René Descartes when science itself was almost a branch of philosophy. Those were also the days of Leeuwenhoek and Hooke, the fathers of microbiology and pioneering inventors of microscopy. One could argue that a scientific journey, from the very top of our day-to-day human scale, down to the depths of the smallest components of the real world, was just getting started. This was the journey of *reductionism*, the quest of understanding complexity by analytically decomposing systems down their fundamental most basic constituents. Fast forward to the dawn of the twentieth century and we reach one of the pinnacles of this approach, quantum physics, which in essence describes the building blocks of matter and energy.

The reductionist approach to particle physics has lead the discipline to incredible levels of precision and detailed understanding of unimaginable phenomena, yielding numerous outcomes both in terms of mathematical understanding of reality and in terms of technological advances that have radically affected society. The journey is far from ending, and so-called string theories might still bring us closer to the dreams of a unified theory of fundamental interactions: a theory built from the study of the most fundamental elements in nature, that would provide answers about the birth of the universe or even the existence of multiple universes. As for biology, however, despite the revolutionary knowledge acquired throughout the path towards the micro (e.g. molecular biology and its open ended possibilities for medicine, molecular genetics and its relation to evolution, etc.), it seems unlikely that the answer to what life is hides in ever smaller parts of the cell. What are we missing?

The reductionist approach has failed to explain how physical phenomena emerge from the interactions of building blocks across scales, that being in inside the cell, the brain, society or rainforests. The recognition that it is the exchange of information that brings complexity to many real-world systems has produced a methodological shift in science. The end goal of the study of complex systems is to find those universal laws and mechanisms that explain in simple mathematical terms how information propagates across non-linear interactions giving rise to emergence, self-organisation and collective phenomena.

In summary, it is worth noting that the journey from top to bottom (reductionism) and back to the top (emergentism) has been far from reversible. Lessons learned during each section of such never-ending trip have a permanent effect in the way science is

conducted, with the composability of these different approaches being a testimony of complexity itself.

1.2.2 Social systems as complex systems

Humans are unpredictable but human collective behaviour can be predicted. Or, at least, that is the main hypothesis of sociophysics [5], [6]. In analogy with (non-equilibrium) thermodynamics, the hope is that collective social phenomena can be described as emerging properties of systems of interacting agents: the aim is to find a description of social phenomena that depend on few fundamental features of the microscopic interaction laws rather than on the idiosyncratic character of single individuals. This point of view has lead to the study of social dynamics using the same tools and models that statistical physics has been developing during the last fifty years [7]. In fact, statistical physics has played a key methodological role in the dawn of complexity science in general, with wide-range adoption and application in biology, computer sciences, urban modelling or medicine, to name a few disciplines beyond the traditional applications of physics such as thermodynamics, electronics, magnetism or superconductivity.

How much do we need to know about individual humans in order to produce useful models of social systems? The approach of sociophysics is often criticised on the basis of being too simple to represent human individuals. This is an important critique, but we should not expect to ground a science of society on perfect knowledge about individuals, for that could be too reductionist. In fact, as in many other real-world systems, the collective properties of society are relatively independent of the internal mechanics of human beings. In this sense, social phenomena exhibits *universality*.

What fundamental points of contact can we find between the particles of statistical physics and humans? Order is perhaps the most important of them. The very pillars of society are ordered structures that emerge from apparent disorder: language, culture, political consensus. They all require the formation of statistical regularities that cannot be explained without interactions. Both for particles and humans, interactions are the real key for understanding the emergence of order from initial disordered states.

In terms of methodology, there are several tools recurrently used in sociophysics. The first is the aforementioned order/disorder transition paradigm. Many models can be seen as extensions of the Ising and Potts models, in that they seek to study the ordering process of some internal state of interacting agents using particular microscopic mechanisms that want to reflect different social contexts. A different approach is that of *sociodynamics*, a subfield focused on describing directly macroscopic societal variables using dynamical and probabilistic models based on the master equation, without a particular description of microscopic mechanics [8]. Agent-based models provide yet another approach, with a computationally-based methodology that studies emergence in social

systems using computer simulations [9]. Agent-based models can naturally deal with complex internal descriptions of agents and heterogeneous interactions, but on the other side they are difficult to use in prediction settings and are prone to over-parametrisation. Finally, the role of the structure and topology of social interactions is of great importance for sociophysics. A dedicated space can be found on Section 1.3 discussing the implications and interfaces between network science and sociophysics.

1.3 The network paradigm

It is a sunny Sunday afternoon in the old beautiful town of Königsberg in 1735 and most of its citizens happily walk the streets, unaware that a Swiss mathematician is tracing their steps. The lore says that the people of Königsberg had a gamble in which they tried to devise a route that would allow someone to walk around the city without crossing any of its seven bridges more than once. None could devise such a path, but even more frustrating was the fact that none could prove that the problem was unsolvable. But along came Leonhard Euler, who in that same year published the solution to the problem. Euler realised that the problem could be abstracted by only considering which land masses where connected to which other, and by how many bridges. He proved that a necessary condition for such a path to exist in any finite graph is that all vertices (land masses) have an even number of connections (bridges). Eulerian cycles provided what is considered to be the first historical use of graph theory.

But network science as a discipline of its own did not emerge until the turn of the twentieth-first century — considering as a rough starting point, for example, the influential *Reviews of Modern Physics* by Albert and Barabási [10], one of the most cited paper in the history of the journal. What pieces were missing so that this shift could happen? In truth, during this two centuries, many developments were achieved in the field of graph theory, including the instrumental work on random graphs by Erdős and Rényi. Other disciplines, actually, started using networks for their own purposes. For instance, ideas from psychology, anthropology, sociology and graph theory began encompassing as early as the nineteen thirties [11]. In fact, the subfield of social networks was already maturing in sociology by nineteen seventies, with very influential papers such as Granovetter's [12]. The reality is,however, that all these efforts were missing large-scale empirical data sources, and the computational power to process them, in order to uncover some of the ideas that would later develop into network science.

The first decade of the twentieth-first century saw an exponential increase in research activity regarding complex networks. Theoretical advances were met by increasingly larger data-collection efforts, such as the development of the first large-scale Internet mapping, the publication of protein-protein interaction databases, the Human Connectome project, or the rapid development and sampling of online social networks such as

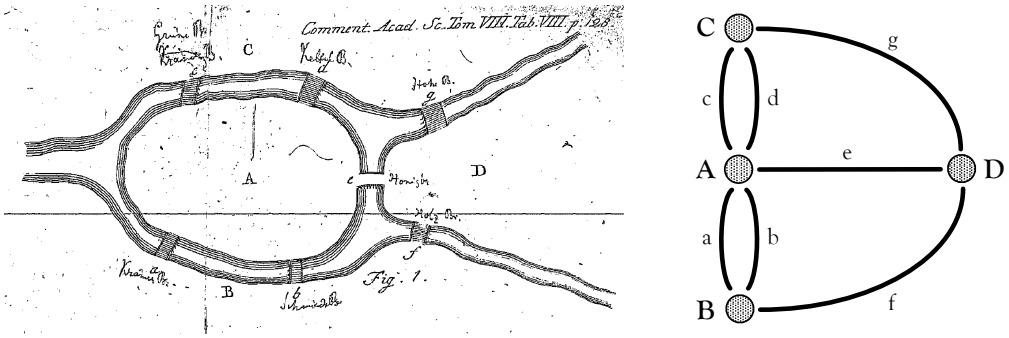


Figure 1.1: The seven bridges of Königsberg. This iconic mathematical problem can be solved using a graph in which nodes (A-D) and edges (a-g) represent, respectively, land masses and bridges.

Facebook. One of the largest steps forward from this recent period was the realisation that networks with vast divergences, in terms of their components (proteins, humans, rail stations, servers, etc.) and their generating process (evolution, social norm and friendship, local urban planning, individuals setting up servers, etc.), share a reduced number of simple organisational principles and internal dynamics, thus resulting in universal mathematical properties that can be validated empirically.

Nowadays, this universality has transformed complex networks into a multidisciplinary methodology that brings together physicists, biologists, sociologists, economists and computer scientist generating a immense body of knowledge. The study of networks is producing very important impact on society in terms of its implications to how we communicate over Internet, how we move and transport goods efficiently, how we understand the emerging properties of the brain, or how we combat diseases and epidemics. From gathering data, building mathematical models, using computer simulations to solve those models, and testing the predictions produced by them with the data gathered initially, it is easy to claim the existence of a science of networks.

1.4 A minimal toolkit for our network analysis

Here we present some of the mathematical concepts related to network analysis we use more frequently throughout the present work. This is not intended as a general introduction to or a comprehensive review of network theory (for which we refer the reader to the extensive body of generalist literature on the topic [13]–[15]), but instead a primer on the concepts needed along the journey of this thesis.

1.4.1 Definitions on graphs

Networks are physical objects measurable in real-world systems. *Graphs* are their mathematical abstraction, conforming the vast field of graph theory, which is the starting point for any numerical or analytical study of network properties. A graph is typically denoted by G and contains two sets $G = (V, E)$, that of *vertices* or nodes V, and that of *edges* or links E. A graph $G' = (V', E')$ is called a *subgraph* of G if $V' \subset V$ and $E' \subset E$ The convention in network science is to consider the size or order of a graph to be the cardinality of the set of nodes $N = |V|$, i.e. the number of nodes in the network. Indeed, the total number of possible edges or maximal cardinality of the edge set is bounded by N in that a graph with $|E| = \binom{N}{2}$ is called a *complete* graph. The *density* of a graph represents the current number of edges divided by the total possible number of edges, $D = |E|/[N(N-1)/2]$. A graph is called *sparse* when $D \ll 1$, a property exhibited by most real-world networks.

Two nodes v_i and v_j (sometimes we will directly use the simpler notation i and j to refer to them) are called adjacent or said to be *neighbours* if there exists an edge (i, j) connecting them. In fact, graphs are usually defined through their *adjacency matrix* $A = \{a_{ij}\}$:

$$a_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{if } (i,j) \notin E \end{cases} . \tag{1.1}$$

Graphs where $a_{ij} = a_{ji}$ are called *undirected*, because their edges do not contain directionality and can be depicted as simple lines. When the adjacency matrix is not necessarily symmetrical, graphs are called *directed* and their edges are usually represented graphically as arrows.

1.4.2 Connectivity

Most network properties are related to how nodes connect and reach to each other. Locally, this is clearly represented by links, but connectivity tends to be related to many different scales beyond locality. Higher-order interactions are crucially represented by *paths* P_{v_0,v_n} between a source node v_0 and a destination node V_n. Paths can be represented by the subset of nodes needed to reach destination from source using existing links, $P_{v_0,v_n} = \{v_0, v_1, \ldots, v_{n-1}, v_n\}$. A class of path where source and destination nodes are the same is called a cycle or *loop*.

A graph is said to be *connected* if there exists a path connecting any two nodes in it. A connected subgraph is known as a *component*, and two components are *disconnected* if we cannot build any path between nodes of such different components. Studying the distribution and sizes of components is an important part of the analysis of real-world networks. In particular, it is common to study the size of the largest component in a

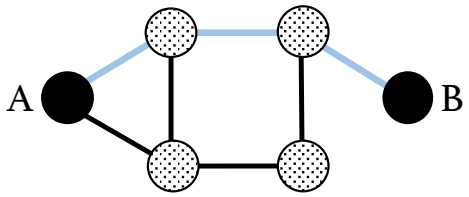


Figure 1.2: The shortest path between nodes A and B, represented by blue lines, contains three edges. Therefore, the distance between A and B is $\ell_{AB} = 3$.

network, which is known as *giant component* when its size scales with the number of nodes, thus diverging in the thermodynamic limit $N \to \infty$.

Paths are also instrumental in defining notions of distance in a graph. The distance ℓ_{ij} between two nodes i and j is typically defined as the minimal number of edges traversed in paths connecting i and j, as shown in Figure 1.2. Note that if a path between i and j does not exists, $\ell_{ij} = \infty$ by convention. Such minimal paths are known as *shortest paths*. Several measures can be readily computed using shortest paths in order to gain an understanding of the shape and scale of a network. For instance, the network *diameter* is defined as:

$$d_G = \max_{i,j} \ell_{ij} \quad . \tag{1.2}$$

We can complement our notion of distance using statistical moments of ℓ_{ij}, such as the average shortest path length:

$$\langle \ell \rangle = \frac{1}{N(N-1)} \sum_{ij} \ell_{ij} \quad . \tag{1.3}$$

1.4.3 Centrality measures

Networks are used to model real-world systems where it is important to understand the position or role of each node with respect to the collective. Numerous definitions with different criteria exist sharing the purpose of defining and discerning important nodes, and they are commonly known as *centrality measures*. Below we define four of the most regularly used measures, which we will employ in different sections of the present thesis.

Degree

One of the most basic and useful centrality measures, *degree* examines how well connected a node is locally in terms of single-step paths. For undirected networks, degree k_i is simply the number of links attached to a given node i:

$$k_i = \sum_j a_{ij} \quad . \tag{1.4}$$

For directed networks, we define in-degree and out-degree as:

$$k_i^{in} = \sum_j a_{ji}, \qquad k_i^{out} = \sum_j a_{ij} \quad . \tag{1.5}$$

In general, examining its degree statistical properties is a recurrent task when analysing a network. In this sense, it is useful to study the probability distribution:

$$p_k = \frac{N_k}{N} \quad , \tag{1.6}$$

where $N_k = \sum_i \delta_{kk_i}$ is the number of nodes with k-degree, and can be expressed in terms of the adjacency matrix through Eq. 1.4. The first-moment or average degree is also a measure of link density in that:

$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2|E|}{N} \tag{1.7}$$

Closeness

Closeness centrality measures the inverse of the average distance from a node i to any other node j:

$$C_i = \frac{1}{\sum_{j \neq i} \ell_{ij}} \quad . \tag{1.8}$$

Not that the expression above is only valid for connected networks. An alternative definition that can be used for unconnected networks is:

$$C_i' = \sum_{j \neq i} \frac{1}{\ell_{ij}} \quad . \tag{1.9}$$

Those nodes that are generally more accessible to the rest of the network via shortest paths will have higher importance in terms of closeness centrality.

Betweenness

Betweenness centrality measures the number of shortest paths traversing a given node:

$$B_i = \sum_{m \neq j \neq i} \frac{\sigma_{mj}(i)}{\sigma_{mj}} \quad , \tag{1.10}$$

where σ_{mj} is the number of possible shortest paths between m and j, and $\sigma_{mj}(i)$ refers to the more restricted set of shortest paths between m and j that go through i. Betweenness highlights those nodes that act as bottlenecks for efficient information flows across the network. In transport networks, for instance, betweenness provides an estimation of the load or traffic expected in a given node, assuming transport occurs through shortest paths.

PageRank

PageRank centrality became highly successful because of its role in the inception of Google [16]. It is recursively defined as:

$$PR_i = \alpha \sum_j \frac{a_{ji}}{k_j^{out}} PR_j + \frac{1-\alpha}{N} \quad , \tag{1.11}$$

where α is known as the dumping factor, which controls the extent by which PageRank centrality of one node depends on others' centrality.

1.4.4 Weighted networks

Link weights are a very important degree of freedom present in most real-world networks [17]. Instead of having binary connections between nodes, networks can have intensities regulating interaction strength. Weights tend to be inherent to the measurement of some real-world networks, where instead of observing a static binary snapshot we sample several observations and infer an interaction probability from them. They can also be representing physical features of the interaction medium, such as bandwith in information or energy transport, passenger capacity in transport networks, amount of trade in international networks or traffic in the Internet.

In any case, adding a weighting structure has a profound impact on all features of a complex network, particularly on centrality measures. The most immediate local measure of weighting structure is *strength* or weighted degree:

$$s_i = \sum_{j \in V(i)} w_{ij} \quad , \tag{1.12}$$

where w_{ij} is the weight of edge (i, j), and $V(i)$ is the set of neighbours of node i. Furthermore, since some centrality measures such as closeness or betweenness depend on the concept of distance, it is important to define the relation between weights and shortest path lengths. The convention we use throughout the thesis is to consider w_{ij}^{-1} as the inherent distance of edge (i, j). Thus, when considering the length of a path we need to sum up the inverse of the weight of all edges conforming such path. Finally, it is straightforward to generalise PageRank definition in Eq. 1.11 to weighted networks by using the weighting matrix w_{ij} instead of adjacency a_{ij}, and strength s_j instead of degree k_j.

1.4.5 Community structure

Communities are groups of nodes that have significantly more internal interactions than external. They are also referred to as clusters or modules. An extreme instance of a community would be a complete subgraph with no edges connecting it with the rest of the

network. But communities are usually defined in much more ambivalent circumstances, where the aforementioned subgraph would have some connections with the rest of the network, and the internal edge density would be lower than in a complete subgraph. Community structures are prevalent in real-world networks, especially in biological (e.g. functional modules in protein-protein interaction networks performing different cellular functions [18]) and social networks (e.g. the famous Zachary's Karate Club network [19]).

The problem of partitioning a network into meaningful communities is referred to in the literature as *community detection*. Given that the number of partitions (combinations of communities) scales super-exponentially with the number of nodes in a network, community detection is an NP-hard problem where inspecting all possible solutions quickly becomes impossible. In the last decades, a large number of heuristic algorithms that do not need to check all possible partitions have emerged, but the most successful ones are methodologically based on hierarchical clustering. Either by agglomerative procedures [20] (start with very small communities and merge them) or divisive methods [21] (start with very large communities and split them), hierarchical clustering requires a quality-function in order to compare partitions and determine the optimal cut of the hierarchy. For this purpose the usual measure is *modularity*, which essentially compares the internal density of communities with the random expectation in the network [22]. However, given that the number of possible partitions is so large, modularity optimisation often leads to over-fitting. This occurs because heuristic algorithms cannot distinguish fluctuations in edge density from real generative mechanisms producing communities: in fact, modularity maximisation algorithms may find optimal partitions even in completely random networks. On the contrary, inferential algorithms based on the stochastic block model [23] are capable of efficiently finding statistically significant communities (see Section 3.1.2 for details), and are gaining popularity in recent years. A further problem arises when considering that nodes in real-world networks usually belong to more than one community, such as in the case of friendship networks where individuals are well connected to several groups of friends (from school, from work, from family, etc.). This is known as *overlapping community* structures. Interesting algorithms have also been devised for this case, including the clique percolation method [24], link clustering [25], and also the inferential stochastic block model approach [26].

Part I

Theoretical Methods

Chapter 2

Emergent herding behaviour

Are important societal events driven by smooth, homogeneous or diffusive changes in the opinion of people? Most frequently not. Financial crashes, unexpected election outcomes or rapidly escalating social polarisation are all events that require particular opinions to spread massively in short time scales — what is known as herding behaviour. This phenomenon is closely related to the emergence of leadership in social systems. What are the minimal influence structures required for herding behaviour to emerge in a social network?

Models of opinion dynamics typically show consensus states where the dynamics is frozen. In many cases, like in the voter [27], [28] or Sznajd models, the (weighted) ensemble average opinion of the population is a conserved quantity. In such cases, the dynamics of the stochastic average opinion is governed by a purely (non-homogeneous) diffusive process without any drift, which eventually leads to one of the possible consensus states. It is therefore difficult to imagine how leadership can emerge in this context. In this chapter, we show that leadership can, in fact, spontaneously arise in a subset of the population when there is a strong heterogeneity in the time scales of the agents coupled with a hierarchical organization of their influence. Heterogeneity of time scales is present, for example, in speculative markets, where noise traders operating at the scale of minutes or hours coexist with fundamentalists, doing so at the scale of weeks or months. Interestingly, we discover a pitchfork bifurcation separating a purely diffusive phase and a phase where the most active agents lead the global state of the entire population. This result can shed light on the dynamics of extreme events driven by human opinion.

2.1 The voter model

The voter model was first introduced in 1973 to model competition between species [27], [28]. Ever since, it has become one of the most paradigmatic and popular models of

opinion dynamics. Its simplicity, analytical tractability, and versatility to introduce new mechanisms make it the perfect model to study many different phenomena in the natural and social sciences, from catalytic reaction models [29], [30] to the evolution of bilingualism [31] or the statistics of the US presidential elections [32]. In its most simple version, the voter model is defined as follows: we have a set of N interacting agents, each endowed with a binary state of opinion (sell or buy, democrat or republican, window or mac, etc). At each time step of the simulation, an agent is randomly chosen to interact with one of her social contacts, after which the agent copies the opinion of her neighbor.

Heterogeneity can be introduced in the population through the activity rate of agents [33], [34]. We assume that agents are given intrinsic activity rates $\{\lambda_i\}$, controlling the frequency at which they interact with their social contacts and, possible change their opinion. In numerical simulations, this is equivalent to chose the next active agent, say agent i, with probability proportional to λ_i. The influence of one agent over others can be modeled by the probability $\text{Prob}(j|i)$ that agent i copies the opinion of agent j when i is activated at rate λ_i. When contacts take place through a fixed social contact graph with adjacency matrix a_{ij}, this probability is given by $\text{Prob}(j|i) = a_{ij}/k_i$, where k_i is the degree of agent i [35]–[38]. In a fully connected graph (equivalent to a mean-field description), this probability is simply $\text{Prob}(j|i) = 1/(N-1)$ for $j \neq i$ and zero otherwise.

The dynamics of the state of the system can be described using a set of N dichotomous stochastic processes $\{n_i(t)\}$ taking values 0 or 1 depending on the opinion state of each agent at time t. If we assume that all temporal processes follow Poisson statistics, the stochastic evolution of $n_i(t)$ after an increment of time dt satisfies the stochastic equation [39], [40]

$$n_i(t+dt) = n_i(t)\,[1-\xi_i(t)] + \eta_i(t)\xi_i(t), \tag{2.1}$$

where $\xi_i(t)$ is a dichotomous random variable taking values

$$\xi_i(t) = \begin{cases} 1 & \text{with probability } \lambda_i dt \\ 0 & \text{with probability } 1-\lambda_i dt \end{cases}. \tag{2.2}$$

Notice that $\xi_i(t)$ controls whether node i is activated during the time interval $(t, t+dt)$. If so, the opinion of a neighbor will be chosen according to $\text{Prob}(j|i)$ so that

$$\eta_i(t) = \begin{cases} 1 & \text{with probability } \displaystyle\sum_{j=1}^{N} \text{Prob}(j|i) n_j(t) \\ 0 & \text{with probability } \displaystyle 1 - \sum_{j=1}^{N} \text{Prob}(j|i) n_j(t) \end{cases}. \tag{2.3}$$

In principle, $\eta_i(t)$ should be realized only when $\xi_i(t) = 1$. However, due to the particular form of Eq. (2.1), the value of $\eta_i(t)$ is only relevant when $\xi_i(t) = 1$. Therefore, we can safely consider $\xi_i(t)$ and $\eta_i(t)$ as statistically independent random variables.

Equation (2.1), supplemented with the definitions of variables $\xi_i(t)$ and $\eta_i(t)$, represents the exact stochastic evolution of the system. For instance, the ensemble average of the opinion of agent i, $\rho_i(t) \equiv \langle n_i(t) \rangle$ can be evaluated by taking first the average of Eq. (2.1) over the variables $\xi_i(t)$ and $\eta_i(t)$ and, then, over the ensemble. This program leads to the exact differential equation

$$\frac{d\rho_i}{dt} = \lambda_i \left[\sum_{j=1}^{N} \text{Prob}(j|i)\rho_j - \rho_i \right]. \tag{2.4}$$

This equation implies the existence of a global conserved magnitude [36], [41] related to the eigenvector $\phi(i)$ of eigenvalue 1 of $\text{Prob}(j|i)$, that is, the solution of the equation $\sum_i \phi(i)\text{Prob}(j|i) = \phi(j)$. Indeed, by multiplying Eq. (2.4) by $\phi(i)/\lambda_i$ and summing over all agents, the right side of the equation vanishes. Therefore, the weighted ensemble average of the population

$$\Phi \equiv \sum_{i=1}^{N} \frac{\phi(i)}{\lambda_i}\rho_i(t) = \sum_{i=1}^{N} \frac{\phi(i)}{\lambda_i}\rho_i(0) \tag{2.5}$$

is conserved by the dynamics and, thus, it is a function only of the initial conditions. This fact can be used to evaluate the probability of the final fate of a realization of the dynamics. The probability to end up absorbed in the "1" consensus state is just given by $\Phi/\sum_i \phi(i)/\lambda_i$.

2.2 Emergence of leadership in the voter model

The results presented so far are valid for an arbitrary distribution of individual rates λ_i. However, the behavior of the system can be very different when there is a strong separation of time scales present in the system, like in speculative markets with noise traders and fundamentalists. To shed light on this problem, hereafter we analyze a simple model with a population segregated in two groups, a fast one of size N_f operating at rate λ_f and a slow one of size N_s doing so at rate λ_s, with $\lambda_f > \lambda_s$. Aside from heterogeneity in their time scales, agents in a real population are also heterogeneous in terms of their influence on others. To model this effect, we assume that the probability of agent i to copy the opinion of agent j is a function of the rate of agent j, that is,

$$\text{Prob}(j|i) = \frac{f(\lambda_j)}{\sum_{i=1}^{N} f(\lambda_i)}, \tag{2.6}$$

where $f(\lambda)$ is an arbitrary function measuring the reputation of agents of rate λ as seen by the population. When $f(\lambda)$ is a monotonic increasing function, the influence of agents is hierarchically organized, with fast agents having higher reputation and, thus, being copied more frequently, both by fast and slow agents. In this work, we use $f(\lambda) = \lambda^\sigma$.

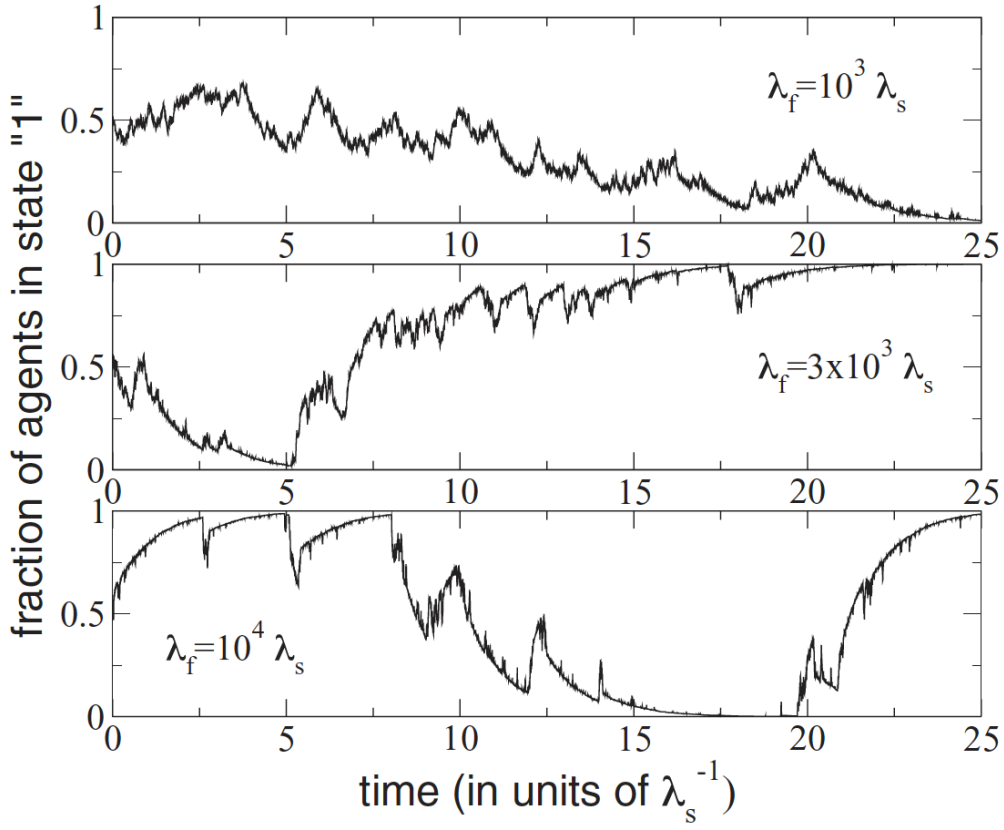


Figure 2.1: Evolution of the fraction of agents in the "1" state of a two-compounded heterogeneous system, in a mean-field random network of N=5000 agents, where 20% of them are fast. λ_f (λ_s) refers to fast (slow) group's activation rate. **Top:** $\lambda_f = 10^3\lambda_s$. **Center:** $\lambda_f = 3 * 10^3\lambda_s$. **Bottom:** $\lambda_f = 10^4\lambda_s$

.

Figure 2.1 shows particular realizations of the process in a system made of a small group of fast agents, $N_f = 1000$, and a large one of slow agents, $N_s = 4000$. In this particular example, we set $\sigma = 1$, a fixed value of $\lambda_f = 1$, and different values of λ_s. When the separation of time scales between the two groups is not very important, the global dynamics is purely diffusive, as in the standard voter model. However, when the separation of time scales exceeds a certain critical value, the behavior changes completely. Periods of quasi-regular growth and decrease alternate, which are suddenly broken by sharp peaks. Although the system ends up absorbed in one of the two absorbing states, the peculiar pathway to reach consensus cannot be observed in the standard voter model.

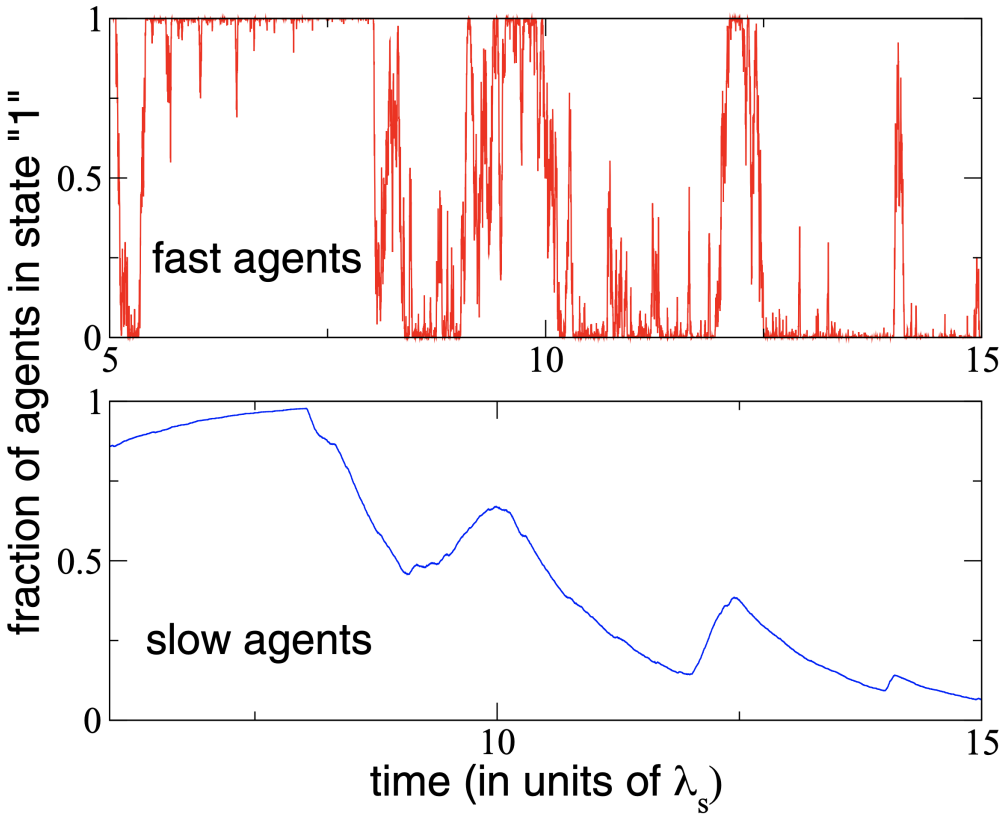


Figure 2.2: Evolution of the fraction of fast (top) and slow (bottom) agents in state "1" of the same system as in Figure 2.1. Plots correspond to the supercritical phase with $\lambda_f = 10^4\lambda_s$.

To understand this phenomenon, in Figure 2.2, we show the temporal evolution of both groups. From this figure, it is clear that the anomalous behavior we observe in Figure 2.1 is the result of a very differentiated dynamics of the fast and slow agents. Due to the huge differences between time scales, from the fast group perspective slow agents will seem as being frozen in their state. However, due to the growing form of function $f(\lambda)$, the effect of slow agents in the dynamics of fast ones is small. In this situation, fast agents evolve as in the simple voter model until they reach one of their consensus states. Nonetheless, unlike in the simple voter model, this consensus state is not an absorbing one. Indeed, despite the small probability of a fast agent to copy a slow one, its time scale is small enough to realize this interactions many times during the evolution of the process. When such events occur, fast agents may copy an opposite opinion from a slow outsider, thus introducing some noise in the small subsystem, preventing it to

be trapped in the consensus state. In other words, the absorbing boundary is replaced by a reflecting one. The same noise induced by slow agents can make the group of fast agents to abruptly change to the opposite state, turning the dynamics into an effective two-state dynamical system.

At the same time, from the slow group perspective fast agents spend long periods of time in the consensus states. Again, due to the growing form of function $f(\lambda)$, slow agents have a higher tendency to copy the fast agents' opinion that, being quasi-frozen, acts as a constant drift that pulls the slow agents' opinion towards the opinion of the fast ones. We can interpret that the group of slow agents has become a herd-like group following the leadership of the group of fast agents. However, this behavior is not observed in the whole range of parameters and it is unclear whether it appears suddenly at a critical value or, instead, it is a crossover effect interpolating continuously from the diffusive behavior of the standard voter model to the herding behavior we observe in Figure 2.1.

2.2.1 Langevin description

The existence of the conserved quantity Φ implies that the dynamics cannot be completely understood only in terms of Eq. (2.4) as such equation does not contain any information about the noise of the system. We are then forced to develop a theory that includes the second order terms of the dynamics. To do so, we take advantage of the homogeneity within each group of agents and define the instantaneous average opinion state of each group as

$$\Gamma_f(t) \equiv \frac{1}{N_f} \sum_{i \in fast} n_i(t) \; ; \; \Gamma_s(t) \equiv \frac{1}{N_s} \sum_{i \in slow} n_i(t). \tag{2.7}$$

In the limit of large systems, $\Gamma_f(t)$ and $\Gamma_s(t)$ can be considered as quasi-continuous stochastic processes in the range $[0, 1]$. Besides, they are the result of a sum of a large number of random variables so that the central limit theorem can be invoked. As a result, we conclude that the stochastic evolution of the vector $\vec{\Gamma}(t) \equiv (\Gamma_f(t), \Gamma_s(t))$ can be described by a Langevin equation. In particular, for the fast group dynamics, we can write

$$\frac{d\Gamma_f(t)}{dt} = A_f\left[\vec{\Gamma}(t)\right] + \sqrt{D_f\left[\vec{\Gamma}(t)\right]}\xi_f(t), \tag{2.8}$$

where $\xi_f(t)$ is a gaussian white noise. The drift and diffusion terms are respectively defined in terms of the infinitesimal moments as

$$A_f = \frac{\langle \Delta\Gamma_f(t) | \vec{\Gamma}(t) \rangle}{dt} \; , \; D_f = \frac{\langle [\Delta\Gamma_f(t)]^2 \, | \vec{\Gamma}(t) \rangle}{dt}, \tag{2.9}$$

where $\Delta\Gamma_f(t) \equiv \Gamma_f(t + dt) - \Gamma_f(t)$ [42]. These two terms can be computed exactly using Eq. (2.1) and read

$$A_f = \alpha_{fs}(\Gamma_s - \Gamma_f) \tag{2.10}$$

$$D_f = \frac{\alpha_{fs}}{N_f}\left(\Gamma_s + \Gamma_f\left[1 + 2\beta_{fs} - 2\Gamma_s - 2\beta_{fs}\Gamma_f\right]\right), \tag{2.11}$$

where we have defined

$$\alpha_{fs} = \frac{\lambda_f}{1+\beta_{fs}} \quad \text{and} \quad \beta_{fs} = \frac{N_f f(\lambda_f)}{N_s f(\lambda_s)}. \tag{2.12}$$

Similar equations can be derived for the slow group by replacing the index $f \leftrightarrow s$ in the previous equations.

2.2.2 Effective potential function

When the separation of time scales is large, the state of the slow group is perceived by the fast group as constant. In this case, we can consider Γ_s in the previous equations as a constant parameter. As a consequence, the diffusion term in Eq. (2.11) does not vanish when $\Gamma_f = 0$, or $\Gamma_f = 1$ and the system reacts at these points as in the presence of a reflecting barrier. Therefore, the system has a well defined steady state controlled by an effective potential that, up to a constant value, takes the form [42]

$$V_{eff}(\Gamma_f) = \ln D_f - 2\int \frac{A_f}{D_f} d\Gamma_f. \tag{2.13}$$

This potential has a single extremum approximately at $\Gamma_f^* = \Gamma_s$, which changes from being a minimum to a maximum when

$$2\frac{f(\lambda_f)}{f(\lambda_s)} > N_s. \tag{2.14}$$

When this condition is met, the combination of a maximum with the two reflecting barriers at $\Gamma_f = 0$ and $\Gamma_f = 1$, transforms the effective potential into a double-well potential with a barrier at $\Gamma_f = \Gamma_s$. This defines a pitchfork bifurcation separating a diffusive phase, where the fast group is dragged by the slow one, and a herding phase, with the fast group behaving effectively as a two-state system, jumping from one state to the other as in an activated process. This is illustrated in Figure 2.3, where we show the effective potential when $\Gamma_s = 0.5$ in the two cases, along with examples of realizations of the slow and fast group dynamics.

We should note that, while this transition is not a true phase transition, as it disappears in the thermodynamic limit $N_s >> 1$, for finite systems it behaves effectively as a first order phase transition. Besides, the strong separation of time scales we find in some real systems, like in speculative markets (which can be of the order of $\lambda_f \sim 10^{4\sim5}\lambda_s$), coupled with a growing preference function $f(\lambda) \sim \lambda^\sigma$ can make condition Eq. (2.14) to hold quite easily even for very large populations, in particular when exponent $\sigma > 1$.

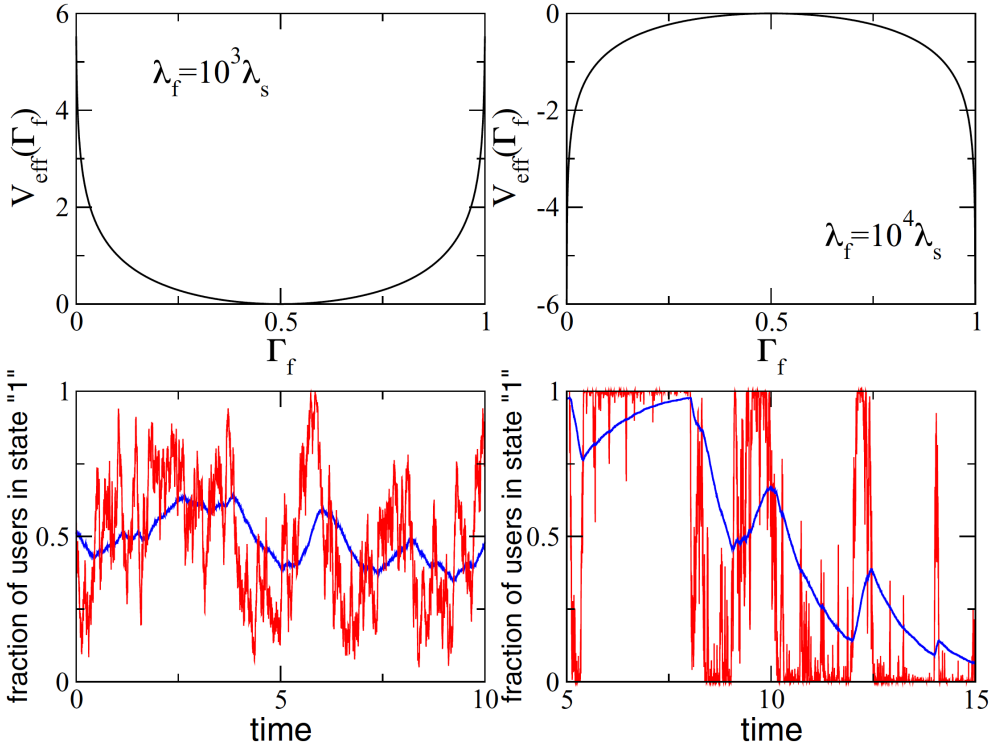


Figure 2.3: **A:** Effective Potential for $N_f = 500$ and $k_{fs} = 25$. **B:** Effective Potential for $N_f = 500$ and $k_{fs} = 2500$. **C:** Simulation of process with conditions in A. **D:** Simulation of process with conditions in B. For this parameters, $k^c_{fs} = 250$.

2.2.3 Consensus time

Voter systems satisfying Eq. 2.14 must be at the herding phase, with fast agents behaving as a two-state system with switching dynamics. But it is important to note that the path towards global consensus will vary depending on the time scales interplay between both fast and slow agents. In order to understand these different dynamics, we can define the order parameter:

$$x \equiv \frac{f(2\lambda_f)}{N_s f(\lambda_s)}. \tag{2.15}$$

As shown in [43], in the limit $N_s > N_f \gg 1$ the Langevin Equation for the average opinion of fast agents (Eq. 2.8) can be rewritten as:

$$\frac{d\Gamma_f(t)}{dt} = \frac{2\lambda_f}{xN_f}\left[\Gamma_s - \Gamma_f\right] + \sqrt{\frac{2\lambda_f}{N_f}\Gamma_f\left(1-\Gamma_s\right)}\xi_f(t), \tag{2.16}$$

which reflects the diffusive dynamics of fast agents Γ_f towards one of the consensus states, modulated by a drift towards slow agents opinion Γ_s controlled by the term $\frac{2\lambda_f}{xN_f}$. Assuming Γ_s remains constant from the perspective of fast agents, we can compute the characteristic switching time of fast agents T_f as the standard mean first passage time for a stochastic process following Eq.2.16, with one reflecting barrier at $\Gamma_f = d\Gamma$ and an absorbing boundary at $\Gamma_f = 1 - d\Gamma$ [43]:

$$T_f = \frac{N_f}{\lambda_f} \int_0^{1-d\Gamma} \frac{B\left(z, \frac{2\Gamma_s}{x}, \frac{2(1-\Gamma_s)}{x}\right)}{z^{\frac{2\Gamma_s}{x}}(1-z)^{\frac{2(1-\Gamma_s)}{x}}} dz, \tag{2.17}$$

where $B(z, a, b)$ refers to the incomplete Beta function. The Langevin equation for slow agents under the same conditions reads:

$$\frac{d\Gamma_s(t)}{dt} = \lambda_s \left[\Gamma_f - \Gamma_s\right], \tag{2.18}$$

leading to an exponential decay of slow agents opinion Γ_s with quasi-deterministic drift towards the fast-group consensus state Γ_f with a characteristic time $T_s = \lambda_s^{-1}$.

Note that when $T_f \gg T_s$, or equivalently $\lambda_s T_f \gg 1$, the group of slow agents will typically reach the quasi-frozen consensus state of fast agents $\Gamma_f = 0, 1$ before the latter can switch their state. This means that global consensus T_{con} will be reached according to decay rate of the slow group:

$$T_{\text{con}} \sim \lambda_s^{-1}. \tag{2.19}$$

On the contrary, when $T_f \ll T_s$, it can be shown [43] that:

$$T_{\text{con}} \sim T_f \exp\left(\frac{1}{\lambda_s T_f}\right). \tag{2.20}$$

Assuming again that $f(\lambda) = \lambda^\sigma$ and $N_f = aN_s$, we can rewrite Eq.2.17 as:

$$\lambda_s T_f = a\left(\frac{2}{x}\right)^{1/\sigma} N_s^{1-1/\sigma} \int_0^{1-d\Gamma} \frac{B\left(z, \frac{2\Gamma_s}{x}, \frac{2(1-\Gamma_s)}{x}\right)}{z^{\frac{2\Gamma_s}{x}}(1-z)^{\frac{2(1-\Gamma_s)}{x}}} dz. \tag{2.21}$$

Finally, we can combine all of the above to show [43]:

$$T_{\text{con}} \sim \begin{cases} \lambda_s^{-1} & \text{if} \quad \sigma \geq 1 \\ \frac{\exp\left(N_s^{1/\sigma-1}\right)}{N_s^{1/\sigma-1}} & \text{if} \quad \sigma < 1 \end{cases}. \tag{2.22}$$

When $\sigma \geq 1$ Eq.2.21 diverges, so that slow agents bring the system to consensus in constant characteristic time. On the contrary, when $\sigma < 1$ consensus time diverges with system size, making the absorbing states unreachable in the thermodynamic limit.

2.3 Discussion

In this chapter, we have uncovered a simple and parsimonious mechanisms giving rise to the emergence of leadership and herding behaviour in a population of interacting agents, namely, a strong separation of time scales coupled with hierarchical structures of influence exerted by some agents on the others.. This bring important differences with respect to the diffusive behaviour and consensus path characteristics of the standard voter model. Despite the simplicity of the toy model that we use in this work, the mechanisms are general enough to be extrapolated to more complex and realistic situations. For instance, the simple segregation of the population in only two groups is not really necessary. Although mathematically more involved, it can be shown that the same phenomenology takes place in systems with a strong heterogeneous distribution of activity rates.

The hierarchical organization can also be induced by different mechanisms, like a hierarchical organization of the network of interactions among the agents.For instance, a large sample of real-world networks present core-periphery structures [44], made of a core of well interconnected agents and a periphery made of agents that are mainly connected to the core. These type of structures are particularly pervasive in online social networks such as Twitter [45], [46]. Finally, one could also argue that the influence that a group of agents have on the others is a stochastic process by itself. In our case, this could be easily modelled by assigning to the parameter σ some stochastic dynamics. This is particularly interesting as, being the transition effectively discontinuous, the dynamics would be a mixture of purely diffusive periods, when σ is such that the condition in Eq. (2.14) is violated, and periods with strong herding behaviour.

Chapter 3

Bridgeness and dynamical centrality

Whether finding influencers in online social networks [47], protecting key stations in a power distribution grid [48] or vaccinating spreaders in an epidemic [49], there is no single recipe to rank the nodes of a complex network according to their importance [50]. Despite the existent variety of centrality measures that have been introduced in recent decades, most of them share in common the need for complete or partial topological information. A natural question is, therefore, can we measure centrality directly from local dynamical observables when network topology is uncertain? That is, can we identify those actors that are most important for the collective functionality or the robustness of a network just by looking at each node's internal state and dynamical behaviour?

In this chapter we propose two dynamical centrality measures, *asynchrony* and *flip-rate*, based respectively on observables from two paradigmatic dynamical systems, that of the Kuramoto model of synchronisation [51] and that of the Potts model of spins [52], both of which have extensive applications in physics, chemistry, biology and the social sciences. In networks with community structure, we find an interplay between these dynamical observables and bridgeness centrality —a measure of the extent to which a node acts as a modular broker, i.e. of its participation into the different communities of a network. This interplay is important as it ensures we can infer topological centrality (bridgeness) by measuring certain observables (asynchrony and flip-rate) directly from network dynamics. In order to describe this relation, we introduce a prototype network model we call Stochastic Block Model with bridgeness (SBMb), which generates graph ensembles with a given bridgeness distribution, while controlling the effect of other properties such as degree or community structure. Using the SBMb, we show how in fact bridgeness induces locally higher values of both asynchrony and flip-rate,

promoting global ordering at the same time. We generalise such interplay showing that bridgeness generates localised patterns in the Laplacian eigenvectors of the SBMb, which are attributable to functional modes performing at different timescales and regions of the network, from inside communities to their boundaries and bridges. Finally, using asynchrony and flip-rate, we propose a novel method for detecting network vulnerabilities even when the underlying topology cannot be accessed. Our conjecture is that some networks can be as efficiently dismantled by targetting certain functional behaviour of their nodes as by using topological targets such as degree, betweenness or bridgeness centrality. We show how this is the case for two synthetic models (SBMb and Random Geometric Graph) and one real network (Western US Power Grid).

3.1 The Stochastic Block Model

3.1.1 Generative model

Like preferential attachment for scale-free networks or the Watts-Strogatz model of small-worlds, modular networks have a well-known generative mechanism based on planted partitions, the Stochastic Block Model (SBM). It has its origin in the social sciences, particularly in the study of social networks [53]. As its name suggests, the high-level purpose of the SBM is to generate a parameterized ensemble of networks that have their nodes somehow grouped into blocks of nodes that have internal statistical similarities.

Given a network with N nodes that are partitioned along B different groups, we represent each node's affiliation through the block or partition vector

$$b = (b_1, \ldots, b_N), \tag{3.1}$$

with entries $b_i \in \{1, \ldots, B\}$. Then, the aim of the SBM is to generate networks using b as a parameter. One way to achieve that is through fully characterizing the probability

$$P(A|b), \tag{3.2}$$

where $A = \{a_{ij}\}$ is the adjacency matrix of the generated network. In this sense, building a SBM is equivalent to coming up with a reasonable $P(A|b)$ that reflects the desired modular structure we are trying to model.

For networks with single-edges (i.e. $a_{ij} \in \{0, 1\}$) and without self-edges, the standard SBM is:

$$P(A|p, b) = \prod_{i<j} p_{b_i b_j}^{a_{ij}} (1 - p_{b_i b_j})^{a_{ij}}, \tag{3.3}$$

where p_{rs} is a matrix parameter accounting for the probability of finding an edge between nodes from blocks r and s respectively. In this case, edges are distributed according to a Bernoulli distribution controlled by parameters b and p. It can be shown [23] that

the particular choice in Eq. 3.3 attests for the maximum indifference towards $P(A|b)$ when only the expected total number of edges between each group is known: that is, it maximizes the entropy function

$$\Omega = -\sum_{A} P(A|b) \ln P(A|b) \tag{3.4}$$

subject to the constraint

$$\langle e_{rs} \rangle = \sum_{ij} a_{ij} \delta_{b_i r} \delta_{b_j s}, \tag{3.5}$$

where e_{rs} is the number of edges between groups r and s. As will be shown in the next section, other variants of $P(A|b)$ can and are typically used for studying the SBM due to their more tractable form.

3.1.2 Bayesian inference of communities

The task of detecting and describing community structures of complex networks has been approached in many different ways (see Section 1.4.5 for details). Most of them are based on non-statistical heuristics, and thus lack a principled method towards the discovery of modular structure. In contrast, the SBM fundamentally is a generative model for modular structures, as explained in the previous section. For this reason it allows to build probabilistic models describing real and synthetic network data, thus being widely used as cornerstone for Bayesian community detection methods. One of the most appealing features of this SBM Bayesian inference framework is the ability to distinguish random structures from statistically significant modules, something that most heuristic methods for community detection cannot do.

Bayesian framework

In an inferential setting, our objective is to work out the conditional probability $P(b|A)$ of observing a partition or block vector $b = (b_1, \ldots, b_N)$ given an empirical network represented by its adjacency matrix A. This can be interpreted as a posterior probability using Bayes' rule:

$$P(b|A) = \frac{P(A|b)P(b)}{P(A)} \tag{3.6}$$

where $P(A|b)$ is called the marginal likelihood and assumes data is generated by a certain SBM. $P(b)$ is the prior probability representing *a priori* assumptions on that SBM. $P(A)$ is called the evidence and normalizes the posterior by counting all possible partitions, although for the purposes of maximizing or sampling from the posterior distribution its computation is not required.

Likelihood functions

For the remaining of this chapter, we will use two of the most well-studied likelihood models: the regular Poisson SBM and the degree corrected SBM [54]. Note that these models differ from the previously described Bernoulli SBM in that they actually allow for multi-edges and self-loops. It can be shown [23], however, that in the sparse limit they introduce corrections of order $O(1/n)$ which are insignificant for large-enough sparse networks.

For networks with degree homogeneity, we approach community-detection using the regular Poisson SBM likelihood function

$$P(A|b,\omega) = \prod_{i<j} \frac{\omega_{b_i b_j}^{a_{ij}} e^{-\omega_{b_i b_j}}}{a_{ij}!} \prod_i \frac{(\omega_{b_i b_j}/2)^{a_{ij}/2} e^{-\omega_{b_i b_j}/2}}{a_{ij}/2!}, \tag{3.7}$$

a model that considers the number of edges between any pair of vertices distributed as independent Poisson variables. Similarly to the Bernoulli SBM, the model has a parameter matrix ω_{rs}, which accounts for the expected number of links between nodes in communities r and s.

The likelihood function in Eq. 3.7 considers each node in each community as statistically identical regarding the expected number of incident edges. However, this is not a realistic assumption for most real-world networks, which have heterogeneous degree distributions. The degree corrected SBM takes into account this by introducing a new vector parameter θ, which controls the expected degree of each node in the network. In this case, the marginal likelihood function reads:

$$P(A|b,\omega,\theta) = \prod_{i<j} \frac{(\theta_i\theta_j\omega_{b_i b_j})^{a_{ij}} e^{-\theta_i\theta_j\omega_{b_i b_j}}}{a_{ij}!} \prod_i \frac{(\theta_i^2\omega_{b_i b_j}/2)^{a_{ij}/2} e^{-\theta_i^2\omega_{b_i b_j}/2}}{a_{ij}/2!} \tag{3.8}$$

Choice of priors

To preserve the unbiased nature of the inference mechanism, we use maximum entropy priors for the parameters of the two considered SBM. Starting with the partition vector b, we consider a prior which is agnostic about the number of communities B and the size n_r of each community r. It can be shown [23] that these assumptions lead to a prior function of the form:

$$P(\omega|b) = \prod_{r\leq s} \frac{\prod_r n_r!}{N!N} \binom{N-1}{B-1}^{-1} \tag{3.9}$$

The prior for the inter-community connectivity matrix ω necessarily takes into account the information considered for b, and it can be shown that its entropy-maximizing dis-

tribution is an exponential function [23] around an average value $\bar{\omega}$, of the form:

$$P(\omega|b) = \prod_{r \leq s} \frac{\exp\left(-\frac{n_r n_s \omega_{rs}}{(1+\delta_{rs})\bar{\omega}}\right)}{(1+\delta_{rs})\bar{\omega}} \tag{3.10}$$

Finally, we also chose for θ its entropy-maximizing prior, which also depends on b and its hyper-parameter n_r, and it can be expressed as [23]:

$$P(\omega|b) = \prod_r (n_r - 1)!\ \delta\left(\sum_i \theta_i \delta_{b_i r} - 1\right) \tag{3.11}$$

where the outer δ indicates a Dirac delta function, and the inner is a Kronecker delta.

Numerical implementation

Although the previous expressions for priors and likelihoods can be combined using Bayes rule, and analytical expressions for posterior distributions can be obtained, such expressions will be complex and in general sampling or deriving their maximum will be an NP-hard problem. Using Monte Carlo Markov Chain (MCMC) methods, however, a numerical algorithm can be devised to achieve this: starting from an initial partition b_0, a Metropolis acceptance-rejection rule produces changes to community assignments until we can ensure a convergence to an equilibrium dominated by the posterior distribution [23], [55]. These sort of algorithms, including many useful modifications to include very general SBM inference settings, are readily implemented in the freely available python library Graph-Tool [56]. Throughout the experiments done in the chapter, we have used this library for every community-inference step, making use of its exhaustive documentation when needed.

3.2 Bridgeness centrality

Many real-world networks found in nature or society have modular structures that are hierarchical and overlapping, such that some nodes may be affiliated to several communities at the same time [24]. Given their ability to connect groups of nodes that otherwise would interact poorly, highly overlapping nodes are typically called bridge nodes. Bridge nodes have been studied in social network analysis since the 1970s [12], [57], [58], focusing on their role as promoters of diffusion and cross-communication in social systems. Modular social networks are also the substrate on which epidemics usually spread, and it has been shown that targeted immunisation or specific social distancing interventions focused on community-bridging agents is even more effective than those strategies based solely on number of contacts (degree) [59]. Furthermore, recent examples in molecular biology research have also shown that protein-protein interaction networks generate

overlapping community-structures closely related to essential cellular functions [60], [61], where some critical nodes integrate different functional modules [62], [63] and can be classified as functional bridges. Bridge nodes are also found in cortical networks [64] and word association networks [65], among other real-world examples.

In this section we review existent methods used to evaluate to which extent nodes are bridges, i.e. to measure so-called bridgeness centrality. In addition, we propose a generative mechanism based on link-rewiring which extends the SBM to include an arbitrary distribution of bridgeness centrality across generated networks.

3.2.1 Measuring bridgeness

Several methods have been proposed to quantify bridgeness, which can be generally divided in two categories. On the one hand, there are methods that do not use directly any community-level information. For example, Hwang *et al.* [66] introduced a measure of bridgeness S^H which combines degree k and betweenness centrality B (see Section 1.4.3 for details on these measures), and is defined as:

$$S_i^H = B_i \frac{k_i^{-1}}{\sum_{j \in N(i)} k_j^{-1}} \quad , \tag{3.12}$$

where $N(i)$ refers to the first neighbours of node v_i. The term B_i above favours nodes with high betweenness whereas the term $\frac{k_i^{-1}}{\sum_{j \in N(i)} k_j^{-1}}$ highlights nodes with low degree that are surrounded by high-degree nodes. Similarly, Jensen *et al.* [67] proposed a bridgeness measure S^J that only differs from betweenness centrality in that it discards the shortest path starting or ending in the first neighbourhood of each node:

$$S_i^J = B_i - \sum_{j \notin N(i), k \notin N(i)} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \quad , \tag{3.13}$$

where σ_{jk} counts the number of shortest paths between nodes j and k and $\sigma_{jk}(i)$ accounts for those same paths only if they traverse node i. The authors concluded that S_i^J is only significantly different from betweenness when bridges are low degree, and even then the difference is generally small. On a further example, Wu *et al.* [68] considered a more restricted notion of bridge, defining it as an edge whose removal increases the number of connected components in a graph. They defined an associated bridgeness measure on edges S_i^W that simply counts the number of nodes disconnected from the largest connected component after the removal of each edge. Through this definition, the authors of this method directly associate the capacity for damaging a network with bridgeness centrality.

On the other hand, there are those methods that use mesoscopic information form community detection to infer bridgeness. In this category we can find methods such

as that introduced by Nepusz *et al.* [69], which is based on overlapping community detection (see Section 1.4.5 for details). Given a network with C detected communities, and assuming a membership vector $u_i = (u_i^1, \ldots, u_i^C)$ can be inferred for each node i, the authors defined bridgeness centrality S_i^{Ne} as the normalised and inverted Euclidean distance to a reference vector $(1/C, \ldots, 1/C)$ representing equally spread membership to all communities:

$$S_i^{Ne} = 1 - \sqrt{\frac{C}{C-1} \sum_{c=1}^{C} \left(u_j^c - \frac{1}{C} \right)^2} \tag{3.14}$$

For the remaining of this chapter we will use one of the earliest bridgeness measures described in the literature, namely the participation coefficient described by Guimera *et al.* [70], which uses mesoscopic information from community detection. Although this method is based on non-overlapping community partitions, we can quantify the participation of node v_i to each of the C communities of a partition with the probability mass function $\pi_i = (\pi_i^1, \ldots, \pi_i^C)$, where:

$$\pi_i^c = \frac{\sum_j a_{ij} \delta_{c,c_j}}{\sum_j a_{ij}} = \frac{k_i^c}{k_i} \quad , \tag{3.15}$$

with $\delta_{c,c_j} = 1, 0$ if v_i's first neighbour v_j belong to community c or not respectively, and $\sum_{c \in C} \pi_i^c = 1$. Note how the fraction of edges connecting a vertex to a given community is used as a proxy for membership strength. Then for each node's v_i the participation coefficient or, what for our purposes we call bridgeness centrality S_i, is defined as:

$$S_i = 1 - \sum_{c \in C} \left(\pi_i^c \right)^2 \quad . \tag{3.16}$$

This measure reaches its minimum at $S_i = 0$ (when a node participates only in one module) and maximum at $S_i = 1 - 1/C$ (participates evenly across the C communities of the partition), and thus accounts for extensiveness and uniformity.

Note that the partition underlying the calculations of π_i is generally unknown, and therefore this method requires a choice of community detection algorithm. Throughout this chapter, we will use the Stochastic Block Model for inferring communities (see Section 3.1.2). Given the probabilistic nature of this framework, instead of a single partition we obtain an ensemble of partitions $\mathcal{B}$. For each partition in the ensemble we can obtain a single bridgeness centrality measure S_i, so that we can define an ensemble-averaged bridgeness:

$$\langle S_i \rangle_{\mathcal{B}} = \sum_{\mathcal{B}} \frac{S_i}{|\mathcal{B}|} \quad . \tag{3.17}$$

3.2.2 A stochastic block model with bridgeness

The model starts from an initial graph $G_0 = (V, E_0)$ consisting of a set of M isolated subgraphs $K \in \{K_1, \ldots, K_M\}$, such that $G_0 = \bigcup_{i=1}^{M} K_i$ but $K_i \bigcap K_j = \emptyset \; \forall i, j \in \{1, \ldots, M\}$.

Then we rewire the incident links of each node v_i with a certain probability p_R^i, keeping v_i at one end and randomly selecting the other from the set of nodes V. Note that in order to induce a homogeneous distribution of degree, we restrict our analysis to the case of having an initial set of isolated complete subgraphs (cliques). Depending on the exact manner p_R is distributed across the network, different patterns of modular interaction will emerge, conforming the final network $G = (V, E)$. Nodes with low rewiring rates keep most incident links inside their original community, but still can promote direct inter-community borders. Differently, nodes with high p_R^i will emerge as mediators between modules. We consider two types of SBMb models.

- **SBMb$_1$**: this model prescribes three different types of nodes: *Bulk nodes*, with $p_R^{\text{bulk}} = 0$; *Border nodes*, with $0 < p_R^{\text{border}} << 1$; *Bridge nodes*, with $p_R^{\text{bridge}} = 1$. Note that the split between bridge and border nodes is not fundamental, although it allows us to study the difference between moderate and high rewiring rates.

- **SBMb$_2$**: in contrast, the second model merges bridge and border nodes into a continuous category of nodes that draw their rewiring probabilities uniformly at random from $p_R \in (0, 1]$. Bulk nodes are still controlled separately by $p_R^{\text{bulk}} = 0$.

Figure 3.1(a) shows a schematic representation of the rewiring process described above. Figure 3.1(b) depicts a particular instance of the SBMb$_1$ with $N = 500$ nodes, including 25 bridge nodes with $p_R^{\text{bridge}} = 1$ and 75 border nodes with $p_R^{\text{border}} = 0.2$, homogeneously distributed across the $M = 25$ initial cliques. Figure 3.1(c) illustrates how bridgeness induces clearly distinctive functional behaviour to each nodal role, as demonstrated in sections below.

Measuring bridgeness in the SBMb

Both in the SBMb$_1$ and SBMb$_2$, to get accurate bridgeness measures we will need to produce R realisations of the rewiring protocol from the initial condition of isolated cliques to obtain an ensemble of model instances $\mathcal{G} = \{G_1, \ldots, G_R\}$. As explained above, we use the Stochastic Block Model to infer a partition ensemble $\mathcal{B}(G)$ in every sampled network G, and derive its corresponding partition ensemble bridgeness $\langle S_i \rangle_{\mathcal{B}(G)}$ using Eq. (3.17). Using partition ensembles for every realisation, we can finally compute the SBMb-ensemble bridgeness $\langle S_i \rangle$ for every node v_i:

$$\langle S_i \rangle = \sum_{\mathcal{G}} \frac{\langle S_i \rangle_{\mathcal{B}(G)}}{R}. \tag{3.18}$$

Shuffling probabilities and bridgeness

The SBMb provides planted partitions with inter-modular connections controlled by the shuffling probability p_R. A first test to the precission of our bridgeness methodology

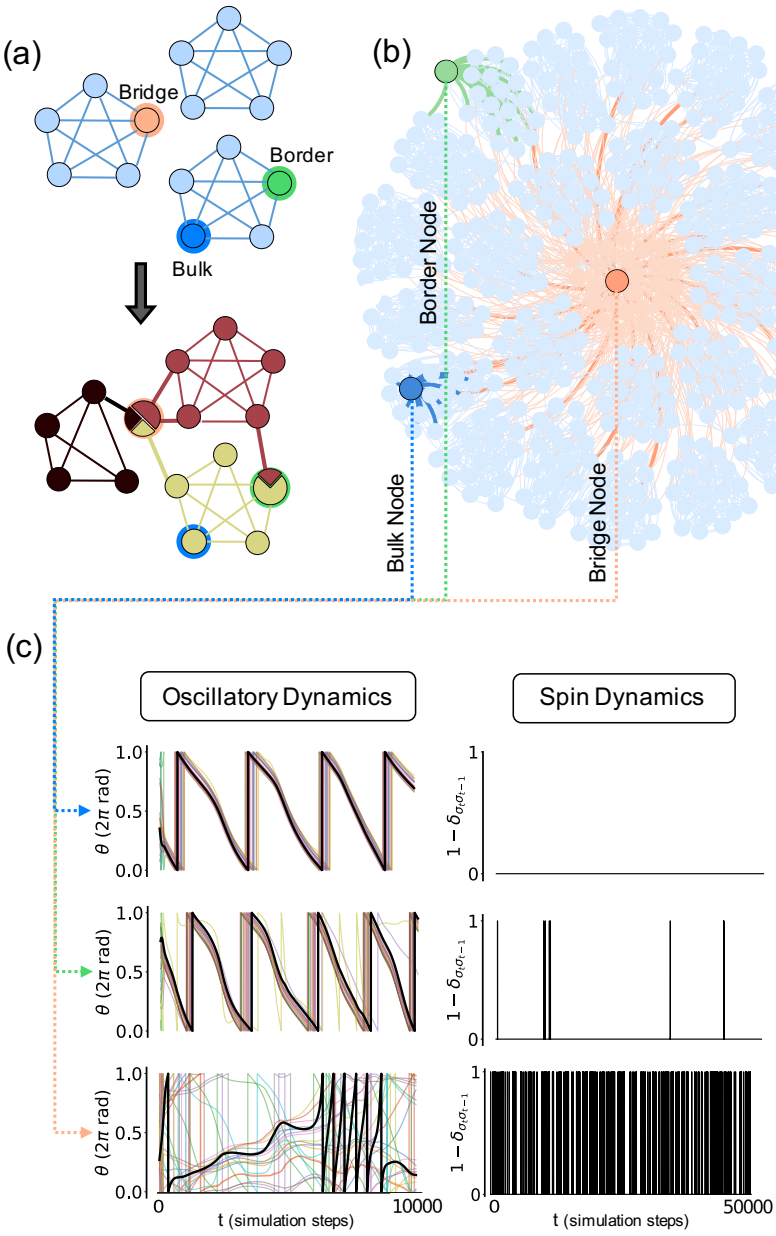


Figure 3.1: **(a)** Schematic link-rewiring mechanism of the SBMb. **(b)** Particular realisation of the SBMb_1 with $N = 500$ nodes, including 25 bridge nodes with $p_R^{\text{bridge}} = 1$ and 75 border nodes with $p_R^{\text{border}} = 0.2$, homogeneously distributed across $M = 25$ initial cliques. **(c)** Characteristic functional behaviour of each structural role: left column represents asynchrony, showing the phase evolution $\theta(t)$ of a given node (black line) and its neighbourhood (coloured lines) under Kuramoto dynamics; right column presents flip rate, using a dichotomous variable $1 - \delta_{\sigma_t,\sigma_{t-1}}$ showing whether the spin has flipped its internal state in the current time step under Potts dynamics.

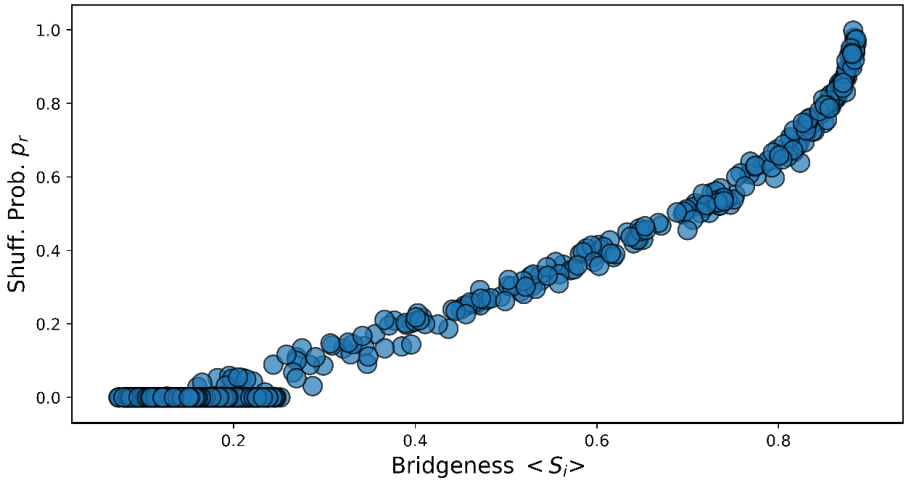


Figure 3.2: Relation between bridgeness $\langle S \rangle$ (Eq. 3.18) and shuffling probability p_R in the SBMb_2 for $R = 100$ rewiring realisations, with N=500 including 100 uniformly rewired nodes with $p_R \in \mathcal{U}(0, 1]$ and 400 bulk nodes with $p_R = 0$.

consists in computing the relation between $\langle S \rangle$ and p_R. To do so, we will use several realizations of the SBMb2 with N=500 and 400 bulk nodes. In fact, as shown in Figure 3.2, there exists a continuous function relating the shuffling probabilities assigned to each node with their bridgeness measure. Note we can see that the set of bulk nodes with $p_R^{\text{bulk}} = 0$ does not always have a bridgeness value of 0: this is natural given the nature of the rewiring process, considering that some nodes from outside their clique have a chance to rewire their links and connect with bulk nodes.

3.3 Dynamical centrality

3.3.1 Dynamical processes on modular networks

Modularity [22] —like scale-free degree distributions [71] and small-world properties [72]— deeply influences any dynamical process occurring on a network. For example, research has shown that modular structure hinders the spread of epidemics regardless of degree heterogeneity, given that the presence of communities favours the natural confinement of outbreaks [73]. Cascading processes also show distinctive patterns of active nodes which are directly related to the modular structure of the underlying network [74]. Similar results are observed in diffusion processes [75], consensus of spin systems [76], [77] and synchronization of coupled oscillators [78], [79]. However, it remains unclear how different patterns of inter-modular connection affect the outcome of such dynamical

processes: does a system behave differently when modules are directly connected than when bridge-nodes connect them indirectly?

3.3.2 Measuring dynamical centrality

Kuramoto model

Under Kuramoto dynamics each node v_i in the network represents an oscillator with an internal angular phase θ_i and natural frequency ω_i. Non-linear couplings are reflected on the instantaneous frequency $\dot{\theta}_i$ of each oscillator:

$$\dot{\theta}_i = \omega_i + K \sum_{j=1}^{N} a_{ij} \sin(\theta_i - \theta_j) \quad . \tag{3.19}$$

The transition towards phase synchronization (where all oscillators pulse with the same phase) is mediated by the global coupling parameter K and has been studied for different topologies [13]. These include modular networks [79] where for sufficiently low coupling communities lock in locally-synchronised phases [80]. Also, previous studies suggest that overlapping interfaces between communities behave differently than the rest of the system when such locally clustered steady-states are reached, exhibiting anomalous distributions of instantaneous frequency $\dot{\theta}_i$ [81]. Following [79] we measure local synchronisation using:

$$\rho_i = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} \cos(\theta_i - \theta_j) \quad . \tag{3.20}$$

In this case, we will actually use $(1 - \rho_i)$, i.e. asynchrony. Averaging over many realisations with different initial conditions for ω_i and $\theta_i(t = 0)$, both drawn from uniform probability distributions $\mathcal{U}(0, 2\pi)$, we finally obtain an asynchrony centrality measure

$$\langle 1 - \rho_i \rangle \tag{3.21}$$

for each node in a given network of oscillators.

Experimentally, in order to simulate the trajectory of a system of oscillators we solve the system of equations for θ_i using a 4^{th}-order Runge-Kutta method for many different initial distributions of internal frequency ω_i. Figure 3.3(a) exemplifies both local and global evolution of asynchrony for the SBMb_1 of a single simulation. Low-bridgeness (bulk) nodes quickly settle down to local synchrony, whilst higher-bridgeness (bridge) nodes reach steadiness later, remaining at higher asynchrony than the network average at all times.

Note that by observing the trajectory of the network average asynchrony we can detect when the system has reached a metastable state of local cluster synchronisation. This is important because our measures of dynamical centrality $(1 - \rho_i)$ are only significant at

this state, which means we discard the transient towards it. A simple way to asses if the system has reached the steady state consists of keeping track of the network-averaged nodal asynchrony. As shown in Figure 3.3(b), we can heuristically find steady-states by setting a condition such as:

$$\left| \frac{d}{dt} \left(\sum_{i=1}^{N} \frac{1-\rho_i}{N} \right) \right| < \epsilon_K \quad , \tag{3.22}$$

where ϵ_K is an arbitrary convergence threshold. It is expectable that the system undergoes several metastable states before settling on the most robust steady state. For this reason, we enforce that the condition in Eq. 3.22 holds for at least τ_K simulation steps in order to declare a steady state.

It is worth noting that, for the normal Kuramoto model described in Eq. 3.19, the network is expected to fall into a (coherent or incoherent) steady state where a single global order parameter can be measured [82], thus ensuring the conditions in Eq. 3.22 can be safely obtained. However, modified versions of the Kuramoto model — including but not limited to those with non-local coupling, second-order dissipation terms or degree-frequency correlations — can exhibit periodic trajectories or limit cycles under some conditions [83]. In these type of situations the condition in Eq. 3.22 could become unfeasible, potentially requiring different criteria to measure asynchrony centrality.

Potts model

In the Q-state Potts model each node v_i contains a spin σ_i with an internal state $q \in \{1, \ldots, Q\}$, and network neighbours interact according to the Hamiltonian:

$$H_\sigma = -\sum_{i,j} J_{i,j} \delta_{\sigma_i, \sigma_j} \quad , \tag{3.23}$$

where $\delta_{\sigma_i,\sigma_j} = 0, 1$ if $\sigma_i \neq \sigma_j$ or $\sigma_i = \sigma_j$ respectively, and we consider $J_{ij} = a_{ij}$ where A is the adjacency matrix of the network.

At equilibrium, we can locally describe the spin probabilities for each node as [84]:

$$p_i^q \equiv P\big(\sigma_i = q\big) = \frac{e^{\beta h_i^q}}{\sum\limits_{q'=1}^{Q} e^{\beta h_i^{q'}}} \quad , \tag{3.24}$$

where

$$h_i^q = \sum_j a_{ij} \delta_{q,\sigma_j} \tag{3.25}$$

quantifies the q-state field at node v_i, and β is the inverse temperature. When running on sufficiently modular networks and low temperature, the system reaches 'frustrated'

states, generating steady spin-structures closely related to underlying topological communities [76]. In those situations, however, some nodes cannot reach steadiness and change state *ad infinitum*: they are known as 'blinkers' —and have only been previously described for lattice topologies, where they appear to be randomly located [85]. The rate at which node v_i changes its internal state —which we simply call 'flip rate'— is:

$$W_i \equiv P\Big(\sigma_i^t \neq \sigma_i^{t-1}\Big) = \sum_{q=1}^{Q} p_i^q\big(1-p_i^q\big) = 1 - \sum_{q=1}^{Q}\big(p_i^q\big)^2 \quad . \tag{3.26}$$

Using Gibbs sampling from an ensemble of uniformly random initial conditions $\sigma_i(t=0)$ we obtain a measure $\langle W_i \rangle$, which we denominate flip-rate centrality, for each node in a given spin network.

Similarly to Kuramoto dynamics, flip-rate centrality measures are only significant at the aforementioned steady states. Figure 3.3(c) illustrates the evolution of a spin system realisation of the SBMb_1. We can see how low-bridgeness nodes quickly settle into the spin state of their local community, conforming frustrated states, whereas high-bridgeness nodes keep on blinking, i.e. jumping between surrounding spin states. As shown in Figure 3.3(d), a simple criterion for detecting the steady state in this case is:

$$\left|\frac{d}{dt}\Big(\mathrm{MA}\big(T, N_\sigma\big)\Big)\right| < \epsilon_P. \tag{3.27}$$

where N_σ refers to the total number of different spin states present in the system at a given time, and $\mathrm{MA}(T, N_\sigma)$ refers to a moving average filter of period T applied to the time series of N_σ. The ordering dynamics of the Potts model will decrease the number of existent spin states until the system reaches a steady state, moment at which N_σ will remain stable. Given that the time series of N_σ is significantly noisy, applying a moving average filter is useful in order to better track convergence.

As in the Kuramoto case above, the Potts model can also be modified to exhibit a more rich phase space with periodic trajectories. This may be the case when introducing particular configurations of anisotropic couplings or external driving fields [86]–[88]. Again, under such circumstances the condition in Eq. 3.27 may become unfeasible, thus requiring the choice of an alternative criterion to sample flip-rate centrality.

Note that, in modular networks, blinkers are not located randomly but instead their location is determined by bridgeness. In fact, when local consensus is reached, each community c will have a characteristic spin state σ_c. In this case, it is reasonable to assume that local fields are non-zero only for such characteristic states σ_c, with a value determined by participation coefficients:

$$h_i^{\sigma_c} = \sum_j a_{ij}\delta_{\sigma_c,\sigma_j} \approx k_i^c = \pi_i^c k_i \quad . \tag{3.28}$$

Combining Eqns. 3.15 and 3.28, we can compute flip rate centrality W_i as a function of degree and community participation:

$$W_i \approx 1 - \sum_{c=1}^{C} \left(\frac{\exp\{\beta \pi_i^c k_i\}}{\sum_{c'=1}^{C} \exp\{\beta \pi_i^{c'} k_i\}} \right)^2 \quad . \tag{3.29}$$

In Figure 3.4(b,f) we test this approximation for both the SBMb_1 and SBMb_2, with further detail and interpretation provided on this result in the next Section.

3.4 Interplay of bridgeness and dynamical centralities

3.4.1 Empirical analysis

Figures 3.4 (a,b) show empirical results for the SBMb_1 with $N = 500$ nodes, including 25 bridge nodes with $p_R^{\text{bridge}} = 1$ and 75 border nodes with $p_R^{\text{border}} = 0.2$, homogeneously distributed across $M = 25$ initial cliques. Figures 3.4 (e,f) studies the SBMb_2 with $N = 500$ and 400 bulk nodes. In both cases, we simulate $R = 100$ realisations of the rewiring protocol from the initial condition of isolated cliques to obtain an ensemble of model instances to compute the SBMb-ensemble bridgeness $\langle S_i \rangle$ (see Eq. 3.18). Using adequate temperature and coupling parameters on SBMb networks, we study the Kuramoto and Potts models at the metastable state where dynamical structures reminiscent of topological communities emerge. As previously described, we obtain Monte-Carlo estimators of dynamical centralities $\langle W_i \rangle$ and $\langle 1 - \rho_i \rangle$ for each SBMb network realisation.

The results in Figures 3.4(a,b) reveal a clear interplay between bridgeness and both dynamical centralities for the SBMb_1. Bridge nodes, which by definition have the highest values of $\langle S \rangle$, also present significantly higher levels of dynamical centrality: under Kuramoto dynamics, bridge nodes store the most unsynchronized regions across the network; under Potts rules, they have distinctively high flip rates W, thus clearly corresponding to so-called blinkers. In Figure 3.4(b), black-cross markers show that flip rate predictions from Eq. 3.29 match correctly our numerical results.

Furthermore, Figures 3.4(e,f) show that similar results apply to the SBMb_2. In this case, the model produces a continuous spectrum of bridgeness and consequently it induces a continuous profile of dynamical centrality, which varies depending on the tuning parameter used. Again, black markers show the analytical predictions of Eq. 3.29 matching our numerical results for flip rates also in this heterogeneous setting.

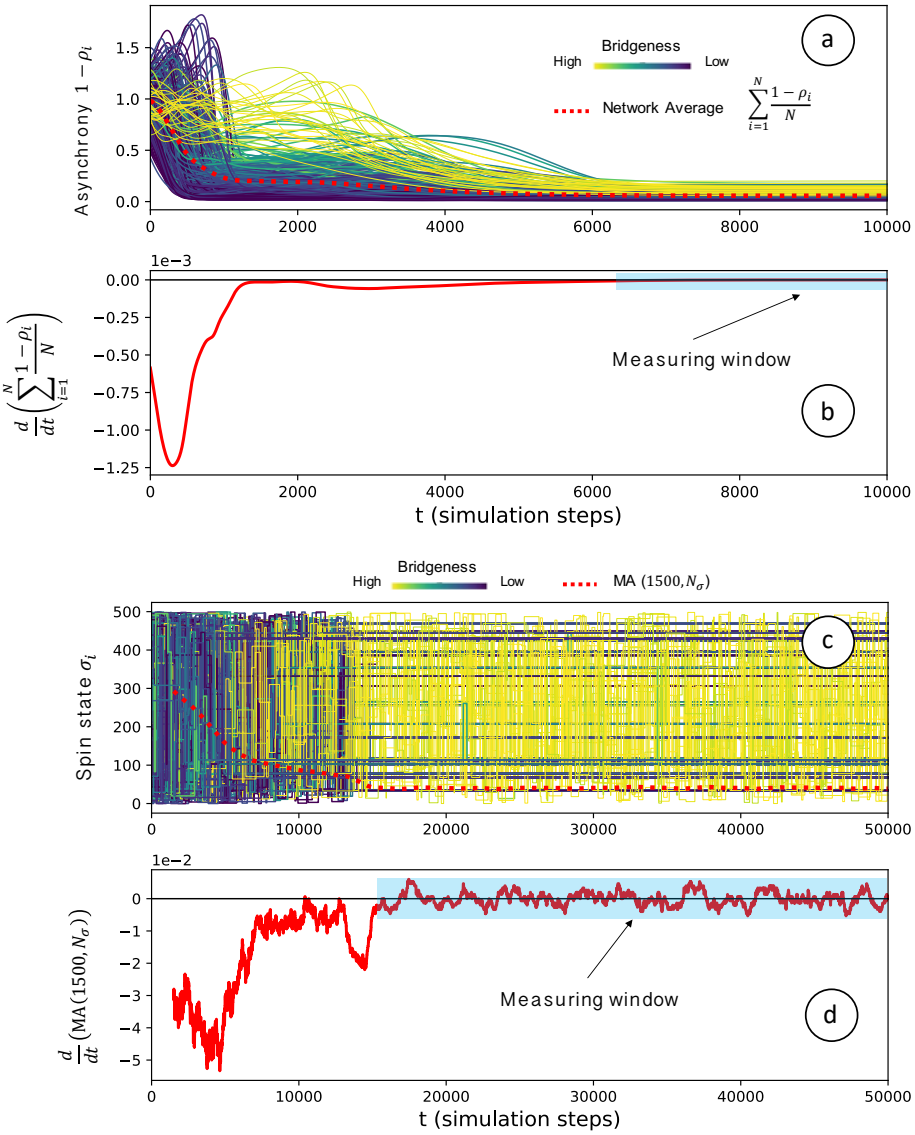


Figure 3.3: Simulated realisations of the Kuramoto (a,b) and the Potts (c,d) models for an underlying SBMb_1 with $N = 500$ nodes, including 25 bridge nodes with $p_R^{\mathrm{bridge}} = 1$ and 75 border nodes with $p_R^{\mathrm{border}} = 0.2$, homogeneously distributed across the $M = 25$ initial cliques. Panel (a) shows the evolution over time of asynchrony for each node-oscillator coloured according to their bridgeness centrality; the dashed red line corresponds to the network average, with its derivative plotted in panel (b) (see convergence condition in Eq. 3.22). Panel (c) shows the evolution of the spin state of each node with colour according to bridgeness; the dashed red line shows the number of unique spins present in the network, with a moving average filter of period 1500, and its derivative over time plotted in panel (d) (see convergence condition in Eq. 3.27).

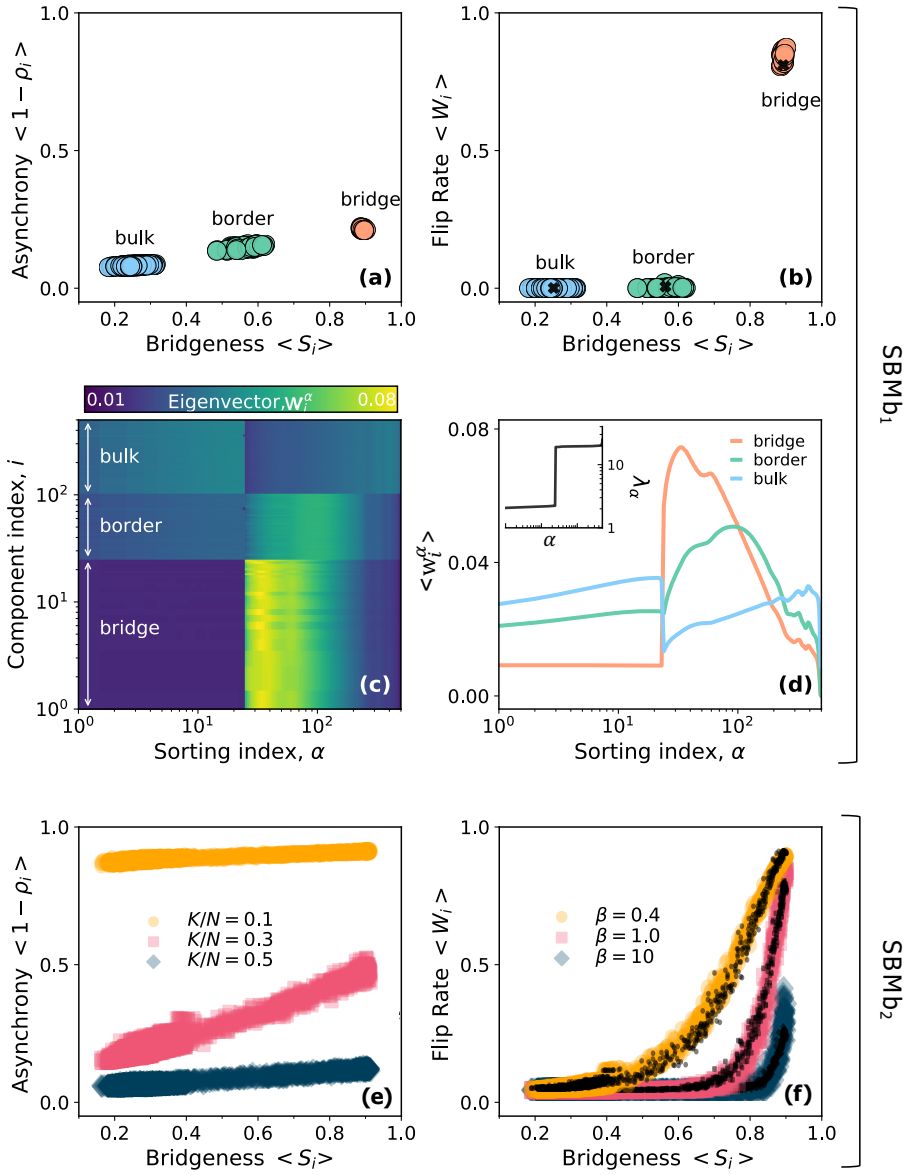


Figure 3.4: **(a)** Bridgeness (Eq. 3.16) and Asynchrony (Eq. 3.21) centrality measures for an SBMb$_1$ ensemble of 100 realisations, parametrised as in Figure 3.3, using Kuramoto dynamics with $K/N = 0.4$ averaged over 1000 simulations. **(b)** Same as previous, showing Flip Rate centrality $\langle W_i \rangle$ using $\beta = 1$ averaged over 1000 simulations. Black markers show predicted flip rates (Eq. 3.29). **(c)** Laplacian eigenspectrum v for the same SBMb$_1$ ensemble: rows show the component i of each eigenvector v^α, sorted by nodal role; columns show eigenvectors sorted by corresponding eigenvalue index α. **(d)** Same as previous, but showing the average eigenvector component value $\langle w_i^\alpha \rangle$ in each nodal category. Inset: sorted eigenvalues λ_α. (e) Same as (a) using an SBMb$_2$ parametrised as in Figure 3.2 using different K/N values. (f) Same as (b) using an SBMb$_2$ parametrised as in Figure 3.2 using different β values.

3.4.2 Effect of tuning parameters

Note that dynamical centrality measures arise from ordering dynamics which are sensitive to their corresponding tuning parameter, namely the inverse temperature β for flip-rate $\langle W_i \rangle$ and the coupling strength K for asynchrony $\langle 1-\rho_i \rangle$. Tuning parameters determine the trade-off between random fluctuations of internal states and the influence of local surrounding states. The effect of K on the evolution of oscillators is explicit from Eq. 3.19, whereas β has an explicit effect on the spin-transition probabilities as shown in Eq. 3.24.

Figure 3.5 exemplifies through the SBMb_1 how tuning parameters have an important effect on the interplay between dynamic centralities and bridgeness. The figure shows node clustering in the horizontal axis, according to their bridgeness category. Depending on the value of tuning parameters, nodes also cluster on the vertical axis, indicating that they can be distinguished by measure of their dynamical centrality.

Lower value (e.g. $\beta = 0$ and $K = 0$) induce dynamics dominated by noise, which prevents the system from reaching the partially ordered states where patterns relating bridgeness and dynamic centrality emerge. For larger values (e.g. $\beta = 10$ and $K = 10$), ordered states dominate the network, although some bridgeness-related patterns persist on dynamical centrality: for the Potts model, bridge nodes retain their blinker behaviour, although border and core nodes are indistinguishable in terms of dynamic centrality; for the Kuramoto model, dynamic centrality still differentiates bridge, border and core nodes at the local level, although the scale of asynchrony is so low that the network would appear as generally synchronised from a macroscopic perspective.

Figures 3.4(e,f) show a similar effect for the SBMb_2. We can see that asynchrony centrality retains the capacity to distinguish nodes according to their bridgeness as the value of the tuning parameter increases. In contrast, low and medium bridgeness node become less distinguishable as we increase the tuning parameter for flip-rate centrality.

3.4.3 Laplacian localisation of dyamical centralities

Graph Laplacian matrices have been widely used to describe the relation between structure and dynamical behaviour in complex networks regarding diffusive processes for random walks, coupled oscillators or epidemic spreading [13]. A common example of Laplacian matrix is the combinatorial Laplacian, which for unweighted networks reads:

$$L_{ij} = \delta_{ij} k_i - a_{ij} \quad . \tag{3.30}$$

In most cases, these matrices are studied through their spectrum of eigenvectors w^α and corresponding eigenvalues λ_α, which relate as follows:

$$\sum_{j=1}^{N} L_{ij} w_j^\alpha = \lambda_\alpha w_i^\alpha \quad . \tag{3.31}$$

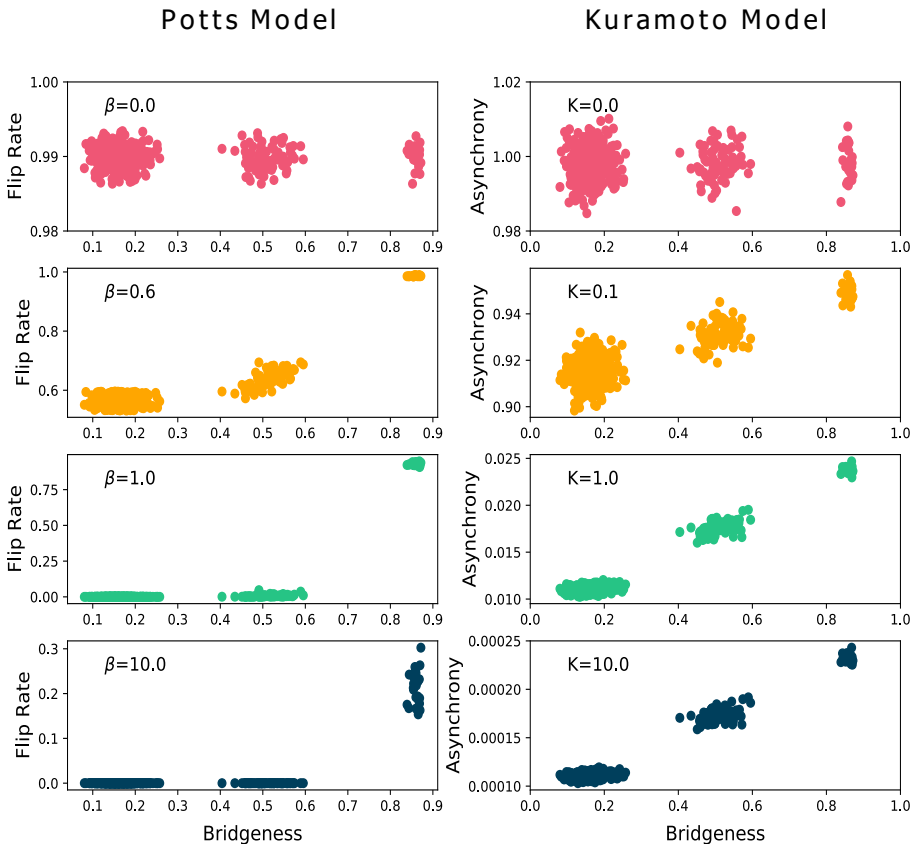


Figure 3.5: Each point represents nodal averages over an SBMb_1 ensemble of 100 realisations, parametrised as in Figure 3.3. Different clusters in the vertical axis indicate groups of nodes that can be distinguished by dynamical centrality. **Right column:** Bridgeness (Eq. 3.16) and Asynchrony (Eq. 3.21) centrality measures for increasing values of K/N in Kuramoto dynamics averaged over 1000 simulations. **Left column:** Same as right column, but showing Flip Rate centrality $\langle W_i \rangle$ using increasing values of β each averaged over 1000 simulations.

Global dynamical properties, such as the stability of the synchronised state for several network topologies, are usually studied through the Master Stability Function formalism [51], [89]–[91]. This typically involves finding expressions for the extreme eigenvalues of the Laplacian, such as the eigenratio:

$$R_\lambda = \frac{\lambda_N}{\lambda_1} \quad , \tag{3.32}$$

where λ_1 and λ_N are the minimal and the maximal non-zero eigenvalues.

A different research path is that of Laplacian localisation. It builds on growing evidence that, for many complex networks, there exists a relation between the components of each eigenvector and the local topological properties of nodes associated with such components. Given that eigenvectors with similar eigenvalues tend to represent differentiated dynamical modes of the process occurring on a network, Laplacian-eigenvector localisation is a great tool to diagnose which modes are dominated by each type of node. For example, in degree-heterogeneous networks, higher degree hub-nodes are known dominate the eigenvectors with largest eigenvalues, exhibiting eigenvector localisation and degree-eigenvalue correspondence, which helps identifying the dynamical role of each degree class [79], [92]. Also for modular networks, it is well known that nodes from the same community dominate the same eigenvectors and have closely similar eigenvalues, providing evidence that each module has its own dynamical modes which can be used to detect communities [93]–[95].

In our case, here we study the Laplacian spectrum of the SBMb_1, showing how localisation phenomena is also applicable to bridgeness centrality. Panels in Figure 3.4(c,d) show numerical results for the combinatorial Laplacian (Eq. 3.30). The heatmap in Figure 3.4(c) represents the eigenvector spectrum w^α, where α has been sorted according to eigenvalues (see inset of Figure 3.4(d)) from smallest ($\alpha = 1$) to largest ($\alpha = 499$), excluding the first trivial null eigenvalue:

$$0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{499} \quad . \tag{3.33}$$

Eigenvector components have also been sorted according to the bridgeness category (of each node (bridge, border and bulk), so that localisation can be more easily visualised. The heatmap shows clear bridgeness-localisation throughout the spectrum, i.e. nodes with similar bridgeness centrality exhibit similar component values in each eigenvector. Localisation is also evident in Figure 3.4(d), where we show the sample mean $\langle w_i^\alpha \rangle$ and confidence interval $\langle w_i^\alpha \rangle \pm 1.96\sigma_{w_i^\alpha}$ at each eigenvector w_i^α for bridge, border and bulk nodes respectively.

As mentioned above, the Laplacian eigenvectors form a basis where to project functional observables (such as phase and frequency in synchronisation) onto a coordinate system of normal modes. In this sense, Figure 3.4(d) reveals two groups of such modes in the spectrum of the SBMb_1. The first group of modes, with the smallest eigenvalues

(see inset), represents internal community dynamics (there are 25 modes, one for each community) and thus are dominated by bulk nodes. The second group of modes, of larger eigenvalues, concerns the global ordering of the network: in fact, after the eigengap we find the modes with strongest localisation, which are dominated by bridge nodes, emphasising their role as promoters of global synchronisation and consensus.

3.5 Detection of critical nodes

Given the interplay between dynamical and topological centralities we have described above, it is reasonable to assume this relation can be used to uncover critical nodes –those which compromise the robustness of their network when attacked or removed– even when physical connections cannot be completely observed. Logically, removing the regions of high bridgeness will quickly lead to the collapse of a modular network, and our framework shows how to target those regions by looking at local functional behaviours such as asynchrony (3.21) and flip rate $\langle W_i \rangle$.

In order to test this idea we use a site percolation process where an increasing fraction of nodes (along with their incident links) is sequentially removed using attacks targeted at dynamical centrality [96]–[98]. We measure network robustness to such attacks using two well-known indicators: size of the largest connected component; and network efficiency, which quantify how efficiently information flows across a network [99], and is defined as:

$$\mathrm{Eff}(G) = \frac{1}{N(N-1)} \sum_{i,j \in G} d_{ij}^{-1} \quad , \tag{3.34}$$

where d_{ij} is the shortest-path distance between nodes v_i and v_j.

We study such functionally-targetted percolation process in three types of networks with increasing complexity, comparing this to targetting degree, betweenness and bridgeness. As a base model, we use the SBMb_1 parametrised as in the previous sections, that is with $N = 500$ nodes, including 25 bridge nodes with $p_R^{\mathrm{bridge}} = 1$ and 75 border nodes with $p_R^{\mathrm{border}} = 0.2$, homogeneously distributed across $M = 25$ initial cliques. Its homogeneous and correlation-free degree structure, coupled with its uniform community-size distribution, controls the interplay between bridgeness, asynchrony and flip rate. Secondly, we consider the Random Geometric Graph (RGG) [100] with $N = 500$ nodes in a unit square with a connection radius $R = 0.07$. These graphs lie just above the percolation threshold and exhibit a distribution of both community sizes and degree broader than the SBMb, thus providing a more general framework where to test our method. Finally, we use the empirical structure of the US Western States Power Grid (WSPG) [72], a large spatial network comprising 4941 nodes representing electricity generation and transformation stations and 6594 links depicting distribution lines amongst them. As many real-world networks, it is more heterogeneous and presents richer correlation

structures than the models considered above. This network has been extensively studied from many perspectives, including its community and bridgeness structure [101], and thus it provides a realistic test for our percolation method.

Figure 3.6 summarises the results of the methodology described above, where nodes are sequentially removed in order of centrality magnitude. Note that we also add a control strategy where we removed nodes randomly. Overall, attacks based on dynamical targetting are effective when compared to attacks based on topology in the three types of networks considered. In addition, we can see how targetting asynchrony is generally more efficient than targetting flip rate, and as in [101] bridgeness more than betweenness. In particular, the SBMb_1 is most and equally vulnerable to dynamical and bridgeness based attacks, whereas attacks based on betweenness and especially degree are less effective. The RGG also dismantles fastest when targetting asynchrony at early stages, whereas flip rate performs slightly worst than bridgeness but better than betweenness. Interestingly, the WSPG is most vulnerable to attacks based on bridgeness. For the WSPG network, however, all strategies have a similar performance. This indicates that the presence of centrality correlations makes hubs also important network bridges, and consequently are far more critical for the robustness of the system but can be similarly detected by most centrality measures. Finally we can see that, on the whole, due to the long-range connections induced by rewiring the SBMb shows higher robustness (requires more node removals) and a notably sharper transition to a dismantled state than the RGG and the WSPG. This can be seen in the graph layouts in the lower section of Figure 3.6. The layouts represent each of the three networks considered when the size of the largest component has decreased to 40% of its original size due to asynchrony-based percolation attacks. We can see that reaching this point has required the removal of 20% of nodes in the SBMb_1, whereas only 10% and 5% node removals are required for the RGG and WSPG respectively.

Figure 3.5 exemplifies how the efficiency of asynchrony and flip-rate based attacks in dismantling a network will depend on the tuning parameters from the underlying dynamical processes. We use the same SBMb_1 as before, with $N = 500$ nodes, including 25 bridge nodes with $p_R^{\text{bridge}} = 1$ and 75 border nodes with $p_R^{\text{border}} = 0.2$, homogeneously distributed across $M = 25$ initial cliques. In fact, similarly to what we have shown in Section 3.4.2, low tuning parameter values (e.g. $\beta = 0$ and $K = 0$) promote very noisy processes where nodes become functionally undistinguishable, therefore yielding ineffective dismantling performance. As tuning parameters are increased (e.g. $\beta = 1$ and $K = 1$), the emergence of partially ordered states allows to distinguish nodes according to their bridgeness category: we can see how both network efficiency and size of largest component decrease sharply after having removed 100 nodes targetting their functional behaviour, mainly because they correspond to all bridge and border nodes which actually hold the network together. For higher tuning parameters (e.g. $\beta = 10$ and $K = 10$) the

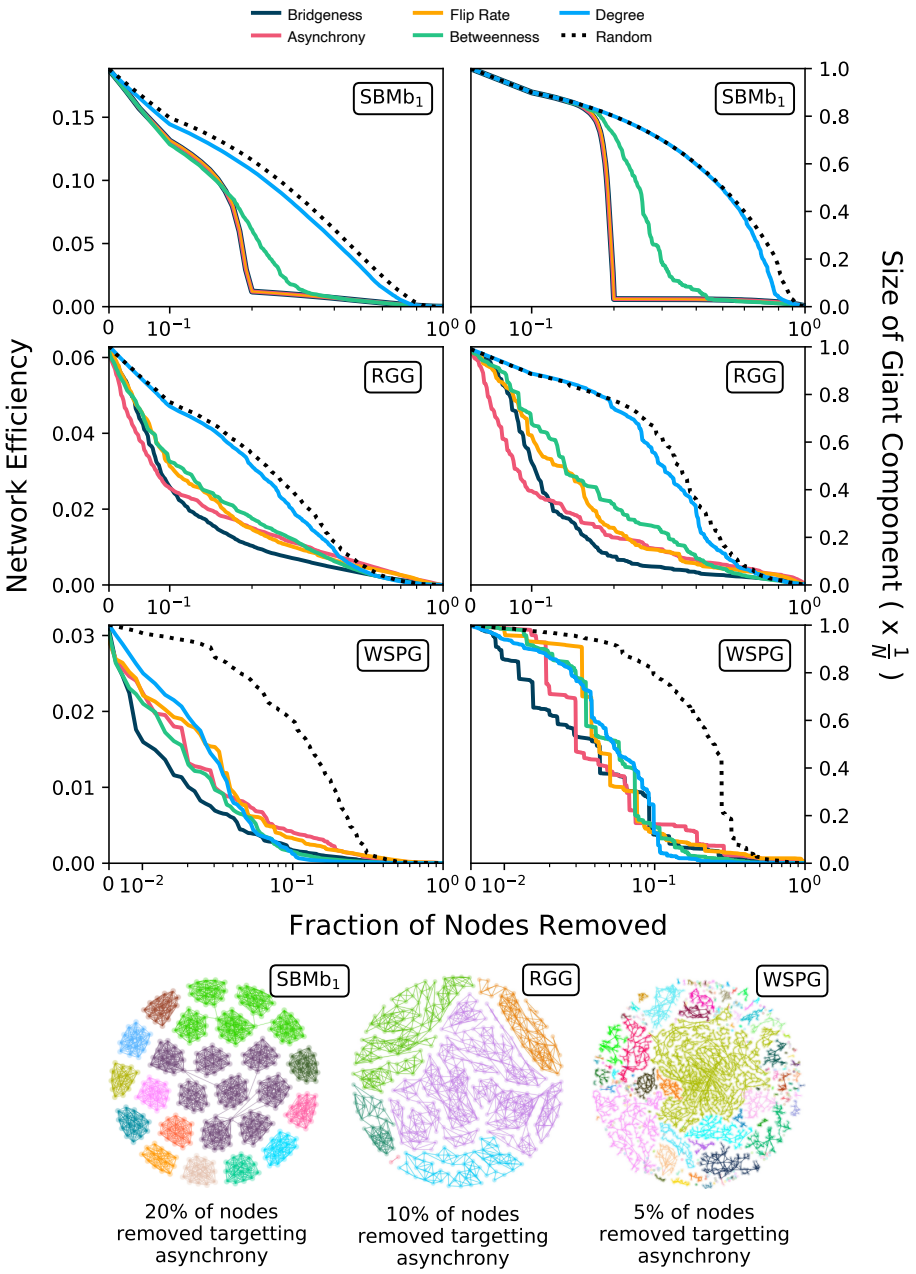


Figure 3.6: Reduction of Network Efficiency (left) and Size of the Largest Component (right) for sequential node removals, in order of centrality magnitude, targetting both topological and dynamical centralities. We use 100 realisations of the SBMb_1 parametrised as in Figure 3.3 (top), 100 realisations of a Random Geometric Graph with connection radius $R = 0.07$ (centre) and the Western United States Power Grid (bottom). Graph layouts at the bottom show the state of each network when the size of the largest component has reached 40% of original size, with each connected component coloured differently. The underlying text shows the amount of node removals needed to reach that state, by targetting asynchrony.

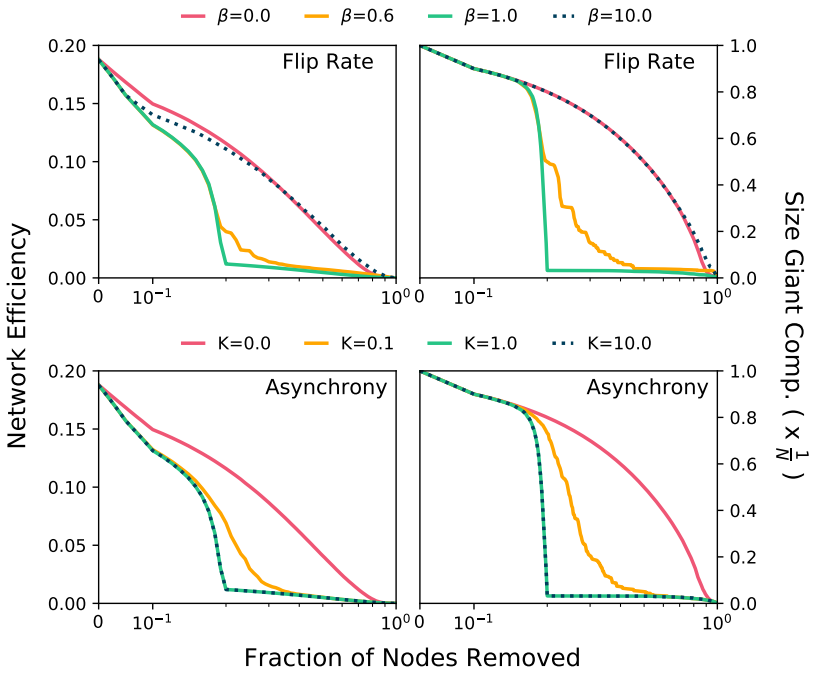


Figure 3.7: Reduction of Network Efficiency (left) and Size of the Largest Component (right) for sequential node removals targetting flip rate (top) and asynchrony (bottom) centralities, using 100 realisations of the SBMb_1 parametrised as in Figure 3.3. Different lines represent different values of the corresponding tuning parameter.

results are again different for both processes: flip-rate becomes ineffective in dismantling because border and core nodes become indistinguishable (see Figure 3.5); on the contrary, asynchrony is still useful to distinguish each category and thus is still an effective target for network attacks.

3.6 Discussion

Throughout this chapter we have reviewed some important topological features of modular complex networks, and have seek to uncover several aspects of their interplay with dynamical system models. After a revision of the generative Stochastic Block Model (SBM) and its applications to community detection, we have presented different ways of measuring and important mesoscopic property of modular networks, namely bridgeness centrality. Bridgeness measures to which extent a given node serves as an intermediary between different communities. We have proposed a simple generative mechanism based on link rewiring, the Stochastic Block Model with bridgeness (SBMb), that can produce networks with arbitrary distributions of bridgeness across its nodes.

Given the important effect that modular structures have on all macroscopic aspects of dynamical processes on networks, it is reasonable to expect that bridgeness may induce changes in the local dynamical behaviour of nodes. That is, since communities tend to produce differentiated internal dynamics, nodes connecting several of such communities should be expected to produce distinguishable functional patterns. Using the SBMb in conjunction with two paradigmatic dynamical system models, that of Potts and that of Kuramoto, we have positively tested this hypothesis. Using information from local observables, we have proposed two dynamical centrality measures, flip-rate and asynchrony. For each spin or oscillator, these measures asses the level of internal state disorder relative to surrounding partially-ordered states. We have shown how that, when the tuning parameters and the modular structure are strong enough to produce such partially-ordered metastable states, bridgeness centrality is highly correlated with dynamical centrality. We have found further evidence of such interplay between topology and dynamics by uncovering Laplacian eigenvector localisation phenomena in the SBMb. In this sense, high-bridgeness nodes contribute to well-differentiated eigenvector modes that influence the ordering dynamics at different network and time scales., given the pervasive character of Laplacian matrices this interplay could be extended to other physical processes, such as spreading [102] or voter systems [32], and more complex topologies such as multilayer networks [103].

Given that high-bridgeness nodes will tend to connect clusters of nodes which are otherwise sparsely connected, exploiting the interplay with bridgeness centrality is particularly important in the context of network robustness to targetted attacks. We show how node-removal strategies targetting flip-rate and asynchrony perform significantly well in comparison with topological centralities such as degree, betweenness and bridgeness. We have positively tested this result in three types of networks: the SBMb, a synthetic model with planted partitions and controllable bridgeness distribution; the Random Geometric Graph, which has heterogeneous community sizes; and the empirical Western US Power Grid, a spatial transportation network with more realistic features.

To what extent can this framework be applied to data gathered from empirical processes? We conjecture that the dynamical centrality measures we have presented here could indeed be extended to realistic situations. For instance, the dissemination of information in social networks has been previously described under the spin-ordering paradigm using the voter and Axelrod models, amongst other processes [7]. It is reasonable to assume that in the context of Online Social Networks (e.g. Twitter or Reddit), the volatility of user-opinion as extracted from posts could provide a measure similar to flip-rate. Under the framework presented here, such measure could be used in the detection of central actors bridging different affinity-clusters [104]. Another example would be neurobiological networks inferred from fMRI or EEG signals, which are modular and spatially embedded, showing evidence of important brain regions bridging functionally-specialized areas [105]: given that Kuramoto models have been previously applied to neuronal networks [106], [107], our framework could also help in the detection of such bridging regions targetting locally asynchronous patterns. Even in protein-protein interaction networks, organised modularity is manifested by date-hub proteins capable of interacting with several functional modules: in fact, date-hubs can be detected from their dynamical behaviour through genetic interaction profiles [61], providing further empirical insights for the framework presented here.

Chapter 4

Network uncertainty propagation

The study of critical phenomena has been, and still is, a fruitful area of research in network science [52]. Critical phenomena in networks include a wide set of aspects, from structural changes in networks, or percolation phenomena [108], to epidemic [73] or synchronization [51] thresholds and many other phase transitions in dynamical processes defined on networks [52], [109]. The estimation of the critical threshold is of upmost importance to predict the onset of the phase transition, and hence a major concern in several applications, such as the containment of an infectious disease [97] or the control of synchronization in the power grid [110], [111]. However, an accurate estimation of the threshold is often elusive and costly because it depends on the particular details of the whole network structure, usually through its eigenvalues.

As network science becomes more and more extended, its potential applications grow fuelled by the necessity of analyzing data produced in diverse fields of research, such as sociology, biology, experimental physics, etc. However, the data collected in any of the former fields is not free from experimental error, induced for example by sampling biases, device accuracy, or mistakes in data entry. Nevertheless, the literature on network science usually dismisses these error sources, and produces results that are only valid if data is error free. Some authors have concentrated their attention on inference of missing data in networks [112]–[115]. However, no similar attention has been paid to the propagation of uncertainty from the structure to the properties of dynamical processes running on it.

The lack of works devoted to the analysis of error propagation in networks is probably due to the fact that many studies consider unweighted networks, where a link is a binary variable denoting its existence or not. However, the vast majority of networks are weighted, i.e. the existence or not is valued by its intensity. The accurate determination

of the weight is unlikely, and therefore, the error in their numerical values will influence any particular measurement of the network properties.

Here, we present a study of error propagation in networks where links are subject to uncertainty in their weights, and wonder about the effect that this uncertainty will have in the determination of the critical threshold. In particular, we focus on those dynamical processes in which the critical point is known to be inversely proportional to the largest eigenvalue of the connectivity matrix. In section II, we present the particularities of our analysis and derive our main results, in section III we study the range of uncertainty in the critical point for different network structures, and finally, in section IV we discuss the implications and limitations of the current study, paving the way for new analysis to come.

4.1 Uncertainty in the critical threshold

We consider a dynamical process running on top of a complex network with N units. We restrict the study to the class of dynamical models in which a phase transition occurs at a critical value of the coupling intensity (the threshold), and where this value is given in terms of the largest eigenvalue λ_{max} of the network connectivity matrix A whose values represent the weighted structure of the network [116]

$$K_c = \frac{K_0}{\lambda_{max}(A)}, \tag{4.1}$$

where K_0 is a constant that depends on the specific details of the particular process. Without loss of generality, we fix $K_0 = 1$. Eq.4.1 estimates the threshold for a wide variety of dynamical processes, including the synchronization of heterogeneous phase-oscillators [51], the onset of endemicity of a disease in epidemic models [73], [117], and the phase transition in the Ising model in networks, to name a few [52], [108], [109]. The aim of this work is to understand how small noise in the entries of A affects the statistical properties of the macroscopic threshold given by Eq.4.1, without looking into the details of a specific dynamical model. For the sake of simplicity, we assume that the noise in the entries is gaussian and uncorrelated (white gaussian noise) where each weight is drawn from a normal distribution $N(\mu, \sigma^2)$, with $\mu > 0$ the average weight and σ^2 its variance. Nevertheless, the proposed analysis can be extended to other distributions of noise, either theoretical or obtained through empirical measurements.

To study the exact statistics of K_c in Eq.4.1 induced by the presence of noise, one could use in principle the available tools from Random Matrix Theory [118], [119] and Spectral Graph Theory [120], [121]. However, it becomes very challenging to study noisy sparse networks with arbitrary degree distributions in these frameworks. Here, we use an alternative approach, based on applying error propagation to the mean-field

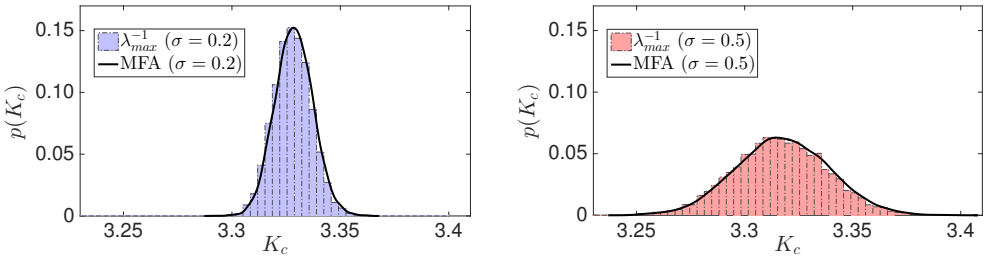


Figure 4.1: Empirical distribution of the critical point K_c governed by Eq.4.1 (boxes) and MFA (solid lines) in an Erdős-Rényi network with $N = 200$, $p = 0.3$, $K_0 = 1$, $\mu = 1$ for two different noise intensities ($\sigma = 0.2$ grey and $\sigma = 0.5$ red). The distribution corresponds to 10^4 independent realizations of the noise.

approximation of Eq.4.1. This approximation obviously restricts the validity range of the analysis, however, the results are found to be very accurate in some scenarios and, more importantly, they provide clear analytical insight on how the uncertainty in the structure affects the determination of the critical threshold.

Our derivation starts assuming a mean–field approach. For simplicity, we restrict to the case of undirected (symmetric) networks. Under the aforementioned conditions, the critical threshold in Eq.4.1 can be approximated [122]–[124] by

$$K_c = \frac{\langle s \rangle}{\langle s^2 \rangle}, \tag{4.2}$$

where $\langle s^n \rangle$ is the n-moment of the strength distribution (the strength of a node is the sum of in-coming/out-going weights). Eq.4.2 can also be obtained directly from the equations of motion of the dynamical process (for instance in the Kuramoto Model [123]) by assuming that the local field in a node is proportional to the global field weighted by the in-strength of the node [51]. Below, we will refer to Eq.4.2 as the Mean-Field approximation (MFA).

First we test the accuracy of the critical threshold in the MFA, Eq.4.2, compared to the exact result, Eq.4.1, in Erdős-Rényi networks with uncertainty in the weights. In Figure 4.1 we plot the threshold distribution for two different values of the intensity of the uncertainty σ. We observe that the MFA accurately determines the distribution, and that the values of the expected critical threshold K_c and its variance are clearly dependent on σ. In general, we expect our results to be accurate in the cases in which the approximation of Eq.4.2 remains valid.

Using Eq.4.2, we can express K_c in terms of the moments of the degree distribution

and noise parameters. The detailed calculations are shown in Section 4.5.1. We obtain

$$K_c = \frac{\sum_{i=1}^{N} \mu_i k_i}{\sum_{i=1}^{N} \mu_i^2 (k_i^2 - k_i) + \sum_{i=1}^{N} \langle w^2 \rangle_i k_i}, \tag{4.3}$$

where μ_i is the average weight of node v_i, and $\langle w^2 \rangle_i$ the average second moment of the weight distribution for node i. In random homogeneous networks, for sufficiently large degree ($k_i \gg 1$), we can approximate $\mu_i = \mu$, and $\langle w^2 \rangle_i = \sigma^2 + \mu^2$ in Eq.4.3. This approximation allows to write down a simple relation between the mean of the critical threshold and the uncertainty of the network as

$$\langle K_c \rangle \approx \frac{\mu \langle k \rangle}{\mu^2 \langle k^2 \rangle + \sigma^2 \langle k \rangle}. \tag{4.4}$$

Interestingly, the naïve approximation in Eq.4.4 already informs that the critical threshold decreases as the noise intensity σ increases. This can be understood because the noise increases the structural heterogeneity of the network, and heterogeneity tends to make the epidemic threshold vanish. Note that for $\mu = 1$ and $\sigma = 0$, we recover the usual threshold for unweighted, undirected networks [124] and for $\sigma \ll 1$, $\langle K_c \rangle \approx \langle k \rangle / \mu \langle k^2 \rangle$.

4.2 Error propagation on the critical threshold

Now, we estimate confidence intervals for the uncertainty of K_c, that is the standard deviation named here δK_c (or the variance $(\delta K_c)^2$). For this purpose, we use the method of error propagation[125], [126], that quantifies how the error in the microscopic variables of a system (the $2N$ random variables in our nodal description) propagate through a macroscopic quantity (the critical threshold K_c). In a first-order expansion, we have

$$(\delta K_c)^2 \approx J_0^T \mathbf{V} J_0, \tag{4.5}$$

with $J \in R^{2N}$ the Jacobian of the system evaluated at the mean values of the random variables $\vec{\mu}$ and $\langle \vec{w^2} \rangle$ and $\mathbf{V} \in R^{2N \times 2N}$ the covariance matrix, which depends on the full connectivity matrix A. The details of these calculations (for white gaussian noise and fixing $K_0 = 1$) are shown in Section 4.5.2. Finally, we obtain the following closed form expression

$$\begin{aligned} (\delta K_c)^2 &\approx a[\mu^4 (2\langle k \rangle \langle k^3 \rangle \\ &- \langle k^2 \rangle^2) - 2\mu^2 \sigma^2 (\langle k \rangle \langle k^2 \rangle - \langle k \rangle^2) + \sigma^4 \langle k \rangle^2] \end{aligned} \tag{4.6}$$

with $a = 2\sigma^2 \langle k \rangle / [N(\mu^2 \langle k^2 \rangle + \sigma^2 \langle k \rangle)^4]$.

Eq.4.6 shows that, beyond the non-linear dependence on the network and noise parameters, the uncertainty in the threshold is a finite-size effect, and decays with $N^{-1/2}$.

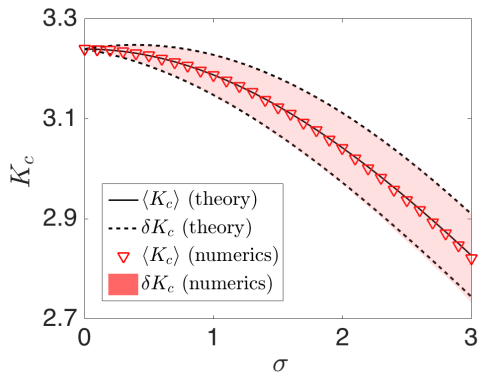


Figure 4.2: Numerics (Eq.4.1) vs theory (Eqs.(4,6)): mean and standard deviation of the threshold K_c depending on the noise intensity σ for an Erdős-Rényi network with $N = 200$, $p = 0.3$, $\mu = 1$, and 5000 independent realizations for each value of the noise intensity σ.

To compare networks of different sizes, we will scale the threshold by the size N in the current analysis.

In Figure 4.2, we show the accuracy of the theoretical expressions for an Erdős-Rényi network, confirming the validity of the approach, at least for small noise and homogeneous structures. Note that the linear approximation used in Eq.4.5 is valid as far as [126]

$$J_0^T \mathbf{V} J_0 \gg \frac{1}{2} \mathrm{Tr}[(\mathbf{H_0 V})^2] \tag{4.7}$$

where $\mathbf{H_0} \in R^{2N \times 2N}$ is the Hessian matrix of the system evaluated at the mean values of the random variables. The detailed calculations of $\mathbf{H_0}$ are shown in 4.5.2. Both terms in Eq.4.7 depend implicitly on the value of the noise, so their scaling with σ will determine the range of validity of Eq.4.6. We numerically examine the goodness of both the linear, Eq.4.5 and the second-order approximation for the uncertainty δK_c

$$(\delta K_c)^2 \approx J_0^T \mathbf{V} J_0 + \frac{1}{2} \mathrm{Tr}[(\mathbf{H_0 V})^2] \tag{4.8}$$

against the numerical results obtained for the Erdős-Rényi network analyzed so far, and also for a real world network with large size and heterogeneous connectivity patterns (the worldwide air transportation network). The air transportation network was constructed using data from the website openflights.org, which has information about the traffic between airports updated to 2012, data available from [97]. This network accounts for the largest connected component, with 3154 nodes and 18,592 edges.

Figure 4.3 shows that the first and second order solutions are practically indistinguishable for small noise, therefore validating the result in Eq.4.6 in this regime. The

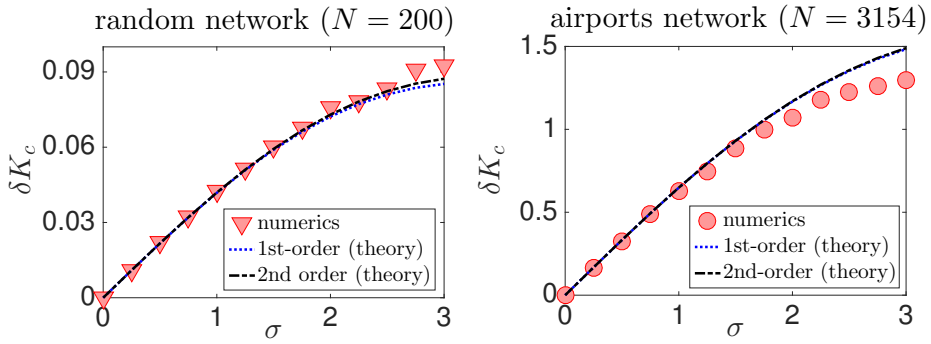


Figure 4.3: Numerics vs theory: standard deviation of the critical threshold δK_c depending on the noise intensity σ with $\mu = 1$ for a (left) fixed Erdős-Rényi network ($N = 200$, $\langle k \rangle = 60$, $p = 0.3$) and (right) the empirical network of airports ($N = 3154$, $\langle k \rangle \approx 6$) for 2000 independent realizations for each value of the noise. Results have been rescaled by N.

deviation of the theory from the actual values in the empirical network (right plot in Figure 4.3) points towards another direction: the goodness of the MFA itself. Basically, the theory is expected to be accurate for networks that deviate from a random structure as long as the MFA in Eq.4.2 holds. We refer the reader to the literature [122], [124], [127] for details on the validity of the MFA. Moreover, it is important to remark that even if the MFA holds, the method of error propagation (at any order) can only be applied in our problem when the mean of the signal μ is sufficiently large compared to the noise.

4.3 The role of the topology in error propagation

Network structure plays an important role in the uncertainty range of K_c. After the finding of Eq.4.6, some interesting questions arise: does the heterogeneity induce an increase of the critical fluctuations with respect to a homogeneous network? Is the behavior of (δK_c) monotonous with the moments of the degree distribution of the network? If not, is there any particular structure that maximizes the uncertainty of the critical point induced by noise in the weights?

To answer these questions, we consider the regime where networks are sufficiently large and $\sigma \ll \mu$. Then, we can approximate Eq.4.6 by its leading term, neglecting terms in σ larger than $\mathcal{O}(\sigma^2)$ as:

$$(\delta K_c)^2 \approx 2\sigma^2 \frac{2\langle k \rangle \langle k^3 \rangle - \langle k^2 \rangle^2}{N \langle k \rangle^3} \langle K_c \rangle^4. \tag{4.9}$$

Note that δK_c increases linearly with the noise intensity and scales with $\langle K_c \rangle^2$. We know that $\langle K_c \rangle^2$ is reduced by the heterogeneity of the degree distribution, and therefore one would expect δK_c to follow the same trend. However, the nonlinear dependence on the moments of the degree distribution could change this intuition.

To understand this effect, we choose first as a reference the most homogeneous network we can consider, a regular network, i.e. $k_i = k, \forall i$. We compute K_c and δK_c for a regular network, obtaining

$$\begin{aligned} \langle K_c \rangle_{\text{reg}} &\approx \frac{1}{\mu k}, \\ (\delta K_c)^2_{\text{reg}} &\approx \frac{2\sigma^2}{N\mu^4 k^3}. \end{aligned} \tag{4.10}$$

The role of the heterogeneity will be detected by comparing $(\delta K_c)^2$ with $(\delta K_c)^2_{\text{reg}}$ for networks with the same size and average degree, and for the same noise parameters μ and σ. After some algebra, the condition for a given network to display higher uncertainty in K_c than a random regular network reads

$$\langle k^3 \rangle > \frac{\langle k^2 \rangle^2}{2\langle k \rangle} \left(1 + \frac{\langle k^2 \rangle^2}{\langle k \rangle^4} \right). \tag{4.11}$$

Now, we can use Eq.4.11 to evaluate the role of heterogeneity. Let us consider a power-law distribution $p(k) \approx k^{-\gamma}$, where the exponent γ controls the tail of the distribution. For the value $\gamma = 3$, one recovers the well-know scale-free network that emerges from preferential attachment [71]. For lower (higher) values of γ, the network becomes more (less) heterogeneous. For a finite power-law network, the moments of the degree distribution are given by

$$\langle k^n \rangle = \frac{(-\gamma + 1)(k_{\text{max}}^{n-\gamma+1} - k_{\text{min}}^{n-\gamma+1})}{(n - \gamma + 1)(k_{\text{max}}^{\gamma+1} - k_{\text{min}}^{\gamma+1})}. \tag{4.12}$$

By fixing the value of k_{min}, we can explore the space of networks with a given (γ, k_{max}), thus revealing the effect of heterogeneity and size. To simplify the visualization, we define

$$q = \log \left[\frac{2\langle k \rangle \langle k^3 \rangle}{\langle k^2 \rangle^2 \left(1 + \frac{\langle k^2 \rangle^2}{\langle k \rangle^4} \right)} \right]. \tag{4.13}$$

This way, when $q = 0$, the uncertainty of the critical threshold of a network is the same than that of the regular one, and for positive (negative) values of q, we are measuring an increase (decrease) of δK with respect to the homogeneous network. In Figure 4.4 we show the theoretical results obtained for the q value of networks in the space (γ, k_{max}). We note that the three horizontal lines correspond to the cases where the network has an integer exponent of 2, 3 or 4. In these cases, the first, second or third moments diverge. It

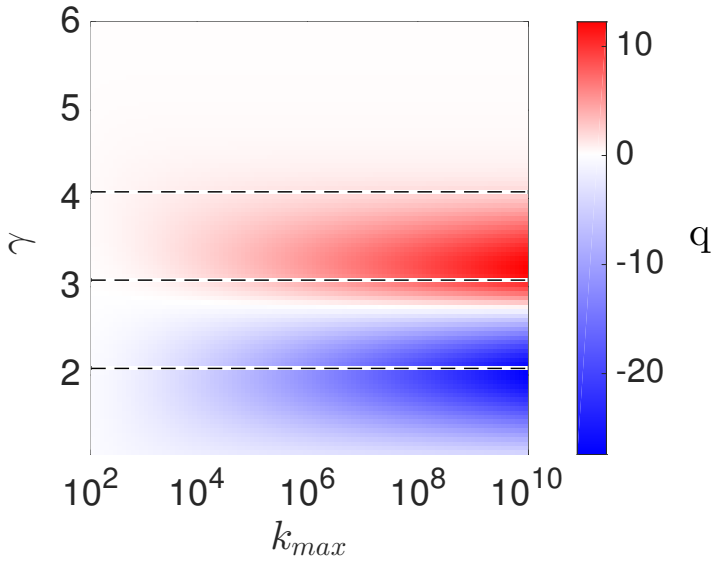


Figure 4.4: Colormap showing the theoretical dependence of q on the exponent γ and the maximum degree of the network $k_{\max}$. The value of $k_{\min}$ is fixed to $k_{\min} = 5$ and the resolution of the map is 100x100.

is also important to remark that below $\gamma = 2$, it is not feasible to generate networks with a pure power-law distribution [108]. Besides these considerations, we observe an interesting result. As expected, for large values of the exponent γ, the networks show similar uncertainty to that of a regular network. However, for $\gamma < 4$, uncertainty significantly increases, reaching a maximum as the exponent approaches $\gamma = 3$, before decreasing again. When approaching the value of $\gamma = 3$, the network maximizes the third moment of the degree distribution, while minimizing its second moment, and therefore emerges as the optimal uncorrelated structure amplifying the uncertainty in the threshold. Conversely, uncertainty is minimal for maximally heterogeneous networks, corresponding to an exponent $\gamma \approx 2$. Interestingly, the non monotonous dependence on γ is amplified as we increase the size of the system (in terms of its maximum degree).

To validate the previous theoretical prediction, we generate synthetic power-law networks using the modified preferential attachment algorithm with an attractiveness parameter that control the exponent[128]. Fixing the value of the minimum degree $k_{\min}$, and tuning the exponent and the size of the network, we detect a maximum in the uncertainty δK_c for the exponent $\gamma = 3$, as shown in Figure 4.5 thus confirming the prediction of the theory. We observe good qualitative agreement for the non monotonous dependency on the heterogeneity, and also that system size reinforces this dependency.

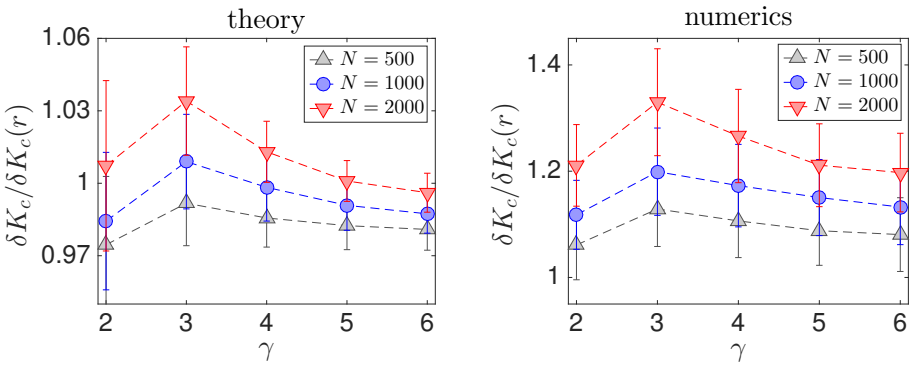


Figure 4.5: Relative value of the theoretical (left) and numerical (right) uncertainty δK_c for scale-free networks in the range $\gamma \in [2, 6]$ for sizes $N = 500$, 1000 and 2000, $\mu = 1$, $\sigma = 0.05$ and minimum degree fixed at $k_{\min} = 5$ compared to regular networks with the same average degree, and the same characteristics of the noise. The results are obtained with 200 realizations of the noise for each network and then averaging with 200 networks for each configuration of the modified preferential attachment algorithm. The high variance at each point shows that the results are very sensitive to the particular structure of the network, although the general trend is captured.

The results point towards the difficulty of accurately determine the critical threshold of scale-free networks, with exponent $\gamma \approx 3$, because δK_c is maximized in the presence of noisy weights for these networks.

4.4 Discussion

The results found in section III are of theoretical and practical relevance for the field of network science and they should be investigated further in detail. We have shown that particular network structures, as power-law degree distribution networks with exponent $\gamma \approx 3$ maximize the uncertainty of the critical threshold in the presence of noisy weights. This fact should be taken into account in the prediction of the critical threshold in empirical networks (which are usually heterogeneous) because, as proven, the accuracy in the estimation crucially depends on the underlying structure of the network. Moreover, the results might have a strong impact in the context of network optimization and adaptation [129]–[131], specially considering the ubiquity and theoretical relevance [71], [108] of power-law networks with exponent $\gamma \approx 3$ and the well-stablished hypothesis that many biological networks are operating near the critical point [132], [133]. In particular, one could wonder to which extent the existence of power-law networks with an exponent close to 3, maximizing the range of critical values has been evolutionary favourable. In

this sense, the current results make a natural connection with the previous work in [134], where it was shown that scale-free networks with exponent $\gamma = 3$ are able to achieve a larger variety of macrostates with respect to homogeneous networks (specifically near the critical threshold) by deterministically tuning the weights of the links.

From the methodological side, the formalism introduced in section II represents a first step in the use of error propagation methods to the analysis of complex networks with dynamical processes on top of them. The formalism is flexible and it can be applied to other network properties and in other scenarios, being of special importance the case of colored noise obtained directly from empirical measurements. We conjecture that this line of research will receive more attention in the future due to the increasing amount of data (not free of errors), that is being collected for a large variety of systems. We remark also that the current method is based on a MFA of the largest eigenvalue of the connectivity matrix, and this approximation neglects strong correlations of the eigenvalues in the presence of noise [135], [136]. While definitely more results are needed, the present formalism provides analytical insight to the studied phenomena, and turns out to give very accurate quantitative predictions if a few assumptions on the network hold.

To summarize, in this work we have studied how noise in the weights of a complex network affects the critical threshold of a dynamical process. We have restricted our study to the wide family of processes where the threshold depends on the largest eigenvalue of the connectivity matrix. In this scenario, and using the well-known MFA, we have applied error propagation to derive analytical expressions for the mean and standard deviation of the threshold depending on the noise parameters and the moments of the degree distribution. We validated our results against numerical simulations, showing good agreement when the initial MFA holds. Moreover, the formalism allowed us to carefully examine the effect that the network structure plays in the amplification of the noise at the critical point. Surprisingly, we found a non-monotonous behavior of the critical uncertainty with respect to the heterogeneity of the underlying network. By considering the paradigmatic case of uncorrelated power-law networks, we found that networks with exponent $\gamma \approx 3$ ($\gamma \approx 2$) emerge as the structures that maximize (minimize) the uncertainty of the threshold, due to an interplay between the second and third moment of the degree distribution.

4.5 Analytical derivations

4.5.1 Calculation of the mean

We can write the degrees and strengths in terms of the binary connections ($a_{ij} = 0$ or 1) and weights ($w_{ij} \in R$) of the connectivity matrix A, i.e $k_i = \sum_{j=1}^{N} a_{ij}$ and $s_i = \sum_{j=1}^{N} a_{ij} w_{ij}$. For the average strength $\langle s \rangle$, we have:

$$\langle s \rangle = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} w_{ij}. \tag{4.14}$$

Note that we can write Eq.4.14 equivalently as $\langle s \rangle = (1/N) \sum_i \mu_i k_i$, where μ_i is the average weight of node v_i. For sufficiently large degree ($k_i \gg 1$), one can approximate $\mu_i = \mu$, and therefore $\langle s \rangle = \mu \langle k \rangle$. However, in general, it is important to keep the contribution of each node because each μ_i has a specific uncertainty depending on the degree of node v_i, and this affects the overall uncertainty on K_c. For the second moment $\langle s^2 \rangle$, we have

$$\begin{aligned} \langle s^2 \rangle &= \frac{1}{N} \sum_{i=1}^{N} (\sum_{j=1}^{N} a_{ij} w_{ij})^2 \\ &= \frac{1}{N} \sum_{i=1}^{N} (\sum_{j=1}^{N} a_{ij} w_{ij}^2 + \sum_{j \neq k}^{N} a_{ij} a_{ik} w_{ij} w_{ik}). \end{aligned} \tag{4.15}$$

Noticing that $\sum_{j \neq k} a_{ij} a_{ik} = k_i^2 - k_i$, we obtain

$$\langle s^2 \rangle = \frac{1}{N} [\sum_{i=1}^{N} \mu_i^2 (k_i^2 - k_i) + \sum_{i=1}^{N} \langle w^2 \rangle_i k_i], \tag{4.16}$$

where $\langle w^2 \rangle_i$ is the average second moment of the i-node. Plugging Eq.4.14 and Eq.4.16 into Eq.4.2 in the main text, we obtain

$$K_c = \frac{\sum_{i=1}^{N} \mu_i k_i}{\sum_{i=1}^{N} \mu_i^2 (k_i^2 - k_i) + \sum_{i=1}^{N} \langle w^2 \rangle_i k_i}, \tag{4.17}$$

which correspond to Eq.4.3 in the main text.

4.5.2 Calculation of the variance

The propagation of uncertainty of a non-linear function of the random variables as Eq.4.3 requires to use a truncated Taylor expansion [126]. Up to second-order, and in the notation used in the main text, the approximate variance of the function is given by

$$(\delta K_c)^2 \approx J_0^T \mathbf{V} J_0 + \frac{1}{2} \operatorname{Tr}[(\mathbf{H_0 V})^2] \tag{4.18}$$

where the Jacobian vector and the Hessian matrix are evaluated at the mean values of the random variables $\vec{\mu}$ and $\langle \vec{w^2} \rangle$. The Jacobian of the system in Eq.4.3 is

$$J = (\frac{\partial K_c}{\partial \mu_1}, ..., \frac{\partial K_c}{\partial \mu_N}, \frac{\partial K_c}{\partial \langle w^2 \rangle_1}, ..., \frac{\partial K_c}{\partial \langle w^2 \rangle_N}). \tag{4.19}$$

First, we compute the partial derivatives in Eq.4.19 explicitly from Eq.4.3, obtaining

$$\begin{aligned} \frac{\partial K_c}{\partial \mu_i} &\approx \frac{1}{N} \frac{k_i(\mu^2 \langle k^2 \rangle + \sigma^2 \langle k \rangle) - 2\mu^2(k_i^2 - k_i)\langle k \rangle}{(\mu^2 \langle k^2 \rangle + \sigma^2 \langle k \rangle)^2}, \\ \frac{\partial K_c}{\partial \langle w^2 \rangle_i} &\approx -\frac{1}{N} \frac{k_i \mu \langle k \rangle}{(\mu^2 \langle k^2 \rangle + \sigma^2 \langle k \rangle)^2}, \end{aligned} \tag{4.20}$$

where the sign $\approx$ stands for assuming, in good approximation, that the input parameters μ and σ^2 are the actual mean values of the random variables $\vec{\mu}$ and $\vec{\sigma^2} = \langle \vec{w^2} \rangle - \vec{\mu}^2$.

The Hessian matrix, the square matrix of the second-order partial derivatives of the function in Eq.4.3 can be directly obtained by taking derivates from Eq.4.20. After some algebra, and defining $Q = \mu^2 \langle k^2 \rangle + \sigma^2 \langle k \rangle$, we obtain

$$\begin{aligned} \frac{\partial^2 K_c}{\partial \mu_i \partial \mu_j} \approx & \frac{1}{N^2 Q^3} [Q(2\mu(k_j^2 - k_j)k_i - (2 + 2\delta_{ij}\mu(k_i^2 - k_i)k_j)) \\ & - (k_i - 8\mu^3 \langle k \rangle (k_i^2 - k_i)(k_j^2 - k_j)). \end{aligned} \tag{4.21}$$

The Hessian matrix of our system is symmetric, such that $\partial^2 K_c / \partial \mu_i \partial \langle w^2 \rangle_j = \partial^2 K_c / \partial \langle w^2 \rangle_i \partial \mu_j$. We obtain

$$\frac{\partial^2 K_c}{\partial \mu_i \partial \langle w^2 \rangle_j} \approx \frac{1}{N^2 Q^3} [-Q k_i k_j + 4\mu^2 \langle k \rangle k_j (k_i^2 - k_i)], \tag{4.22}$$

and for the last term we have

$$\frac{\partial K_c}{\partial \langle w^2 \rangle_i \partial \langle w^2 \rangle_j} \approx \frac{2\mu k_i k_j \langle k \rangle}{N^2 Q^3}. \tag{4.23}$$

For the covariance matrix, we can obtain explicit expression for the entries $(\mathbf{V})_{ij}$ when the noise in the weights is assumed gaussian and uncorrelated. By assumption, the network is symmetric and so it will be the covariance matrix, which can be written in block form as

$$\mathbf{V} = \left(\begin{array}{c|c} \mathbf{v_\mu}^2 & \mathbf{v_{\mu,\langle w^2 \rangle}} \\ \hline \mathbf{v_{\mu,\langle w^2 \rangle}} & \mathbf{v_{\langle w^2 \rangle}}^2 \end{array} \right),$$

where $\mathbf{v_\mu}^2$, $\mathbf{v_{\mu,\langle w^2 \rangle}}$ and $\mathbf{v_{\langle w^2 \rangle}}^2$ are symmetric matrices in $R^{N \times N}$ that capture each

covariance term between the two random variables $(\mu_i, \langle w^2 \rangle_i)$ of all nodes. Explicitly

$$(\mathbf{v}_\mu{}^2)_{ij} = \frac{\sigma^2}{k_i}(\delta_{ij} + \frac{a_{ij}}{k_j}), \tag{4.24}$$

$$(\mathbf{v}_{\mu,\langle \mathbf{w^2} \rangle})_{ij} = \frac{2\mu\sigma^2}{k_i}(\delta_{ij} + \frac{a_{ij}}{k_j}), \tag{4.25}$$

$$(\mathbf{v}_{\langle \mathbf{w^2} \rangle}{}^2)_{ij} = \frac{2\sigma^2(2\mu^2 + \sigma^2)}{k_i}(\delta_{ij} + \frac{a_{ij}}{k_j}). \tag{4.26}$$

The first term in the sums is the contribution of the diagonal entries. The gaussian variances (σ^2 and $2\sigma^2(2\mu^2 + \sigma^2)$) and covariance ($2\mu\sigma^2$) of a single weight w_{ij} drawn from (μ, σ^2) are divided by the number of elements (the degree k_i) involved in computing the averages μ_i and $\langle w^2 \rangle_i$. The second term accounts for the non-diagonal entries. If two nodes (i, j) are neighbours, i.e. $a_{ij} = 1$, then we have to add an additional correlation due to the presence of the shared weight, which is divided by the product of their degrees (k_i and k_j).

For the first order expansion, we can compute explicitly $(\delta K_c)^2$ in terms of the noise parameters (μ, σ) and the moments of the degree distribution. We can write Eq.4.5 as

$$(\delta K_c)^2 \approx \sum_{i=1}^{N}\sum_{j=1}^{N}[(\frac{\partial K_c}{\partial \mu_i})(\frac{\partial K_c}{\partial \mu_j})(\sigma_\mu{}^2)_{ij}, \tag{4.27}$$

$$+ (\frac{\partial K_c}{\partial \langle w^2 \rangle_i})(\frac{\partial K_c}{\partial \langle w^2 \rangle_j})(\sigma_{\langle \mathbf{w^2} \rangle}{}^2)_{ij}, \tag{4.28}$$

$$+ 2(\frac{\partial K_c}{\partial \mu_i})(\frac{\partial K_c}{\partial \langle w^2 \rangle_j})(\sigma_{\mu,\langle \mathbf{w^2} \rangle})_{ij}], \tag{4.29}$$

and after some algebra, we obtain

$$\begin{aligned}(\delta K_c)^2 \approx \frac{2\sigma^2\langle k \rangle}{NQ^4}[&Q^2 - 4\mu^2\langle k^2 \rangle Q + 2\mu^2(2\mu^2 + \sigma^2)\langle k \rangle^2 \\ &+2\mu^4(\langle k \rangle\langle k^3 \rangle + \langle k^2 \rangle(\langle k^2 \rangle - 4\langle k \rangle) + 2\langle k \rangle^2) \\ &+8\mu^4\langle k \rangle(\langle k^2 \rangle - \langle k \rangle)],\end{aligned} \tag{4.30}$$

where we have used that $\sum_i \sum_j a_{ij}k_ik_j = N\langle k^2 \rangle^2/\langle k \rangle$. Simplifying further, we get the resulting Eq.4.6 in the main text. Explicitly,

$$\begin{aligned}(\delta K_c)^2 \quad &\approx \quad a[\mu^4(2\langle k \rangle\langle k^3 \rangle \\ &- \quad \langle k^2 \rangle^2) - 2\mu^2\sigma^2(\langle k \rangle\langle k^2 \rangle - \langle k \rangle^2) + \sigma^4\langle k \rangle^2]\end{aligned} \tag{4.31}$$

with $a = 2\sigma^2\langle k \rangle/[N(\mu^2\langle k^2 \rangle + \sigma^2\langle k \rangle)^4]$.

Part II

Applications

Chapter 5

Vulnerabilities in rail networks

Many critical infrastructure systems have network structures and are under stress. Despite their national importance, the complexity of large-scale transport networks means that we do not fully understand their vulnerabilities to cascade failures. The research conducted through this chapter examines the interdependent rail networks in Greater London and surrounding commuter area. We focus on the morning commuter hours, where the system is under the most demand stress. There is increasing evidence that the topological shape of the network plays an important role in dynamic cascades. Here, we examine whether the different topological measures of resilience (stability) or robustness (failure) are more appropriate for understanding poor railway performance. The results show that resilience, not robustness, has a strong correlation with the consumer experience statistics. Our results are a way of describing the complexity of cascade dynamics on networks without the involvement of detailed agent-based models, showing that cascade effects are more responsible for poor performance than failures. The network science analysis hints at pathways towards making the network structure more resilient by reducing feedback loops.

5.1 Introduction to rail transport networks

Cascade delays and cancellations on rail transport can cause devastating economic damage and dent consumer satisfaction. Existing knowledge either focuses on improving operational practices or considers a pure topological analysis. However, by considering both real passenger travel flows and the network topology together, in this chapter we obtain a stronger understanding of its dynamic vulnerability and resilience. In earlier years, research largely focused on improving specific functionalities in rail systems; and more recent research has focused on the relationship between the general network topology and whether this has macroscopic bearing on the overall system performance [137].

The efficiency of transport networks has been related with their resilience [138] and the different types of topologies have been analysed, comparing the network geometry and the level of connectivity. However, these studies predominantly focus on the pure topological characteristics of a graph [139], [140].

5.1.1 Identifying vulnerabilities in rail networks

The concept of vulnerability in transportation network, introduced in the literature by Berdica [141], is generally defined as the susceptibility to disruptions that could cause considerable reductions in network service or the ability to use a particular network link or route at a given time. Many have applied general network science disruption analysis. For example, several studies [142]–[144] have been conducted for modelling railway vulnerability with promising predictive results. Bababeik *et al.* [145] recently proposed a mathematical programming model that is able to identify critical links with consideration of supply and demand interactions under different disruption scenarios. Recent work has also used graph properties to infer interaction strengths and use an epidemic spreading model to predict delays in railway networks [146].

In the current literature, most of the proposed studies consider natural or man-made disasters, but they do not consider the stress of the network during the peak-hours and how the structure of the network created by the massive flows of people can influence their ability to maintain a good service. For example, several graph-based approaches have been proposed to improve the performances by revising the design and maintenance of the rail networks [147], but do not consider dynamic passenger flows. Other studies focus on specific extreme scenarios [148] or unfavourable conditions [149] that cause disruptions.

The UK rail network transports more than 1.7 billion passengers per year, of which 1.1 billion passengers commute in and around London.[1] According to the Office of Rail and Road,[2] last year in and around London only 86.9% of passenger trains arrived on time and 4.8% of the journeys were cancelled or significantly late. Often these delays are interrelated and the relationship between cascade effects and network dynamics is not well understood.

The data used for this chapter (see Section 5.4) indicates that under the same external conditions, the major rail companies in and around London show dramatically different performance levels. In this work, we hypothesize that this difference can, in part, be attributed to the peak passenger demand. The interplay between flow and network structure can tease out which structural measures correlate strongly with overall

[1] Passenger rail usage. Office of Rail and Road. See `http://dataportal.orr.gov.uk/browsereports/12`. (10 September 2018)

[2] Passenger and freight rail performance. Office of Rail and Road. See `http://dataportal.orr.gov.uk/browsereports/3`. (10 September 2018).

performance.

We take a systems-of-systems approach by applying a complex network analysis to transport networks. Unlike prior studies that focus only on the topological aspects of the network, we consider several important additional aspects which attempt to match our analysis to reality. First, we consider passenger volumes during morning commuter or rush-hour, which weights the network and adds directionality. The morning rush-hour is important because most of the delays and the highest economic impact of delays occur during this time. Second, we consider a multiplex of different urban overground, regional and national rail services (both together and separately). As a result, we have a weighted and directed multiplex network, which requires more sophisticated network analysis methods to uncover its resilience and robustness to cascade failures. Finally, we map our network resilience and robustness results to actual railway performance figures of delay and cancellation statistics and consumer satisfaction.

5.2 Theoretical framework

Vulnerability is a major problem in the study of complex networks and it can be regarded as the susceptibility of a networked system to suffer important changes in its structure and dynamic functions under any form of disruption. When such disruptions affect the internal state of the nodes (e.g. stations) or links (e.g. train lines) of the network, it becomes important to predict the extent of such perturbations under the perspective of dynamical systems (e.g. linear stability analysis); throughout this chapter, we refer to this problem as the study of *resilience*. Resilience is important for understanding cascade effects that suppress the performance of the network, such as cascade delays due to signal failures or poor scheduling. Resilience is related to the type of problem where a train going from A to B that is running late, which affects the ensuing service B back to A using the same train. But, when the perturbations involve some sort of attack or out-right failure (e.g. a disruption in a station due to someone walking on to the tracks or a signal failure), the challenge tends to be in studying the resulting connectivity loss and secondary loss of functionality in neighbouring stations. We refer to this as the *robustness* problem, which is describes different situations from the aforementioned resilience. In plain terms, robustness considers when a train from A to B will be halted if the track in between is blocked or station B is closed.

5.2.1 Measuring resilience

The concept of resilience on networks admits various interpretations and definitions [150], [151]. A generally accepted definition of stability is applicable when the system performance returns to a desirable state. For homogeneous linear stability, one might equate

resilience with equilibrium points and look at the leading eigenvalue of the Jacobian matrix [152]. When linear stability is not suitable due to complex dynamics, many authors [153]–[156] have studied system resilience from different perspectives. Some consider the dynamic response (e.g. time to recovery) of the whole system after a specific disruption, while others use random perturbations to numerically quantify system response [157].

However, such approaches depend strongly on assumptions about the system, such as details of the dynamics or the number of neighbours required for a node to function. In this work, we use instead recent advances in the ecological system analysis to study resilience, namely the framework of *trophic coherence* [158]. While there are obviously differences between ecosystems and rail systems, both are essentially transport networks in which either biomass or passengers flow from sources (plants or home towns) through various intermediary nodes, and end in sinks (top predators or work places).

Trophic coherence is a property of directed graphs that defines how much a graph is hierarchically structured. The rationale is that hierarchical systems have fewer feedback loops and are less likely to suffer from cascade effects. When networks are modelled as a discrete linear time invariant (LTI) system with a defined input and output [159], the dynamic response stability is defined by the location of roots of its transfer function (negative domain). In such a case, the absence of feedback loops ensures stability. The presence of feedback loops will cause non-zero roots and risk instability. When we consider a complex network with N^2 input-output combinations, the transfer function cannot be defined. As such, we measure the overall network incoherence, which is a compressed figure of merit for how many feedback loops exist [158], [160]. Johnson *et al.* [158] proved that 'a maximally coherent network with constant interaction strengths will always be linearly stable', and that it is a better statistical predictor of linear stability than size or complexity. We measure network coherence through the incoherence parameter (see Figure 5.1(c)), a measure of how tightly the trophic distance associated with edges is concentrated around its mean value (see Section 5.3.1).

In order to define trophic coherence in a directed network, the first step is to define its basal nodes (i.e. nodes that predominantly supply energy—high out-degree and low in-degree). That is to say, stations with a high trophic level receive passengers while stations with a low trophic level provide passengers. Thus, basal nodes are likely to be home train stations of commuters. Unlike networks studied in previous works (e.g. food webs [160]–[162]), the London urban rail network in peak-hours does not have predefined basal nodes (i.e. nodes with in-degree 0). In transportation, this means that there is always a non-zero passenger counter-flow travelling from urban to the countryside stations during the morning rush hour. To distil the basal nodes from the data, we developed and tested two different approaches (see Section 5.3.2) to approximately define basal nodes in networks where they do not naturally exist. In the first proposed approach, we apply basal node enforcement, whereby basal nodes are selected from those with lowest ratio

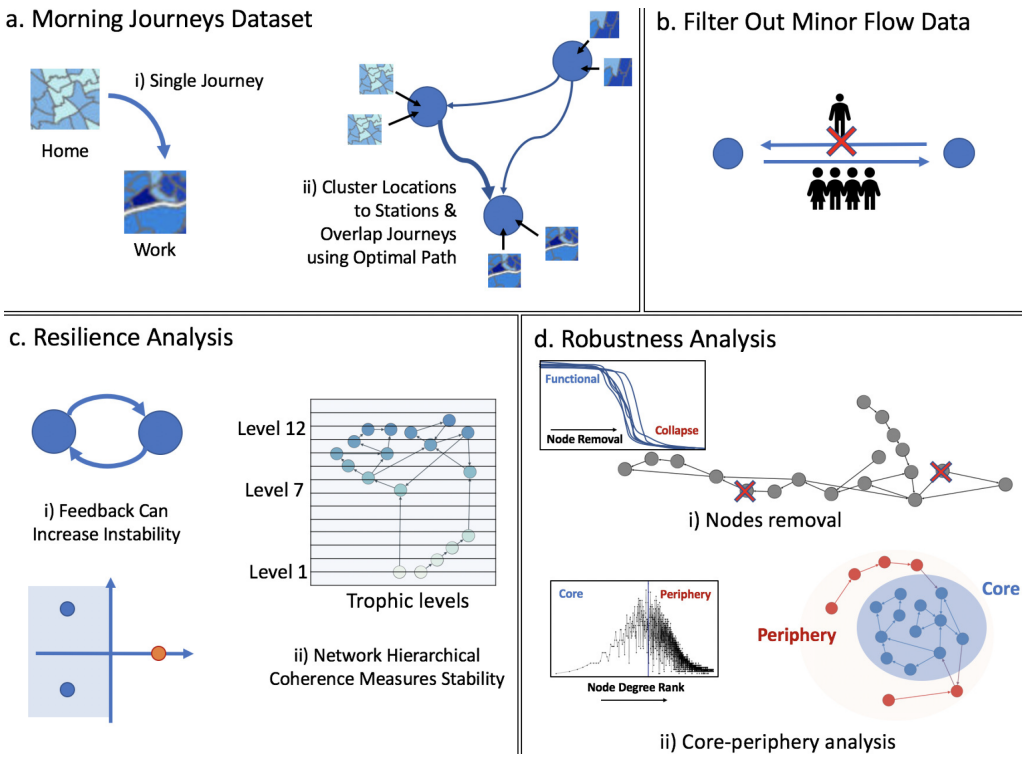


Figure 5.1: We reconstruct the major rail networks under stress conditions considering the morning journeys (a) and we measure the topological characteristics of these networks, removing the uninteresting flows (b). Then, the resilience (c) and robustness (d) of these networks are analysed using the framework described in Section 5.2.

between incoming and outgoing edges. The trophic level of the remaining nodes is then computed using the standard formula (Eq. 5.1). In the second proposed approach we apply passenger flow filtering, a method by which redundant edges are removed until basal nodes naturally emerge (see Figure 5.1(b)).

5.2.2 Measuring robustness

The objective in this case is to use both proxy and direct measures of robustness. Direct measures include random or targeted node removal. However, as robustness is not uniquely defined, proxy measures may yield more holistic insights. As such, here we use a variety of robustness measures to establish a wider evidence base. Regarding the first approach, we directly evaluate network robustness by performing sequential node removal [15]: nodes of the rail networks are randomly removed whilst evaluating network connectivity, computing the size of the largest strongly connected component [163], [164].

On the other approach, as a first proxy we evaluate the core and periphery meso-scale structure of the rail network (see Figure 5.1(d)). The core-periphery ratio (see Section 5.3.3) gives a scalable and compressed understanding of robustness, and the argument is formalized by Borgatti *et al.* [165]. As a second proxy measure we use the rich-club coefficient (see Eq. 5.4) [161], [162], [166], [167].

5.3 Methods

5.3.1 Computing trophic coherence

The trophic level of a node i, denominated by s_i, is recursively defined as the average trophic level of its in-neighbours, plus 1:

$$s_i = 1 + \frac{1}{k_i^{in}} \sum_j a_{ij} s_j \quad , \tag{5.1}$$

where a_{ij} is the adjacency matrix of the graph and $k_i^{in} = \sum_j a_{ij}$ is the number of in-neighbours (in degree) of node i. Basal nodes, i.e. those with $k_i^{in} = 0$ have trophic level $s_i = 1$ by convention. Note that Eq. 5.1 can have non-integer solutions. By solving the system of equations in 5.1, it is always possible to assign a unique trophic level to each node as long as there is at least one basal node, and every node is on a directed path which includes a basal node [158]. In our study, the trophic level of a station is the average level of all the stations from which it receives passengers plus 1. For this reason, stations near residential areas in the suburbs will have lower trophic level than those close to business areas and those in the centre.

Each edge has an associated trophic difference: $x_{ij} = s_i - s_j$. The probability distribution function of trophic differences, $p(x)$, always has mean 1. The smaller the variance of this distribution is, the more a network is considered to be trophically coherent. We can measure *trophic coherence* with the incoherence parameter q, which is simply defined as the standard deviation of $p(x)$ [158]:

$$q = \sqrt{\frac{1}{L} \sum_{ij} a_{ij} x_{ij}^2 - 1} \quad , \tag{5.2}$$

where $L = \sum_{ij} a_{ij}$ is the total number of connections (edges) between the stations (nodes) in the network. A perfectly coherent network will have $q = 0$, while $q > 0$ indicates less coherent networks.

The degree to which empirical networks are trophically coherent can be investigated by comparison with a null model. The basal ensemble expectation $\tilde{q}$ can be considered a good approximation to a null model for finite random networks [160]:

$$\tilde{q} = \sqrt{\frac{L}{L_b} - 1} \quad , \tag{5.3}$$

where L_b is the number of edges connected to basal nodes. The ratio $q/\tilde{q}$ is used to analyse the coherence of the network: a value close to 1 shows a network with a trophic coherence similar to a random expectation. Values lower than 1 reveal significant coherence, while values greater than 1 reveal significant incoherence. For example, Johnson & Jones [160] found that food webs are significantly coherent ($q/\tilde{q} = 0.44 \pm 0.17$), metabolic networks are significantly incoherent ($q/\tilde{q} = 1.81 \pm 0.11$) and gene regulatory networks are close to their random expectation ($q/\tilde{q} = 0.99 \pm 0.05$).

5.3.2 Finding basal nodes

In our study of the morning peak-hour rail networks, there are not natural basal nodes. In order to be able to solve the equations and compute trophic levels, we define two methodologies to identify them: the basal nodes enforcement and the flows filtering.

Basal nodes enforcement

The first technique used to select the basal nodes revolves around the enforcement of the desired number of basal nodes, selecting them according to some properties of the nodes. This technique enforces a predefined number EN of nodes to be basal nodes (their trophic level is imposed to be 1). The nodes to be enforced are selected according to their similarity to real basal nodes, namely the nodes with the lowest ratio between incoming and outgoing edges. More formally, the $k_{\text{out}}/k_{\text{in}}$ ratio is computed for all the nodes, then the trophic level of the EN nodes with the lower ratio is enforced to 1 ($s_i = 1$). If parts of the network are not connected to basal nodes, only the largest strongly connected component will be considered. This technique maintains the structure of the network intact (it does not add/remove nodes or edges) but, instead, it does not take into account its natural topology when selecting the basal nodes, making the selection artificial: the selection of the number of basal nodes is artificially defined by the user.

Flows filtering

In the analysis of the morning peak-hour commute, the factors that determine the stability of the network depend on the major flows of people (from home to work commute). The paths with just a small portion of commuters can thus be ignored. To remove these paths, a threshold T for the detection of major flows is defined: when two nodes i and j are connected with two reciprocal edges ($a_{ij} = 1$ and $a_{ji} = 1$), the edges e_{ij} whose weight ratio is below the threshold T, i.e. $\omega_{ij}/\omega_{ji} < T$, are deleted. With this technique, basal nodes are not enforced but rather naturally emerge from the change in the structure of the network (i.e. the edges with a low impact on the study are removed from the network).

For example, if there are 100 people going from node i to j and only 1 going from j to i, the edge e_{ji} can be removed without degrading the quality of the peak-hour flows study. If $T \geq 1$, for each pair of nodes the edge with smaller weight is always removed; the edge with the highest weight is preserved only if it is sufficiently greater than its reciprocal. Note that larger values of T will require higher directionality unbalance in order to keep the edge in the dominant direction. If $T < 1$, the edge with the highest weight is always preserved, whereas the lower-weight edge could potentially be preserved if flow directionality is sufficiently balanced. Note that for $T < 1$, the lower the value of T, the easier it is to preserve edges with unbalanced flows.

5.3.3 Core-periphery and robustness

The study of the core–periphery structure of the network is used to identify the densely connected stations where people can choose more than one path to reach the destination in contrast to sparsely connected stations which can cause a major interruption of the service in case of disruptions.

Finding the core of a network

The core of a network is computed ranking all the nodes in a network according to a predefined centrality measure (in our case total degree and trophic coherence) and then counting the number of connections they have with higher ranked nodes. The node with the highest number of high-level connections is the core-border. All the nodes with a higher ranking than the core-border node along with the border node itself compose the core of the network, whilst the other nodes are its periphery. A big core suggests several different ways to reach the majority of the nodes and accordingly a more robust network.

Rich-club coefficient

To study the robustness of the networks, we analysed the rich-club phenomenon [168]. This structural characteristic appears when nodes of higher degree are more interconnected than nodes with lower degree. The presence of this phenomenon may be indicative of several interesting high-level network properties, such as its robustness. More precisely, this behaviour appears when nodes with degree larger than k are more densely connected among themselves than the nodes with degree smaller than k [169]. This is quantified by computing the rich-club coefficient across a range of k values, and if this value is greater than 1 for some k the network is considered to exhibit rich-club phenomenon.

The rich-club coefficient is usually defined using the degree of nodes, but it can be generalized to other richness metrics (in our case, the trophic level). Note that, in order to compute it, we need to convert the morning peak-hours directed graph to an

undirected one so that it is consistent with the standard rich-club definition. The formula to compute the rich-club coefficient for a generalised richness measure is as follows:

$$\phi(r) = \frac{2E_{>r}}{N_{>r}(N_{>r} - 1)} \quad , \tag{5.4}$$

where $E_{>r}$ refers to the number of edges present between nodes with richness measure above r and $N_{>r}$ refers to the number of nodes with richness measure above r.

5.4 Data

5.4.1 UK rail network

In this study, we analyse a real-world rail network under demand stress conditions (morning rush-hour). The commuter paths are computed considering the information relative to places where people live and work provided by the UK National Census Transformation Programme.[3] The optimal travel paths were provided by the National Rail (including rail services through underground tunnels, but not including the underground/subway system) through their TransportApi service.[4] Given an origin station and a destination station, the TransportApi service provides all the information about the travel, including the intermediate stop stations. We first check if rail travel is required for a person to go from home to work, and if so, we compute their optimal journey and use these data to weight the network (see Figure 5.1(a)). In the current study, only the travels that start and end in a bounding area of 80 km from central London have been taken into account (this approximately covers Cambridge to the north, Oxford to the northwest, Reading to the west and Brighton to the south). It roughly represents all 1 h commuter paths, which is the national standard according to UK's Office for National Statistics.

The resulting dataset represents the flows of people in morning peak-hours on the rail network (available on Dryad [170]), when they travel from their homes to their places of work. Each journey is defined as a set of two or more stations (in case of intermediate stops of the train all the intermediate stations are included). The dataset is transformed in a directed weighted graph (see Section 1.4.1) where the nodes are the train stations, the edges are the weighted flows of passengers and a journey is an ordered set of nodes that includes the departure station, the arrival station and any intermediate station (if the train stops, as we consider the service class of the train).

When, in our graph, one or more passengers are going from node i to node j (or these two nodes are intermediate stations of the travel), an edge e_{ij} is added to the graph. The weight ω_{ij} of this edge is the sum of all the passengers of the journeys that include

[3] *Census transformation programme*, Office for National Statistics. See `https://www.ons.gov.uk/census/censustransformationprogramme`. (May 2018).

[4] *transportApi*, National Rail. See `https://www.transportapi.com`. (May 2018).

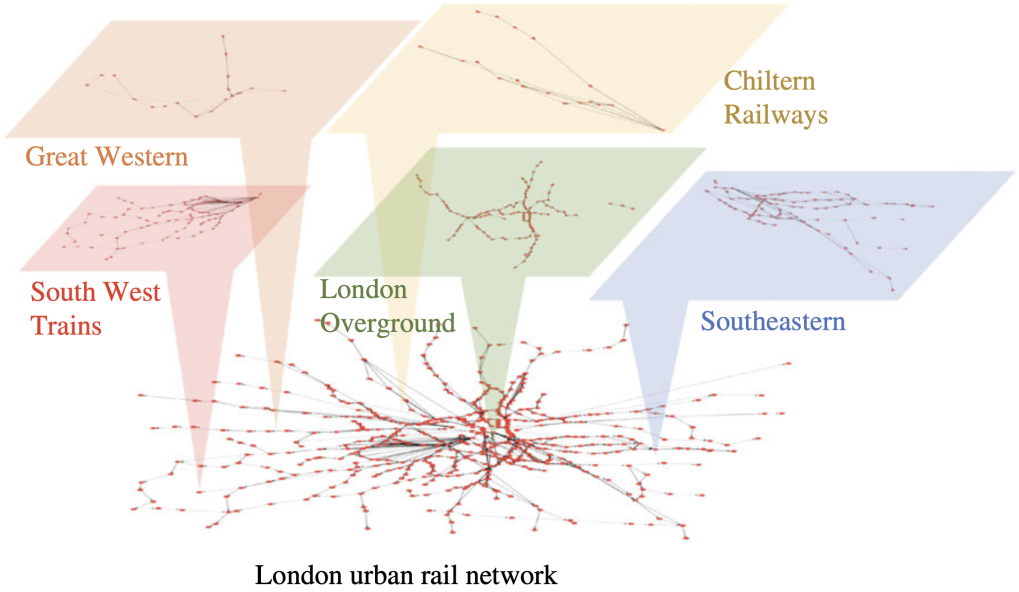


Figure 5.2: Directed graph representing passenger flows during morning peak-hours in the urban rail network of London and its surroundings, built as detailed in Section 5.4.

travels from node i to node j. The directed graph of the passenger flows during morning peak-hours is shown in Figure 5.2. We show the whole multiplexed network, as well as some examples of the individual sub-networks comprising urban overground (London Overground), regional links (Thameslink) and national services (e.g. Southern rail).

5.4.2 Service performance measures

, http://orr.gov.uk/about-orr/who-we-are. (May 2018).

We use data from the Public Performance Measures provided by the ORR (Office of Rail and Road) [5], an independent regulator that monitors the rail industry's health and safety performance. ORR holds Network Rail [6], the company that with 20.000 miles of track owns, operates and develops Britain's railway infrastructure. In particular, two performance measures are used in our comparison:

- **PPM**. The *Public Performance Measure* combines figures for punctuality and reliability into a single performance measure. Usually, it shows the percentage of trains which arrive at their terminating station within 5 min (for London and South East

[5] *Office of Rail and Road - who we are*, Office of Rail and Road. See `http://orr.gov.uk/about-orr/who-we-are`. (May 2018).

[6] *Public performance measure*, National Rail. See `https://www.networkrail.co.uk/who-we-are/how-we-work/performance/public-performance-measure/`. (May 2018).

and regional services) or 10 min (for long distance services) [7]. Here, for the sake of clarity, we define oPPM as the opposite value of PPM (oPPM=100%−PPM). oPPM is the percentage of trains which do not arrive at their terminating station within 5 or 10 min (depending on the distance).

- **CaSL**. The *Cancellation and Significant Lateness* is a percentage measure of scheduled passenger trains which are either cancelled (including those cancelled en route) or arrive at their scheduled destination more than 30 min late [8].

We use performance measures from the year 2017 (key statistics by train operating company (TOC)—2016–2017 [9]). To provide statistically significant results (small networks are more sensitive to local functional effects than macroscopic topological structure), we considered the five companies with the highest number of nodes in the network, excluding companies with very simple network structures (e.g. Heathrow Express has only one line). The companies taken into account and the number of stations are shown in Table 5.1.

Operator name	**Number of stations**
London Overground	109
Great Western Railway	18
Chiltern Railways	18
South West Trains	91
Southeastern	64

Table 5.1: Number of stations (nodes) per company in the morning peak-hours network.

5.5 Results

Our hypothesis is that the delays in a rail network and, more generally, the performance of the services are influenced by the topological structure of the network. The intuition we seek to validate is that a more resilient and/or robust network should guarantee lower cascade delays and faster recovery in case of disruptions.

[7] *Public performance measure*, National Rail. See `https://www.networkrail.co.uk/who-we-are/how-we-work/performance/public-performance-measure/`. (May 2018).

[8] *Public performance measure*, National Rail. See `https://www.networkrail.co.uk/who-we-are/how-we-work/performance/public-performance-measure/`. (May 2018).

[9] *Statistical releases,*, Office of Rail and Road. See `http://orr.gov.uk/statistics/published-stats/statistical-releases`. (May 2018).

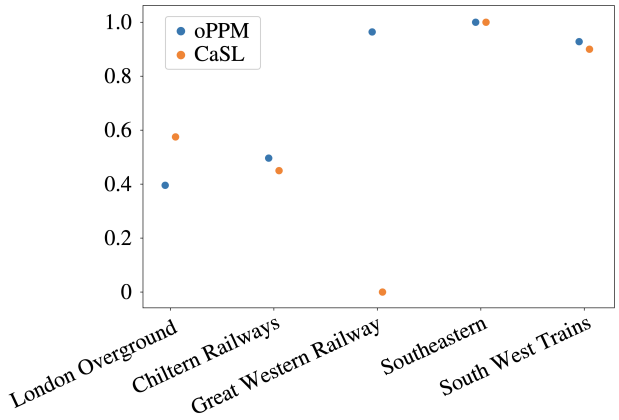


Figure 5.3: oPPM versus CaSL Person correlation coefficient for each different network operator.

5.5.1 Performance metrics correlation

Figure 5.3 shows how four out of five of the rail companies analysed show a strong correlation between the two performance measures oPPM and CaSL (see definitions in Section 5.4), while in one case (Great Western Railway) these values are not correlated, possibly meaning that this company often has little delays (low resilience) but generally does not have major disruptions (high robustness). The Pearson correlation coefficient (PCC) [171] is used to establish if there is a correlation between the topology parameters of the network and the performance measures. PCC has a value between $+1$ and -1, where $+1$ is total positive linear correlation, 0 is no linear correlation and -1 is total negative linear correlation. As a rule of thumb, variables with a correlation coefficient greater than 0.7 are considered highly correlated, while they are considered moderately correlated when the PCC coefficient is between 0.3 and 0.7.

5.5.2 Choosing a method to find basal nodes

Our analysis crucially relies on the filtering parameter values NE and T defined in Section 5.3.2 in order to reasonably reconstruct some underlying network structure. In this section, we analyse empirically the properties of both basal node selection methods presented before, namely the *basal nodes enforcement* and the *flows filtering* methods. We're looking for the minimum filtering value range (higher values may remove too much data) such that our measures of interest (e.g. trophic coherence for resilience or core size for robustness) remains invariant to further increases in filtering parameter values. To do so, we apply a methodology consisting on constructing the morning peak-hours rail network for each separated provider and for each filtering value. Then we compute the

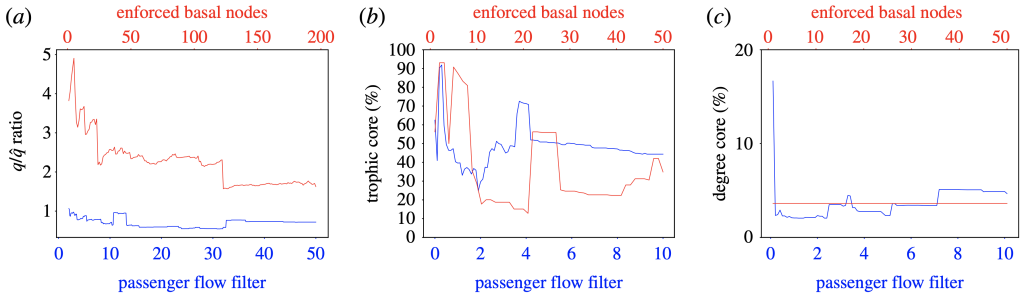


Figure 5.4: The figure shows several ranges of filtering parameter values for the two techniques proposed: the basal nodes enforcement parameter NE is shown in red, and the flows filtering parameter T is shown in blue. It analyses the behaviour of three major resilience and robustness measures used in this work: (a) incoherence of the network; (b) trophic core–periphery ratio; (c) Degree core–periphery ratio.

average measure of interest from all provider networks for each filtering value.

Regarding resilience, as shown in Figure 5.4(a) the node enforcement method produces incoherent networks throughout all the range of filter parameters — even with a large number of stations enforced (e.g. 100) the networks remain highly incoherent on average with $q/\tilde{q} > 2$. On the contrary, the passenger filtering technique can achieve stable values of low incoherence even eliminating few links at low filtering values. Regarding robustness, Figure 5.4(b) shows that the measure of trophic-core requires larger filtering values in order to stabilise for both method, although it reaches more consistent stability with the flow filtering technique. Figure 5.4(c) shows that the measure of degree-core remains consistently stable for both methods and any range of filtering parameters.

In the light of this results, we choose to work with the flow filtering method because, besides being more intuitive, it provides more stable results across a range of lower filtering parameters. In particular, throughout the subsequent analysis we choose to work with filtering parameters between $T = 1$ and $T = 4$. Across this range, the flow filtering method removes the small counter-flow (e.g. people that live in the centre and work in the suburbs) in order to evidence the mass commute that causes the major stress on the network.

5.5.3 Topology-performance correlation

Trophic incoherence analysis

We compute the degree to which each of the specific provider rail networks are incoherent by comparing them with the basal ensemble expectation as a null model, using the trophic incoherence measure $q/\tilde{q}$. This measure has a value close to 1 when a network

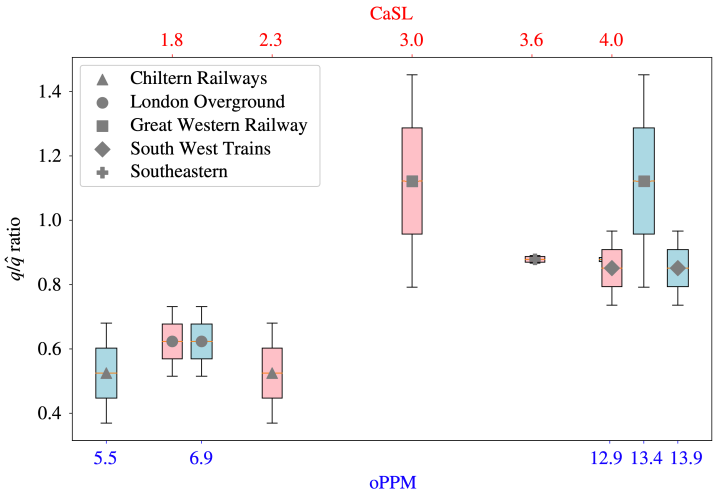


Figure 5.5: Box-plot distribution of trophic incoherence $q/\tilde{q}$ for a range of filtering threshold $T \in [1, 4]$ showing the mean (marker) and the standard deviation (box limit), for each service provider. The x-axis represents service performance through oPPM (lower blue x-labels, blue boxes) and CaSL (upper red x-labels, red boxes).

has a trophic coherence similar to a random expectation, it has a value lower than 1 when the network is coherent, and it has a value greater than 1 when the network is incoherent (the details of this computation are provided in Section 5.3.1).

As described in the previous section, the morning peak-hour network is computed using the passenger flow filter method with different flow filtering thresholds, between $T = 1$ and $T = 4$, with a granularity of $\Delta T = 0.5$. Figure 5.5 presents the distribution of incoherence across all considered filtering thresholds in the form of a box-plot showing the average and standard deviation of $q/\tilde{q}$, for each of the service providers ordered according to their corresponding performance metrics oPPM and CaSL.

We can see that more coherent networks (low $q/\tilde{q}$) are generally associated with lower delays (oPPM) and cancellations (CaSL). In particular, the results exhibit a highly positive correlation between the trophic incoherence of the network and the Public Performance Measure (PCC = 0.98), suggesting that there is a high correlation between the resilience of a rail network and the probability of its trains to arrive at their terminating station on time. There is also a high positive correlation between the trophic incoherence and the Cancellation and Significant Lateness measure (PCC = 0.92), evidencing also a correlation between low resilience and the percentage of trains either cancelled or that arrive to their destination with more than 30 min late.

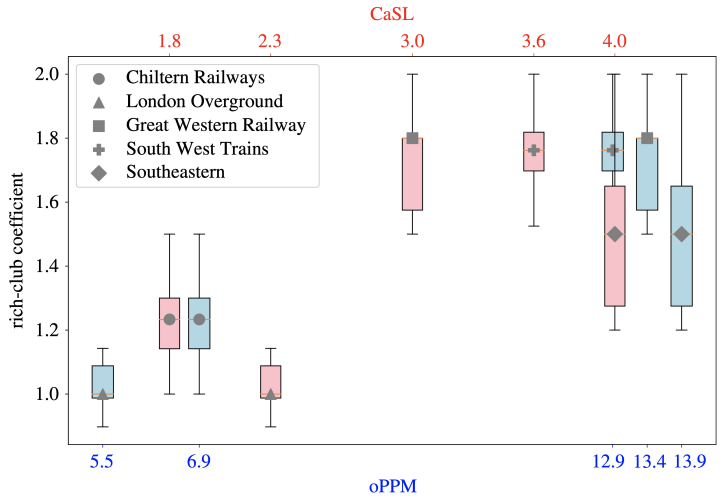


Figure 5.6: Box-plot distribution of rich-cub coefficient (see Eq. 5.4) for a range of filtering threshold $T \in [1, 4]$ showing the mean (marker) and the standard deviation (box limit), for each service provider. The x-axis represents service performance through oPPM (lower blue x-labels, blue boxes) and CaSL (upper red x-labels, red boxes).

Rich-club coefficient analysis

In Figure 5.6 we compare the highest rich-club coefficient observed considering all the possible k (degree) values for each service provider with its performances metrics. Our results show that even if there is a moderate correlation between the value of the rich-club coefficient and the performances (PPM has PCC= 0.62 and CaSLhas PCC= 0.55), there is no evidence of significant correlations between the presence of rich-club phenomenon and service performances.

Core-periphery analysis

The ratio between the size of the core of a network over the size of its periphery represents the percentage of well-connected core stations, versus the sparse periphery stations — intuitively a network with a larger core has more connections between stations and, thus, higher robustness to disruptions. In this section, we compare the percentage of core nodes of each provider network, computed ranking nodes according to degree on the one hand and trophic level on the other. We compare this to the oPPM and CaSL measures. As shown in Figure 5.7, our findings suggest that there is a moderate positive correlation between the size of the degree-core (PCC= 0.38) and the trophic-core (PCC= 0.59) of a provider network and the oPPM. However, there is no correlation with the CaSL (degree-core PCC= -0.09, trophic-core PCC= 0.28).

Removal of random nodes

We attack the networks by removing nodes and analysing the size of the remaining largest component. These experiments are repeated through several random simulations. The upper panel in Figure 5.8 shows the average size of the largest component of each provider and its standard deviation for increasing quantities of nodes removed. We set a threshold value for the largest component of 50% of network size (dashed red line), representing the connectivity limit below which the network is considered non-functional. We measure the percentage of nodes required to disrupt each of the networks, representing its robustness to attacks. The robustness to attacks is then compared with the performance measures of the companies in the lower panel of Figure. Our results show a strong correlation between robustness to attacks and CaSL measures (PCC= 0.83) and a moderate correlation (PCC= 0.58) with oPPM measures.

Extension to other provider statistics

In Figure 5.9, the correlation analysis has been extended to other significant provider-related statistics (number of employees, stations, trains and passengers), showing how oPPM and CaSL are related to these metrics. Note that the incoherence ratio $q/\tilde{q}$ is indeed the most significant correlate. Figure 5.9 also provides a synthesised view of the correlation space described above. It shows that robustness to attacks is a good indicator for cancellations and significant delays (CaSL). The size of the core (both degree and trophic cores) and the rich-club phenomenon do not provide significant correlation with performances. The size of the rail network in terms of the number of employees and stations also has a strong correlation to oPPM, which is probably indicating that larger networks are more likely to have feedback loops and incur cascade effects.

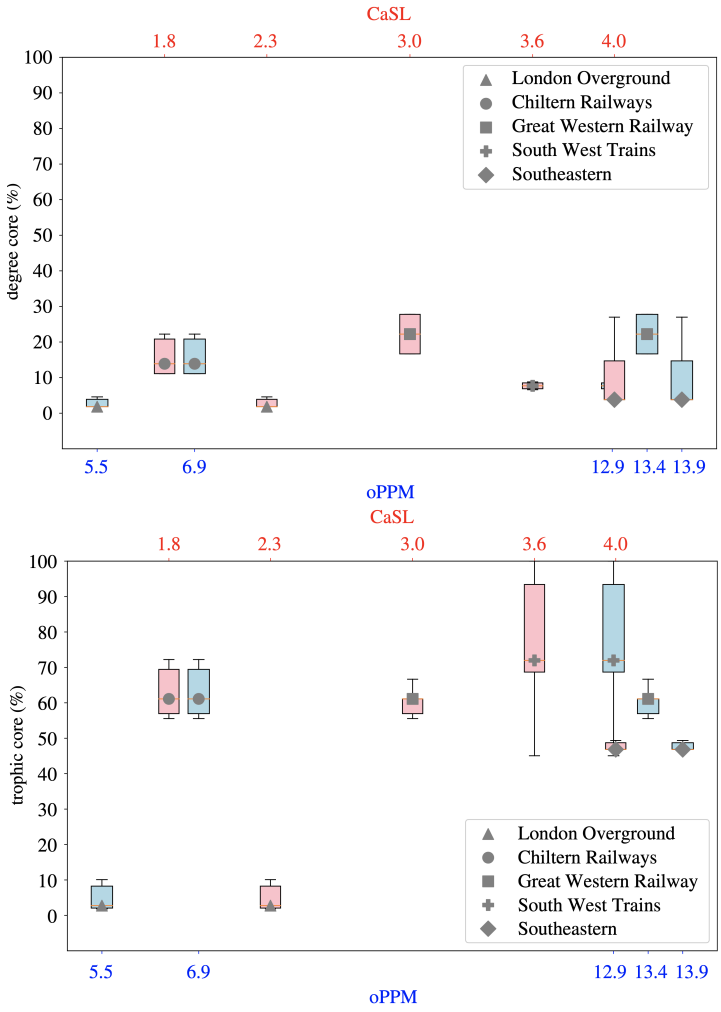


Figure 5.7: Box-plot distribution of core size for nodes ranked by degree (upper panel) and trophic coherence (lower panel) for a range of filtering threshold $T \in [1, 4]$ showing the mean (marker) and the standard deviation (box limit), for each service provider. The x-axis represents service performance through oPPM (lower blue x-labels, blue boxes) and CaSL (upper red x-labels, red boxes).

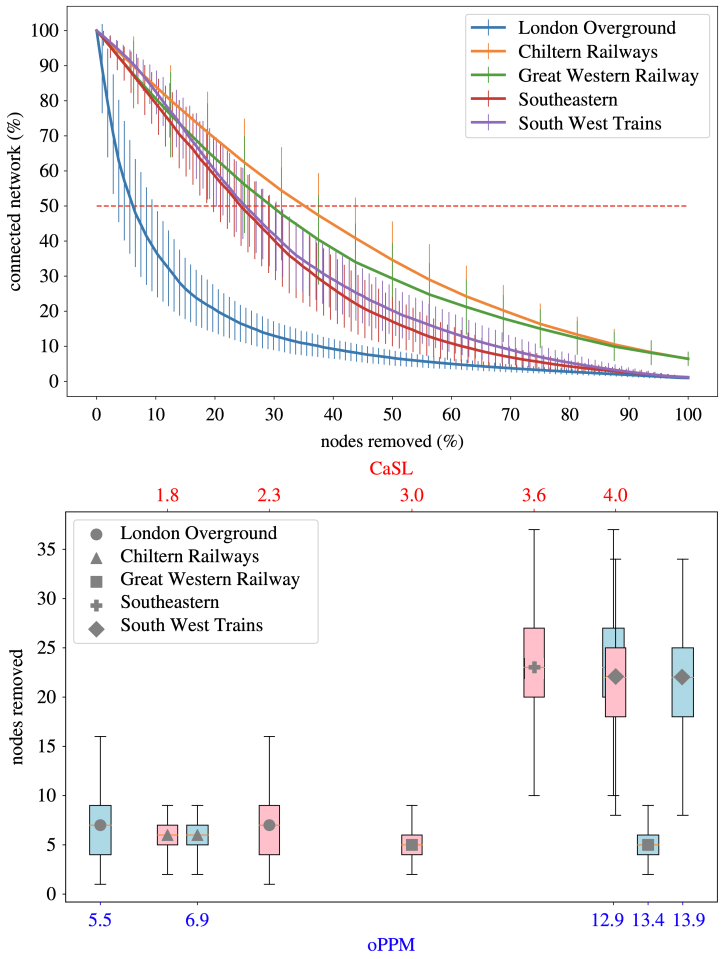


Figure 5.8: *Upper panel*: size of the largest strongly connected component for random node removal for each service provider. The horizontal line indicates when more than 50% of the network is compromised. *Lower panel*: box-plot distribution of node removal percentage needed to lower size of the largest component by 50% for a range of filtering threshold $T \in [1, 4]$, showing the mean (marker) and the standard deviation (box limit), for each service provider. The x-axis represents service performance through oPPM (lower blue x-labels, blue boxes) and CaSL (upper red x-labels, red boxes).

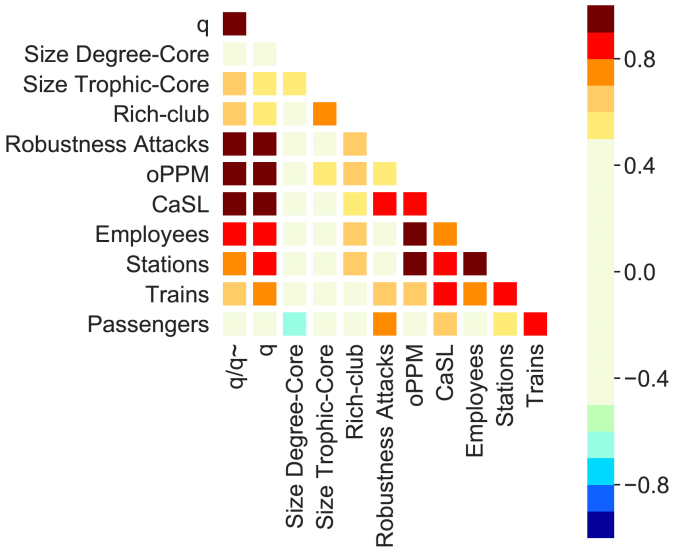


Figure 5.9: Pearson correlation coefficient between: topological measures, including normalized incoherence parameter ($q/\tilde{q}$), incoherence parameter (q), size of degree-core (size degree-core), size of trophic-core (size trophic-core), rich-club phenomenon (rich-club), robustness to attacks (attacks); and operator-related metrics, including public performance measure (oPPM), cancellations and significant lateness (CaSL), number of employees, number of stations, number of trains and number of passengers.

5.6 Discussion

In this chapter, we have proposed a topology-driven data study of London's urban rail network under stress conditions during morning peak-hours. We have represented the rail networks of major service providers as weighted directed graphs,where nodes indicate stations, edges represent commute flows of people, and edge weights count the number of people travelling on that segment. Note that if two stations are connected but there are no passengers travelling through them in the morning peak-hours, these stations are considered disconnected (there is not an edge between these nodes).

We studied the resilience and robustness of these networks drawing inspiration from techniques used in the study of natural complex networks, such as food webs. Our results suggest that network resilience, as measured by trophic incoherence ($q/\tilde{q}$), is strongly correlated with the performance parameters PPM (Public Performance Measure) and CaSL (Cancellations and Significant Lateness) of the underlying service provider of the network. In contrast, most of the different network robustness indicators considered (size of the core and rich-club phenomenon) are not significantly correlated with performance measures, except for robustness to attacks (random percolation) which is correlated with CaSL measurements.

There is interesting research remaining with regards to the dataset we have built for this study, especially regarding the network-related methods we have presented here. For example, it would be interesting to model the flow of passengers using a biased diffusive process on top of the rail networks (e.g. biased random-walks [172]). This could help assess the role of noise — deviations from the shortest path routing between origin and destination we have considered here — in the design of more resilient passenger flows of complex rail networks. Note that strictly shortest path protocols will tend to overload links when the network is attacked [173], [174], an effect that we hypothesise could be mitigated by adding randomness in the routing of passengers. Intuitively, artificially inducing a certain degree of random behaviour in passenger trajectories may alleviate overall congestion and thus achieve lower global travel times. Given that we have shown that trophic coherence is a desirable property of rail networks because it implies a lower number of feedback loops, another interesting open question is: what is the optimal set of changes (train rescheduling or path removal) that a service provider should perform in order to achieve a positive step change in coherence? A problematic in this case would be that the operator will be facing a number of trade-offs when modifying the network, because there will usually be local economic incentives for the presence of loops in transport network.

Chapter 6

Gravity-networks for conflict prediction

6.1 The scientific study of peace and conflict

The scientific study of peace and armed conflict is not a development of recent decades, but rather an old discipline of political sciences. The goal of a researcher in this field is usually related to the discovery of mechanisms promoting sustainable peace across human populations around the world. But peace is a rather elusive term which might be difficult to quantify by itself. Ironically, armed conflict is an objective phenomena with evident and catastrophic effects, which can be readily quantified, modelled and to some extend predicted.

The origin of conflict research can be traced back to the post-World War I era. Before that, theories that influenced international relations were based on the ethics related to the circumstances under which it was morally acceptable and thus legal for countries to attack each other. But the atrocities of World War I, and the even larger-scale catastrophe of World War II moved the inquiry about conflict to the settings of realism, and to the question of why and how does armed conflict emerge. Psychologists, sociologists, economists but also physicists and mathematicians put in motion several academic efforts to understand the mechanisms underlying war with a clear mindset of preventing new global escalations. By the decade of the 1960s many different theoretical frameworks had already emerged [175]–[177]. Such large variety of theories required a process of selection and validation using empirical tests. That could only be accomplished through the use of data.

Several large-scale data-gathering efforts and techniques were developed in the following two decades, including projects like the World Event/Interaction Survey (WEIS)

[178], the Conflict and Peace Data Bank (COPDAB) [179] or the Correlates of War (COW) [180] databases, some of them still being in active development nowadays. Most of these databases contain either individual country attributes (GDP, population, regime types, minority groups, etc.), interaction events between nations or political actors (alliances, international trade, militarized disputes, attacks to civilians, embargoes, etc.) or both. The availability of data, together with the development of accessible computational statistical techniques, led the field of conflict research into a model-testing mindset during the 1980s and 90s. Classical hypothesis testing and p-values analysis became prevalent in the literature, with the aim of finding explanatory variables for conflict data backed by theoretical underpinnings. However, fundamental limitations recurrent in many social sciences, like the impossibility of isolating causal factors or measuring them with precision, limited the applicability of such research efforts. A too strong focus on p-value significance without careful out-of-sample prediction evaluation brought the development of many significant but over-fitted models without much prediction accuracy [181].

In recent years, the field of conflict research has benefited extraordinarily from an increasing attention towards out-of-sample predictive performance, and from the models and best practices of machine learning in general [182]. Although non-parametric machine-learning models offer wider flexibility to capture non-linear effects and higher-order relationships between large sets of features, this usually comes with the price of reduced interpretability. For this reason, there is an increasing need for holistic approaches where simpler and more interpretable statistical models are combined with more sophisticated predictive models, conforming different steps in a scientific process towards theory building [183]. Finally it is worth mentioning that even in the current predictive paradigm it is difficult to bridge the gap between conflict research and international policy making. Existing research is already addressing the issue of discerning the effects of actions triggered by national and international decision makers with regards to peace-keeping [184]–[186], but realistically current forecasting tools can at their best inform policymaking of what is the likelihood of future events if no actions are taken.

6.1.1 Levels of analysis

The concept of level of analysis [187] is an important methodological factor driving different approaches to conflict research and international relations in general [188]. The level of analysis in a given study relates to the scale of the object causally associated with the phenomena under examination. The most microscopic scale would be the individual level, which in the case of conflict research would study the influence of individuals on particular wars. For example, one could study the individual actions and motivations of dictator Francisco Franco as causal effects for the Spanish Civil War. On the opposite

side of the spectrum, the most macroscopic perspective is the systemic level of analysis, where actors are individual states but the focus is on the emergence of international processes from the different interaction structures between nations. For example, one could search causal explanations for World War I on the absence of strong intergovernmental organizations (IGOs) such as United Nations or the European Union. Most studies lie somewhere in between these two scales. For example, the monadic level of analysis considers domestic factors of nations such as their economic model, political system or religious distribution. One step further, the dyadic level of analysis is possibly the most common approach [189] and studies bilateral interactions amongst states (trade, alliances, vetoes, embargoes, etc.) as explanatory factors for conflict amongst them. Note that levels of analysis do not necessarily need to be discrete choices, and most models will work with features across different level factors.

New generations of studies are increasingly moving towards subnational levels of analysis, which has been reciprocated by the emergence of disaggregated datasets tracking conflict events to precise geographic locations [190]–[192]. This new trend is opening the field to a larger set of explanatory factors related to precise demographic variables, climate, natural resources and other geographic factors or local political unrest, to name a few.

6.1.2 Networks in conflict research

Research efforts trying to transcend the dyad paradigm appear early on in the literature using the framework of social networks, proposing methodologies to measure interactions amongst nations in order to construct higher-level explanations of cooperation and conflict at a network system level [193]–[195]. Interesting complex network techniques such as community detection are used in some of these studies [196]–[199], which tend to be concerned with the formation of groups (communities or clusters) of nations through trade, alliance and conflict. Some others study the concept of centrality (mainly closeness and degree), finding that highly central countries in trade networks tend to be associated with lower levels of conflict [200], [201]. Besides some rare cases where out-of-sample prediction is used as evaluation tool [202], [203], most studies on international networks are purely descriptive and based on significance claims.

6.1.3 Novelty of our study

The present study compounds some of newest trends in conflict research mentioned above. Our unit of analysis is the urban settlement or city. Although defining a city is challenging even from the most basic physical or spatial perspective, cities are interesting subnational actors because they usually have significant political idiosyncrasy and bring large groups of people together into relatively confined spaces promoting com-

mon cultural identities. They have well known geographic locations and in many cases well known population records, and can be reasonably well combined with disaggregated conflict datasets. To the best of our knowledge, this is the first study regarding conflict research using cities as level of analysis.

Importantly, cities interact locally with one another through a multiplicity of dimensions, usually involving the flow of people, goods or information. As a result of such interactions, cities around the world are connected through global complex networks via all sorts of infrastructures such as roads, railways, sea or air routes, but also power distribution lines or telephone and internet connections. Combining all of these networks is challenging from a practical data collection and processing perspective. We use the gravity law (see Section 6.3.1) as an approximation for the amount of flow between cities, which helps us building a global network of interactions. As described above, networks are not new in the field of conflict research, but they have always been used at the state level. Therefore, the study of a global network of cities is another contribution of the present study.

Finally, we capitalise on these models of global networks of cities to derive a set of centrality measures attributed to each city. We test a larger variety of centrality measures than in previous (state-based) network studies, including hybrid measures that combine topological information with metadata on ethnic groups and international borders. We use these measures as factors for a predictive analysis evaluated out-of-sample on disaggregated conflict data, altogether generating a novel set of conflict predictors.

6.1.4 Research outline

Section 6.2 describes the three datasets we use in our analysis: one describing location, population and state membership of cities; another containing spatial representation of politically relevant ethnic groups; and a third containing a global comprehensive set of geographically tagged conflict events from 1989 to 2018.

Section 6.3 describes the set of methods we use to build our global network of interactions between cities. We present two different methods to build such networks, one deriving purely from spatial proximity and the other directly from thresholds on the gravity law. Section 6.3 also describes the set of centrality measures we use as predictors in our predictive analysis. Such measures include degree, betweenness, closeness and pageRank, as well as their weighted versions. They also include three bridgeness measures based on topological communities, ethnic communities and national communities respectively.

Also in Section 6.3 we provide detail on the statistical methodology used for our predictive analysis. We present the prediction objective as a classification task where we want to predict which cities will be under a state of conflict in a given year. We

also describe the algorithms that will be used for such task, namely a logistic regression and random forests. We present the statistical models we will make predictions with: we describe our baseline model, which excludes all network features and is only based on the autoregressive component of conflict history and the population of each city; we also describe the set of full models containing both baseline and network-based features. Still in the same Section, we elaborate on our data partition scheme, which is based on rolling forecasting cross-validation. Finally, we describe the performance metrics used in the analysis and the measures of variable importance.

Section 6.4 presents all the results of our predictive analysis, which are finally discussed in-depth throughout Section 6.5.

6.2 Data

6.2.1 City data

We use data from the National Geospatial Intelligence Agency, containing 7322 settlements around the world with their latitude, longitude, population, country and province affiliation [204]. For the purpose of this chapter, we shall call all settlements cities. The geospatial data includes cities that vary in population from mega-cities (several millions) to small towns. The data represents around 25% of the world's total population. We only use cities with a population above 10,000, of which we find more than 5,900 in the dataset, yielding high city resolution.

6.2.2 Ethnic data

As shown in Section 6.3.3, some of our network features are enriched with metadata representing ethnic groups. For this purpose, we use the Geocoded Ethnic Power Relations (GeoEPR) dataset [192], which provides polygon data for the spatial distributions of politically relevant ethnic groups around the globe. Polygon data is a type of vector data that includes 3 or more vertices with coordinates in the latitude and longitude space, forming closed figures that, in this case, represents boundaries of territory dominated by a particular ethnic group. In this dataset, ethnicity is generally defined as any set of subjective views that bring individuals towards the belief of a common cultural ancestry. Allowing for this subjective variable is a key trait of GeoEPR which differentiates it from other ethnic datasets such as Geo Referencing of Ethnic Groups (GREG) dataset [205] which do not consider aspects such as religion, thus grouping together very distinct ethnic groups such as Hutus and Tutsis or Sunni and Shi'a Arabs. In addition, GeoEPR is focused on 'politically relevant' ethnic groups, which are defined as those having political organizations in the public arena, or those being publicly excluded or discriminated based on ethnicity. Finally, GeoEPR is a dynamic dataset that registers

the spatial evolution of ethnic groups through time. For the purposes of the current work, however, we have only used a snapshot fixed at the year 2017.

6.2.3 Conflict data

We use data from the Uppsala Conflict Data Program (UCDP). In particular, we use the UCDP Georeferenced Event Dataset (GED). This is the most comprehensive event dataset on organised violence existing up to this date. Crucially for our purposes, this dataset is geographically disaggregated below state level, meaning we have access to the geographic coordinates of each event. It contains 179,130 events occurred between 1989 and 2018 around the globe.

UCDP-GED is an event-based dataset. According to UCDP, an event is defined as *"The incidence of the use of armed force that was used by an organised actor against another organized actor, or against civilians, resulting in at least 1 direct death at a specific location and a specific date"*[191]. Therefore, there are clear criteria to qualify an event as such.

Note that organised actors refers to either governments of independent states, formally organised groups which have publicly advertised their name and purpose, and informally organised groups which have not publicly advertised a name or purpose but are recurrently involved in armed violent patterns. Accordingly, this dataset includes three types of events: state-based conflict (violence against state representatives committed by another state or group), non-state conflict (violence between non-state groups) and one-sided conflict (violence against unarmed civilians).

Some of the events registered in the GED dataset may geographically occur in locations which are not registered in our city dataset. For this reason, we have processed GED data so that each event is attributed to the closest city registered in our dataset.

6.3 Methods

6.3.1 The gravity law

One method to infer the volume of flow between any two given cities is the widely used gravity law [100]. The gravity law has been employed in various forms and disciplines in the social sciences for over a century [206], [207], but as with many such laws, its theoretical underpinning comes in many forms. Gravity laws generally describe the attractive force between two social entities and has been used to describe the flow of a wide variety of goods (e.g. vehicles, goods, disease, and human beings) [208]–[211], and information (e.g. telephone calls and social media messages) [212]–[214] between cities and countries. In fact, they have also been used in the conflict research literature [215]–[217], especially in the context of bilateral trade study and its relation to conflict.

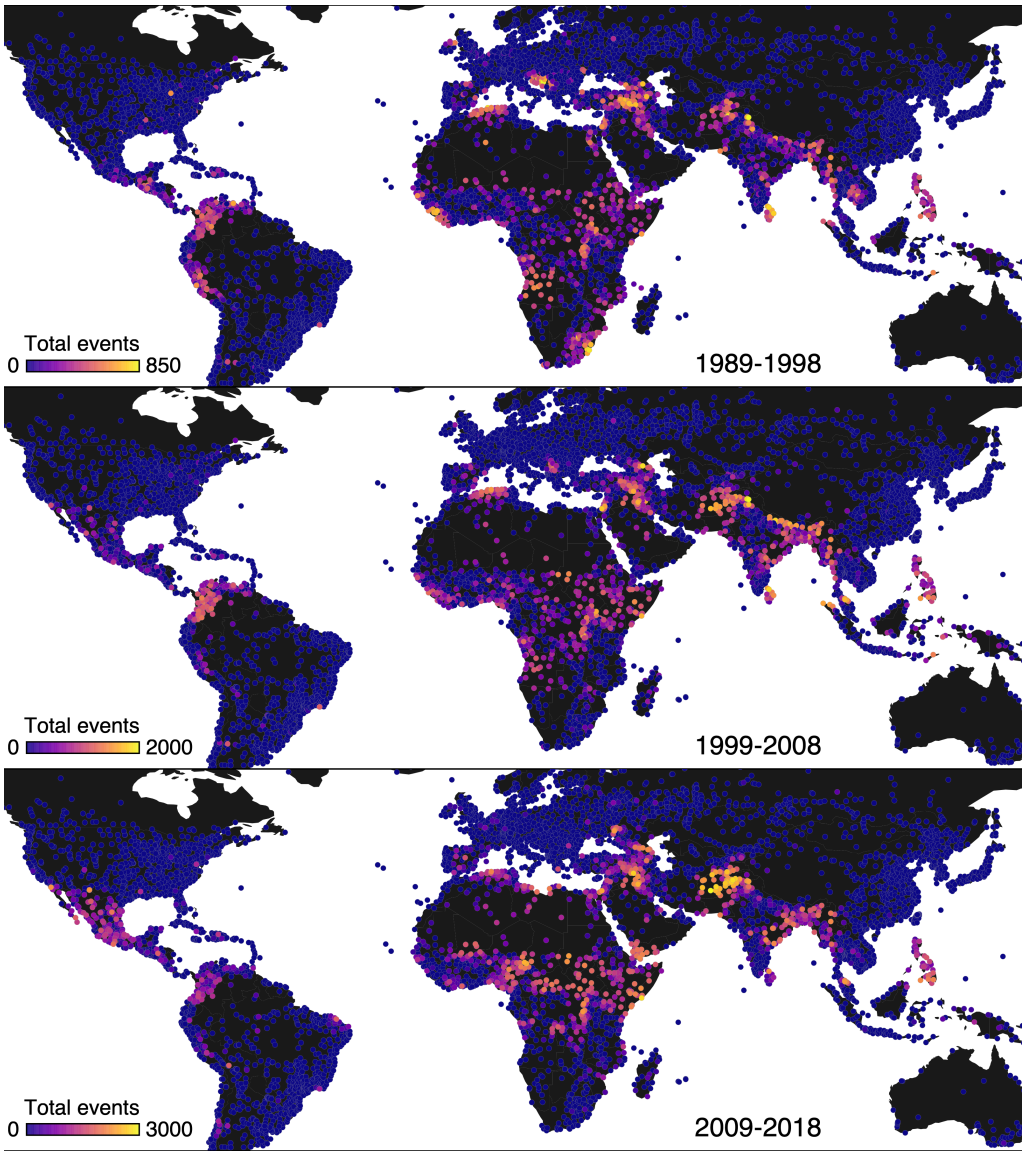


Figure 6.1: Partial temporal aggregation of the GED dataset (Section 6.2.3) across 3 different time periods. Data has also been aggregated to the closest city registered in our city dataset (Section 6.2.1). The colormap shows the total number of events attributed to each city across each time period, using a logarithmic scale.

The law consists of two main dependencies: the importance or fitness of each of the two social nodes (i.e. usually their population P, given that most research agrees that population is a significant factor determining the flow of goods or people [218], [219] in social systems) and the rate of flow-decay dependent on the (typically Euclidean) distance d that separates them. It is typically expressed as:

$$F_{ij} \propto (P_i P_j)^{\alpha} \, d_{ij}^{-\gamma} \quad , \tag{6.1}$$

where the exponent α and γ are the parameters of the model that can take different forms depending on the context of application. In the classical gravity law, $\alpha = 1$ and $\gamma = 2$. The discrepancy between different models lies in what form the gravity law takes, i.e. the value of parameters weighting population and distance, α and γ.

As a model of interactions, it can be shown that the gravity law arises from an entropy maximisation principle where there are constraints on the cost and benefit of interaction amongst a system of actors [220]. Therefore, it can be argued that in the absence of information, the gravity law represents the most likely set of interactions in a system. A thorough review of gravity laws and complex networks can be found in [100].

6.3.2 Network construction

Both physical and intangible networks have always permeated the way cities interact. Although there exists data on several of such networks (e.g. roads, rail or flight networks), such data is typically (i) hard to collect and to process and (ii) very asymmetrical between regions at different development stages. At the same time, it would be hard to find criteria to evaluate which of such networks is more representative of the overall connectivity between cities. For this reasons, here we take a different approach and instead build our networks using spatial interactions models.

The first step of the construction of the network consists on inferring the amount of flow connecting any two cities in our dataset. A natural candidate for this task is the gravity law, as shown in Eq. (6.1). This yields a pairwise interaction matrix that needs to be constrained in order to obtain a network. The density of such network can be controlled using a threshold hyper-parameter $\mathcal{T}$ so that edges will only exist between cities i and j if their pairwise interaction strength is above a certain level, that is:

$$a_{ij} = \begin{cases} 1 & \text{if} \quad F_{ij} > \mathcal{T} \\ 0 & \text{if} \quad F_{ij} < \mathcal{T} \end{cases} \quad . \tag{6.2}$$

Besides using the gravity flow F_{ij} for our connection rule, we also use it to weight the edges that end up created. Note that this method yields three hyper-parameters $(\alpha, \gamma, \mathcal{T})$ that will need to be accounted for when building our statistical analysis. Depending on the values of these, we will consider two types of network models, described below.

Geographic network

In this case, we consider $\alpha = 0$ and $\gamma = 1$, so that the connection rule simply becomes:

$$a_{ij} = \begin{cases} 1 & \text{if} \quad d_{ij} < \mathcal{R} \\ 0 & \text{if} \quad d_{ij} > \mathcal{R} \end{cases} , \tag{6.3}$$

where $\mathcal{R} = 1/\mathcal{T}$. This criterion simply states that those cities that are closer than $\mathcal{R}$ kilometres must be connected by a network edge. This yields a network which is highly constrained by distance, effectively providing a model of land-based connections between cities. Assuming this, we will further restrict edges that connect islands with continental areas, as well as edges connecting continental areas separated by sea (e.g. southern Europe and northern Africa). Furthermore, it is reasonable to model separately the connection rule and the weighting of edges. In practice, this means we will first draw the edges of the network using the condition in Eq. 6.3, and then weigh those existing edges using the gravity law in Eq. 6.1 with general exponents (α, γ).

As shown in Figure 6.2, $\mathcal{R}$ both affects the range of connections and the density of the network. Lower $\mathcal{R}$ values (e.g. $\mathcal{R} = 200$ km) generate very short range connections, consequently producing a sparse and relatively disconnected graph. Highly dense regions such as central Europe or the western coast of North America quickly produce large connected components, whilst less dense regions in Africa and South America remain more isolated. For increasingly larger $\mathcal{R}$ values (e.g. $\mathcal{R} = 500$ km) the network becomes globally percolated, meaning Europe, Africa and Asia are merged into a large single component, separated from another large component connecting all America.

Note that this connection rule produces networks with similarities to Random Geometric Graphs, which are a type of spatially restricted networks with relatively homogeneous degree distributions [221]. In fact, the spatial constraints introduced by Eq. 6.3 prevent the network from forming hubs, although in this case this is offset by the heterogeneity in city density around the globe. In any case, it is clear that the set of parameters $(\mathcal{R}, \alpha, \gamma)$ have a significant impact on the structure of the resulting network and, as we show throughout the rest of the chapter, they have an effect in the ability of the network to produce useful features in the context of a predictive analysis of armed conflict.

Gravity network

In this case, we consider $\alpha > 0$ and $\gamma > 0$, so that the connection rule is:

$$a_{ij} = \begin{cases} 1 & \text{if} \quad (P_i P_j)^{\alpha} \, d_{ij}^{-\gamma} > \mathcal{T} \\ 0 & \text{if} \quad (P_i P_j)^{\alpha} \, d_{ij}^{-\gamma} < \mathcal{T} \end{cases} . \tag{6.4}$$

This connection rule transcends the purely cost-based geographic network described in Eq. 6.3, balancing instead both the costs (distance) and the benefits (population) of

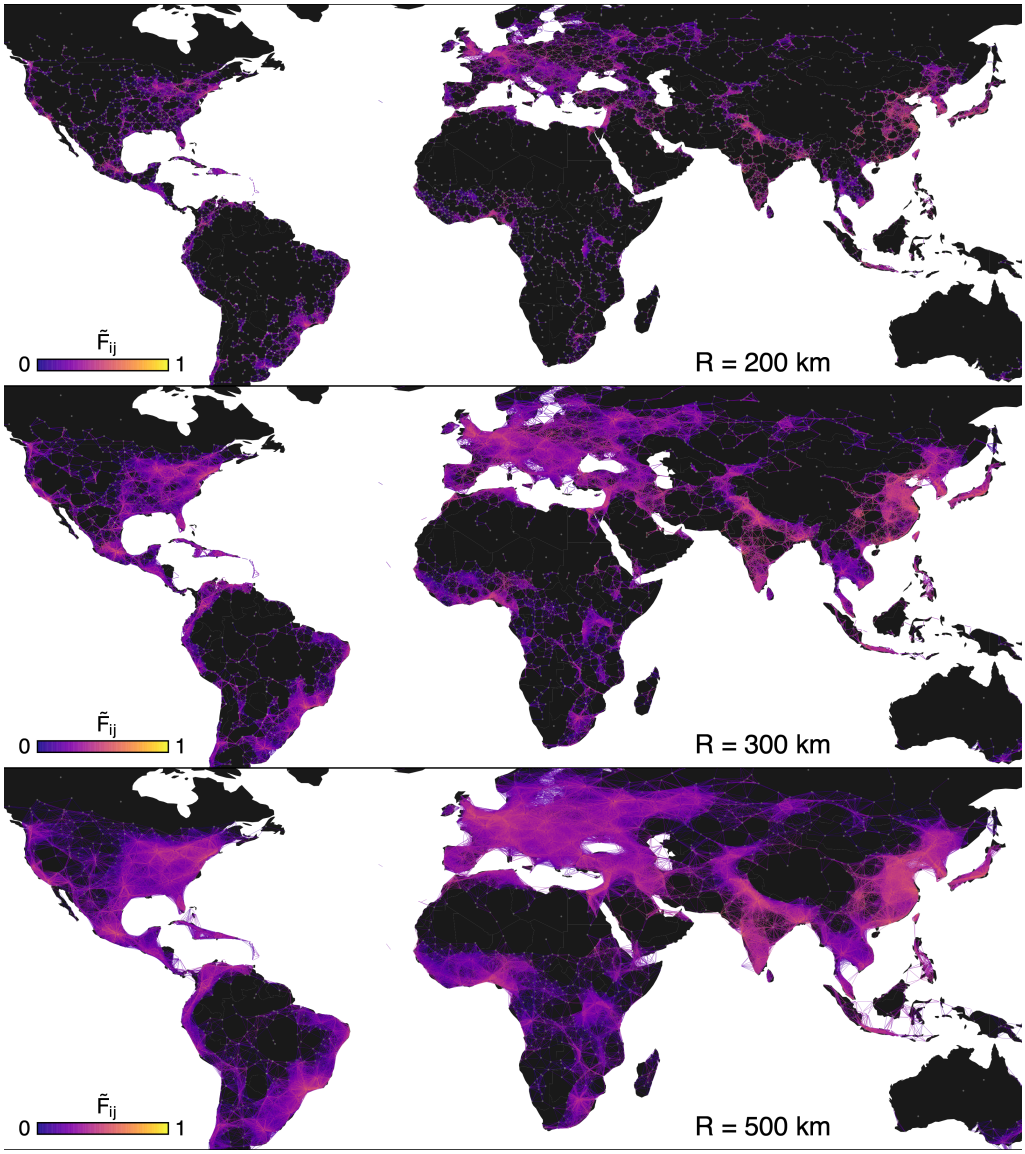


Figure 6.2: Geographic networks derived from the criterion in Eq. 6.3 for different values of $\mathcal{R}$. We are using the standard gravity law (see Eq. 6.1) with $\alpha = 1$ and $\gamma = 2$ for edge weights. The colormap represents the weights of each edge in logarithmic normalised scale so that $\tilde{F}_{ij} = \frac{\log(F_{ij}) - \min \log(F)}{\max \log(F) - \min \log(F)}$.

each potential edge. Distant cities now have the chance to connect if their populations are significant enough to offset the cost involved. Note that in this case, the exponents (α, γ) determine the balance between cost and benefit and thus the final structure of the network. An alternative procedure consists on fixing the number of edges $\mathcal{E}_g$ in the gravity layer and simply chose the $\mathcal{E}_g$ links with largest flow from all possible connections. This yields effectively the same structures as the method in Eq. 6.4, but it provides an easier control over edge density. In fact, this enhanced control over density becomes very useful when searching for optimal hyper-parameters for the predictive analysis presented in the rest of the chapter. Also note that in this model the network has the potential to represent not only land-based routes but also sea or air paths. For this reason, we relax the restrictions of the previous model and allow continental regions to connect to islands or other regions separated by sea. Finally note that in this case, we use the same exponents (α, γ) both for the connection rule and the weights of existing edges.

In Figure 6.3 we show the connection rule in Eq. 6.4 using the edge-density selection method. The resulting networks promote significantly longer-range edges which leads to network structures differing from than the previous model. With $\alpha = 1$ and $\gamma = 2$, the particular configurations in Figure 6.3 still have a significant bias towards distance costs, specially for lower edge densities ($\mathcal{E}_g = 7000$ and $\mathcal{E}_g = 1500$). Despite this, we can see hubs forming around highly populous mega-cities, triggering the emergence of radial structures where peripheral cities connect to their regional hub, which at its turn connects to other distant regional hubs.

6.3.3 Centrality measures

We derive our set of independent predictors from the networks constructed using the methods in Section 6.3.2. We compute two types of network features. On the one hand, we derive a set of standard centrality measures that range from local to global network scales. These are measures that can be deducted from the topology of interactions exclusively, i.e. the adjacency matrix of the underlying graphs. On the other hand, we also compute a set of custom centrality measures that are based on bridgeness centrality.

Standard measures

We consider the following centrality measures:

- *Degree*, k (Eq. 1.4): counts the number of edges of each city. This quantity is highly related to the surrounding density of urban settlements.
- *Weighted degree*, s (Eq. 1.12): also known as strength, counts the number of edges of each city adjusted by their weight. This quantity is highly related to the gravity flows surrounding each city.

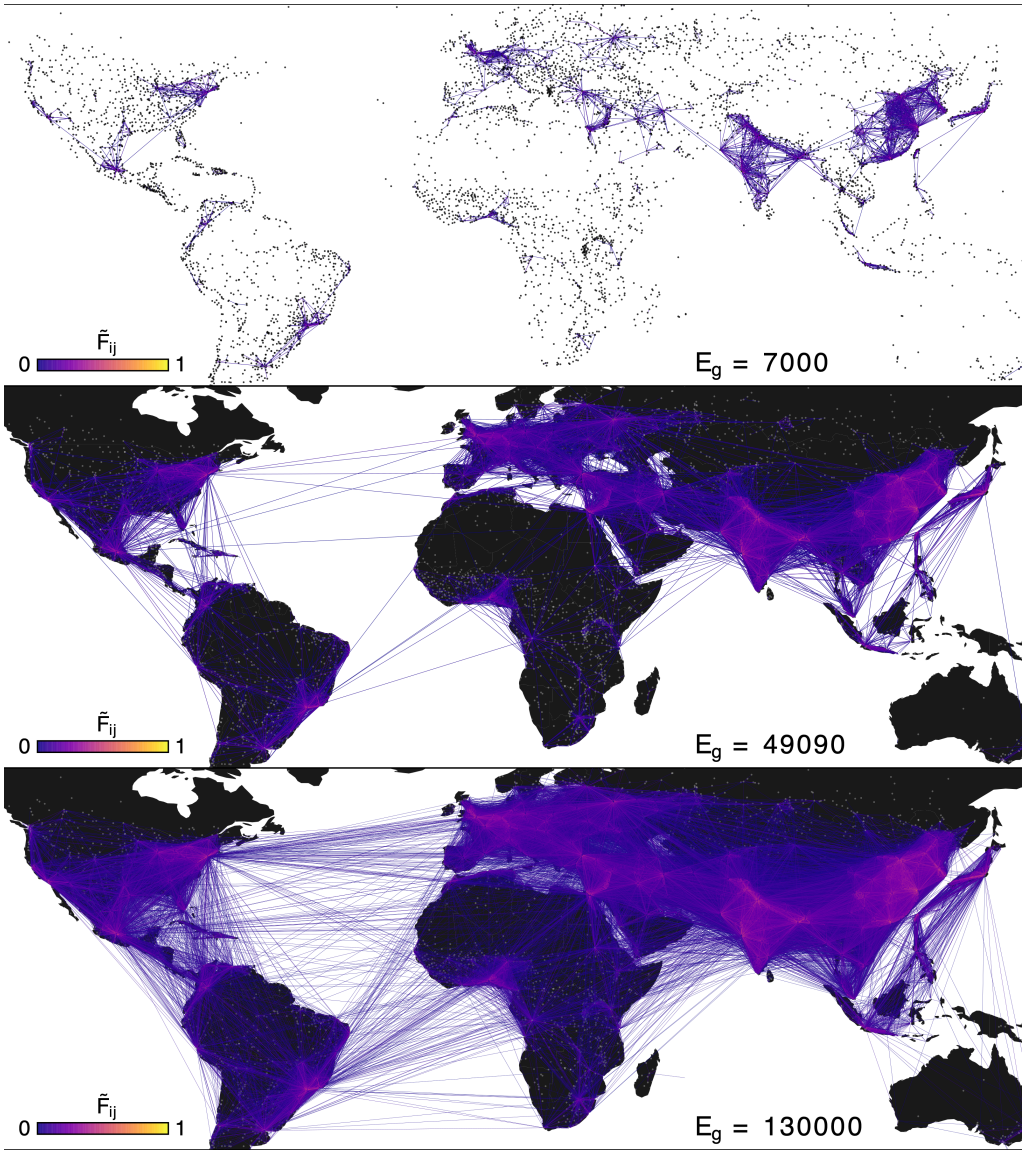


Figure 6.3: Gravity-based networks derived from the criterion in Eq. 6.4 for different values of edge-density $\mathcal{E}_g$. We are using the standard gravity law (see Eq. 6.1) with $\alpha = 1$ and $\gamma = 2$ for both Eq. 6.4 and edge weights. The colormap represents the weights of each edge in logarithmic normalised scale so that $\tilde{F}_{ij} = \frac{\log(F_{ij}) - \min\log(F)}{\max\log(F) - \min\log(F)}$.

- *Betweenness*, B (Eq. 1.10): by counting the number of shortest paths that cross each node, it indicates to which extent each city is a bottleneck for efficient global flows.

- *Weighted betweenness*, B_W: by taking into account gravity weights, shortest paths are biased towards routes traversing populous cities, even at the cost of covering larger distances.

- *Closeness*, C (Eq. 1.8): is the reciprocal sum of the shortest distance form a city to any other city in the world, and thus it measures how well communicated a city is with its environment.

- *Weighted closeness*, C_W: again, by taking into account gravity weights, shortest paths are biased towards routes traversing populous cities, even at the cost of covering larger distances.

- *PageRank*, PR (Eq. 1.11): it is based on the eigenvector concept of recursively defining central nodes as those which are most connected to other central nodes, but it also takes into consideration the degree of the node (highly linked nodes are more central) as well as the degree of neighbouring nodes (links from more parsimonious nodes are more valuable).

- *Weighted pageRank*, PR_W: also takes into consideration the weight of each link in attributing its importance.

In Figure 6.4, we illustrate the weighted variants of these centrality measures for a given instance of geographic network.

Bridgeness measures

We include another set of independent network predictors, based on bridgeness centrality (see Eq. 3.16) and some extensions of it, based on ethnic and country-boundary metadata.

- *Community bridgeness*, $cBridg$: equivalent to the participation coefficient in Eq. 3.16, it measures the participation of each node in each of the communities in a modular partition of the network. We use the degree-corrected Stochastic Block Model (SBM) (see Section 3.1.2) to find such partition and the corresponding bridgeness centrality.

- *Weighted community bridgeness*, $cBridg_W$: in this case, we modify the way we quantify the participation of each node into a given community c (previously de-

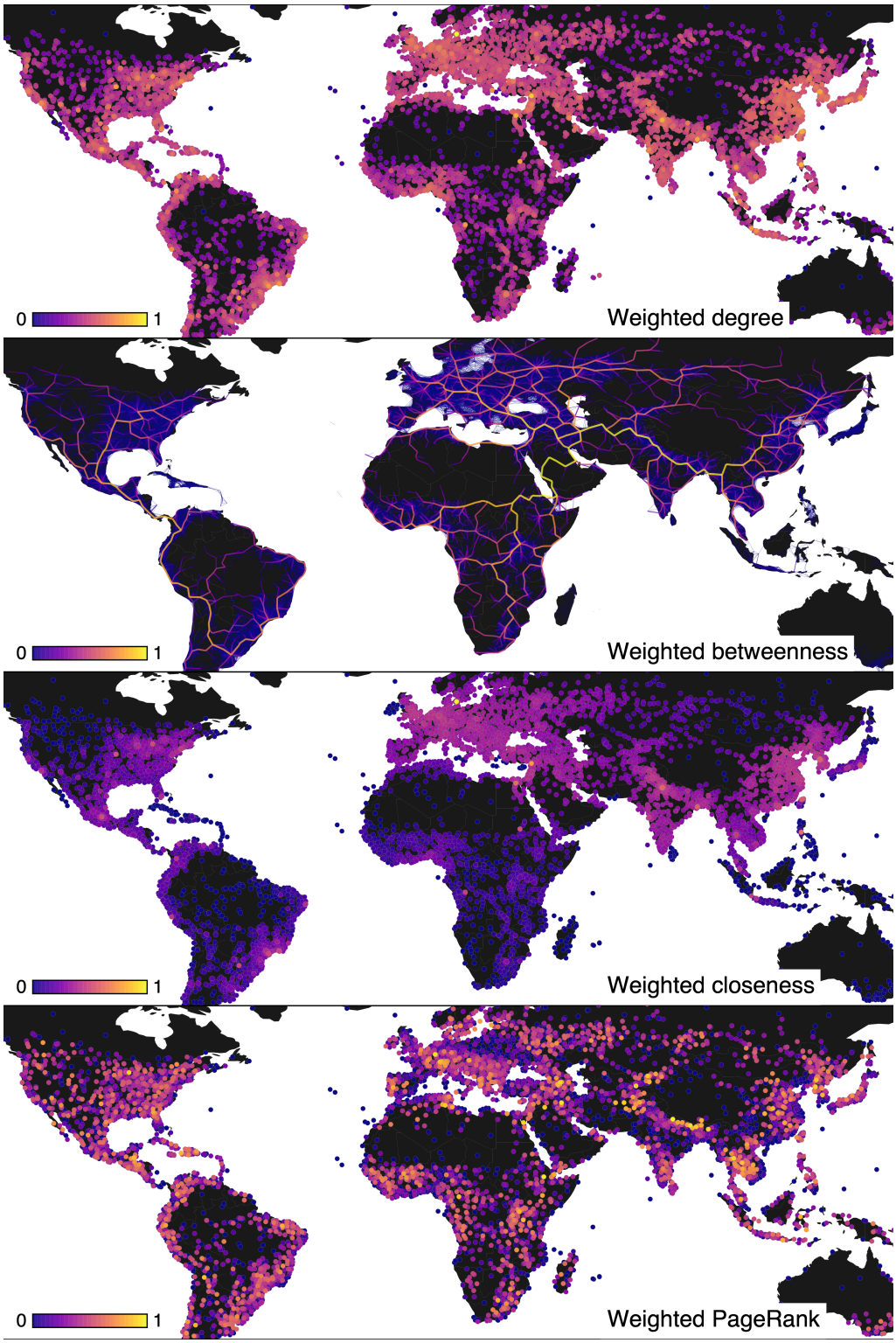


Figure 6.4: Centrality measures in a geographic network derived from Eq. 6.3 $\mathcal{R} = 400$ and gravity law (see Eq. 6.1) weighting with $\alpha = 0.5$ and $\gamma = 1.5$. The colormap represents centrality of each node (edge in the case of betweenness) in logarithmic normalised scale.

fined in Eq. 3.15) so that:

$$\pi^c_W(i) = \frac{\sum_j F_{ij}\delta_{c,c_j}}{\sum_j F_{ij}} \quad , \tag{6.5}$$

where F_{ij} is the gravity flow between cities i and j defined by the gravity law in Eq. 6.1. We calculate weighted community bridgeness by using Eq. 6.5 in Eq. 3.16.

- *Ethnic bridgeness*, $eBridg$: in this case, we seek a bridgeness measure that is based on a community partition arising from the ethnicity-related metadata described in Section 6.2.2. We modify Eq. 3.15 to reflect this, so that:

$$\pi^e_i = \frac{\sum_j a_{ij}\delta_{e,e_j}}{\sum_j a_{ij}} \quad , \tag{6.6}$$

where e represents each of the ethnic groups contained in our dataset. We calculate ethnic bridgeness by using Eq. 6.6 in Eq. 3.16.

- *Weighted ethnic bridgeness*, $eBridg_W$: we modify the previous definition by including gravity weights, so that:

$$\pi^e_W(i) = \frac{\sum_j F_{ij}\delta_{e,e_j}}{\sum_j F_{ij}} \quad , \tag{6.7}$$

where F_{ij} is the gravity flow between cities i and j defined by the gravity law in Eq. 6.1. We calculate weighted ethnic bridgeness by using Eq. 6.7 in Eq. 3.16.

- *International bridgeness*, $iBridg$: in this case, we seek a bridgeness measure that is based on a community partition arising from the country-membership metadata of each city. We modify Eq. 3.15 to reflect this, so that:

$$\pi^s_i = \frac{\sum_j a_{ij}\delta_{s,s_j}}{\sum_j a_{ij}} \quad , \tag{6.8}$$

where s represents each of the states or countries contained in our dataset. We calculate International bridgeness by using Eq. 6.8 in Eq. 3.16.

- *Weighted international bridgeness*, $iBridg_W$: again, we modify the previous definition by including gravity weights, so that:

$$\pi^s_W(i) = \frac{\sum_j F_{ij}\delta_{c,c_j}}{\sum_j F_{ij}} \quad , \tag{6.9}$$

where F_{ij} is the gravity flow between cities i and j defined by the gravity law in Eq. 6.1. We calculate weighted ethnic bridgeness by using Eq. 6.9 in Eq. 3.16.

In Figure 6.5, we illustrate the weighted variants of these bridgeness measures for a given instance of geographic network.

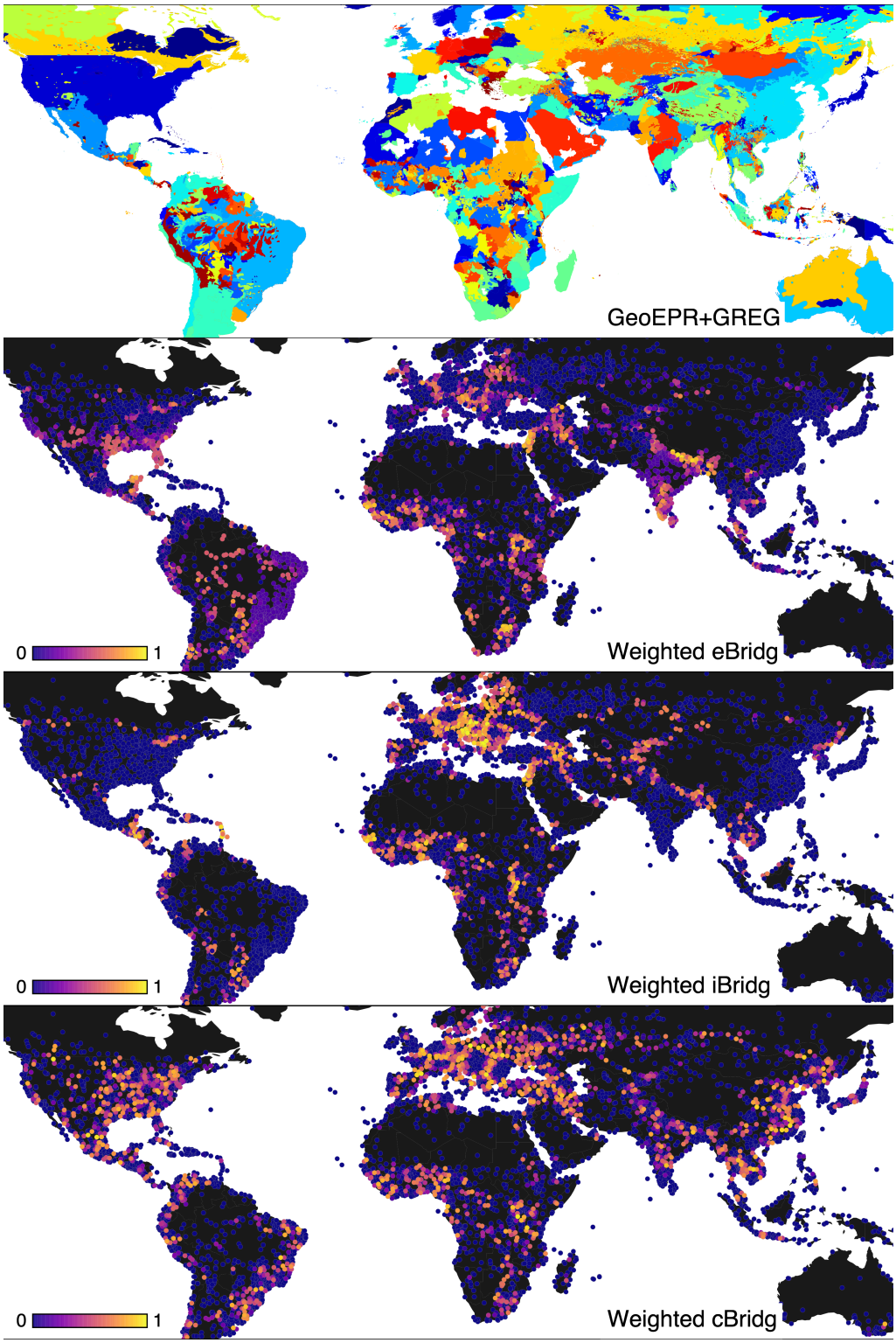


Figure 6.5: **Top panel:** GeoEPR and GREG datasets, with colour-coded ethnic group boundaries. **Lower panels:** Bridgeness measures in a geographic network derived from Eq. 6.3 $\mathcal{R} = 400$ and gravity law (see Eq. 6.1) weighting with $\alpha = 0.5$ and $\gamma = 1.5$. The colormap represents centrality of each node in logarithmic normalised scale.

6.3.4 Predictive modelling

We want to construct a statistical framework that allows us to systematically test whether the gravity-network predictors described above are useful for forecasting armed conflict. More precisely, by forecasting we mean producing a set of predictions about events in the future given estimations from a model trained using events in the past. Note that this approach moves away from correlation analysis or goodness-of-fit tests, and instead seeks to maximize predictive performance under a set of reproducible out-of-sample conditions. By useful predictors, we mean that we want to test whether our network-derived features lead to an increment in predictive performance with respect to a baseline model that does not contain them.

Classification task

Our conflict dataset contains great temporal resolution and describes the occurrence of individual events of political violence. In our case, however, we narrow the scope of our predictive analysis down to a binary classification task which sets out to predict whether individual cities will be involved in some armed conflict at some point in the future. For this, we need to set a criterion that defines whether a city is involved in conflict or not. We begin by temporally aggregating events in one-year periods, and then fixing an event threshold $\mathcal{T}_E$ in the number of events per year: if the number of events $N_E(i,t)$ in city i at year t is greater than some threshold (i.e. $N_E(i,t) > \mathcal{T}_E$) we declare that a conflict occurred. Otherwise, the city is declared peaceful.

Classification algorithms

We use two different statistical models in our analysis. On the one hand, we use logistic regression in the form of a Generalised Linear Model (GLM) with logit link function. On the other hand, we use the Random Forest (RF) model. Both are useful for the classification task we have defined above, but taking into account their respective trade-offs they are used at different stages of the analysis, as we will show below.

Logistic regression is widely used to model binary-outcome dependent variables (such as the conflict/peace dichotomy defined above) by attributing to each event a probability following a logistic function. Given a set of binary outcome variables $\{Y_i\}$ and predictors $\{x_{1,i}, \ldots, x_{m,i}\}$, the logistic regression model can be expressed in terms of a GLM by defining the probability of success ($p_i = \mathbb{E}\left[Y_i | x_{1,i}, \ldots, x_{m,i}\right]$) as a logit (inverse logistic) link function [222]:

$$\text{logit}\left(p_i\right) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_m x_{m,i} \tag{6.10}$$

The logit model tends to perform well when compared with many other machine learning

techniques [223], but it has significantly lower computational costs and is robust to overfitting. It has the additional benefit of having higher interpretability through its linear coefficients $\beta_0, \ldots, \beta_m$, which reflect the impact of each predictor on the response variable when the former have been normalised to a unit range. Given its characteristics, we will use this model to search for optimal combinations in the network-hyperparameter space $(\mathcal{R}, \mathcal{E}_g, \alpha, \gamma)$ described in Section 6.3.2.

The random forest model [224] is a very popular ensemble learning method based on decision trees. Our random forest model is implemented as in the original model introduced by Breiman [225]. It uses classification decision trees combined with bootstrap-aggregating (bagging) and random feature selection [226]. The random forest is a very versatile model that has been shown to perform better than logistic regression under some circumstances [227], but there is no guarantee of superior performance in our particular application. Given its higher computational costs, we will only use the random forest to benchmark logistic regression once an optimal combination of network-hyperparameters have been found using GLMs.

Predictive models

Generally speaking, we use two different sets of models. The first set is based on two simple network-independent metrics that are used in conjunction as a baseline model, including:

- Conflict history: it has been shown that conflict patterns are significantly persistent through time and that past accounts of violence in a region are the best predictors for future conflict [228]. This feature is computed by aggregating all the events occurred in a city during the corresponding training set.
- Population: we have already shown how our network features fundamentally use the gravity law, which in our case derives from city population measures (see Eq. 6.1). Therefore, it is reasonable to include city population as a control variable that helps us discern between purely demographic and network effects.

For clarity, we can express these features in logistic regression terms:

$$\text{logit}(p_i) = \beta_0 + \beta_H \text{History}_i + \beta_P \text{Pop}_i \tag{6.11}$$

where p_i refers to the probability of classifying city i as conflictive, History refers to the conflict history variable and Pop is the population of the city.

The second set uses network centrality measures and constitutes the bulk of our research. As described in Section 1.4.3, its features include:

- Standard centrality measures, namely degree, betweenness, closeness and pageRank (and their weighted representations).

- Bridgeness measures, namely ethnic, international and community bridgeness (and their weighted representations).

In terms of the logistic model, these network models can be expressed as

$$\begin{aligned}\text{logit}(p_i) = \beta_0 + \beta_H \text{History}_i + \beta_P \text{Pop}_i + \beta_k k_i + \beta_B B_i + \beta_C C_i + \beta_{PR} PR_i + \\ \beta_{eB} eBridg_i + \beta_{iB} iBridg_i + \beta_{cB} + cBridg_i \quad ,\end{aligned} \tag{6.12}$$

for the case of unweighted centrality measures; or as

$$\begin{aligned}\text{logit}(p_i) = \beta_0 + \beta_H \text{History}_i + \beta_P \text{Pop}_i + \beta_s s_i + \beta_{B_W} B_{W\,i} + \beta_{C_W} C_{W\,i} + \beta_{PR_W} PR_{W\,i} + \\ \beta_{eB_W} eBridg_{W\,i} + \beta_{iB_W} iBridg_{W\,i} + \beta_{cB_W} + cBridg_{W\,i} \quad ,\end{aligned} \tag{6.13}$$

for weighted centrality measures, as described in Section 1.4.3.

Note that in many situations network features are very heterogeneous and sharply distributed in space, meaning that one city may have several orders of magnitude higher centrality than the nodes in its neighbourhood. For this reason, when deriving the final set of network features we apply a local smoothing filter. This consists on building a purely geographical network using Eq. 6.3 with $\mathcal{R} = 300km$ and, for every node, use the centrality average in the local subset of nodes consisting of the union of itself and its neighbourhood.

Data partition and cross validation

Random sampling cross-validation is not applicable in our case, given that we're dealing with time series with significant autocorrelation structures. For this reason, we partition our data into in-sample training sets and out-of-sample evaluating sets, using rolling forecasting cross-validation [229]. Models are trained using in-sample data and evaluated through data unseen by the model, emulating forecasting. As illustrated in Figure 6.6, we use a fixed-size rolling window to define data splits. Note that growing window size can be used to reflect growth in data availability, but in our case we retain a fixed window in order to lower computation costs. Furthermore, in the present analysis we use data from 4 consecutive years to train our models, which are then evaluated in the immediately subsequent year. Potentially, the training windows could have other sizes, and we could evaluate our models in longer horizons (i.e. more than one year in the future).

For every window in the roll we obtain a predictive performance metric. We average metrics from every window in order to obtain a single representative performance measure for our models. We end up selecting the model with those parameters that maximize the average performance metric.

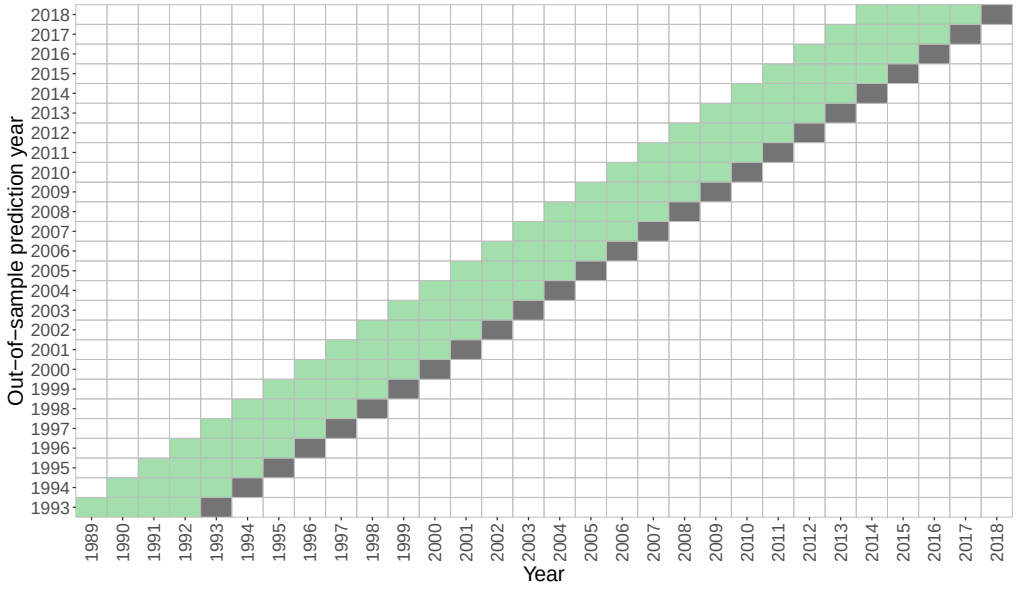


Figure 6.6: Illustration of our data partitioning method. Each row represents a different training/evaluating split, with green cells representing training years, grey cells evaluation years, and white cells years unused.

Performance metrics

There are several performance metrics that can be used for classification tasks. Here we give an overview of the most important of them, and argue which one is the most suitable for our particular task. At the most fundamental level, classification performance is usually measured through the concepts of True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). The terms *positive* or *negative* refer to the outcome of the binary dependent variable under study: in our case, a positive would refer to a city that has been declared as in conflict in a given year t and event threshold $\mathcal{T}_E$ (i.e.$N_E(i,t) > \mathcal{T}_E$), whereas a negative is a city declared peaceful under the same conditions. The terms *true* or *false* refer to whether our classification model produces a correct or incorrect assessment of an observation being positive or negative.

The simplest set of metrics using the aforementioned terms that underlie any classification task are the following [223]:

- Precision (Pr):

$$Pr = \frac{TP}{TP + FP} \tag{6.14}$$

- Recall, sensitivity or true positive rate (R):

$$R = \frac{TP}{TP + FN} \tag{6.15}$$

- Specificity or true negative rate (S):

$$S = \frac{TN}{TN + FP} \tag{6.16}$$

- False alarm or false positive rate (FPR):

$$FPR = 1 - S = \frac{FP}{TN + FP} \tag{6.17}$$

The predictive models used for classification typically yield a probability of an event being positive, and so all of the measures above are dependent upon the probability cutoff used for taking classification decision. Precision, recall or specificity can be used for single cutoff values, but typically they are more informative if their value is used in combination of all possible probability cutoffs, generating aggregated metrics. Two of the most well-established from such aggregated metrics are:

- Area Under Receiver-Operator Curve (AUROC): is based on the ROC curve, which presents the false positive rate FPR on the x-axis versus the recall R on the y-axis for all probability cutoffs.

- Area Under Precision-Recall Curve (AUPRC): is based on the PR curve, which presents the recall R on the x-axis versus the precision Pr on the y-axis for all probability cutoffs.

For a given model, the AUROC reflects a trade-off between the ability to classify positive outcomes correctly and the cost of generating false alarms. For this reason, AUROC rewards models that are good at classifying negative outcome events. This becomes problematic for highly imbalanced datasets such as ours, where the number of peaceful cities largely outnumbers the ones with conflict. Using AUROC in our analysis would favour the selection of models that are good at predicting peace, but we are much more interested in models that predict conflict. On the contrary, the AUPRC only takes into account positive outcomes, only favouring in our case those models that can accurately predict conflict. For all of this, we will only use AUPRC in the following analysis to assess model predictive performance.

Variable importance

The absolute value of the t-statistic is a common measure of variable importance for General Linear Models. It is simply obtained by dividing the absolute value of a predictor's linear coefficient β (see Eq. 6.10) by its estimated standard error, so intuitively

it expresses the magnitude of influence for each predictor on the outcome of the model. In order to have comparable coefficients that provide meaningful variable importance measures, it is important to normalize or standardize the predictors. Note that the absolute value of the t-statistic becomes less significant when there are correlation structures present amongst the predictors. GLM variable importance is scaled to percentage, with values between the minimum importance (0 score) and the maximum importance (100 score).

For the random forest model, we use permutation importance [225]. This is a two-step approach done for each predictor variable, where we capture the loss of predictive performance occurred when the internal data of the predictor under study is randomized and thus its connection with the response variable disappears. For every decision tree in the ensemble, in the first step we compute the predictive performance (AUPRC in our case) using the original predictors. In the second step, we sequentially randomize each predictor and compute the predictive performance in that scenario. For every predictor, the difference between the two predictive performances is used as a variable importance indicator for that tree. Permutation importance is finally obtained by averaging variable importance over all trees in the ensemble. Permutation importance is also scaled to percentage, with values between the minimum importance (0 score) and the maximum importance (100 score).

6.4 Results

6.4.1 Baseline models

We begin analysing the predictive performance of the logistic regression baseline model described in Eq. 6.11. As explained above, the baseline model contains city population data and past conflict history data. Figure 6.7 shows that keeping a history window of 1 year (i.e. using the events produced in the previous year as predictive feature) produces the optimal out-of-sample performance for the logistic regression baseline model. We also compute the baseline random forest using 1-year conflict history and population as features. As we can see in Table 6.1, baseline GLM performs significantly worst than baseline random forest (-11.31%).

6.4.2 Geographic network models

Here we study the geographic network model derived from Eq. 6.3. We explore the performance of this model in two steps of increasing complexity. The first consists on using unweighted networks, so that the gravity law in Eq. 6.1 is not at all used to derive network features. The second step consists on weighting the network using the

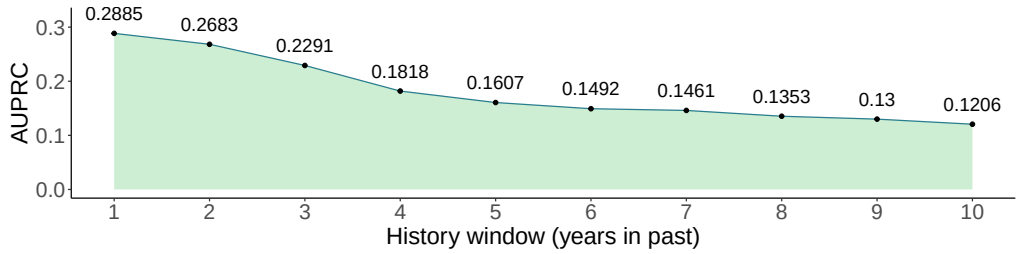


Figure 6.7: Grid search for the conflict history window size that maximizes out-of-sample AUPRC for the baseline GLM model.

Model	AUPRC	Δ_{RF}AUPRC
Baseline RF	0.3252366	-
Baseline GLM	0.2884589	-11.30796%

Table 6.1: Predictive performance (AUPRC) for the basline random forest model (Baseline RF) and the baseline logistic regression model (Baseline GLM). Δ_{RF}AUPRC is calculated as the difference between each model's AUPRC and the random forest baseline model.

gravity law. This allows us to differentiate the predictive performance arising purely from network effects, from that arising from the convolution of the network with a gravity law.

Unweighted network

Here we explore the unweighted geographic network model, described in terms of the logistic regression in Eq. 6.12. In this case there's only one network hyperparameter, namely the connectivity radius $\mathcal{R}$. As mentioned above, the logit GLM is much less computationally costly. For this reason, we use it to search for the hyperparemter value in the grid $\mathcal{R} \in [100, 1000]$km which maximizes out-of-sample prediction performance as measured by AUPRC. We use a step of 100km for our grid search, which provides enough resolution whilst keeping the computational costs tractable. In fact, for every value of $\mathcal{R}$ we need to derive a network and compute its centrality measures, which becomes more costly as $\mathcal{R}$ increases due to the increase in edge density. Furthermore, for every network, we train and test the model in Eq. 6.12 using every rolling window available as explained in Section 6.3.4.

As shown in Figure 6.8, we find a clear AUPRC-maximizing model at $\mathcal{R} = 300$km. We can see that in this case there is an increase in AUPRC of 18.63% with respect to the baseline model. The coefficients of the optimal model are shown in Table 6.2. Using such coefficients we can see which features have a positive effect on violence-propensity (history, population, betweenness, closeness, ethnic and international bridgeness) and

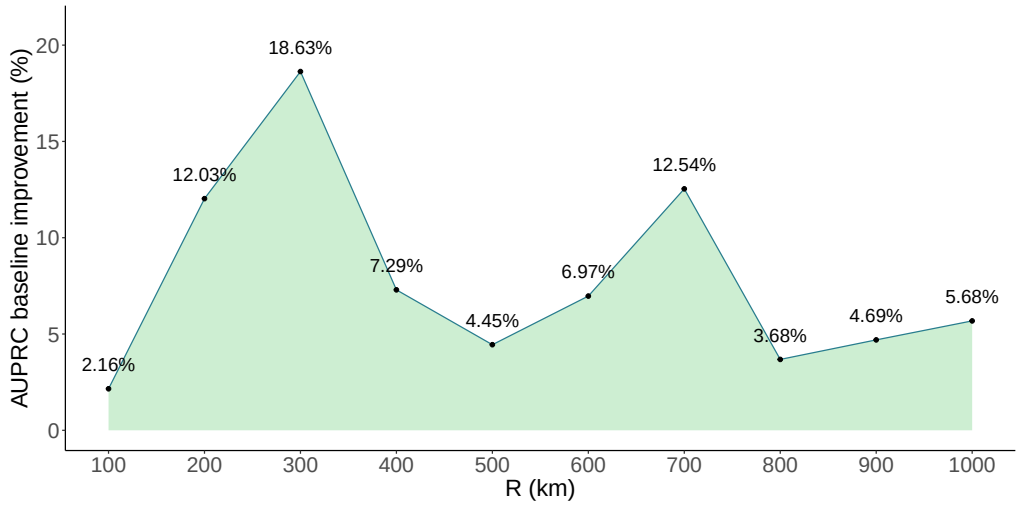


Figure 6.8: Grid search for the unweighted geographic network model. We show the AUPRC-improvement of the model in Eq. 6.12 with respect to the logistic regression baseline model for different values of the connectivity radius $\mathcal{R}$.

which have a negative effect (degree, pageRank and community bridgeness). As described above, we can also use this coefficients to derived variable importance: Figure 6.9 shows that conflict history is the largest contributor to the model, but some network metrics such as betweenness and international bridgeness also have a significant contribution in predicting conflict patterns through logistic regression.

Once the optimal network hyperparameter has been located, we can apply the more computationally costly random forest model for this specific network at $\mathcal{R} = 300$km using the same features as in Eq. 6.12. Table 6.3 summarises the predictive performance of the best logistic regression and random forest models at $\mathcal{R} = 300$km, and shows the increase or decrease of performance of each of them with respect to the random forest baseline model. The optimised network-based GLM performs better than the random forest baseline (5.21%). However, the optimised network-based random forest performs significantly better than the GLM model and than the random forest baseline (12.62%).

Finally, we use the permutation importance method described in 6.3.4 for the optimal random forest at $\mathcal{R} = 300$km. The results are shown in Figure 6.10. In this case, the variables with the highest impact on the predictive performance of the random forest are conflict history, degree, betweenness and ethnic bridgeness, although the rest of network variables also seem to have a non-negligible effect.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.2850	0.2101	-34.68	0.0000
Lag_is_Conflict	6.7748	0.1772	38.23	0.0000
pop	4.1189	1.2134	3.39	0.0007
degree	-8.6870	3.0214	-2.88	0.0040
betweenness	3.1855	0.3818	8.34	0.0000
closeness	2.2491	1.9656	1.14	0.2525
pageRank	-6.2189	2.2693	-2.74	0.0061
eBridg	4.4427	1.6443	2.70	0.0069
iBridg	6.0354	1.4321	4.21	0.0000
cBridg	-1.8100	1.1187	-1.62	0.1057

Table 6.2: Model coefficients for the logistic regression GLM maximising AUPRC for the optimal unweighted geographic network with $\mathcal{R} = 300$ km.

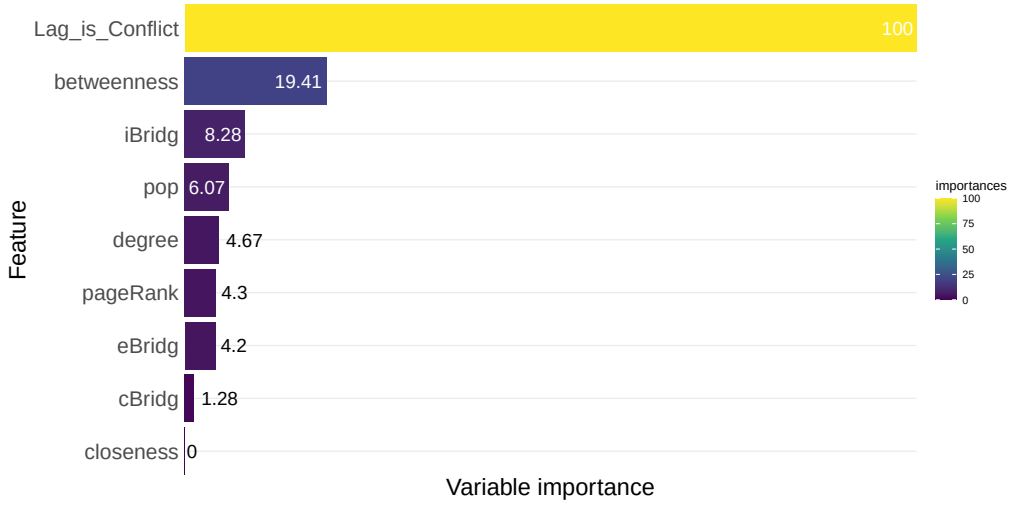


Figure 6.9: Variable importance plot for the out-of-sample AUPRC-maximising logistic model (Eq. 6.12) in the weighted geographic network with $\mathcal{R} = 300$ km. As described in Section 6.3.4, we use the GLM t-statistic of each coefficient as a measure of importance.

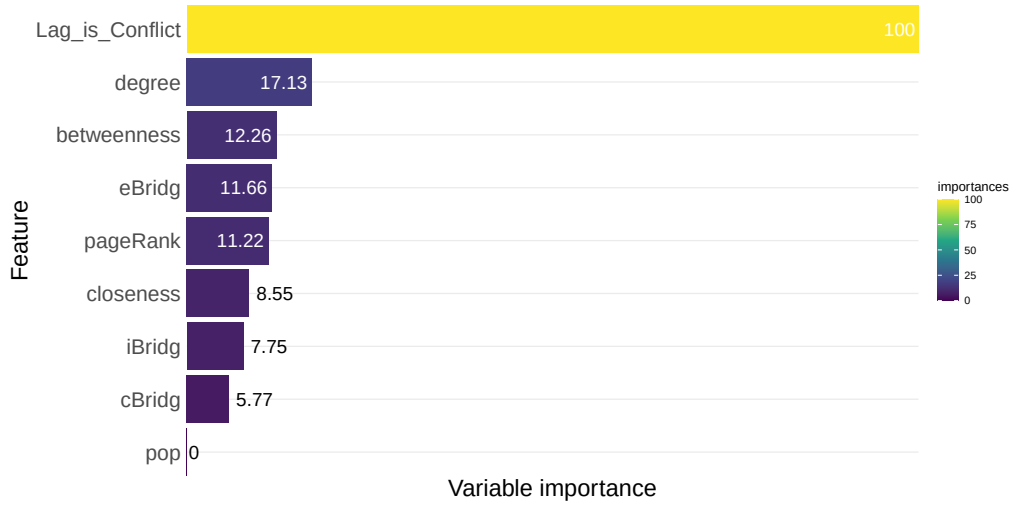


Figure 6.10: Variable importance plot for the out-of-sample AUPRC-maximising random forest model in the weighted geographic network with $\mathcal{R} = 300$ km. As described in Section 6.3.4, we use AUPRC-decrease through permutation as a measure of importance of each variable.

Model	AUPRC	Δ_{RF}AUPRC	Δ_{GLM}AUPRC
uwRF (R=300km)	0.3662935	12.62371%	26.98291%
uwGLM (R=300km)	0.3421942	5.213935%	18.62841%

Table 6.3: Predictive performance (AUPRC) for the optimal unweighted-network random forest (uwRF) at $\mathcal{R} = 300$ km, and the optimal unweighted-network logistic regression GLM (uwGLM) at $\mathcal{R} = 300$ km. Δ_{RF}AUPRC is calculated as the difference between each model's AUPRC and the random forest baseline model, whereas Δ_{GLM}AUPRC is the difference with the logistic regression baseline model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.4871	0.1595	-46.95	0.0000
Lag_is_Conflict	6.6975	0.1849	36.23	0.0000
pop	3.4367	1.4745	2.33	0.0198
degree_W	-3662.3440	1935.6889	-1.89	0.0585
betweenness_W	3.8358	0.3258	11.77	0.0000
closeness_W	1436.2946	3884.5762	0.37	0.7116
pageRank_W	-8.4657	1.5455	-5.48	0.0000
eBridg_W	-0.6055	0.8263	-0.73	0.4636
iBridg_W	6.0985	1.1433	5.33	0.0000
cBridg_W	-1.1317	0.6002	-1.89	0.0594

Table 6.4: Model coefficients for the logistic regression GLM maximising AUPRC for the optimal weighted geographic network with $\mathcal{R} = 700$ km, $\alpha = 0.8$ and $\gamma = 3$.

Weighted network

Here we explore the weighted geographic network model, described in terms of the logistic regression in Eq. 6.13. In this case there are three hyperparameters, namely the connectivity radius $\mathcal{R}$, and the gravity population exponent α, and gravity distance exponent γ. Again, we use logistic regression to search for the out-of-sample AUPRC-maximising hyperparemter values, which in this case are constrained to the grid $\mathcal{R} \in [100, 1000]$ km, $\alpha \in [0, 1]$ and $\gamma \in [0, 5]$.

As shown in Figure 6.11, in this case we identify an AUPRC-maximizing model at $\mathcal{R} = 700$ km, $\alpha = 0.8$ and $\gamma = 3$. We can see that in this case there is an increase in AUPRC of 20.54% with respect to the baseline model. The coefficients of the optimal model are shown in Table 6.4. Note how in this case, some coefficients (weighted degree and closeness) are much larger in absolute value than the rest. This is due to the fact that gravity weights at $\alpha = 0.8$ and $\gamma = 3$ produce large imbalances amongst cities for these centrality measures. Using such coefficients we can see which features have a positive effect on violence-propensity (history, population, weighted betweenness, ethnic and international bridgeness) and which have a negative effect (weighted degree,closeness, pageRank and community bridgeness). As before, we can use these coefficients to derived variable importance: Figure 6.12 shows that conflict history is again the largest contributor to the model, but weighted betweenness, pageRank and international bridgeness also have a significant contribution in predicting conflict patterns through logistic regression.

Having derived the optimal hyperparameters, we can apply the random forest model for this specific network using the same features as in Eq. 6.13. Table 6.5 summarises the predictive performance of the best logistic regression and random forest models at

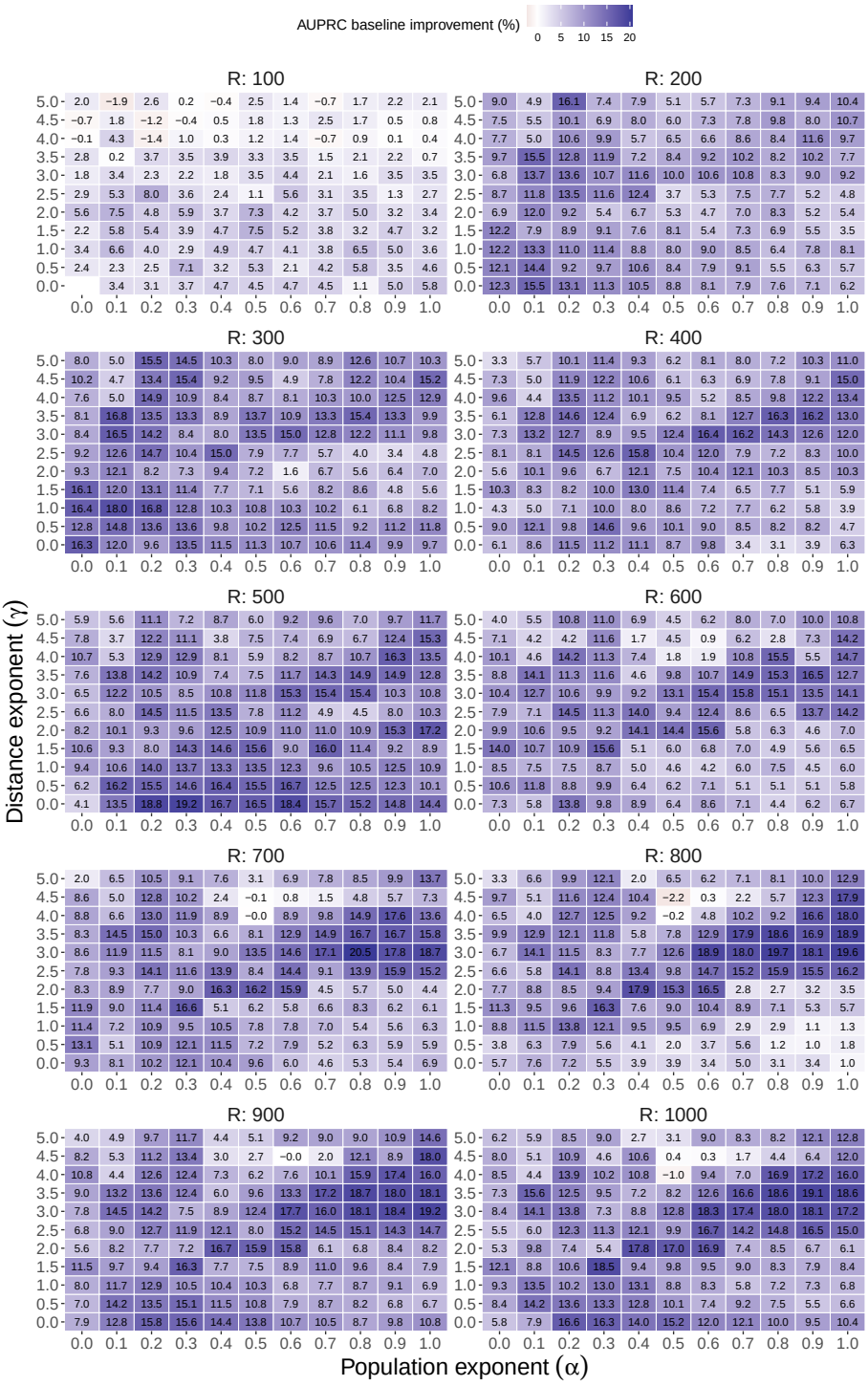


Figure 6.11: Grid search for the weighted geographic network model. We show the AUPRC-improvement of the model in Eq. 6.13 with respect to the logistic regression baseline model for different values of the connectivity radius $\mathcal{R}$ (panels), population exponent α (x-axis) and distance exponent γ (y-axis).

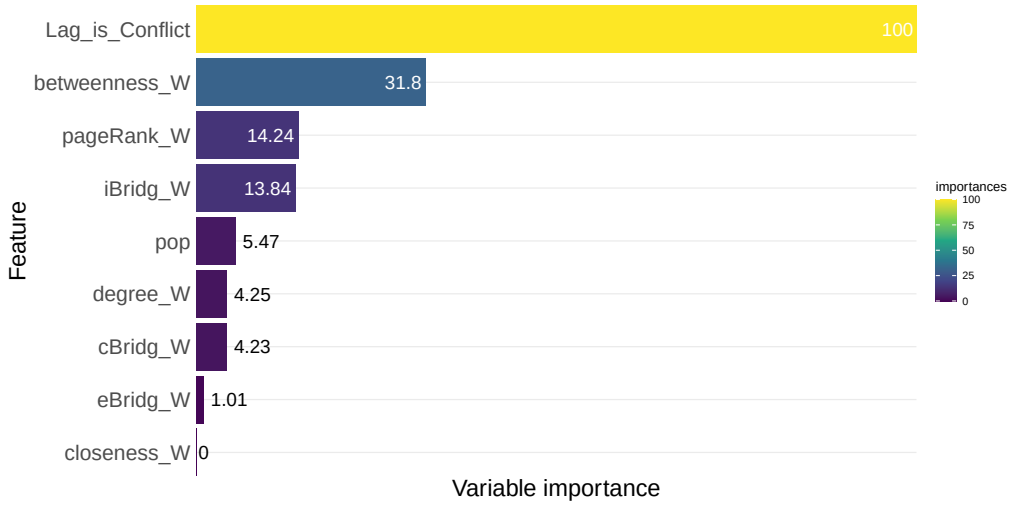


Figure 6.12: Variable importance plot for the out-of-sample AUPRC-maximising logistic model (Eq. 6.13) in the weighted geographic network with $\mathcal{R} = 700$ km, $\alpha = 0.8$ and $\gamma = 3$. As described in Section 6.3.4, we use the GLM t-statistic of each coefficient as a measure of importance.

$\mathcal{R} = 700$ km, $\alpha = 0.8$ and $\gamma = 3$, and shows the increase or decrease of performance of each of them with respect to the random forest baseline model. The optimised weighted network GLM performs better than the random forest baseline (6.91%). However, the optimised weighted network random forest performs again significantly better than the GLM model and than the random forest baseline (15.43%). Finally, we also use the permutation importance method described in 6.3.4 for the optimal weighted network random forest. The results in Figure 6.13 show that all weighted centrality measures (except community bridgeness) have an important effect on predictive performance.

Model	AUPRC	Δ_{RF}AUPRC	Δ_{GLM}AUPRC
wRF (R=700,$\alpha = 0.8, \gamma = 3$)	0.3754271	15.43199%	30.14924%
wGLM (R=700,$\alpha = 0.8, \gamma = 3$)	0.3477251	6.914509%	20.54581%

Table 6.5: Predictive performance (AUPRC) for the optimal weighted-network random forest (wRF) at $\mathcal{R} = 700$ km, $\alpha = 0.8$ and $\gamma = 3$, and the optimal weighted-network logistic regression GLM (wGLM) at $\mathcal{R} = 700$ km, $\alpha = 0.8$ and $\gamma = 3$. Δ_{RF}AUPRC is calculated as the difference between each model's AUPRC and the random forest baseline model, whereas Δ_{GLM}AUPRC is the difference with the logistic regression baseline model.

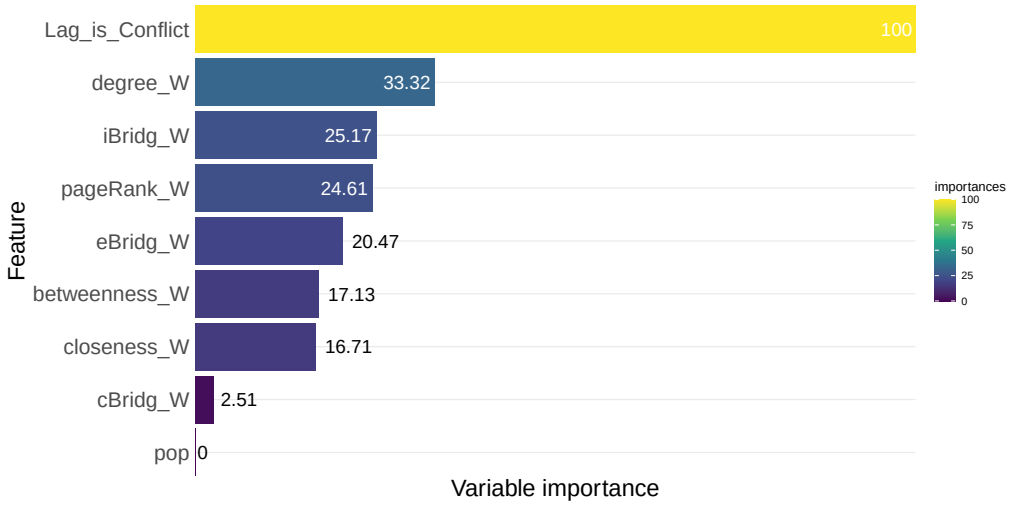


Figure 6.13: Variable importance plot for the out-of-sample AUPRC-maximising random forest model in the weighted geographic network with $\mathcal{R} = 700$ km, $\alpha = 0.8$ and $\gamma = 3$. As described in Section 6.3.4, we use AUPRC-decrease through permutation as a measure of importance of each variable.

6.4.3 Gravity network models

We now move the analysis towards the gravity network model derived from Eq. 6.4. As described before, we use the number of edges $\mathcal{E}_g$ as hyperparameter, as well as the gravity law exponents α and γ. Given that in this case the creation of edges in the network and their weights are dependent on the same gravity law, we directly proceed to study the weighted network model. As before, we use logistic regression to search for the out-of-sample AUPRC-maximising hyperparemter values in the grid $\mathcal{E}_g \in [7000, 130000]$, $\alpha \in [0, 1]$ and $\gamma \in [0, 5]$.

As shown in Figure 6.14, we find an AUPRC-maximizing model at $\mathcal{E}_g = 49090$, $\alpha = 0.8$ and $\gamma = 3$. We can see that in this case there is an increase in AUPRC of 24.07% with respect to the baseline model. The coefficients of the optimal model are shown in Table 6.6, and the variable importance metrics are shown in Figure 6.15. Weighted betweenness, community and international bridgeness and pageRank have significant contribution in predicting conflict patterns through logistic regression.

Again, we fit a random forest model to the optimal set of hyperparameters. Table 6.5 summarises the predictive performance of the best logistic regression and random forest models at $\mathcal{E}_g = 49090$, $\alpha = 0.8$ and $\gamma = 3$. Interestingly, we can see that in this case there is a much lower performance difference between the GLM and the random forest. The random forest performs slightly worst in the gravity network than in the

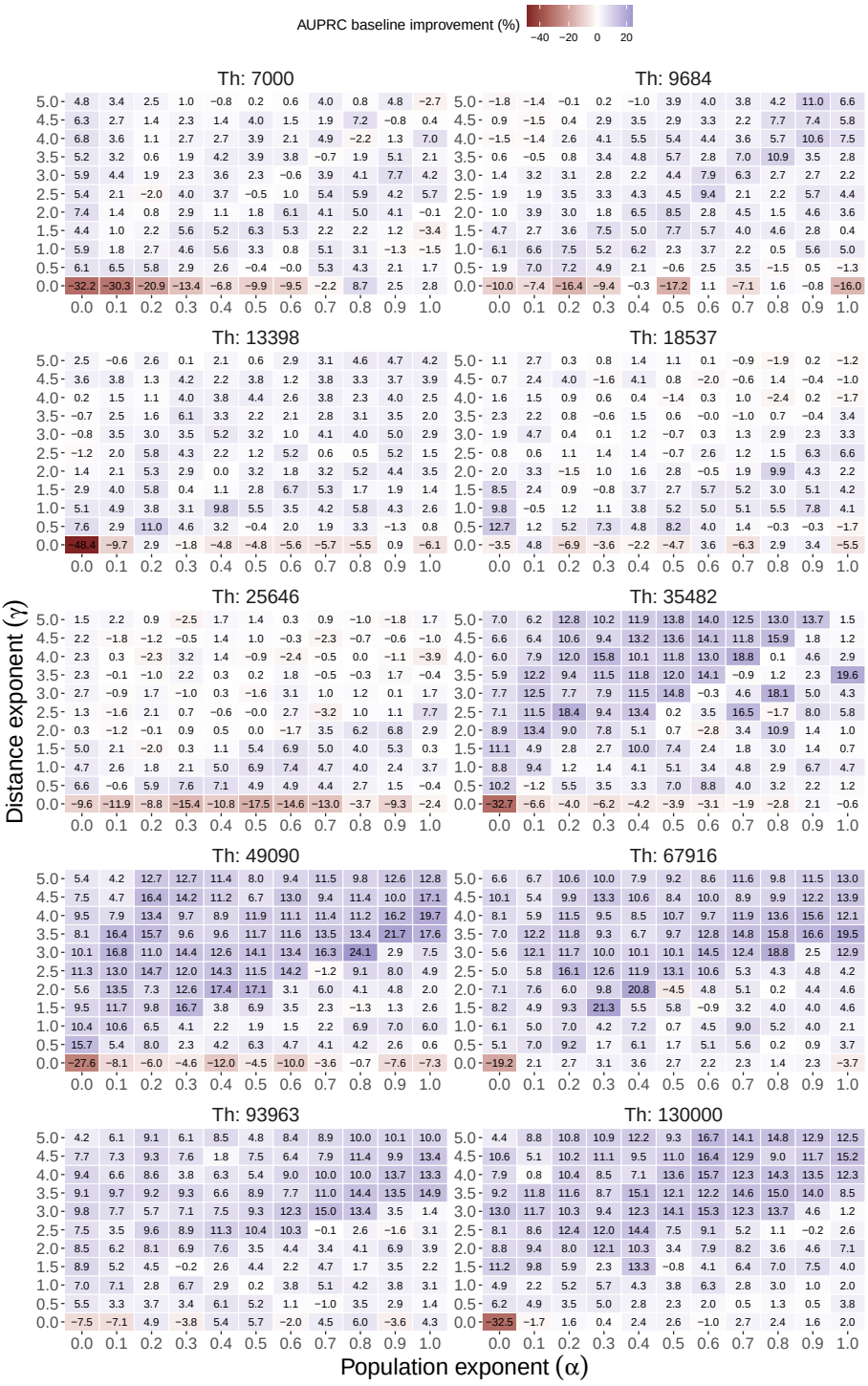


Figure 6.14: Grid search for the weighted gravity network model. We show the AUPRC-improvement of the model in Eq. 6.12 with respect to the logistic regression baseline model for different values of the connectivity radius $\mathcal{R}$.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.3804	0.1177	-62.72	0.0000
Lag_is_Conflict	6.5721	0.1883	34.90	0.0000
pop	4.6970	2.3909	1.96	0.0495
degree_W	-202.9371	168.3299	-1.21	0.2280
closeness_W	-2032.3356	3497.5063	-0.58	0.5612
betweenness_W	9.4505	0.9476	9.97	0.0000
pageRank_W	-25.7742	5.8431	-4.41	0.0000
eBridg_W	1.1110	3.3249	0.33	0.7383
iBridg_W	12.0320	2.0374	5.91	0.0000
cBridg_W	-13.1164	2.1365	-6.14	0.0000

Table 6.6: Model coefficients for the logistic regression GLM maximising AUPRC in the weighted gravity network with $\mathcal{E}_g = 49090$, $\alpha = 0.8$ and $\gamma = 3$.

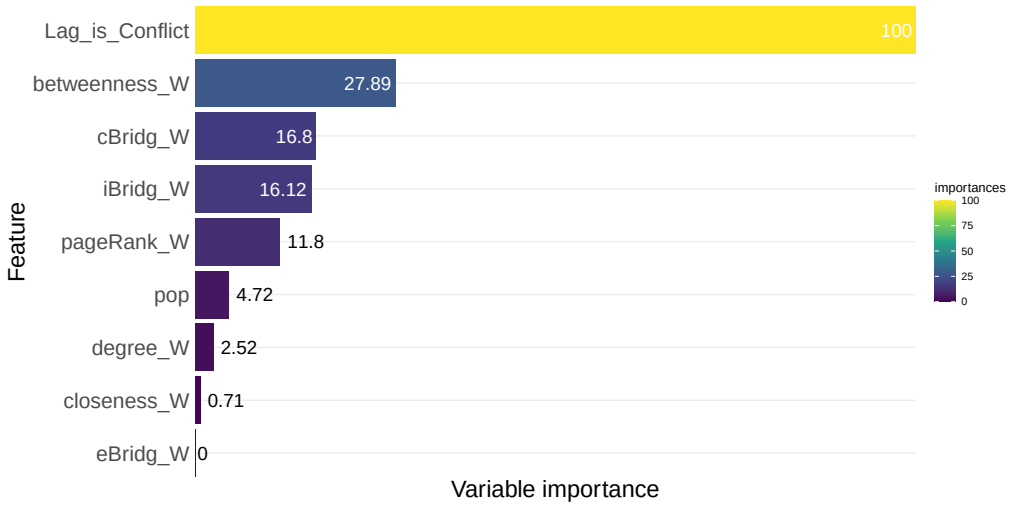


Figure 6.15: Variable importance plot for the out-of-sample AUPRC-maximising logistic model (Eq. 6.13) in the weighted gravity network with $\mathcal{E}_g = 49090$, $\alpha = 0.8$ and $\gamma = 3$. As described in Section 6.3.4, we use the GLM t-statistic of each coefficient as a measure of importance.

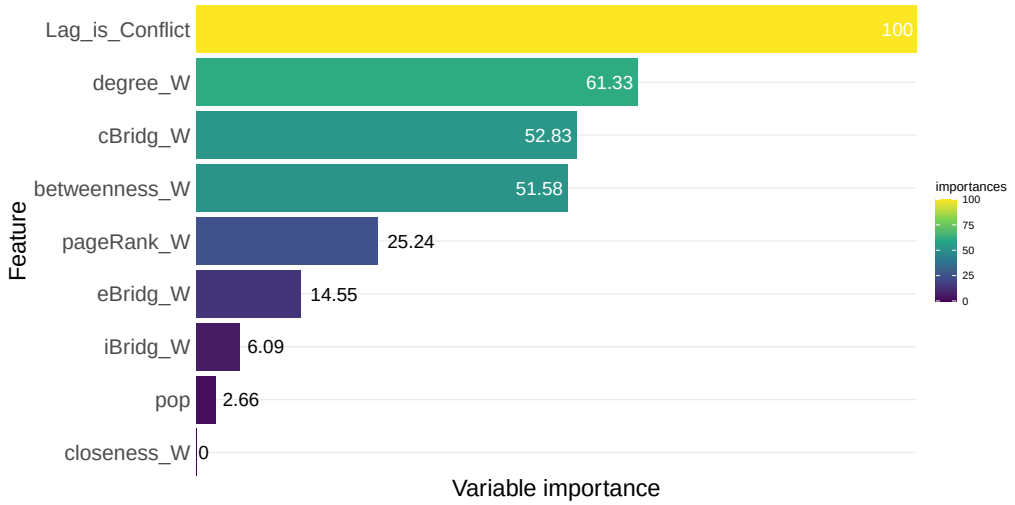


Figure 6.16: Variable importance plot for the out-of-sample AUPRC-maximising random forest model in the weighted gravity network with $\mathcal{E}_g = 49090$, $\alpha = 0.8$ and $\gamma = 3$. As described in Section 6.3.4, we use AUPRC-decrease through permutation as a measure of importance of each variable.

weighted geographic network (12.24% over RF baseline). On the contrary, the GLM performs significantly better in the gravity network than in the weighted geographic network (10.04% over RF baseline). Finally, we use permutation importance again for the optimal gravity network random forest. The results are shown in Figure 6.16. For this gravity-network case, all weighted centrality measures (except closeness) have higher importance scores than in the geographic networks studied before.

Model	AUPRC	Δ_{RF}AUPRC	Δ_{GLM}AUPRC
wRF ($E_g = 49090$,$\alpha = 0.8, \gamma = 3$)	0.3650467	12.24034%	26.55067%
wGLM ($E_g = 49090$,$\alpha = 0.8, \gamma = 3$)	0.357905	10.04451%	24.07488%

Table 6.7: Predictive performance (AUPRC) for the optimal weighted-network random forest (wRF) at $\mathcal{E}_g = 49090$, $\alpha = 0.8$ and $\gamma = 3$, and the optimal weighted-network logistic regression GLM (wGLM) at $\mathcal{E}_g = 49090$, $\alpha = 0.8$ and $\gamma = 3$. Δ_{RF}AUPRC is calculated as the difference between each model's AUPRC and the random forest baseline model, whereas Δ_{GLM}AUPRC is the difference with the logistic regression baseline model.

6.5 Discussion

This chapter presents evidence and results for a disaggregated network-driven forecasting system of political violence events and armed conflict. The system is based on the construction of networks that connect a set of nodes representing major cities around the world. Connections in these networks abstractly represent interactions amongst cities, which may take the form of commercial exchange, population commutes, immigration, cultural relations or infrastructure such as roads and flight routes, amongst other factors. Such networks are then used to derive centrality measures that are attributable as features of each city in our dataset. These features allow to construct predictive models, which are then evaluated out-of-sample using performance metrics suitable for our classification task, namely predicting conflict prevalence amongst the cities under study.

Our predictive models are based on two fundamental hypothesis, namely: (*i*) that complex network analysis is a useful mathematical tool for representing, at the city level, socio-geographical data relevant for conflict prediction; (*ii*) that meaningful centrality measures can be derived from such complex networks, and used as statistical features with significant predictive performance. Below we discuss to which extent such hypothesis are validated from our analysis, and what steps can be taken in the future to refine and generalise the framework exposed above.

Regarding the first hypothesis, we have shown in Section 6.3.2 how interaction networks can be inferred solely using data on the geolocation of cities and their population, using a combination of connection rules (see Eqs. 6.3 and 6.4) and the gravity law (see Eq. 6.1). Importantly, the topology of these networks depends on the value of three hyperparameters: one of them governs the density of edges in the network, while the other two are associated with the gravity law and control the distance cost and population benefit terms respectively. This network framework represents a methodological step forward for conflict prediction that bridges together spatial disaggregation on the one hand, and systemic generalisation of dyadic analysis on the other.

With reference to the second hypothesis, we have also shown how standard centrality measures can be used to rank cities by importance under various criteria (see Section 6.3.3). Such criteria may refer to local features such as the number of interactions (related to the presence of populous settlements in the proximity of a city) or global features such as how close to important settlements one city is, or how strategic it is in terms of efficient shortest-path flows. In addition, we have proposed three novel centrality measures based on the concept of bridgeness. One of them uses topological communities derived from the Stochastic Block Model (see Section 3.2), and the other two use metadata-driven communities from countries and ethnic groups respectively.

In order to test the predictive performance of these centrality measures we have used baseline models to quantify the relative improvement provided by network features. Such

baseline models use conflict history and population as only features. Regarding conflict history, we have shown in Figure 6.7 how using the history of the previous year brings the highest predictive performance. Table 6.1 summarises the performance of both a logistic regression (0.288 AUPRC) and a random forest (0.325 AUPRC) baseline models.

It is not easy to put these performance values in context, given that the present work is (to the best of the author's knowledge) the first predictive study done at the city level of analysis. However, we can use recent scores published by the ViEWS project [228], which is one of the most advanced conflict prediction systems to date. ViEWS uses two levels of analysis: at country level, their baseline model (which only considers conflict history, but not population) scores 0.675 AUPRC; at grid level, which studies conflict patterns at quadratic grid cells that cover a resolution of 0.5 x 0.5 decimal degrees, their baseline model scores 0.225 AUPRC. The difference in baseline scores comes intuitively from the fact that the coarser the level of analysis, the easier it is for a model to classify conflictive zones. Thus it is reasonable to see that our city-network baseline, although slightly coarser than the grid, scores similarly to it.

Beyond the particular baseline performance values, an interesting way to compare our models with others (e.g. ViEWS) consists on reporting performance improvements relative to their baseline. For instance, the ViEWS model, which combines a sophisticated ensemble of constituent models using logistic regression and random forest based on a large variety of predictors (related to natural geography, social geography, economy and social unrest), reports a $\sim 2\%$ improve at country level and a $\sim 23\%$ improve at grid level, relative to their respective baselines.

Focusing on our study, Table 6.3 shows how our simplest unweighted geographical network model (see Section 6.4.2) produces a 18.62% improvement when using logistic regression, and a 12.62% improvement when using random forests. When adding gravity weights to such geographic network, these improvements scale up to 20.54% for logistic regression and 15.43% for random forest. For the case of gravity-based networks, the baseline improvements are 24.07% for logistic regression and 12.24% for random forest.

One conclusion to derive from the results above is that geolocation alone is one of the most important factors driving the predictive performance of our networks. In fact, the unweighted geographical network only uses the location of cities to make predictions and still captures most predictive improvements — the weighted geographical network only scores 1.75% (logistic regression) and 2.46% (random forest) higher than the unweighted in absolute AUPRC terms. Note, however, that the gravity law remains an important factor influencing conflict prediction. In the gravity-based network model, for instance, we classify conflictive cities using logistic regression up to 4.64% better than the unweighted geographic network in absolute AUPRC performance.

Another interesting conclusion, drawn by looking at the regression coefficients of the logistic models, is that some network features seem to have a consistently positive

effect on conflict whereas some others have a negative impact. Considering only GLM coefficients with high significance (i.e. the p-value $\Pr(>|z|) < 0.05$) in Tables 6.2, 6.4 and 6.6, we can see that degree and pageRank (and their weighted versions) have negative coefficients, meaning that cities with higher degree and pageRank centrality (i.e. cities with high connectivity and connected to important cities) will be less prone to conflict. On the contrary, given their positive coefficients, betweenness, international bridgeness and ethnic bridgeness will tend to increase the probability of observing conflict in a city. This suggests that being located in geostrategic crossroads for global flows (e.g. for international trade or migration) and being connected to a larger variety of borders and ethnic groups are risk factors for cities to experiment conflict. In fact, our two bridgeness measures can be related to some extent with existent predictors in social geography: distance to and number of borders have been regarded as a risk factor since the early days of the field [176], [230], whereas measures related to ethnic diversity have also been frequently studied [192], [231]. Also note that community bridgeness is not significant enough in any of the logistic regression t-tests, so no judgement on its net impact can be made.

Besides the directionality of each predictor, we can also conclude that some network features are more important than others in terms of predictive performance. The most important centralities are betweenness, degree, pageRank, international and ethnic bridgeness, although their exact ranking depend on the particular network and statistical model used. In most cases, however, closeness and community bridgeness seem to have smaller contributions.

In terms of future extensions of this framework, there are four main branches with significant room for improvement. The first branch relates to composability and levels of analysis: our city-level analysis could be simply projected down to grid levels such as those used by ViEWS [228] or aggregated up to the country level, providing a path towards integration and benchmarking with existing conflict prediction systems. The second branch relates to the use of ensemble prediction methods. We have provided a variety of models: unweighted, weighted, geographic and gravity-based networks; logistic regression, and random forest statistical models. As a consequence, we have produced different sets of predictions, which could be calibrated and combined using ensemble Bayesian model averaging techniques [232], in order to extract the best contributions from each constituent model. The third branch concerns network construction: using real-world data (e.g. for global infrastructure or global trading patterns) we could reconstruct our networks directly from real world observations, instead of using gravity law inference. The fourth branch refers to the ability of the model to predict new conflicts: currently we are measuring predictive performance for both new and recurrent conflicts, but the UCDP-GED dataset [191] contains information distinguishing both types of conflict.

Chapter 7

Outlook and Conclusions

The first part of this thesis opens with three chapters containing theoretical methods that investigate the emergence of unexpected dynamical behaviour as a result of an interplay with structural properties in complex systems. Chapter 2 is set among the framework of opinion dynamics, where we have shown a parsimonious mechanisms giving rise to the emergence of leadership and herding behaviour in a population of interacting agents of a voter model. These mechanisms consist on a strong separation of activity time-scales coupled with a hierarchical organization of the influence exerted by some agents on others. Herding behaviour and leadership is expressed dynamically by large groups of the population quickly adopting the opinion of a small minority of leading agents, producing observable sharp shifts in global opinion that are much more pronounced than the typical diffusive fluctuations observed in voter models.

Importantly, our results are very general and apply on a wide variety of real-world social systems. For instance, we argue that any social network containing core-periphery structures can potentially express self-organised herding behaviour if the right circumstances apply. These circumstances basically consist on a coupling between topological features (e.g. the number of connections) and the popularity and activity of agents. These are assumptions likely to occur in systems such as the stock market or online social networks, where feedback loops exist reinforcing the coupling between connectivity and social influence. We conjecture that herding behaviour could shed light in emerging social phenomena such as stock market crashes or rapid opinion polarisation in social networks. Interesting empirical studies of our results could be conducted using online social networks such as Twitter, where topology can be easily inferred by friendship, like and retweet structures, and the opinion state of nodes can be retrieved using sentiment analysis.

Communities are an essential feature present in both natural and human-made complex networks. In Chapter 3 we have reviewed one of the most successful generative models for modularity, the Stochastic Block Model (SBM), making use of the Bayesian approach to community detection. The SBM works well for defining and finding communities in networks, but does not explain the different connectivity patterns amongst modules. We have shown how bridgeness centrality is a useful measure to determine how important a node is, in terms of its capacity to mediate and integrate different communities. Combining both ideas, we have introduced the SBM with bridgeness (SBMb), a generative model that allows us to build networks with arbitrary community structures and arbitrary bridgeness distributions.

In fact, the development of the SBMb has been an instrumental step to test hypothesis on the interplay between bridgeness and functional behaviour. In particular, we have asked the question of whether the position of a node with respect to the community-interfaces present in the network has an effect on its dynamical behaviour. In other words, we have looked for a universal functional effect induced by bridgeness. For this, we have used the Potts model of spins and the Kuramoto model of oscillators. These two models have different applications and descriptions, but both reveal a very similar conclusion with regards to bridgeness. Namely, that bridgeness induces what we could call dynamical disorder: in the case of Potts, bridgeness prevents spins to settle down to a single state but instead keep on flipping ad infinitum; in the case of Kuramoto, bridgeness prevents oscillators to synchronise normally. Additionally, we have shown how bridgeness induces special patterns in the Laplacian matrix of the network, producing localisation phenomena, where nodes with similar bridgeness tend to exhibit similar Laplacian eigenvector components. This observation adds generality to our results, given that the shape of the Laplacian matrix has important effects on the dynamical processes on networks. Further work should be done to illustrate the effects of bridgeness on other paradigmatic dynamical processes such as epidemic spreading models.

These observations show a clear interplay between topology and dynamics in modular networks. We use such interplay to define the concept of dynamical centrality. The idea is that when we know how a topological measure (in our case bridgeness) relates to a particular dynamical behaviour (in our case dynamical disorder in spins or oscillators), we can use functional observations to infer topological centrality even when topology itself cannot be observed. Further work could be done to understand to which extent dynamical centrality can be extrapolated to other interplaying factors between topology and dynamics. We finish Chapter 3 illustrating how bridgeness and dynamical centrality are all useful measures to dismantle both synthetic and real-world modular networks, even when the underlying topology is unknown. Further empirical tests on the dismantling performance of dynamical centrality would be useful to confirm our results.

In Chapter 4, we have moved our analysis towards weighted networks. In particular, we have asked the question of how does uncertainty in weights affect critical onsets of phase transitions in dynamical systems occurring on such fluctuating networks. Uncertainty in weights is pervasive in many network settings, either due to measurement error or due to some intrinsic behaviour that generates fluctuations in interaction strength amongst nodes. We have presented a mathematical framework that can analytically propagate uncertainty from weights to what we call critical range, which is the uncertainty in the critical threshold.

We have used such analytical framework to study how uncertainty propagation is affected by network structure. In particular, we have focused in the effects of degree heterogeneity. We have found that scale-free networks with exponent $\gamma \simeq 3$ maximise noise amplification, a result which may shed light to existing evolutionary arguments on the prevalence of scale-free networks with such exponent in the real world. In fact, it can be argued that having a wider critical range provides a network with larger adaptability in that, by producing very small changes in the weights of its constituents, the collective behaviour of the network can change dramatically in response to evolving environmental conditions.

The second part of the thesis contains two research applications of network analysis to real-world systems. We expose the first of such data studies in Chapter 5. The experimental setting is that of the rail network of London and its surrounding area, and our general aim has been to show how the interplay between topological features of this infrastructure and the human mobility patterns occurring on top of it affect the performance of this particular transport network. We have proposed to first measure the stability and robustness of the network from a theoretical point of view, and then compare our measures with empirical data on the performance of train operators on such network.

We have approached the measurement of stability and robustness drawing novel parallels between ecological networks and rail networks. In particular, we have built on well-known results from ecology regarding the abundance of feedback loops and its relation to lack of stability. In fact, by studying the distribution of feedback loops in our rail network, we have found that their abundance is related to lower performance metrics. Our results could be further generalised using larger datasets concerning entire countries, and other network metrics could be used to asses theoretically the stability and robustness of such networks. Larger datasets could also help moving our analysis from correlations to out-of-sample predictions, increasing the impact for policy-making and infrastructure planning of our results.

Finally, Chapter 6 presents perhaps the most ambitious results of this thesis. Here we have developed a theory of armed conflict based on spatial networks of urban settlements interacting across the globe. In its essence this is a theory that studies, at the city level of analysis, the effect of network-driven features of social geography on the likelihood of developing political violence. In methodological terms, we have proposed two network models that can be used to infer the patterns of interactions of cities around the world: one of them uses purely geographical proximity arguments to derive the connectivity of the network, whereas the other uses the social gravity law for the same purpose. From these networks, we have derived several centrality measures that inform us on how geostrategic each city is under different graph-theoretical criteria. It is important to note that, beyond standard centrality measures, we use a set of bridgeness-related measures that are intended to extrapolate the results of Chapter 3 to this particular case. After presenting the theory for deriving such network features, we have proposed a predictive framework to test their forecasting performance using one of the most comprehensive datasets containing armed conflict events occurred globally in the last thirty years. In order to isolate the predictive power of the network approach, we have used autoregressive baseline models to benchmark against full statistical models that use our centrality measures as features.

Out-of-sample predictions show a very significant increase deriving from the inclusion of network information in conflict forecasting models. We discuss several extensions of the statistical modelling framework and justify how our current work could be easily extended to higher or lower levels of analysis such as the country-level or the spatial grid-level. These extensions would make our model compatible with existing sophisticated bayesian ensemble prediction systems that are currently being used to inform policy-making with regards to international relations and peace-keeping. Further extensions include building a city-network that changes with real-world data dynamically, which would allow studying Granger-causality in relation to conflict events. All in all, we conjecture that in the future geographical network analysis can be an important tool to study the stability of the international relations system, perhaps developing into urban-planning methodologies at the global scale.

The work presented here takes us, at best, a very small step closer to a mathematical understanding on how some social systems work. An important collateral outcome of the thesis, however, is the illustration that the methods of network science, combined with the extraordinary availability of human-generated data, are bringing the social sciences to a new paradigm of understanding and applicability. By using statistical mechanics and complex networks, our results (especially from Chapters 3 and 4) enjoy greater generality because they can be easily interpreted and applied in many different scientific areas. We

conclude remarking that in the subjective opinion of the author, the universality and interdisciplinary mindset required for the study of complex networks, altogether build one of the most rewarding and exciting experiences a scientist can aspire to today.

Bibliography

[1] G. Mosquera-Doñate and M. Boguñá, "Follow the leader: Herding behavior in heterogeneous populations," *Physical Review E*, vol. 91, 2015.

[2] L. Arola-Fernández, G. Mosquera-Doñate, B. Steinegger, and A. Arenas, "Uncertainty propagation in complex networks: From noisy links to critical properties," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, 023 129, 2020.

[3] A. Pagani, G. Mosquera, A. Alturki, S. Johnson, S. Jarvis, A. Wilson, W. Guo, and L. Varga, "Resilience or robustness: Identifying topological vulnerabilities in rail networks," *Royal Society Open Science*, vol. 6, 181 301, 2019.

[4] W. Outhwaite, *The Blackwell Dictionary of Modern Social Thought*. Wiley, 2003.

[5] P. Arnopoulos, *Sociophysics: Cosmos and Chaos in Nature and Culture*. Nova Science Publishers, Inc., 1993.

[6] S. Galam, *Sociophysics: A Physicist's Modeling of Psycho-Political Phenomena*. Springer, New York Dordrecht Heidelberg London, 2012.

[7] C. Castellano, S. Fortunato, and V. Loreto, "Statistical Physics of Social Dynamics," *Reviews of Modern Physics*, vol. 81, 591–646, 2009.

[8] W. Weidlich, *Sociodynamics: A Systematic Approach to Mathematical Modelling in the Social Sciences*. Dover Publications, 2006.

[9] R. Axelrod, "Chapter 33: Agent-based Modeling as a Bridge Between Disciplines," in *Handbook of Computational Economics*, vol. 2, Elsevier, 2006, 1565–1584.

[10] R. Albert and A. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, 47–97, 2002.

[11] P. J. Carrington and J. Scott, "Introduction," in *The SAGE Handbook of Social Network Analysis*. SAGE Publications Ltd, 2011, 1–8.

[12] M. S. Granovetter, "The Strength of Weak Ties," *American Journal of Sociology*, vol. 78, 1360–1380, 1973.

[13] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Porcesses on Complex Networks*. Cambridge: Cambridge University Press, 2008.

[14] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.

[15] A.-L. Barabási and M. Posfai, *Network Science*. Cambridge University Press, 2016.

[16] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks*, vol. 30, 107–117, 1998.

[17] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proc. Natl. Acad. Sci. USA*, vol. 101 (11), 3747–3752, 2004.

[18] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, "Hierarchical Organization of Modularity in Metabolic Networks," *Science*, vol. 297, 1551–1555, 2002.

[19] W. W. Zachary, "An Information Flow Model for Conflict and Fission in Small Groups," *Journal of Anthropological Research*, vol. 33, 452–473, 1977.

[20] E. Ravasz and A. L. Barabási, "Hierarchical Organization in Complex Networks," *Phys Rev E*, vol. 67, 026 112, 2003.

[21] M. Girvan and M. E. J. Newman, "Community Structure in Social and Biological Networks," *Proc Natl Acad Sci USA*, vol. 99, 7821–7826, 2002.

[22] M. E. J. Newman, "Modularity and Community Structure in Networks," *Proc Natl Acad Sci USA*, vol. 103, 8577–8582, 2006.

[23] T. P. Peixoto, "Bayesian Stochastic Blockmodeling," *Advances in Network Clustering and Blockmodeling*, 289–332, 2019.

[24] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, 814–818, 2005.

[25] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, 761–764, 2010.

[26] T. P. Peixoto, "Model Selection and Hypothesis Testing for Large-Scale Network Models with Overlapping Groups," *Physical Review X*, vol. 5, 2015.

[27] P. Clifford and A. Sudbury, "A model for spatial conflict," *Biometrika*, vol. 60, 581–588, 1973.

[28] R. A. Holley and T. M. Liggett, "Ergodic Theorems for Weakly Interacting Infinite Systems and the Voter Model," *The Annals of Probability*, vol. 3, 643–663, 1975.

[29] J. W. Evans, "Kinetic phase transitions in catalytic reaction models," *Langmuir*, vol. 7, 2514–2519, 1991.

[30] J. W. Evans and T. R. Ray, "Kinetics of the monomer-monomer surface reaction model," *Phys. Rev. E*, vol. 47, 1018–1025, 1993.

[31] X. Castelló, V. M. Eguíluz, and M. San Miguel, "Ordering dynamics with two non-excluding options: Bilingualism in language competition," *New Journal of Physics*, vol. 8, 308, 2006.

[32] J. Fernández-Gracia, K. Suchecki, J. J. Ramasco, M. San Miguel, and V. M. Eguíluz, "Is the Voter Model a Model for Voters?" *Phys. Rev. Lett.*, vol. 112, 158 701, 2014.

[33] N. Masuda, N. Gibert, and S. Redner, "Heterogeneous voter models," *Phys. Rev. E*, vol. 82, 010 103, 2010.

[34] J. Fernández-Gracia, V. M. Eguíluz, and M. San Miguel, "Update rules and interevent time distributions: Slow ordering versus no ordering in the voter model," *Phys. Rev. E*, vol. 84, 015 103, 2011.

[35] V. Sood and S. Redner, "Voter Model on Heterogeneous Graphs," *Phys. Rev. Lett.*, vol. 94, 178 701, 2005.

[36] K. Suchecki, V. M. Eguíluz, and M. S. Miguel, "Conservation laws for the voter model in complex networks," *EPL (Europhysics Letters)*, vol. 69, 228, 2005.

[37] F. Vazquez and V. M. Eguiluz, "Analytical solution of the voter model on uncorrelated networks," *New Journal of Physics*, vol. 10, 063 011, 2008.

[38] V. Sood, T. Antal, and S. Redner, "Voter models on heterogeneous networks," *Physical Review E*, vol. 77, 041 121, 2008.

[39] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, "Diffusion-annihilation processes in complex networks," *Phys. Rev. E*, vol. 71, 056 104, 2005.

[40] M. Boguñá, C. Castellano, and R. Pastor-Satorras, "Langevin approach for the dynamics of the contact process on annealed scale-free networks," *Phys. Rev. E*, vol. 79, 036 110, 2009.

[41] M. Á. Serrano, K. Klemm, F. Vazquez, V. M. Eguíluz, and M. S. Miguel, "Conservation laws for voter-like models on random directed networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, P10024, 2009.

[42] G. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Springer-Verlag, Berlin Heidelberg New York, 2004.

[43] L. Rozanova and M. Boguñá, "Dynamical properties of the herding voter model with and without noise," *Physical Review E*, vol. 96, 2017.

[44] P. Csermely, A. London, L.-Y. Wu, and B. Uzzi, "Structure and dynamics of core-periphery networks," *Journal of Complex Networks*, vol. 1, 93–123, 2013.

[45] R. A. Baños, J. Borge-Holthoefer, N. Wang, Y. Moreno, and S. González-Bailón, "Diffusion Dynamics with Changing Network Composition," *Entropy*, vol. 15, 4553–4568, 2013.

[46] V. Kandiah and D. L. Shepelyansky, "PageRank model of opinion formation on social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 391, 5779–5793, 2012.

[47] G. F. de Arruda, A. L. Barbieri, P. M. Rodríguez, F. A. Rodrigues, Y. Moreno, and L. d. F. Costa, "Role of Centrality for the Identification of Influential Spreaders in Complex Networks," *Physical Review E*, vol. 90, 032 812, 2014.

[48] Z. Wang, A. Scaglione, and R. J. Thomas, "Electrical Centrality Measures for Electric Power Grid Vulnerability Analysis," in *49th IEEE Conference on Decision and Control (CDC)*, 2010, 5792–5797.

[49] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of Influential Spreaders in Complex Networks," *Nature Physics*, vol. 6, 888–893, 2010.

[50] S. P. Borgatti, "Centrality and network flow," *Social Networks*, vol. 27, 55–71, 2005.

[51] A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou, "Synchronization in Complex Networks," *Physics Reports*, vol. 469, 93–153, 2008.

[52] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, "Critical phenomena in complex networks," *Rev. Mod. Phys.*, vol. 80, 1275–1335, 2008.

[53] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic Blockmodels: First Steps," *Social Networks*, vol. 5, 109–137, 1983.

[54] B. Karrer and M. E. J. Newman, "Stochastic Blockmodels and Community Structure in Networks," *Physical Review E*, vol. 83, 2011.

[55] T. P. Peixoto, "Efficient Monte Carlo and Greedy Heuristic for the Inference of Stochastic Block Models," *Physical Review E*, vol. 89, 2014.

[56] ——, "The graph-tool python library," *figshare*, vol. https://dx.doi.org/10.6084/m9.figshare.1164194, 2014.

[57] R. S. Burt, *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1992.

[58] P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol, and V. M. Eguiluz, "Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media," *PLoS ONE*, vol. 7, 2012.

[59] M. Salathé and J. H. Jones, "Dynamics and Control of Diseases in Networks with Community Structure," *PLoS Computational Biology*, vol. 6, e1000736, 2010.

[60] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From Molecular to Modular Cell Biology," *Nature*, vol. 402, C47–C52, 1999.

[61] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal, "Evidence for Dynamically Organized Modularity in the Yeast Protein–Protein Interaction Network," *Nature*, vol. 430, 88–93, 2004.

[62] C. J. Jeffery, "Moonlighting Proteins," *Trends in Biochemical Sciences*, vol. 24, 8–11, 1999.

[63] C. E. Chapple, B. Robisson, L. Spinelli, C. Guien, E. Becker, and C. Brun, "Extreme Multifunctional Proteins Identified from a Human Protein Interaction Network," *Nature Communications*, vol. 6, 2015.

[64] M. Najafi, B. W. McMenamin, J. Z. Simon, and L. Pessoa, "Overlapping communities reveal rich structure in large-scale brain networks during rest and task conditions," *NeuroImage*, vol. 135, 92–106, 2016.

[65] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, "Detect overlapping and hierarchical community structure in networks," *Physica A: Statistical Mechanics and its Applications*, vol. 388, 1706–1712, 2009.

[66] W. Hwang, Y.-r. Cho, A. Zhang, and M. Ramanathan, "Bridging Centrality: Identifying Bridging Nodes in Scale-Free Networks," in *Proc. 12th ACM SIGKDD*, 2006, 20–23.

[67] P. Jensen, M. Morini, M. Karsai, T. Venturini, A. Vespignani, M. Jacomy, J.-P. Cointet, P. Merckle, and E. Fleury, "Detecting Global Bridges in Networks," *Journal of Complex Networks*, vol. 4, 319–329, 2016, Comment: Journal of Complex Networks Preprint; 14 pages; 6 figures.

[68] A.-K. Wu, L. Tian, and Y.-Y. Liu, "Bridges in Complex Networks," *Physical Review E*, vol. 97, 2018.

[69] T. Nepusz, A. Petróczi, L. Négyessy, and F. Bazsó, "Fuzzy Communities and the Concept of Bridgeness in Complex Networks," *Physical Review E*, vol. 77, 2008.

[70] R. Guimerà and L. A. Nunes Amaral, "Functional Cartography of Complex Metabolic Networks (Supplementary)," *Nature*, vol. 433, 895–900, 2005.

[71] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, 509–512, 1999.

[72] D. J. Watts and S. H. Strogatz, "Collective Dynamics of "Small-World" Networks," *Nature*, vol. 393, 440–442, 1998.

[73] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Rev. Mod. Phys.*, vol. 87, 925–979, 2015.

[74] J. P. Gleeson, "Cascades on Correlated and Modular Random Networks," *Physical Review E*, vol. 77, 2008.

[75] D. Gfeller and P. De los Rios, "Spectral Coarse Graining of Complex Networks," *Phys. Rev. Lett.*, vol. 99, 038 701, 2007.

[76] J. Reichardt and S. Bornholdt, "Detecting Fuzzy Community Structures in Complex Networks with a Potts Model," *Physical Review Letters*, vol. 93, 2004.

[77] R Lambiotte and M Ausloos, "Coexistence of Opposite Opinions in a Network with Communities," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2007, 2007.

[78] E. Oh, K. Rho, H. Hong, and B. Kahng, "Modular Synchronization in Complex Networks," *Physical Review E*, vol. 72, 2005.

[79] A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente, "Synchronization Reveals Topological Scales in Complex Networks," *Physical Review Letters*, vol. 96, 2006.

[80] J. Gómez-Gardeñes, Y. Moreno, and A. Arenas, "Paths to Synchronization on Complex Networks," *Physical Review Letters*, vol. 98, 2007.

[81] D. Li, I. Leyva, J. A. Almendral, I. Sendiña-Nadal, J. M. Buldú, S. Havlin, and S. Boccaletti, "Synchronization Interfaces and Overlapping Communities in Complex Networks," *Physical Review Letters*, vol. 101, 2008.

[82] J. A. Acebrón, L. L. Bonilla, C. J. Pérez Vicente, F. Ritort, and R. Spigler, "The Kuramoto Model: A Simple Paradigm for Synchronization Phenomena," *Reviews of Modern Physics*, vol. 77, 137–185, 2005.

[83] F. A. Rodrigues, T. K. D. Peron, P. Ji, and J. Kurths, "The Kuramoto Model in Complex Networks," *Physics Reports*, vol. 610, 1–98, 2016.

[84] L. Landau and E. Lifshitz, *Statistical Physics*, Fifth. Elsevier Science, 2013.

[85] J Olejarz, P. L. Krapivsky, and S Redner, "Zero-Temperature Coarsening in the 2d Potts Model," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2013, 06–18, 2013.

[86] J. F. F. Mendes and E. J. S. Lage, "Dynamics of the infinite-ranged Potts model," *Journal of Statistical Physics*, vol. 64, 653–672, 1991.

[87] C.-K. Chan, T. E. Lee, and S. Gopalakrishnan, "Limit-cycle phase in driven-dissipative spin systems," *Physical Review A*, vol. 91, 051 601, 2015.

[88] T. Herpich and M. Esposito, "Universality in driven Potts models," *Physical Review E*, vol. 99, 022 135, 2019.

[89] L. M. Pecora and T. L. Carroll, "Master Stability Functions for Synchronized Coupled Systems," *Physical Review Letters*, vol. 80, 4, 1998.

[90] M. Barahona and L. M. Pecora, "Synchronization in Small-World Systems," *Physical Review Letters*, vol. 89, 2002.

[91] K. Park, Y.-C. Lai, S. Gupte, and J.-W. Kim, "Synchronization in Complex Networks with a Modular Structure," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 16, 015 105, 2006.

[92] S. Hata and H. Nakao, "Localization of Laplacian Eigenvectors on Random Networks," *Scientific Reports*, vol. 7, 2017.

[93] L. Donetti and M. A. Muñoz, "Detecting Network Communities: A New Systematic and Efficient Algorithm," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2004, P10012, 2004.

[94] A. Arenas, A. Fernández, and S. Gómez, "An optimization approach to the structure of the neuronal layout of C. elegans. In , editors," in *Handbook of Biological Networks. World Scientific*, 2009.

[95] P. N. McGraw and M. Menzinger, "Laplacian Spectra as a Diagnostic Tool for Network Structure and Dynamics," *Physical Review E*, vol. 77, 2008.

[96] X.-L. Ren, N. Gleinig, D. Helbing, and N. Antulov-Fantulin, "Generalized Network Dismantling," *Proceedings of the National Academy of Sciences*, vol. 116, 6554–6559, 2019.

[97] J. T. Matamalas, A. Arenas, and S. Gómez, "Effective Approach to Epidemic Containment Using Link Equations in Complex Networks," *Science Advances*, vol. 4, 2018.

[98] M. A. Serrano and M. Boguñá, "Percolation and epidemic thresholds in clustered networks," *Phys. Rev. Lett.*, vol. 97, 088 701, 2006.

[99] P. Crucitti, V. Latora, M. Marchiori, and A. Rapisarda, "Efficiency of Scale-Free Networks: Error and Attack Tolerance," *Physica A*, 21, 2003.

[100] M. Barthélemy, "Spatial networks," *Physics Reports*, vol. 499, 1–101, 2011.

[101] I. A. Kovács, R. Palotai, M. S. Szalay, and P. Csermely, "Community Landscapes: An Integrative Approach to Determine Overlapping Network Module Hierarchy, Identify Key Nodes and Predict Network Dynamics," *PLoS ONE*, vol. 5, e12528, 2010.

[102] P. Van Mieghem, K. Devriendt, and H. Cetinay, "Pseudoinverse of the Laplacian and Best Spreader Node in a Network," *Physical Review E*, vol. 96, 032 311, 2017.

[103] M. De Domenico, C. Granell, M. A. Porter, and A. Arenas, "The Physics of Spreading Processes in Multilayer Networks," *Nature Physics*, vol. 12, 901–906, 2016.

[104] B. R. C. Amor, S. I. Vuik, R. Callahan, A. Darzi, S. N. Yaliraki, and M. Barahona, "Community Detection and Role Identification in Directed Networks: Understanding the Twitter Network of the Care.Data Debate," in *Dynamic Networks and Cyber-Security*, vol. Volume 1, World Scientific, 2015, 111–136.

[105] D. Meunier, R. Lambiotte, and E. T. Bullmore, "Modular and Hierarchically Modular Organization of Brain Networks," *Frontiers in Neuroscience*, vol. 4, 2010.

[106] J. Gómez-Gardeñes, G. Zamora-López, Y. Moreno, and A. Arenas, "From Modular to Centralized Organization of Synchronization in Functional Areas of the Cat Cerebral Cortex," *PLoS ONE*, vol. 5, e12313, 2010.

[107] J. Hizanidis, N. E. Kouvaris, G. Zamora-López, A. Díaz-Guilera, and C. G. Antonopoulos, "Chimera-like States in Modular Neural Networks," *Scientific Reports*, vol. 6, 2016.

[108] M. E. J. Newman, *Networks: An Introduction*. Oxford University Press, Oxford, 2010.

[109] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwanga, "Complex Networks: Structure and Dynamics," *Phys Rep*, vol. 424, 175–308, 2006.

[110] I. Dobson, B. Carreras, V. Lynch, and D. Newman, "Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 17, 026 103, 2007.

[111] P. S. Skardal and A. Arenas, "Control of coupled oscillator networks with application to microgrid technologies," *Science Advances*, vol. 1, e1500339, 2015.

[112] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical Structure and the Prediction of Missing Links in Networks," *Nature*, vol. 453, 98–101, 2008.

[113] R. Guimerà and M. Sales-Pardo9, "Missing and spurious interactions and the reconstruction of complex networks," *Proc. Natl. Acad. Sci. USA*, vol. 106, 22 073–22 078, 2009.

[114] D. Hric, T. P. Peixoto, and S. Fortunato, "Network structure, metadata, and the prediction of missing nodes and annotations," *Physical Review X*, vol. 6, 031 038, 2016.

[115] T. Hoffmann, L. Peel, R. Lambiotte, and N. S. Jones, "Community detection in networks with unobserved edges," *arXiv preprint arXiv:1808.06079*, 2018.

[116] M. E. J. Newman, "Analysis of weighted networks," *Phys. Rev. E*, vol. 70, 056 131, 2004.

[117] S. Gómez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno, "Discrete-time Markov chain approach to contact-based disease spreading in complex networks," *EPL (Europhysics Letters)*, vol. 89, 38 009, 2010.

[118] E. Wigner, "Characteristic vectors of bordered matrices with infinite dimensions," *Annals of Mathematics*, vol. 62, 548–564, 1955.

[119] M. L. Mehta, *Random Matrices*. Elsevier Academic Press, 2004.

[120] D. A. Spielman, "Spectral graph theory and its applications," *Proc. 48th Annual IEEE Symposium on Foundations of Computer Science*, 29–38, 2007.

[121] P. V. Mieghem, *Graph Spectra for Complex Networks*. Cambridge University Press, 2011.

[122] F. Chung, L. Lu, and V. Vu, "Spectra of random graphs with given expected degrees," *Proceedings of the National Academy of Sciences*, vol. 100, 6313–6318, 2003.

[123] J. G. Restrepo, E. Ott, and B. R. Hunt, "Onset of synchronization in large networks of coupled oscillators," *Phys. Rev. E*, vol. 71, 036 151, 2005.

[124] J. G. Restrepo, E. Ott, and B. R. Hunt, "Approximating the largest eigenvalue of network adjacency matrices," *Phys. Rev. E*, vol. 76, 056 119, 2007.

[125] H. Ku, "Notes on the Use of Propagation of Error Formulas," *Journal of Research of the National Bureau of Standards - C. Enginnering and Instrumentation*, vol. 70, 0–6, 1966.

[126] G. Mana and F. Pennecchi, "Uncertainty propagation in non-linear measurement equations," *Metrologia*, vol. 44, 246, 2007.

[127] J. P. Gleeson, S. Melnik, J. A. Ward, M. A. Porter, and P. J. Mucha, "Accuracy of mean-field theory for dynamics on real-world networks," *Phys. Rev. E*, vol. 85, 026 106, 2012.

[128] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, "Structure of growing networks with preferential linking," *Phys. Rev. Lett.*, vol. 85, 4633–4636, 2000.

[129] S. Valverde, R. Ferrer-i-Cancho, and R. V. Sole, "Scale-Free Networks from Optimal Design," *EPL (Europhysics Letters)*, vol. 60, 512–517, 2002.

[130] D. B. Larremore, W. L. Shew, and J. G. Restrepo, "Predicting Criticality and Dynamic Range in Complex Networks: Effects of Topology," *Phys. Rev. Lett.*, vol. 106, 058 101, 2011.

[131] B. Corominas-Murtra, J. Goñi, R. V. Solé, and C. Rodríguez-Caso, "On the origins of hierarchy in complex networks," *Proceedings of the National Academy of Sciences*, vol. 110, 13 316–13 321, 2013.

[132] O. Kinouchi and M. Copelli, "Optimal dynamical range of excitable networks at criticality," *Nature Physics*, vol. 2, 348–351, 2006.

[133] D. R. Chialvo, "Emergent complex neural dynamics," *Nature Physics*, vol. 6, 744–750, 2010.

[134] L. Arola-Fernández, A. Díaz-Guilera, and A. Arenas, "Synchronization invariance under network structural transformations," *Phys. Rev. E*, vol. 97, 060 301, 2018.

[135] S. Jalan and J. Bandyopdhyay, "Random matrix analysis of complex networks," *Phys. Rev. E*, vol. 76, 046 107, 2007.

[136] C. Sarkar and S. Jalan, "Spectral properties of complex networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, 102 101, 2018.

[137] J. Lin and Y. Ban, "Complex Network Topology of Transportation Systems," *Transport Reviews*, vol. 33, 658–685, 2013.

[138] J.-P. Rodrigue, *The Geography of Transport Systems*. Taylor & Francis, 2016.

[139] L. Guo and X. Cai, "Degree and weighted properties of the directed china railway network," *International Journal of Modern Physics C*, vol. 19, 1909–1918, 2008.

[140] M. P. M. Cañizares, A. L. Pita, and A. G. Álvarez, "Structure and topology of high-speed rail networks," *Proceedings of the Institution of Civil Engineers - Transport*, vol. 168, 415–424, 2015.

[141] K. Berdica, "An introduction to road vulnerability: What has been done, is done and should be done," *Transport Policy*, vol. 9, 117–127, 2002.

[142] R. Gedik, H. Medal, C. Rainwater, E. A. Pohl, and S. J. Mason, "Vulnerability assessment and re-routing of freight trains under disruptions: A coal supply chain network application," *Transportation Research Part E: Logistics and Transportation Review*, vol. 71, 45–57, 2014.

[143] A. A. Khaled, M. Jin, D. B. Clarke, and M. A. Hoque, "Train design and routing optimization for evaluating criticality of freight railroad infrastructures," *Transportation Research Part B: Methodological*, vol. 71, 71–84, 2015.

[144] Z. Zhang, X. Li, and H. Li, "A quantitative approach for assessing the critical nodal and linear elements of a railway infrastructure," *International Journal of Critical Infrastructure Protection*, vol. 8, 3–15, 2015.

[145] M. Bababeik, N. Khademi, A. Chen, and M. M. Nasiri, "Vulnerability Analysis of Railway Networks in Case of Multi-Link Blockage," *Transportation Research Procedia*, vol. 22, 275–284, 2017.

[146] B. Monechi, P. Gravino, R. Di Clemente, and V. D. P. Servedio, "Complex delay dynamics on railway networks from universal laws to realistic modelling," *EPJ Data Science*, vol. 7, 35, 2018.

[147] Y. K. Al-Douri, P. Tretten, and R. Karim, "Improvement of railway performance: A study of Swedish railway infrastructure," *Journal of Modern Transportation*, vol. 24, 22–37, 2016.

[148] R. Faturechi and E. Miller-Hooks, "Measuring the Performance of Transportation Infrastructure Systems in Disasters: A Comprehensive Review," *Journal of Infrastructure Systems*, vol. 21, 04 014 025, 2015.

[149] P. Norrbin, J. Lin, and a. A. Parida, "Infrastructure Robustness for Railway Systems," *International Journal of Performability Engineering*, vol. 12, 249, 2016.

[150] G. W. Klau and R. Weiskircher, "Robustness and Resilience," in *Network Analysis: Methodological Foundations*, Berlin, Heidelberg: Springer, 2005, 417–437.

[151] V. Grimm and J. M. Calabrese, "What Is Resilience? A Short Introduction," in *Viability and Resilience of Complex Systems: Concepts, Methods and Case Studies from Ecology and Society*, Berlin, Heidelberg: Springer, 2011, 3–13.

[152] R. M. May, "Will a large complex system be stable?" *Nature*, vol. 238, 413–414, 1972.

[153] A. Reggiani, T. De Graaff, and P. Nijkamp, "Resilience: An Evolutionary Approach to Spatial Economic Systems," *Networks and Spatial Economics*, vol. 2, 211–229, 2002.

[154] M. Bruneau, S. E. Chang, R. T. Eguchi, G. C. Lee, T. D. O'Rourke, A. M. Reinhorn, M. Shinozuka, K. Tierney, W. A. Wallace, and D. von Winterfeldt, "A Framework to Quantitatively Assess and Enhance the Seismic Resilience of Communities," *Earthquake Spectra*, vol. 19, 733–752, 2003.

[155] A. Rose, "Economic resilience to natural and man-made disasters: Multidisciplinary origins and contextual dimensions," *Environmental Hazards*, vol. 7, 383–398, 2007.

[156] S. E. Chang and M. Shinozuka, "Measuring Improvements in the Disaster Resilience of Communities," *Earthquake Spectra*, vol. 20, 739–755, 2004.

[157] M. D'Lima and F. Medda, "A new measure of resilience: An application to the London Underground," *Transportation Research Part A: Policy and Practice*, vol. 81, 35–46, 2015.

[158] S. Johnson, V. Domínguez-García, L. Donetti, and M. A. Muñoz, "Trophic coherence determines food-web stability," *Proceedings of the National Academy of Sciences*, vol. 111, 17 923–17 928, 2014.

[159] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*. New York, NY: Wiley, 2006.

[160] S. Johnson and N. S. Jones, "Looplessness in networks is linked to trophic coherence," *Proceedings of the National Academy of Sciences*, vol. 114, 5618–5623, 2017.

[161] A. Ma and R. J. Mondragón, "Rich-Cores in Networks," *PLOS ONE*, vol. 10, e0119678, 2015.

[162] X. Lu, C. Gray, L. E. Brown, M. E. Ledger, A. M. Milner, R. J. Mondragón, G. Woodward, and A. Ma, "Drought rewires the cores of food webs," *Nature Climate Change*, vol. 6, 875–878, 2016.

[163] P. Boldi, M. Rosa, and S. Vigna, "Robustness of social and web graphs to node removal," *Social Network Analysis and Mining*, vol. 3, 829–842, 2013.

[164] M. S. Dehghani, G. Flintsch, and S. McNeil, "Impact of Road Conditions and Disruption Uncertainties on Network Vulnerability," *Journal of Infrastructure Systems*, vol. 20, 04 014 015, 2014.

[165] S. P. Borgatti and M. G. Everett, "Models of core/periphery structures," *Social Networks*, vol. 21, 375–395, 2000.

[166] V. Zlatic, G. Bianconi, A. Díaz-Guilera, D. Garlaschelli, F. Rao, and G. Caldarelli, "On the rich-club effect in dense and weighted networks," *The European Physical Journal B*, vol. 67, 271–275, 2009.

[167] M. A. Serrano, "Rich-club vs rich-multipolarization phenomena in weighted networks," *Phys. Rev. E*, vol. 78, 026 101, 2008.

[168] M. Csigi, A. Kőrösi, J. Bíró, Z. Heszberger, Y. Malkov, and A. Gulyás, "Geometric explanation of the rich-club phenomenon in complex networks," *Scientific Reports*, vol. 7, 1730, 2017.

[169] S. Zhou and R. J. Mondragon, "The rich-club phenomenon in the Internet topology," *IEEE Commun. Lett.*, vol. 8, 180–182, 2004.

[170] A. Pagani, G. Mosquera, A. Alturki, S. Johnson, S. Jarvis, A. Wilson, W. Guo, and L. Varga, *Data from: Resilience or robustness: Identifying topological vulnerabilities in rail networks*, 2019.

[171] S. M. Stigler, "Francis Galton's Account of the Invention of Correlation," *Statistical Science*, vol. 4, 73–79, 1989.

[172] I. Kivimäki, B. Lebichot, J. Saramäki, and M. Saerens, "Two Betweenness Centrality Measures Based on Randomized Shortest Paths," *Scientific Reports*, vol. 6, 1–15, 2016.

[173] Y. Kornbluth, G. Barach, Y. Tuchman, B. Kadish, G. Cwilich, and S. V. Buldyrev, "Network overload due to massive attacks," *Physical Review E*, vol. 97, 052 309, 2018.

[174] X. Yao, P. Zhao, and K. Qiao, "Simulation and evaluation of urban rail transit network based on multi-agent approach," *Journal of Industrial Engineering and Management*, vol. 6, 367–379, 2013.

[175] M. A. Kaplan, *System and Process in International Politics*. Wiley, 1957.

[176] L. F. Richardson, *Statistics of Deadly Quarrels*. Boxwood Press, 1960.

[177] R. C. Snyder, *Foreign Policy Decision-Making: An Approach to the Study of International Politics, Edited by Richard C.Snyder, H.W.Bruck and Burton Sapin*. Free Press of Glencoe, 1962.

[178] C. McClelland, *World Event/Interaction Survey (WEIS) Project, 1966-1978*, 2006.

[179] E. E. Azar, "The Conflict and Peace Data Bank (COPDAB) Project," *The Journal of Conflict Resolution*, vol. 24, 143–152, 1980.

[180] J. D. Singer, S. Bremer, and J. Stuckey, "Capability distribution, uncertainty, and major power war, 1820-1965," *Peace, war and numbers*, 1972.

[181] M. D. Ward, B. D. Greenhill, and K. M. Bakke, "The perils of policy by p-value: Predicting civil conflicts," *Journal of Peace Research*, vol. 47, 363–375, 2010.

[182] M. Colaresi and Z. Mahmood, "Do the robot: Lessons from machine learning to improve conflict forecasting," *Journal of Peace Research*, vol. 54, 193–214, 2017.

[183] H. Hegre, N. W. Metternich, H. M. Nygård, and J. Wucherpfennig, "Introduction: Forecasting in peace research," *Journal of Peace Research*, vol. 54, 113–124, 2017.

[184] G. Clayton and K. S. Gleditsch, "Will we see helping hands? Predicting civil war mediation and likely success:" *Conflict Management and Peace Science*, 2013.

[185] L.-E. Cederman, K. S. Gleditsch, and J. Wucherpfennig, "Predicting the decline of ethnic civil war: Was Gurr right and for the right reasons?" *Journal of Peace Research*, vol. 54, 262–274, 2017.

[186] H. Hegre, L. Hultman, and H. M. Nygård, "Evaluating the Conflict-Reducing Effect of UN Peacekeeping Operations," *The Journal of Politics*, vol. 81, 215–232, 2018.

[187] S. M. Mitchell, P. F. Diehl, and J. D. Morrow, *Guide to the Scientific Study of International Processes*. Wiley, 2012.

[188] J. D. Singer, "The Level-of-Analysis Problem in International Relations," *World Politics*, vol. 14, 77–92, 1961.

[189] S. M. Mitchell and J. A. Vasquez, *Conflict, War, and Peace: An Introduction to Scientific Research*. CQ Press, 2013.

[190] C. Raleigh, A. Linke, H. Hegre, and J. Karlsen, "Introducing ACLED: An Armed Conflict Location and Event Dataset," *Journal of Peace Research*, 2010.

[191] R. Sundberg and E. Melander, "Introducing the UCDP Georeferenced Event Dataset:" *Journal of Peace Research*, 2013.

[192] J. Wucherpfennig, N. B. Weidmann, L. Girardin, L.-E. Cederman, and A. Wimmer, "Politically Relevant Ethnic Groups across Space and Time: Introducing the GeoEPR Dataset," *Conflict Management and Peace Science*, vol. 28, 423–437, 2011.

[193] J. Faber, "Measuring Cooperation, Conflict, and the Social Network of Nations," *Journal of Conflict Resolution*, vol. 31, 438–464, 1987.

[194] P. D. Hoff and M. D. Ward, "Modeling Dependencies in International Relations Networks," *Political Analysis*, vol. 12, 160–175, 2004.

[195] M. R. Frank, N. Obradovich, L. Sun, W. L. Woon, B. L. LeVeck, and I. Rahwan, "Detecting reciprocity at a global scale," *Science Advances*, vol. 4, eaao5348, 2018.

[196] V. A. Traag and J. Bruggeman, "Community detection in networks with positive and negative links," *Physical Review E*, vol. 80, 2009.

[197] Y. Lupu and V. A. Traag, "Trading Communities, the Networked Structure of International Relations, and the Kantian Peace," *Journal of Conflict Resolution*, vol. 57, 1011–1042, 2013.

[198] B. Greenhill and Y. Lupu, "Clubs of Clubs: Fragmentation in the Network of Intergovernmental Organizations," *International Studies Quarterly*, vol. 61, 181–195, 2017.

[199] Y. Lupu and B. Greenhill, "The networked peace: Intergovernmental organizations and international conflict," *Journal of Peace Research*, vol. 54, 833–848, 2017.

[200] Z. Maoz, *Networks of Nations: The Evolution, Structure, and Impact of International Networks, 1816–2001*. Cambridge University Press, 2010.

[201] B. J. Kinne, "Multilateral Trade and Militarized Conflict: Centrality, Openness, and Asymmetry in the Global Trade Network," *The Journal of Politics*, vol. 74, 308–322, 2012.

[202] M. D. Ward, R. M. Siverson, and X. Cao, "Disputes, Democracies, and Dependencies: A Reexamination of the Kantian Peace," *American Journal of Political Science*, vol. 51, 583–601, 2007.

[203] S. J. Cranmer, E. J. Menninga, and P. J. Mucha, "Kantian fractionalization predicts the conflict propensity of the international system," *Proceedings of the National Academy of Sciences*, vol. 112, 11 812–11 816, 2015.

[204] NGI, "Cities location and population data," *National geospatial-intelligence agency*, 2007.

[205] N. B. Weidmann, J. K. Rød, and L.-E. Cederman, "Representing ethnic groups in space: A new dataset," *Journal of Peace Research*, vol. 47, 491–499, 2010.

[206] J. E. Anderson, "A Theoretical Foundation for the Gravity Equation," *American Economic Review*, vol. 69, 106–116, 1979.

[207] J. H. Bergstrand, "The gravity equation in international trade: Some microeconomic foundations and empirical evidence," *The Review of Economics and Statistics*, vol. 67, 474–481, 1985.

[208] P. Pöyhönen, "A Tentative Model for the Volume of Trade between Countries," *Weltwirtschaftliches Archiv*, vol. 90, 93–100, 1963.

[209] D. Balcan, V. Colizza, B. Goncalves, H. Hu, J. J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," *Proceedings of the National Academy of Sciences*, vol. 106, 21 484–21 489, 2009.

[210] P. Kaluza, A. Kölzsch, M. T. Gastner, and B. Blasius, "The complex network of global cargo ship movements," *Journal of The Royal Society Interface*, vol. 7, 1093–1103, 2010.

[211] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, 96–100, 2012.

[212] R. Lambiotte, V. D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren, "Geographical dispersal of mobile communication networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, 5317–5325, 2008.

[213] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel, "Urban gravity: A model for inter-city telecommunication flows," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, L07003, 2009.

[214] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins, "Geographic Routing in Social Networks," *PNAS*, vol. 102, 11 623–11 628, 2005.

[215] A. K. Rose, "Do We Really Know That the WTO Increases Trade?" *The American Economic Review*, vol. 94, 98–114, 2004.

[216] H. Hegre, "Gravitating toward War: Preponderance May Pacify, but Power Kills," *Journal of Conflict Resolution*, vol. 52, 566–589, 2008.

[217] H. Hegre, "Trade Dependence or Size Dependence?: The Gravity Model of Trade and the Liberal Peace," *Conflict Management and Peace Science*, vol. 26, 26–45, 2009.

[218] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, 779–782, 2008.

[219] W.-S. Jung, F. Wang, and H. E. Stanley, "Gravity model in the Korean highway," *EPL (Europhysics Letters)*, vol. 81, 48 005, 2008.

[220] A. Wilson, *Entropy in Urban and Regional Modelling*. Routledge, 2011.

[221] M. Barthélemy, "Spatial Networks," *Physics Reports*, vol. 499, 1–101, 2011.

[222] T. Hastie, R. Tibshirani, and J. Friedman, "Linear Methods for Classification," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, NY: Springer, 2009, 101–137.

[223] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st. O'Reilly Media, Inc., 2017.

[224] D. Muchlinski, D. Siroky, J. He, and M. Kocher, "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data," *Political Analysis*, vol. 24, 87–103, 2016.

[225] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, 5–32, 2001.

[226] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, 1995, 278–282 vol.1.

[227] R. Couronné, P. Probst, and A.-L. Boulesteix, "Random forest versus logistic regression: A large-scale benchmark experiment," *BMC Bioinformatics*, vol. 19, 270, 2018.

[228] H. Hegre, M. Allansson, M. Basedau, M. Colaresi, M. Croicu, H. Fjelde, F. Hoyles, L. Hultman, S. Högbladh, R. Jansen, N. Mouhleb, S. A. Muhammad, D. Nilsson, H. M. Nygård, G. Olafsdottir, K. Petrova, D. Randahl, E. G. Rød, G. Schneider, N. von Uexkull, and J. Vestby, "ViEWS: A political violence early-warning system," *Journal of Peace Research*, vol. 56, 155–174, 2019.

[229] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2018.

[230] J. A. Vasquez, "Why Do Neighbors Fight? Proximity, Interaction, or Territoriality," *Journal of Peace Research*, vol. 32, 277–293, 1995, (Available on Moodle).

[231] J. Esteban, L. Mayoral, and D. Ray, "Ethnicity and Conflict: An Empirical Study," *American Economic Review*, vol. 102, 1310–1342, 2012.

[232] J. M. Montgomery, F. M. Hollenbach, and M. D. Ward, "Improving Predictions using Ensemble Bayesian Model Averaging," *Political Analysis*, vol. 20, 271–291, 2012.