Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation

Doctoral Program in Electrical, Electronics and Communications Engineering (34^{th}cycle)

Unveiling human interactions

Approaches and techniques toward the discovery and representation of interactions in networks

By

Francesco Vincenzo Surano

******

Supervisor(s):

Prof. Alessandro Rizzo, Supervisor

Prof. Maurizio Porfiri, Co-Supervisor

Doctoral Examination Committee:

Prof. G. Ruffo, Referee, Università degli Studi del Piemonte Orientale 'A. Avogadro', Italy

Prof. M. Frasca, Referee, Università degli Studi di Catania, Italy

Prof. R. Sinatra, IT University of Copenhagen, Denmark

Prof. S. Jalan, Indian Institute of Technology Indore, India

Prof. C. Novara, Politecnico di Torino, Italy

Politecnico di Torino

2023

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Francesco Vincenzo Surano
2023

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

We've come too far to give up who we are

Acknowledgements

First of all, I would like to thank my supervisor Prof. A. Rizzo for guiding me through my Ph.D., and my co-supervisor Prof. M. Porfiri, for welcoming me to his lab across the sea. Their support has been essential in this long journey, helping me learn and understand many different aspects of research.

I want to mention the support of the "Joint research project with the top 50 universities in the world", funded by Compagnia di San Paolo, which made my studies possible, and the collaboration between the Complex Systems Laboratory of the Politecnico di Torino and the Dynamical Systems Laboratory of New York University.

Thanks to C. Bongiorno and L. Zino, who helped me take my first steps in academia. To all my colleagues on this side of the ocean, Francesco, Michele, Mattia, Fiorella, David, Stefano, and those on the other side, Alain, Roni, Raghu, Mert, Matthieu, Daniel, Agnieszka, Anna, Jalil, you have made the lab a fun and enjoyable place to work, a place to meet and grow.

I would like to thank all my friends, the distant ones, Alex, Kim, Shahar, Armand, Fabrizio, Prasant, Ines, and Matteo, for making the long stay in America feel like a new home, and all my friends in Italy, "le compagnie", those "intrincea", the "bollenti" and "ADI", for welcoming me back from my travels.

Special thanks to my parents and family, Antonella, Giangi, Rita, Marcello, Chiara, and Beatrice, for always supporting me in all my endeavors.

Finally, a special and loving thought to the person who has allowed me to make sense of all these crazy and, at times, difficult years - Cassandra, who has always been by my side with Nala.

Abstract

What does it mean to unveil a network of interactions? In this dissertation, we will enrich the field of network theory by studying innovative approaches and ideas to reconstruct missing elements and connections in networks. Starting from theoretical approaches and methods to infer uncaptured links and nodes, we will contribute with our probabilistic techniques. We will implement state-of-the-art analysis on big data extracted from social networks to discover unexpected connections in the opinion dynamics of a population in relation to epidemiological events. Finally, open-source software will be developed to perform intuitive and affordable experiments about face-to-face interactions, a field of research almost untapped. Guiding us in this journey, there will be the common thread of network reconstruction to unveil human interactions at different levels of the scientific process.

Contents

List of Figures

List of Tables

Nomenclature

Acronyms / Abbreviations

ADN Activity driven networks

BLE Bluetooth low energy

COVID-19 Corona virus disease discovered in 2019

FDR False discovery rate

fMRI Functional magnetic resonance Imaging

GDPR General data protection regulation

GPS Global positioning system

MRI Magnetic resonance imaging

PET Positron emission tomography

RADN Routed activity driven networks

RSSI Received signal strength indicator

SIR Susceptible-infected-recovered epidemic model

SIS Susceptible-infected-susceptible epidemic model

TPR True positive rate

URL Uniform resource locator

UUID Universally unique identifier

Mathematical Symbols

A Adjacency matrix of a graph

E Edges of a graph

G Graph

N Number of nodes in a graph

V Vertices of a graph

Chapter 1

Introduction

1.1 Origins of network theory

The field of network theory may have gained a lot of attention in recent years, but its origin dates back to the beginning of the industrial revolution. We are in Eastern Prussia, precisely in its capital Königsberg around 1735. This city was a thriving merchant city, supported by the trades of its busy fleet sailing on the river Pregel. The city was rich enough that its officials could build several bridges, seven exactly, to connect the small island of Kneiphof to the mainland and the two sides of the city together. An artistic representation of such an arrangement can be seen in figure 1.1. Thanks to this unique arrangement, a contemporary riddle was born in the city: could someone cross all the seven bridges of Königsberg exactly once? A rigorous proof that such a path did not exist came in the form of a rigorous demonstration from the Swiss mathematician L. Euler [1]. He schematically represented the situation as four points, representing the four stretches of disconnected land, and seven arcs to picture the bridges. This was indeed a graph representation, and thanks to it Euler noted that, in order to cross each arc once and only once, nodes with an odd number of arcs must be either starting points or end points. If more than two nodes have an odd number of links, no unique path can be identified. In Königsberg, each landmass had an odd number of bridges connected to it, so a path satisfying the riddle was impossible to find.

The field of graph theory has been thriving in mathematics, but the field of network theory, which shares a lot with the former but relies on data and models to explain reality, has only recently emerged. Why this is the case can be seen in some of the technological and cultural advantages that struck our society at the beginning of the century. A cornerstone connection

Fig. 1.1 Artistic representation of the Seven bridges of Königsberg, Creative Commons, Wikimedia Foundation

that opened the use of network theory to other fields of research was the paper by Granovetter [2] in 1973. In this study a strong argument is made on the importance of weak ties, all those social connections and encounters that, for the reason of being at the boundary of a person's group core, make them the driving motion of so many processes and dynamics. This finding originated from empirical work on how people were able to find new jobs, rooted in data and understood thanks to the idea of networks. From there on, the idea of collecting data and mapping it using the techniques of network theory took ample support in many social sciences, opening a new era of research and discovery.

A parallel and quite astonishing revolution happened in the field of biology around the same time, with the paper from White [3] in 1986. For the first time, and after tireless work, the complete diagram of an animal's brain was available for researchers to study. Albeit the animal was a modest C. Elegans, with a nervous system comprised of only 302 neurons, the details on the structure of the network of functioning cells could finally be studied. This research focused less on the network approach but was the telltale of an opening era, more and more data on interactions were becoming available in many research fields. A kind of data that couldn't be fully understood just with statistics or heuristic models, but needed a new approach focused on the relationship among the system's constituents.

When academia was adopting the network theory, the big dotcom revolution was starting to embrace it. At the core of one of the biggest companies of our times there is an interesting work, the PageRank algorithm [4]. This algorithm was invented by Brin and Page, at the time two Stanford associates who came on to found Google LLC. The PageRank algorithm leveraged stochastic ideas and network properties to identify, or rather predict, the best web page to show to a user based on very simple inputs. Cleverly the calculations leveraged the connections among webpages suggested by the creators of the pages themselves, as those could be incorporated into the bigger picture of the world wide web. Nowadays, we take for granted the simplicity of browsing online, but that is the result of applied network science at its finest. Google LLC is not the only company leveraging the power of networks; think of all the different social media available to you at the touch of a finger. Those are mapping and exploring the intricate connections and exchanges happening among people all around the world, enhancing and inhibiting messages and information.

Around the same years in which Brin and Page were revolutionizing Information Technology, two of the most important studies on network theory were coming into being. The work of Watts and Strogatz and that of Barabasi and Albert remain to this day two of the most cited in the field of network science and outside of it. Small World [5] is the name given by Watts and

Strogatz to a generative model that is at the same time simple in its basic concept yet able to describe an intriguing emerging property of societies. It seems striking at times that, no matter how far we are from home, our newly met friends know someone that, more or less directly, we also know. In their model, Watts and Strogatz obtained this emergent property by adopting a rewiring technique to regular-lattice networks. From the other side of the research field, that of the data and its analysis, Barabasi and Albert uncovered the scale-free properties of networks [6]. Starting from the well-cataloged data on scientific collaborations, it was evident that not all nodes and connections were created equally. Influential professors in a field were more likely to have more collaborations, and they were more likely to collaborate with other estimated scholars, inside and outside their area of expertise. The authors explained this with a generative model of networks that grow in steps, following simple probabilistic rules. Emerging from this model is that common properties arise no matter the size of the network, hence the name: scale-free. Actors rich in connections grow even richer, and those without connections struggle to gain more. It can be shown that the distribution of connections in many different fields and applications, from the power grid to airports, from friends at a school to web pages, follows a power-law, as the model proposed by Barabasi and Albert explains. These works unveiled the power of network theory. With some simple rules, the complex phenomena we experience in everyday life finally appear naturally from the system's dynamics, and can get a comprehensive mathematical explanation.

Nowadays, network theory has branched from graph theory and seeded numerous fields of research that have emerged to tackle specific questions and explain some peculiar phenomena. Defining and identifying a community in a network has been the focus of those interested in community-detection [7, 8], which still is a difficult task. Evaluating the robustness of a network to random failures [9] or targeted attacks [10] emerged as a field of interest due to the practical needs in modern infrastructures. Developing models that could explain contagion phenomena both of pathogens [11, 12] or information [13, 14] is seen nowadays as essential for the safeguarding of our communities. Alongside those practical applications, more theoretical analyses have been explored, such as the study on how to model and represent networks that evolve in time [15, 16], or the description of networks of higher-order [17, 18]. Among those, the field of network reconstruction [19–21] has steadily risen in interest in the academic community, as more and more data availability poses questions on the observability of networks in the real world. The latter subject of research will be of particular interest to this dissertation.

In the last couple of years, network theory has become, unwillingly, even more of public domain. The power of this research field has been put to the service of the community at

large and for the benefit of many. Its application to the study of epidemics has allowed for an unprecedented understanding and prediction on the diffusion of viruses and pathogens. For example, the H1N1 influenza pandemic around 2009 was one of the most accurately predicted phenomena of such kind [22]. Applying network theory to epidemic predictions allows for a new paradigm in how we think about diffusive phenomena and medicine, the role of mass and long-range transportation, and the interconnected nature of our world. A network approach may be one of the few that offer a solution for rampant and deadly diseases such as Ebola [23], where having a behavioral model interlaced with a network description is not just useful but essential. Sadly we all remember the recent COVID-19 pandemic. The effects it had on our society are still to be fully understood and will unravel for years to come. But the tools that this field of research, our field of research, gave us in fighting this disease helped in understanding its diffusion in the early stages [24], in defining the policies to control and curb it [25], guided us in designing the most effective vaccination campaigns [26], and allowed us to understand the citizens' reaction to the aforementioned policies [27]. All of this work allowed us to approach what could have been an insurmountable obstacle as a tough, yet defeatable, enemy.

Network theory is not just a toolbox or a lens for our society; it is a new paradigm to describe, understand and shape the world around us.

1.2 The problem of reconstructing a network

The problem of reconstructing a network has many different real-world applications and, therefore, different meanings associated with it. First of all, we have to define which part of the network we aim to reconstruct. That could be the links between nodes, which could help infer the diffusion pattern of a disease [28–30] or information [31], the activation pattern of a biological system [32–34] or the formation of social groups [21]. We could also be interested in nodes, or more generally the size of the network, to understand the population involved in a specific phenomenon [35, 36]. Moreover, we could be focusing on the problem of identifying communities, tightly linked groups within a network, which tend to have common behaviors or properties [37]. More generally, we could frame the problem of reconstructing a network as a practical one, developing new tools and techniques in fields where no previous data-collection has been done [38, 39].

Often the problem of reconstructing a network arises from observing a dynamical process taking place on it. Taking as an example disease spreading, the most evident trace left on a network is the number of agents infected by the disease traveling through the population. Yet

this process leaves little to no trace of the exact path of diffusion of the pathogen through society, and this inhibits our ability to stop it [40–42]. Another example could entail the structure of a criminal organization [43, 44], of which we see the devastating effects on society, but little we can do to stop it until we have reconstructed the size of its members and the network of its management. Therefore, new techniques are needed to link the effects of the processes we see unraveling on the network to its core structure in order to reconstruct the parts we are interested in studying. Finally, the reconstruction of a network is essential to assess whether the models, techniques, and ideas we develop are useful and realistic explanations of empirical behaviors or observed social structures.

Network reconstruction in biology

One of the fields that has devoted much energy to network reconstruction is medicine, particularly the fields of genomics, connectomics, and more generally systems biology. At the intersection of biology and network science, these research fields are exploring new approaches with a practical application: understanding the relational chains among different biological elements to cure pathologies and discover the secrets of life.

The first area of research to mention is that of connectomics. Connectomics can be broadly defined as the effort of reconstructing the functional diagram of the human brain. Knowing the biological functioning of cells and neurons is not enough to explain the complex behavior of the brain [45], as for this organ the complex capability of processing data resides not in its basic constituents but in their highly complex interactions. Neurons are assembled in a network of hundreds of billions of nodes [46], whose pattern of connection determines the different and astonishing capability of reasoning that characterize humans. The first tool that enabled connectomics [47–49] were those of magnetic resonance (MRI), functional magnetic resonance (fMRI) [50], diffusion magnetic resonance [51], and tomography (PET). Yet those technical imaging tools are not enough alone [45], analytic and modeling counterparts are needed to fruitfully obtain a viable description of the human brain [52].

The two main computational methods for studying brain connectivity are functional connectivity, and effective connectivity [53, 54]. The first provides information about temporal correlations between events in different areas of the brain that can be spatially far, the latter explores direct influences that different areas of the brain have on each other. Most of the measures applied to the reconstructed networks of the brain are obtained with the aforementioned techniques and aimed at one of four objectives: revealing functional segregation or integration

of information flows, highlighting small-world properties of different areas of the brain, and exploring the brain's resilience against failure[55, 56].

The second big area of research in which network theory had an important role is that of genomics. In humans, only about 20% of DNA contains genes that directly encode proteins, the building blocks of our bodies. The remaining genes play a complex role in the organism, often being responsible for regulating other genes' expression or suppression [57]. This complex interaction of genes is called the gene regulatory network and can vary in complexity from one organism to another [58], and understanding how these interactions work can be pivotal in fighting a vast number of diseases [59].

As a general approach, the reconstruction of a gene regulatory network can take two possible paths [60]: either researcher rely on a physics-based model, which describes how known processes play together in the network, or they allow an influence model, which simply describes the network in terms of statistical properties, without providing an exact physical explanation. We intuitively understand that physics-based models are specific to the biochemical elements involved [61–63], whereas influence models can provide generalized tools that may be used in other fields, as we will explore in the following sections of this chapter. To represent such networks, it is common to represent genes, proteins, or other metabolites as nodes in a graph, whereas interactions, reactions or influences are represented as edges [64]. Model architectures [65] of gene regulatory networks can be separated by studying the activity levels of each component, the type of relationship between each constituent, such as directed or undirected, and the type of model, such as dynamic or static.

Reconstruction in social networks

Thanks to the diffusion of the world wide web, social networks have had a disruptive effect on our society. Their diffusion in the population is undoubted, and their effects are avidly being studied [66–71]. However, privacy is still a concern when analyzing such collections of information, and for that reason many companies do not fully disclose the data on their users. In this domain, network reconstruction can therefore be seen as a successful attack on these data. Furthermore, many social networks are inherently mutable, therefore the accessed data may not reflect a previous stage of the network or fully capture the ongoing interactions among users.

One unexpected application of network reconstruction in social networks is that of connection recommendation. As users join a new platform, they may have some friends already

enrolled that they would like to know about. In this case, predicting a missing link can improve the user experience, and a variety of techniques have been put forward to reach that goal. Some approaches include leveraging homophilia among users [72], learning methods with feedback [73], or applying machine learning techniques [74, 75].

Furthermore, link prediction is of interest for social networks as it can be used by researchers to infer the complete network of which the data only portrays a portion [76]. The partiality of a social network can be due to its intrinsic dynamical nature, as new users join the platform and others leave it, new connections are formed and some severed. To the end of analyzing these processes, numerous approaches have been proposed, some with strict mathematical formulation taking into account the main properties of the network [77], others relying more on local topological features [78]. Understanding the evolutionary dynamics of a graph can also be achieved by mining the features of a graph and propagating them over time with a mathematical approach [79], or with semi-supervised machine-learning techniques [68].

Reconstruction with models and attributes

One of the most intuitive approaches to network reconstruction is to look at the properties and measures on links and nodes. For example, by analyzing centrality measures [80, 81] or other direct measures [82], one could identify deviations from common measurements and infer the hidden presence of nodes or links in a network. Simple approaches like this may be intuitive and informative but may lack the ability to be generalized onto different sets of problems or to properly work in presence of noise and complex dynamics. A practical approach in real-case scenarios is that of relying on metadata [83] connected to the nodes to extract information about the missing links. Although not always implemented in models, metadata are all the ancillary information that can be extracted when analyzing data from real cases such as social networks or face-to-face interactions.

Face-to-face interactions are very difficult to study, because of the need for innovative hardware to properly record the events taking place in the real world [38]. If the problem of collecting data is overcome, one could reconstruct the network by the characteristics of the participants [84], by comparing and integrating with other sources of information [85], or by defining characteristic times of interaction [38]. In either case, ground truth can be extracted by prior knowledge of the participants [86] or by asking the participants to fill out questionnaires [87]. All these methods are very case-specific, but they have an immediate match with the underlying network of interaction.

Reconstruction in deterministic dynamics and time-series

An element that appears in many areas of the literature [20, 88–92] is that in order to unveil the hidden structure of a network, there has to be a dynamical process evolving over time on it. This assumption may seem trivial, but as the idea of recovering hidden variables needs to be supported by the addition of an equation for an algebraic system, so it is in network reconstruction.

We need to reconstruct networks not only to understand static structures, but often we need to match the underlying structure with a dynamical process happening on it. With this aim, techniques have been developed to leverage the temporal data and the information carried by the dynamic elements of the system to infer the structure of the network. Leveraging the dynamics often means having a large set of assumptions on the structure and process under study [93], and some intrinsic limitations [94]. These assumptions can be on the dynamical process itself, or on the network properties and their mathematical representation [89]. Once the boundaries and rules of the reconstruction process are set, techniques can be developed to properly reconstruct the fine connectivity details of the system.

One way of using the dynamical properties of a system to infer its structure is that of perturbing a well-known dynamical process [88, 90, 95], deviations from the stable state propagate through the system in a way that can help discover the connections we are looking for. Although the assumption of a stable state is widely used for network reconstruction, it is not strictly needed and instead transient states can be explored [96, 97], as in certain conditions their behavior can shed a light on the structure of networks.

Among the most common dynamical process studied on networks there is that of epidemics, which have been at the center of the field of research of network theory for a long time [98]. The knowledge on epidemic dynamics can be used to reversely infer the properties of the network, even in presence of scarce data [30, 99]. Furthermore, one could go beyond the dynamical properties of the epidemic to leverage directly the statistical properties of diffusion dynamic, as we did contributing to the development of a technique that allows for network reconstruction only with the pairing of probabilistic distribution within an epidemic scenario [27].

In a similar way, time-series of nodes' states can be informative of the connectivity structure under investigation. A great number of techniques have been developed to study and understand time-series correlations in different scenarios [100–102]. Those same techniques can be applied to dynamical processes happening on networks, and correlations interpreted as the presence of links in a network [103]. Time-series expressing Boolean states [92, 104] can be interpreted

and used intuitively to address simple epidemic scenarios. When analyzing time-series, one could even explore causality between events [105], and as such apply it to dynamical systems.

Model-free reconstruction

The last approaches to network reconstruction that we are going to explore are those that make no assumptions about the structure of the system or data, and are therefore called model-free approaches. Such approaches are useful if only a few details are known about the system under study, and they allow for a blind analysis of networks that may express unexpected and counter-intuitive structures.

One such approach is built upon the stochastic block model [106] and Bayesian probability [107]. By using non-parametric Bayesian inference for network reconstruction [108, 109] it is possible to evaluate a posterior probability of network reconstruction even without any kind of primary error estimate. These methods are therefore extremely flexible and useful and can be used also for jointly estimating communities and other modular properties of a network [8].

In the previous section we briefly touched on the idea of inferring causality among time-series, but the core techniques described previously can be furthermore extended to explore and reconstruct whole networks [110]. By leveraging Granger Causality [111, 112], conditional mutual information [113] and transfer entropy [114, 115], inferences can be made among the elements of a network such that the reconstruction is the pure result of detected entropy changes in the system [116]. Analyzing and measuring entropy means to approach the problem of network reconstruction from an information-theoretic perspective [117], an idea that has been successful in different areas of research [118, 119]. The information-theoretic approach does not need any assumption on the model and has been shown to be capable of not only estimating correlations, but also causality among the elements under study [120, 121]. In this aim, we applied with novelty the tools of transfer entropy to an analysis of sentiment dynamics and contributed in exploring the geographical diffusion of opinions as perceived in social media [122].

Finally, some works are focusing on reconstructing networks using compressed sensing [123–125]. With compressed sensing, signals can be reconstructed with fewer samples than required by Nyquist-Shannon sampling theorem [126, 127], withstanding two assumptions: sparsity and incoherence of the signal. This technique relies on the sparsity of networks under study and does not need the system to be in a steady state, all properties that differentiate this approach from the others previously seen, with interesting results [97, 128–130]. The

bottleneck of this technique is that the amount of data needed to fully reconstruct the number of links insisting on a node is proportional to the number of links, which can be a problem in the presence of hubs or in scale-free networks.

1.3 Models of networks

In this section, we will show some of the most important models of networks, and their graph representation [131]. Generally, we will refer to a network as the representation of the system under study, it being a physical set of elements and their interactions, or a schematic depiction of interconnected concepts. We usually call nodes the representation of entities, and links the representation of interactions among nodes. The graph associated with a network will be its mathematical representation, composed of vertices, in place of nodes, and edges, in place of links.

Basic concepts

A graph [132] is an ordered pair $G = (V, E)$ where V is the set of all vertices and E is the set of edges.

Generally we can say that the set of edges is $E \subseteq \{\{i, j\} | i, j \in V\}$. If we add the condition that $\{i \neq j\}$ we say that no self-loop is present, meaning a vertex cannot be connected to itself. A graph is said to be undirected if the set $\{i, j\} | i, j \in V$ is unordered, otherwise the graph is directed.

Different mappings of $E \mapsto \mathbb{R}$ or $V \mapsto \mathbb{R}$ can exist. These mapping may represent different properties and qualities of vertices and edges. Notably, when one mapping of $E \mapsto \mathbb{R}$ exists, we say that the associated graph is weighted.

Moreover, we could have that a pair $\{i, j\}$ is present more than once in E. If more than one instance of $\{i, j\} | i, j \in V\}$ is present in E, then $G = (V, E)$ is a multigraph.

Static Networks

Statics networks are those directly derived from graph theory, and they represent a network as an object that does not evolve in time, of which here we present some of the most prominent models in the field.

Erdös-Rényi model

The first model we are going to describe is the so-called random graph [133]. It is an undirected graph $G = (V, E)$, with a fixed number of nodes N, and where the edges are chosen randomly from the $\binom{N}{2}$ possible in the graph if the edges have an equal probability of being realized. If we define the parameter p for the probability of realizing a single edge, the expected number of edges will be $p \cdot \binom{N}{2}$. Because of the importance of this parameter, this class of graphs is often represented with the notation $G = (N, p)$, to emphasize the probabilistic realization of each possible edge. Each graph $G = (N, p)$ has a binomial likelihood of having exactly e edges realized,

$$\mathbb{P}(|E| = e) = \binom{\frac{N(N-1)}{2}}{e} \cdot p^e \cdot (1-p)^{\frac{N(N-1)}{2}} - e \tag{1.1}$$

In their work, Erdös and Rényi explored in detail all the properties of the networks $G = (N, p)$, for $p = [0, 1]$. The key to the asymptotic behavior of this class of graphs is the parameter $\lambda = p \cdot N$. A phase change is noted for the system when $\lambda = 1$. When this threshold is overcome a giant connected component emerges in the graph, whereas below it only small disconnected components are present. In detail:

- If $\lambda < 1$ no connected component is present in the graph that is bigger than $O(\log N)$, for $N \to \infty$
- If $\lambda = 1$ then the largest component will be of size $O(N^{2/3})$, for $N \to \infty$
- If $\lambda > 1$ a giant connected component will be present, where other components will have less then $O(\log N)$ nodes, for $N \to \infty$.

In this model, the average edges incident on a vertex is similar for all vertices, as the probability of each edge to be realized is equal. Although this property of the Erdös-Rényi graph is rarely found in real networks, the simple mathematical formulation, and the interesting phase-change, have made for a great body of research on this class of graphs [134–136].

Small-world model

In section 1.1 we talked about one of the revolutionary models in network theory, proposed by Watts and Strogatz [5]. This model is of interest because it evolves drastically the properties of regular graphs, via the integration of a simple mechanism. The author starts from a regular lattice of $|V| = N$ vertices, each one having exactly k edges insisting on itself. Then, with

probability p each edge gets rewired, meaning that one of the ends (i, j) of the edge gets substituted with another vertex l, drawn randomly from all vertex in the network, avoiding self-loops. Of the $\frac{k \cdot N}{2}$ edges in the network, $p \cdot \frac{k \cdot N}{2}$ will be originating from this rewiring process, and $(1-p) \cdot \frac{k \cdot N}{2}$ will be left from the original lattice structure. Intuitively we can expect the lattice structure to be disrupted by the rewiring process, introducing new properties in the graph. Random rewiring will connect otherwise distant parts of the graph lattice, effectively shortening the distances among vertices.

In the limit case of $\lim p \to 1$, the obtained small-world graph shows the same properties of an Erdös-Réyni graph having the same number of vertices and a parameter $\lim p \to \frac{k}{N-1}$. If the small-world graph has $\lim p \to 0$ the graph is a regular lattice. The greatest improvement brought by the Watts-Strogatz model is that two properties of the graph, the clustering coefficient and the average path length, vary in a specific way following parameter p.

For a simple lattice the average path length $l(0)$ is

$$l(0) \approx \frac{N}{2 \cdot k} >> 1 \tag{1.2}$$

For a Erdös-Réyni model the average path length $l(0)$ is

$$l(0) \approx \frac{\ln N}{\ln k} \tag{1.3}$$

In the Watts-Strogatz model, we have that the average path length $l(0)$ falls sharply for a small increase of p while keeping a relatively high clustering coefficient and a somewhat regular structure. This property is remarkable, as we said before, because it could explain the 'small-world' effect we experience in a society, where groups apparently disconnected share an element connecting them to one another [137].

Barabási-Albert model

The scale-free model proposed by Barabási and Albert, that we already introduced in section 1.1, takes a different approach to graph modeling, as it introduces a growth technique that incorporates some of the qualities of a vertex. This technique has some similarities with the generative process proposed by De Solla Price in 1965 [138], used to explain the growth of scientific citations. The process of growing a network is obtained by defining an algorithm that works in different steps, starting from $|V| = 0$ to $|V| = N$. Defining k_i as the degree of vertex i,

and m_0 as the maximum number of edges that can be created at each iteration, the network is initialized with a fixed number of vertices, usually one. At each interaction, a new vertex is added, until all the N vertices are in the graph. Each time a vertex l is added, $m \leq m_0$ edges are created, each one stemming from the newly introduced vertex and insisting on another vertex i already in the graph at that iteration. The probability that vertex l connects to a node i is

$$p_i = \frac{k_i}{\sum_j k_j} \tag{1.4}$$

where k_i is the degree of a vertex i present in the graph and j are all the nodes present at that iteration, each one with degree k_j. This model generates a so-called 'rich-get-richer' dynamic, where vertices with already a lot of edges are more likely to get even more.

The most interesting property of the graph obtained by following this model is that the degree distribution follows a power-law

$$P(k) \sim k^{-3} \tag{1.5}$$

This result is extremely important, as it is a property commonly found in many real-life networks [139–143], an achievement not previously reached with such simplicity from other models.

Temporal Networks

Static networks can be characterized by a variety of measures, based on the connections between neighboring nodes, sets of nodes, or related metadata on nodes and links. When we add the degree of freedom of time into the picture, some of these measurements need a rethinking. Although classical measures can be adapted and still be meaningful over aggregated representations of temporal networks, some properties of the system directly rely on the order of appearance of nodes and links, therefore needing a deeper reformulation. Taking as an example the path that connects two nodes across a network, we can immediately see that adding the dimension of time to the order in which we cross one or more of the links of this path drastically changes the reachability of one node to the other, as many properties cease to be symmetrical.

Time-respecting paths assess that only some vertices can be accessed by others within a specific time-window $t \in [t_0, T]$. The set of vertices that can be reached from a vertex i is

named the set of influence of i [144]. This property becomes pivotal when studying diffusion dynamics on networks, such as epidemic spreading, where an agent can pass the infection onto another only after having contracted the infection itself. Conversely, we define the source set of i as the set of vertices that, through time-preserving paths, can reach vertex i [145].

Recognizing that the literature on static networks is many times larger than that on temporal networks, one approach that has been developed is to reduce or derive static graphs from temporal networks. Of such approaches two are the most used, the first is to derive a different static graph for each modification of the temporal network, such as the addition or deletion of a node or a link [146]. Such an approach may be seen as technically akin to studying persistent homologies in time [147, 148]. The second approach is to define a characteristic time over which the contacts that occurred in the network are aggregated and therefore represented as static graphs [149]. Similarly one could also try to aggregate temporal dynamics to static networks [150], but the loss of information and the deviation of the model from reality could be great. Therefore, specific models to describe temporal networks, along with specific tools to analyze them, have been explored.

Temporal exponential random graphs

Temporal exponential random graphs have been proposed [151, 152] as the temporal counterpart of exponential random graphs [153]. Just like their static counterpart, a connection with the Ising model can be found, and therefore a time-varying partition function can be defined. This may be useful either to define a reference model for measuring biases or as a generative model to use in simulations. For such a model the probability distribution function can be written as:

$$\mathscr{P}(A^t|A^{t-1},\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta},A^{t-1})} exp\{\boldsymbol{\theta}\cdot\Phi(A^t,A^{t-1}\} \tag{1.6}$$

Defining A^t as the weight matrix representation of the network at time t, and making a Markov assumption that A^t is independent of $A^1,...,A^{t-1}$. We can then specify the function $\Phi : \mathbb{R}_{n\times n}\times\mathbb{R}_{n\times n}\to\mathbb{R}^k$, the parameter vector $\boldsymbol{\theta}\in\mathbb{R}^k$ and a normalizing function Z.

Activity driven networks

Activity driven networks [154] seldom find a specific counterpart in static models, increasing our interest in them. Many of the most common models are connectivity-driven, meaning that the network's topology is at the core of them and their algorithms. Those models are well

suited to describe systems in which connections are long-lasting or, more precisely, where the timescale of the persistence of a connection is far larger than any dynamics evolving on the network. The assumption of long-lasting connections is not well suited for the study of many real-life scenarios, for example in the case of a disease spreading through face-to-face interactions [155, 156]. In such a case the timescale of the spreading dynamic and that of the contact pattern on which it is happening is comparable, and the more fitting paradigm of activity driven networks is better suited.

In activity driven networks a new parameter is introduced, the activity potential $a_i = \nu \cdot x_i$ for each one i of the N nodes in the network. The coefficient ν can be chosen so that the average number of active nodes at each time-step is $\langle x \rangle \cdot N$. The activity potential is in fact a connection probability over a characteristic time $\Delta_t = \tau_i$ equal for all nodes. The algorithm to model an activity driven network is as follows:

- at each time-step t the network G_t has N disconnected nodes
- with probability $a_i \cdot \Delta_t$ the node i becomes active, generating m links to randomly selected nodes in G_t
- at the next time-step $t + \Delta_t$ all links are deleted from G_t

Such a model is both random and Markovian, meaning that each time-step is agnostic of its past, and links are created with no preference in the choice of the node to connect to. Yet we see that the paradigm of activity potentials has drastic consequences in the resulting network, affecting the dynamics on it, and therefore providing for better explanations of complex behaviors and dynamics [11, 156–160].

Beyond classic networks

In this section, we will briefly introduce some notions of networks with more generalized definitions of edges, vertices, and graphs. These new models are interesting because they show peculiar and useful properties [148, 161–164], that cannot be found within the classical definition of networks, where an edge connects two vertices in a graph.

Bipartite networks

In bipartite networks, we have two kinds of nodes, one representing the agents, and the other representing the properties they're associated with. Such networks are useful for example to study the grouping of customers for an online platform [165, 166]. Each user will be connected to one or more products, and each product with one or more users. No links among products or customers will be present. From these networks, we can derive two different graphs: one linking customers with the same preferences, and one linking products with the same buyers. Similar examples can be generalized, but the practical implications for this model appear significant.

A bipartite graph is a triple $G = (\top, \bot, E)$, where $\top$ and $\bot$ are two disjoint sets of vertices, $E \subseteq \top \times \bot$ is the set of edges [167]. Vertices in $\top$ are called top-vertices, and those in $\bot$ are called bottom-vertices. The top-vertices degree distribution is defined as $\top_k = \frac{|t \in \top | d(t)=k|}{|\top|}$, and the bottom-vertices degree distribution is defined as $\bot_k = \frac{|t \in \bot | d(t)=k|}{|\bot|}$. For a bipartite graph a classical representation can be obtained, called the one-mode or $\bot$-projection, defined as $G' = (\bot, E')$ for $\{i, j\} \in E'$ if i and j are connected to the same top-vertex in G. A mirror definition can be made for $\top$-projection. It is interesting to note that a $\top(\bot)$-vertex induces a clique in the converse $\bot(\top)$-projection.

Multilayer networks

As before, let's start with a simple example. In our everyday life there are many people that we can consider friends: colleagues, old classmates, sports partners, and people we met online or at a bar. At first, we could assume that all these people are part of our network of friends. But due to the characteristics of the relationship we have with them, and the place and context of the encounter, the connections and interactions shared do carry a different set of information. Each context is more correlated with a specific set of friends, and rarely do we share the same exact behavior and ideas with all of the people we know. Some dynamics, like getting infected with the common cold, may be impossible if said friends connect with us mainly online or via the phone. This is only one of the examples [168, 169] that are well suited to be represented with the use of multilayer networks.

Multilayer networks [170–172] can be seen as a further generalization of the bipartite networks, and in general, of all static networks we have previously seen. But multilayer networks can also be leveraged to represent temporal networks, where the multiple grouping,

or layers, are associated with different characteristic times to be explored. A multilayer network is an ordered pair $\mathscr{M} = (\mathscr{G}, \mathscr{E})$ where $\mathscr{G} = \{G_\alpha;\ \alpha \in \{1,...,M\}\}$ is a family of graphs $G_\alpha = (X_\alpha, E_\alpha)$ each one called a layer of $\mathscr{M}$, and that can be directed, undirected, weighted or unweighted. Moreover, $\mathscr{E} = \{E_{\alpha,\beta} \subseteq X_\alpha \times X_\beta;\ \alpha,\beta \in \{1,...,M\},\ \alpha \neq \beta\}$ is the set of edges between nodes of different layers G_α, G_β when $\alpha \neq \beta$. The elements of $\mathscr{E}$ are called crossed layers, in that way there can be intra-layer and inter-layers edges. As previously defined graphs, each G_α has a set of vertices X_α and a set of edges E_α. If $X_\alpha = X_\beta$, $\forall \alpha,\beta \in \{1,...,M\}$ then the multilayer takes on the name of multiplex.

1.4 Main measures on networks

Many measures have been devised to characterize, classify and explore different networks. The most important measures concern the two main elements of the network: nodes and links. Others measurements can emerge if ensembles such as communities are taken into account, or if other dimensions, such as time or layers, are considered.

Basic concepts

The distance, or geodesic $d(i,j)$ on a graph between two vertices i,j is the smallest set of adjacent edges starting and ending on the nodes i and j. Two edges are considered adjacent if they insist on one common vertex [131].

The adjacency matrix is defined as a square matrix, A associated with a graph, of size $\mathbb{R}^{N\times N}$, where N is the number of vertices in the graph. In unweighted graphs the value of the matrix $a_{i,j} = 1$ if an edge exist $i \to j$, and zero otherwise [131].

Degree Centrality

One of the first measures to be devised and used is that of how many links insist on a node [173]. It is one of the most intuitive and simplest properties of a node in a network, yet is greatly informative. In an indirect and unweighted network that measure is usually noted with k_i for a node i having exactly k links insisting on it. This measure can be normalized by defining $k'_i = \frac{k_i}{N-1}$ where N is the number of nodes in the network. More comprehensive information on the network is obtained from the degree distribution $P(k)$, which represents the fraction of

nodes in a network having degree k. It has been proved that real-life networks follow specific degree distribution [6], making this first measure one of the most important in network theory.

Moreover, the study of the degree distribution can go into the correlation between the degrees of different vertices. This is usually expressed using the joint degree distribution $P(k,k')$, that is the probability of an arbitrary node with degree k connecting with a node of degree k', and has led to some practical analysis [153, 174] and theoretical advances [175, 176] in the field.

Degree centrality can also be defined for directed networks [177]. Being edges in directed graph ordered pairs of vertices, we have that the out-degree for a node i is the number of edges having i as the first element, usually written as $k_{i,out}$. The number of edges having i as the second element is the in-degree, written as $k_{i,in}$.

Closeness centrality

The closeness centrality [178] is defined as the inverse sum of geodesic distances to every other vertices from each vertex within the network. It is $C_i = \frac{1}{\sum_{j \in V} d(i,j)}$ where V is the set of vertices in the graph and $d(i,j)$ is the geodesic distance between i and j on the graph. This measure has been extended fruitfully on weighted graph thanks to Dijkstra [179] and applied in different optimization and mobility problems [180, 181].

Betweenness centrality

This measure has been defined [173, 182] to highlight the importance of a vertex not for the number of edges but the quality of those connected to it. In particular, the betweenness is higher for vertices that more often are in the shortest path between all other vertices. Given a graph $G = (V,E)$ the definition of betweenness centrality is $B_i = \sum_{i \neq j \neq l \in V} \frac{\sigma_{j,l}(i)}{\sigma_{j,l}}$ where $\sigma_{j,l} = d(j,l)$ is the geodesic distance and $\sigma_{j,l}(i)$ is a geodesic distance in which two edges insist on i.

Betweenness centrality can also be computed as a measure on edges [183, 184]. It is defined as the sum of fraction of all shortest-paths that pass trough and edge $e \in E$, specifically $B_e = \sum_{i \neq j \in V} \frac{\sigma i,j(e)}{\sigma_{i,j}}$. As before $\sigma_{i,j}$ is a geodesic distance, and $\sigma_{i,j}(e)$ is a geodesic distance that contains the edge e.

Eigenvector centrality

Based on the adjacency matrix, a new measure of centrality is developed [185] on the computation of matrix eigenvalues and eigenvectors. Being A the adjacency matrix of a unweighted graph with elements a_{ij}, The eigenvector centrality of node i is nothing else than $x_i = \frac{1}{\lambda}\sum_j a_{ji}x_j$, for any vertex j in the graph. The main concept behind this centrality measure is to give importance to a node if it is connected to important nodes. Even if more mathematically complex, this centrality has seen theoretical development [186] as well as practical use in the analysis of real-world scenarios [187, 188].

K-shell centrality

The k-shell centrality focus on the concept of ranking vertices to the membership of sub-graphs derived from the main graph under study [189]. Starting from the lowest degree present in the graph G, k_{min} we remove all vertices i where $k_i = 1$, and assign to the remaining vertices in the subgraph G' the k-shell centrality $KS = 1$. Then we proceed iteratively, removing all vertices where $k_i = 2$, and assigning to the remaining vertices in the subgraph G'' a rank $KS = 2$. This process can highlight important nodes, as well as communities, in classic graphs [190] or in weighted networks [191].

Eccentricity

The eccentricity [192] of a vertex i is the largest shortest-path between it and any other reachable vertex in the graph $C_e(i) = \frac{1}{max_j d(i,j)}$. The smallest eccentricity of a graph defines the quantity called the radius of the graph, whereas the biggest eccentricity is called the diameter of the graph [193].

1.5 Thesis contribution

This thesis aims to approach the problem of network reconstruction from different perspectives, contributing to the field with different targeted contributions. Approaching the problem from different angles allows for a wider study of the problem, which benefits the training of the candidate, and explores flexible solutions that could benefit from the implementation of heterogeneous techniques. Research questions in different areas of study of the same field can be at

times tackled thanks to the inherent interconnectedness of the scientific approach, with a deep understanding of the experimental phase, the analysis of data and the theoretical formulation. This document is split into four parts, to better guide the readers into the work.

Starting in part I from the theoretical approach to the problem, in chapter 2 we define a technique to differentiate between weak and strong links in activity driven networks, being able to identify the backbone structure of them; in chapter 3 we present a technique to infer whether hidden nodes activity driven networks are designed, starting from the observation of an epidemiological event unraveling on the system.

In part II we explore the tools of data-analysis for network reconstruction, working in chapter 4 with a large data-set of online posts used to reconstruct opinion dynamics in the United States during the recent pandemic

Then, part III is dedicated to an experimental platform to collect network formation data toward reconstruction; in chapter 5 we describe the development and characterization of two smartphone applications that, leveraging Bluetooth® technology, have the potential to make face-to-face experiments within reach of many more researchers in different fields.

Finally, we draw our conclusions in part IV.

Part I

A theoretical approach

Chapter 2

Backbone reconstruction in temporal networks from epidemic data

In this chapter, we start approaching the problem of reconstructing the nature of links in a network from a theoretical point of view. In order to do that, we first have to define the scope and breadth of our approach. As we described before, activity driven networks [154] present a unique set of features and properties that are well suited to describe real phenomena [23]. Being a relatively new model to represent time-varying networks, very few tools are available to explore and analyze activity driven networks, therefore we decided to develop ourselves those we need. The main goal of our work is focused on how to tackle corrupt or noisy information, such as that originating from an experiment or a real-life data-set, to reconstruct all the actors and the interactions that generated such data.

We decided to start our research by approaching the problem of reconstructing unknown links in a graph. Many complex systems are characterized by time-varying patterns of interactions. These interactions comprise strong ties, driven by dyadic relationships, and weak ties, based on node-specific attributes. The interplay between strong and weak ties plays an important role in dynamical processes that could unfold on complex systems. However, we rarely have access to precise information about the time-varying topology of interaction patterns. A particularly elusive question is to distinguish strong from weak ties, on the basis of the sole node dynamics. Strong ties represent the dyadic interactions that followed an underlying structure of the network, called backbone, whereas weak ties represent the random connections that may happen in real scenarios. Building upon rigorous analytical results, we explore a statistically-principled algorithm to reconstruct the backbone of strong ties from data

of a spreading process, consisting of the time-series of individuals' states. To validate our approach, we validate it numerically, over a range of synthetic datasets, encapsulating salient features of real-world systems. In real-life scenarios, a spreading process can be associated with an infodemic or an epidemic, and having the ability to reconstruct the quality of interactions enables the researchers to better recognize the structure of the underlying network of contacts. Better knowledge of the underlying interactions could, in turn, inform better-designed policies, to prove this we propose the integration of our algorithm in a targeted immunization strategy that prioritizes influential nodes in the inferred backbone. Through Monte Carlo simulations on synthetic networks and a real-world case study [194], we demonstrate the viability of our approach.

2.1 Background

In the last few decades, network science has experienced significant developments, providing researchers with an array of powerful tools to represent and analyze complex biological, social, and technological systems [195]. Besides improving our knowledge of the very structure of complex systems, network science has contributed new paradigms to study dynamical processes unfolding on a complex system. These paradigms have shed light on the intertwining between structure and dynamics in the spread of epidemic diseases [196], diffusion of innovation [197], and opinion formation [198].

Empirical studies suggest that patterns of interactions between nodes in many complex networks evolve ceaselessly in time [16, 199]. These interactions can be categorized into two main classes [2]. One class corresponds to interactions that are recurrently formed between node pairs, following dyadic relationships that are called *strong ties* [200]. Interactions in the workplace or family ties belong to this class, which forms the *backbone* of the network [201, 202]. The second class encompasses interactions that are based on features of the nodes, which are not attributable to dyadic ties with other nodes. For instance, interactions among people queuing in a line or sitting on a plane belong to this class, whereby interactions are triggered by individual attributes such as extroversion in talking to strangers. These relationships are called *weak ties* [200]. Strong and weak ties concur in shaping the dynamic behavior of complex networks [203–205].

Activity driven networks (ADNs) have emerged as a valuable framework for temporal networks [154], allowing for modeling the co-evolution of the network structure and the unfolding nodal dynamics at comparable time-scales. The temporal nature of the network is

captured through a single parameter that measures the node propensity to generate interactions. The distribution of this parameter, called *activity*, can be inferred from real-world data [154]. The potential of ADNs has been demonstrated through the study of several network problems, including epidemics [147, 155, 156, 12, 206], diffusion of innovation [157], opinion formation [159], and percolation [207].

In their fundamental incarnation, ADNs are an ideal tool to model weak ties, whereby the whole process of network assembly is driven by a node-specific attribute, the activity. Routed ADNs (RADNs) have been recently proposed to include strong ties within the ADN paradigm [208, 209]. In this model, temporal connections are wired according to a stochastic rule that encapsulates both the topological information of strong ties and the unstructured connections of weak ties. RADNs share similarities with other approaches to include strong ties in ADNs, such as the superimposition of a static network [11, 210], and the inclusion of memory mechanisms in the link wiring process [211, 212].

The use of RADNs in real-world scenarios relies on accurate knowledge of the activity distribution and the topology of the backbone. While activities can be estimated following the literature on ADNs [23, 154], the inference of the backbone of strong ties remains an open challenge. Preliminary efforts in this direction can be found in [213]. Therein, the authors have proposed a method to reconstruct the backbone of a temporal network from the direct observation of the pattern of interactions over an accessible time-window. Particularly elusive is the problem of distinguishing strong from weak ties from observations of node dynamics, which is typically the only knowledge available in real epidemiological settings [28].

In the technical literature, the problem of link reconstruction and prediction has been studied from a variety of angles, mostly relying on the direct observations of contacts [93, 214, 215]. Dealing with observations of nodal dynamics, several methods have been proposed to reconstruct patterns of interactions [94], including the use of similarity [216], information theory [217], belief propagation [30], likelihood maximization [218], compressed sensing [130, 219], optimization [19], nonparametric Bayesian methods [8], and data-driven approaches [92, 220]. However, these strategies are of limited use when strong and weak ties coexist, thereby presently challenging the inference of backbone networks from observations of node dynamics.

Drawing inspiration from [164, 221], here we design a backbone detection algorithm that identifies strong ties from node dynamics, in the form of empirical data about a spreading process. Because of its widespread use in the study of epidemic outbreaks, we adopt the epidemiological lexicon throughout the work when referring to the spreading dynamics. However, the application of our algorithm should not be considered limited to the epidemiological

field, since spreading processes in temporal networks are widely used to model other phenomena, including diffusion of innovation in social groups [157] and information flow in brain networks [222–224].

2.2 Mathematical foundation

Our algorithm is based on the intuition that strong ties should leave a distinguishable footprint on the temporal evolution of an epidemic outbreak. We analytically characterize such a footprint in terms of the probability for a node to contract the disease, given knowledge about the health state of other nodes. In this section, we provide mathematical details of the models herein used to study temporal networks with a backbone structure of strong ties, along with the dynamical process.

2.2.1 Routed ADNs

We consider a network of n nodes, each belonging to the node set $V = \{1, \dots, n\}$. Temporal undirected links are represented through time-varying adjacency matrix $A_t \in \{0,1\}^{n\times n}$, where $t \in \mathbb{Z}_+$ is the discrete time index. The adjacency matrix is assembled so that $(A_t)_{ij} = 1$ if and only if node i is connected with node j at time t. We denote by N_t^i the neighborhood of node i at time t, that is, the set of other nodes to which i is connected at time t.

Both strong and weak ties contribute to the evolution of A_t. Strong ties are described by an undirected and time-invariant adjacency matrix $G \in \{0,1\}^{n\times n}$. We indicate with d_i the degree of node i in the backbone network. Degrees are gathered in the degree vector $d \in \mathbb{N}^n$. Empirical evidence from real-world observations suggests that real-world backbones are often sparse [195] and nodes have bounded degree [225]. Without loss of generality, we assume that the backbone network does not contain isolated nodes, that is, $d_i \geq 1$, for all $i \in V$ [1].

Following [209], each node $i \in V$ is characterized by an activity parameter $a_i \in [0,1]$. At each time, node i activates with probability a_i and generates an undirected link with another node. The selection of which node to connect to is probabilistically dictated by a

[1] Similar to [209], the assumption $d_i \geq 1$, for all $i \in V$, can be removed with a slight modification of (2.1).

row-stochastic [2] matrix $P \in \mathbb{R}^{n\times n}_{\geq 0}$ such that

$$P = (1-\gamma)\frac{1}{n-1}J + \gamma \mathrm{diag}(d)^{-1}G, \tag{2.1}$$

where $\gamma \in [0,1]$ is a constant parameter and J is the $n \times n$ matrix of all ones, except the diagonal entries, which are set to 0. The generic entry P_{ij} represents the probability that i connects with j. The first term on the right-hand side of (2.1) accounts for the weak ties, while the second summand models strong ties in the backbone. The parameter $\gamma \in [0,1]$ weights the role of strong versus weak ties in the formation of temporal links. When $\gamma = 0$, the model reduces to a standard ADN [154] such that strong ties are uninfluential; when $\gamma = 1$, the probability of a connection mirrors the adjacency matrix of the backbone network. A realization of an RADN is shown in Fig. 2.1.

[2] A matrix is said to be row-stochastic if it is nonnegative (entrywise) and each row sums to 1.

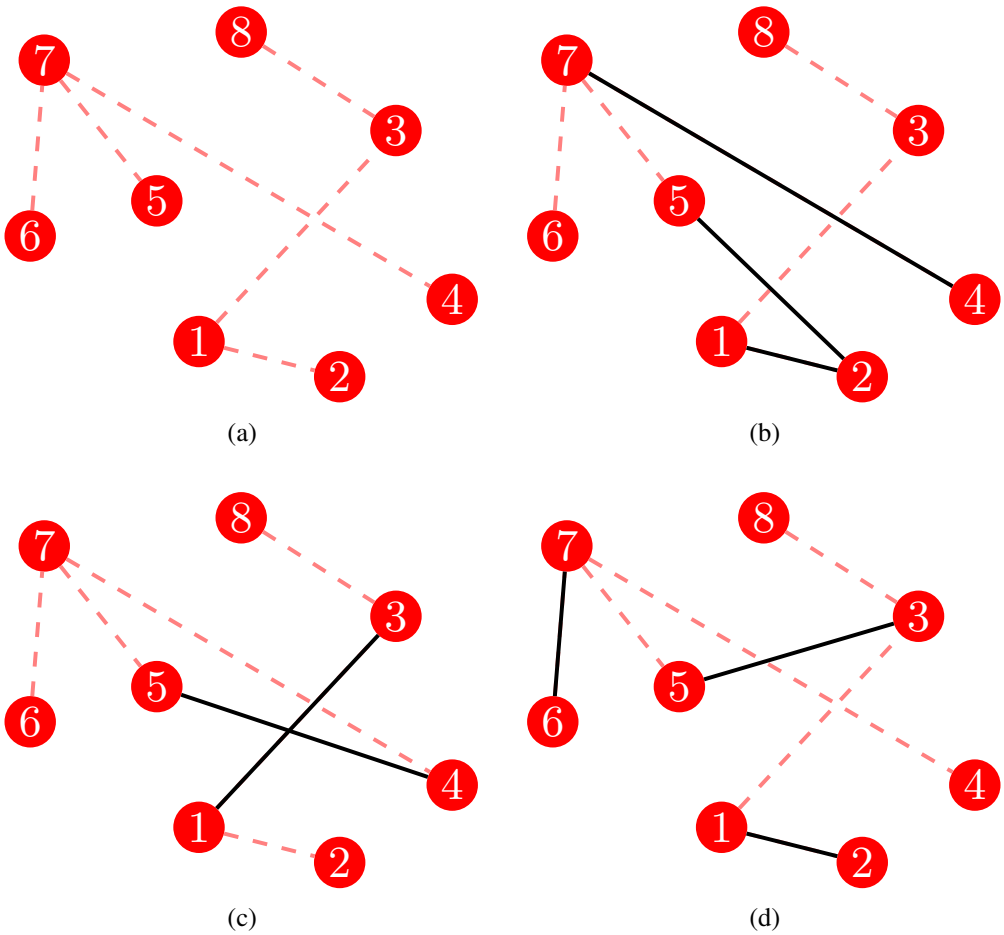


Fig. 2.1 Illustration of a backbone network (a) along with three consecutive realizations of an RADN (b–d) at time $t = 0, 1, 2$, respectively. Red dashed links are the strong ties in the backbone, and black solid links are temporal links generated from nodes' activity.

To generate a temporal network from $t = 0$, up to time T, we implement the following steps:

1. the temporal adjacency matrix is initialized as $(A_t)_{ij} = 0$, for all $i, j \in V$;
2. each node $i \in V$ activates with probability a_i, independent of the others;
3. for each node i that is active, a node j is selected with probability P_{ij}, and we set $(A_t)_{ij} = (A_t)_{ji} = 1$; and
4. the time index t is incremented by 1; if $t \geq T$, the algorithm is terminated, otherwise it is resumed to step 1.

2.2.2 Susceptible–infected–susceptible model

We focus on a susceptible–infected–susceptible (SIS) epidemic model [226]. In an SIS model, each node of the network is characterized by a binary health state. Specifically, at time t, node $i \in V$ is either susceptible to the disease ($X_t^i = 0$) or infected ($X_t^i = 1$). At each time, two contrasting mechanisms govern the evolution of the epidemic process: propagation and recovery. Each susceptible node can contract the disease through interactions with infected nodes.

The propagation of the disease may occur with probability $\lambda \in [0,1]$ along each link of the RADN independently of the others, such that

$$\mathbb{P}(X_{t+1}^i = 1 \,|\, X_t^i = 0) = 1 - (1-\lambda)^{\sum_{j \in N_t^i} X_t^j}. \tag{2.2}$$

Following the recovery mechanism, instead, each node i that is infected at time t, recovers at time $t+1$ with probability $\mu \in [0,1]$, becoming again susceptible to the epidemics. The generality of our theoretical approach suggests that our algorithm could be extended to more complex epidemic models on ADNs [23, 227].

2.3 Backbone detection algorithm

We present here the main technical innovation, which consists of an algorithm to detect the backbone of strong ties in a temporal network from epidemic data. Our method is based on the exact computation of the probability of a node to contract the disease given the health states of other nodes. Building on the knowledge about neighbors, we are able to pinpoint the effect of the presence of strong ties through a statistical test.

2.3.1 Conditional probabilities for RADNs

Given two nodes, i and j, observed from the initial time 0 over a time-window of duration T, we define the following quantity:

$$\begin{aligned} \mathscr{P}_{j\to i} \;:=\; & \frac{1}{T}\sum_{t=0}^{T-1}\Big[\mathbb{P}(X_{t+1}^i = 1 \,|\, X_t^i = 0, X_t^j = 1) \\ & -\mathbb{P}(X_{t+1}^i = 1 \,|\, X_t^i = 0)\Big]. \end{aligned} \tag{2.3}$$

The quantity $\mathscr{P}_{j\to i}$ summarizes the extent by which the infection of node i over the time-window $0,\dots,T$ is explained by the disease propagation from node j [3]. Intuition suggests that such a quantity is larger when i and j are connected by a strong tie, such that the infection of nodes connected by the backbone network will increase the chance of contracting the infection.

Let's compute the infection probability for node i at time instant t, for either the case in which we include or exclude knowledge about node j. Let $x_1,\dots,x_n$ be the state of the system at time t, then the RADN model indicates that

$$\mathbb{P}(X_{t+1}^i = 1 \,|\, X_t^i = 0) = 1 - \prod_{k\in V\setminus\{i\}} (1-\lambda a_i P_{ik} x_k)\,(1-\lambda a_k P_{ki} x_k)\,. \tag{2.4}$$

Upon conditioning on $X_t^j = 1$, we factor the term associated with j out of the multiplication to obtain

$$\begin{aligned}\mathbb{P}(X_{t+1}^i = 1 \,|\, X_t^i = 0, X_t^j = 1) = 1 - \left(1-\lambda a_i P_{ij}\right)\left(1-\lambda a_j P_{ji}\right) \\ \times \prod_{k\in V\setminus\{i,j\}} (1-\lambda a_i P_{ik} x_k)\,(1-\lambda a_k P_{ki} x_k)\,.\end{aligned} \tag{2.5}$$

First, we consider the case in which nodes i and j do not share a strong tie, that is $G_{ij} = G_{ji} = 0$. In this case, from (2.1) we derive $P_{ij} = P_{ji} = (1-\gamma)/(n-1)$. We substitute P_{ij} and P_{ji} in (2.4) and (2.5), and we compute the limit for $n\to\infty$ of their difference as

[3] In principle, for an arbitrary network model, this quantity might also attain negative values.

$$
\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}(X^i_{t+1}=1\,|\,X^i_t=0,X^j_t=1)-\mathbb{P}(X^i_{t+1}=1\,|\,X^i_t=0)\\
&\quad=\lim_{n\to\infty}\Bigg[\left(1-\frac{\lambda(1-\gamma)a_i x_j}{n-1}\right)\left(1-\frac{\lambda(1-\gamma)a_j x_j}{n-1}\right)\\
&\qquad-\left(1-\frac{\lambda(1-\gamma)a_i}{n-1}\right)\left(1-\frac{\lambda(1-\gamma)a_j}{n-1}\right)\Bigg]\\
&\qquad\times\prod_{k\in V\setminus\{i,j\}}(1-\lambda a_i P_{ik}x_k)\,(1-\lambda a_k P_{ki}x_k)\\
&=\lim_{n\to\infty}\left(\frac{\lambda(1-\gamma)(a_i+a_j)(1-x_j)}{n-1}-\frac{\lambda^2(1-\gamma)^2 a_i a_j(1-x_j)}{(n-1)^2}\right)\\
&\qquad\times\prod_{k\in V\setminus\{i,j\}}(1-\lambda a_i P_{ik}x_k)\,(1-\lambda a_k P_{ki}x_k)\\
&\qquad\le\lim_{n\to\infty}\frac{\lambda(1-\gamma)(a_i+a_j)}{n-1}=0.
\end{aligned}
\tag{2.6}
$$

We note that (2.6) is the generic summand of $\mathscr{P}_{j\to i}$ in (2.3), from which the claim in (2.10b) follows. We further observe that each of the summands of $\mathscr{P}_{j\to i}$ is a nonnegative random variable, which is bounded from above by the estimation in (2.6). Even though these random variables are not independent and not identically distributed (since they depend on the time-series of the nodes' health state that are self-correlated) they are bounded and their correlation tends to 0 in the long-time. Hence, a central limit theorem applies to $\mathscr{P}_{j\to i}$, according to [228]. Such an observation guarantees that $\mathscr{P}_{j\to i}$ converges to a Gaussian distribution, as supported by the numerics in Fig. 2.2. However, an explicit statement of the central limit theorem cannot be readily formulated, since it requires the computation of the variance.

We now consider the case in which nodes i and j share a strong tie, that is, $G_{ij}=G_{ji}=1$. Similar to the previous analysis, from (2.1) we derive $P_{ij}=(1-\gamma)/(n-1)+\gamma/d_i$ and $P_{ji}=(1-\gamma)/(n-1)+\gamma/d_j$. Defining the neighborhood of node i in the backbone $N^i_G:=\{j\in V: G_{ij}=1\}$, we proceed specializing to the present case the difference between (2.4) and (2.5) at time t. Considering that $(1-k/x)^{x-1}\ge 1/e^k$, for any $x\ge 1$ and $k>0$, and that $d_i\le n-1$, for any $i\in V$, we compute

$$
\begin{aligned}
&\mathbb{P}(X_{t+1}^i = 1 \,|\, X_t^i = 0, X_t^j = 1) - \mathbb{P}(X_{t+1}^i = 1 \,|\, X_t^i = 0) = \\
&= \left[\left(1 - \lambda a_i x_j \left(\frac{\gamma}{d_i} + \frac{1-\gamma}{n-1}\right)\right) \left(1 - \lambda a_j x_j \left(\frac{\gamma}{d_j} + \frac{1-\gamma}{n-1}\right)\right) \right. \\
&\quad \left. - \left(1 - \lambda a_i \left(\frac{\gamma}{d_i} + \frac{1-\gamma}{n-1}\right)\right) \left(1 - \lambda a_j \left(\frac{\gamma}{d_j} + \frac{1-\gamma}{n-1}\right)\right) \right] \\
&\quad \times \prod_{k \in V \setminus \{i,j\}} (1 - \lambda a_i P_{ik} x_k)(1 - \lambda a_k P_{ki} x_k) \\
&\geq \lambda\gamma(1-x_j) \left(\frac{a_i}{d_i} + \frac{a_j}{d_j} - \lambda\gamma \frac{a_i a_j}{d_i d_j}\right) \prod_{k \in N_G^i \setminus \{j\}} (1 - \lambda a_i P_{ik} x_k)(1 - \lambda a_k P_{ki} x_k) \\
&\quad \times \prod_{h \notin N_G^i \cup \{i\}} (1 - \lambda a_i P_{ih} x_h)(1 - \lambda a_h P_{hi} x_h) \\
&\geq \lambda\gamma(1-x_j) \left(\frac{a_i}{d_i} + \frac{a_j}{d_j} - \lambda\gamma \frac{a_i a_j}{d_i d_j}\right) \prod_{k \in N_G^i \setminus \{j\}} \left(1 - \frac{\lambda a_i}{d_i}\right)\left(1 - \frac{\lambda a_k}{d_k}\right) \\
&\quad \times \prod_{h \notin N_G^i \cup \{i\}} \left(1 - \frac{\lambda(1-\gamma)a_i}{n-1}\right)\left(1 - \frac{\lambda(1-\gamma)a_k}{n-1}\right) \\
&\geq \lambda\gamma(1-x_j) \left(\frac{a_i}{d_i} + \frac{a_j}{d_j} - \lambda\gamma \frac{a_i a_j}{d_i d_j}\right) \left[\left(1 - \frac{\lambda a_i}{d_i}\right)\left(1 - \frac{\lambda a_M}{d_m}\right)\right]^{d_i - 1} \\
&\quad \times \left[\left(1 - \frac{\lambda(1-\gamma)a_i}{n-1}\right)\left(1 - \frac{\lambda(1-\gamma)a_M}{n-1}\right)\right]^{n-1-d_i} \\
&\geq \frac{\lambda\gamma}{\exp\{\lambda(1-\gamma)(a_i + a_M)\}} \left(1 - \lambda \frac{a_i}{d_i}\right)^{d_i - 1} \\
&\quad \times \left(1 - \lambda \frac{a_M}{d_m}\right)^{d_i - 1} \left(\frac{a_i}{d_i} + \frac{a_j}{d_j} - \lambda\gamma \frac{a_i a_j}{d_i d_j}\right)(1 - x_j) := F(x_j).
\end{aligned}
\tag{2.7}
$$

where a_M is the maximum node activity and d_m is the minimum degree in the backbone. The bounding function $F(x_j)$ is such that $F(1) = 0$, and $F(0) > 0$, for any $\gamma > 0$.

We now focus on the variable X_t^j. According to the SIS dynamics described in 2.2.2, X_t^j changes from 1 to 0 with probability equal to μ, while the probability of switching from 0 to 1 depends on the health state of the other nodes, according to (2.4). However, it can be bounded

from above as follows:

$$\begin{aligned}
&\mathbb{P}(X_{t+1}^i = 1 \,|\, X_t^i = 0) = \\
&\mathbb{P}\left(\bigcup_{k \in V \setminus \{i\}} \{i \text{ is infected by } k\} \right) \\
&\leq \sum_{k \in V \setminus \{i\}} \mathbb{P}(\{i \text{ is infected by } k\}) \\
&= \sum_{k \in V \setminus \{i\}} \lambda a_i P_{ik} x_k + \lambda a_k P_{ki} x_k - \lambda^2 a_i a_k P_{ik} P_{ki} x_k \\
&\leq \lambda \sum_{k \in V \setminus \{i\}} (a_i P_{ik} + a_M P_k) = \lambda \left[a_i + \left[1 - \gamma \left(1 - \frac{d_i}{d_m} \right) \right] \right] a_M .
\end{aligned} \tag{2.8}$$

Hence, the frequency of $X_t^j = 0$ converges almost surely to at least $\mu/(\lambda(a_i + (1-\gamma(1-d_i/d_m))a_M) + \mu)$ for $T \to \infty$. Hence, using (2.7) and the definition of $\mathscr{P}_{j\to i}$ in (2.3), the latter quantity can be bounded from below as follows:

$$\begin{aligned}
&\lim_{T\to\infty} \mathscr{P}_{j\to i} = \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \left[\mathbb{P}(X_{t+1}^i = 1 \,|\, X_t^i = 0, X_t^j = 1) - \mathbb{P}(X_{t+1}^i = 1 \,|\, X_t^i = 0) \right] \\
&\geq \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} F(X_t^i) \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t \in \{0,\ldots,T-1\} : X_t^i = 0} F(0) \geq \frac{\mu}{\lambda [a_i + (1 - \gamma(1 - \frac{d_i}{d_m}))a_M] + \mu} F(0) \\
&\geq \frac{\mu \lambda \gamma}{e^{\lambda(1-\gamma)(a_i+a_M)} (\lambda (a_i + (1 - \gamma(1 - d_i/d_m))a_M) + \mu)} \\
&\times \left(1 - \lambda \frac{a_i}{d_i} \right)^{d_i - 1} \left(1 - \lambda \frac{a_M}{d_m} \right)^{d_i - 1} \left(\frac{a_i}{d_i} + \frac{a_j}{d_j} - \gamma \lambda \frac{a_i a_j}{d_i d_j} \right) > 0.
\end{aligned} \tag{2.9}$$

As shown in Fig. 2.2, our bound is accurate, albeit conservative. The main bottlenecks for improving the bound are in the substitution of the random variables X_t^j with 1 in (2.7) and in the estimation of the time elapsed with $X_t^j = 0$ in the derivation of (2.9). To obtain a tighter

bound, one should rigorously compute the endemic state of an SIS model over a RADN, which is a nontrivial open problem [206].

Similar to our observations following (2.6), we should note that a central limit theorem could in principle be established here as well, since $\mathscr{P}_{j\to i}$ is a temporal average of the transition probabilities. However, the derivation of its explicit statement is not possible, since it requires the exact computation of mean and variance of the summands.

Specifically, we demonstrate that, in the asymptotic limit of large time-windows, if there exists a strong tie between i and j, that is, if $G_{ij}=1$, then

$$\lim_{T\to\infty}\mathscr{P}_{j\to i}\geq\frac{\mu\lambda\gamma\left[\left(1-\lambda\frac{a_i}{d_i}\right)\left(1-\lambda\frac{a_M}{d_m}\right)\right]^{d_i-1}}{\lambda(a_i+(1-\gamma(1-\frac{d_i}{d_m}))a_M)+\mu}\times \\ \times\frac{1}{e^{\lambda(1-\gamma)(a_i+a_M)}}\left(\frac{a_i}{d_i}+\frac{a_j}{d_j}-\gamma\lambda\frac{a_ia_j}{d_id_j}\right)>0, \tag{2.10a}$$

almost surely, for any network size, where a_M and d_m are the maximum activity and the minimum backbone degree over the node set, respectively. On the other hand, if the two nodes are disconnected in the backbone, that is, if $G_{ij}=0$, we find that in the asymptotic limit of large networks,

$$\lim_{n\to\infty}\mathscr{P}_{j\to i}=0. \tag{2.10b}$$

As a consequence, if the size of the network is sufficiently large, the probability that a node becomes infected is not influenced by the health state of another, unless they share a strong tie. Based on this analytical result, we construct our identification algorithm, which starts from empirical observations of the disease dynamics to detect strong ties.

Figure 2.2 compares the empirical estimation of $\mathscr{P}_{j\to i}$ for pairs of nodes that share (orange) or not (blue) a strong tie. These simulations validate our analytical results and suggest that $\mathscr{P}_{j\to i}$ is close to its asymptotic expressions in (2.10), also for reasonably small population size (that is, starting from $200-300$ nodes, according to our numerical simulations) and an observation window of limited duration. In fact, while the empirical distribution of the entries of $\mathscr{P}_{j\to i}$ that correspond to strong ties (in orange) is shifted and bounded away from 0, the empirical distribution of the entries that do not correspond to strong ties is centered at 0.

By comparing our analytical bound from (2.10a) (dotted red line) with the empirical observation, we propose that our estimation, albeit conservative, yields an accurate estimate of the order of magnitude of $\mathscr{P}_{j\to i}$. The two empirical distributions are well separated and both of

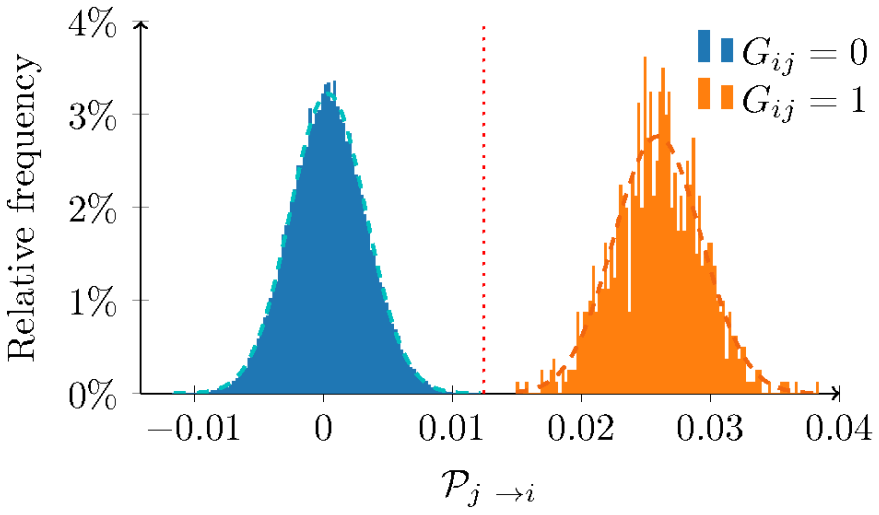


Fig. 2.2 Empirical estimation of $\mathscr{P}_{j\to i}$ in a realization of an RADN with $n = 200$ nodes, $\gamma = 0.95$, $\lambda = 0.9$, $\mu = 0.1$, and $a_i = 0.3$ for all nodes, over all the pairs of nodes $(i, j) \in V \times V$. The orange distribution relates to nodes that share a strong tie and the blue one to the opposite case. The backbone network is a 4-regular random graph. The network is simulated for $35,000$ time-steps. The figure suggests that conditioning on the state of node j affects the infection probability for nodes that share a strong tie with j, confirming our analytical results. The red dotted line is the lower bound on $\mathscr{P}_{j\to i}$ in the presence of the strong tie $\{i, j\}$, computed using (2.10a).

them can be accurately fitted by a Gaussian distribution (dashed blue and orange, respectively) with mean equal to 0.000 and 0.026, respectively, and variances both equal to 0.003. This evidence suggests that a central limit theorem should hold for $\mathscr{P}_{j\to i}$, which is defined as an average over T. As a consequence, we may conjecture that the length of the time-window T plays a key role in shaping the two distributions and, consequently, in determining whether strong and weak ties are statistically distinguishable. More details to support our conjecture can be found in Section 2.4.

We conclude this section by noting that our derivation is performed by using specific properties of the SIS epidemic model. We believe that a similar argument could be pursued to establish rigorous bounds on the transition probabilities for other dynamics, including more complex and realistic epidemics processes, or opinion dynamics, such as the voter model. In fact, the key properties of our argument is that the state transitions (from susceptible to infected) are triggered by the interactions and that they occur multiple times, due to the spontaneous recovery process. The former leaves the footprint of strong ties on the nodal dynamics, the latter affords the use of statistical tests to ensure significance to our results.

2.3.2 Statistical test

Building on our analytical results, we put forward a statistically-principled analysis to determine the presence of a strong tie between the two nodes for a network of conveniently large size. To perform such an analysis, for any pair of nodes i and j, we measure the following four quantities over the observation time-window of duration T:

- the number of time-steps in which node i is susceptible, denoted as s_i;
- the number of transitions of node i from susceptible to infected, denoted as i_i;
- the number of time-steps in which node i is susceptible and node j is infected, denoted as n_{ij}; and
- the number of transitions of node i from susceptible to infected with node j being infected at the previous time, denoted as q_{ij}.

From the first two quantities, we compute the ratio $r_i = i_i/s_i$, which measures the sampling probability that a susceptible node i at time t becomes infected at $t+1$.

According to (2.10b), if i and j do not share a strong tie, then the probability that i contracts the infection should not be influenced by j, that is, q_{ij} should be a realization of a Bernoulli trial with expected value equal to $r_i n_{ij}$. We set this as the null hypothesis of our statistical test, which is rejected if q_{ij} is significantly larger than $r_i n_{ij}$. We associate with the node pair a p-value, coming from the binomial cumulative distribution, equal to

$$\pi_{ij} = 1 - \sum_{h=0}^{q_{ij}-1} \binom{n_{ij}}{h} r_i^h \, (1-r_i)^{n_{ij}-h} \,. \tag{2.11}$$

This procedure generates a set of $n-1$ statistical tests for each node, that is, $n(n-1)$ tests, overall. Hence, a multiple comparison correction should be implemented to assess whether each one of the null hypotheses can be rejected. We adopt the Benjamini–Hochberg procedure to control the false discovery rate, which offers a less conservative criterion with respect to the standard Bonferroni criterion [229]. This method is implemented as follows.

First, we set the level of significance $\alpha \in [0,1]$. The quantity α measures the largest admissible probability that at least one of the null hypotheses is erroneously rejected and it is typically a small quantity, to ensure the test significance. Then, the $n(n-1)$ p-values are sorted in ascending order and denoted as $\pi^{(1)} < \pi^{(2)} < \cdots < \pi^{((n-1)n)}$. Let L be the largest integer

for which it holds $\pi^{(L)} < L\alpha/(n-1)n$. Then, the null hypothesis is rejected for all the pairs of nodes associated with a p-value smaller than $\pi^{(L)}$. If the null hypothesis is rejected for i and j, then we estimate that there is a link in the backbone network between nodes i and j. Hence, we set the corresponding element of the estimated backbone adjacency matrix $\hat{G}$ as $\hat{G}_{ij} = \hat{G}_{ji} = 1$. We note that this is the step that requires the highest computational effort since the $n(n-1)$ p-values should be computed and sorted in ascending order. The algorithm can be implemented according to the pseudo-code below.

Algorithm 1: Backbone detection algorithm

Data: empirical observations r_i, n_{ij}, q_{ij}, $\forall i, j \in V$
Result: estimation of the adjacency matrix $\hat{G}$
$\hat{G} \longleftarrow 0$;
for $i \in V,\ j \in V,\ j \neq i$ **do**
    compute π_{ij} using (2.11);
sort π_{ij} in ascending order $\pi^{(1)} \leq \pi^{(2)} \leq \dots$;
$L \longleftarrow \max\{k \in \mathbb{N} : \pi^{(L)} < L\alpha/(n-1)n\}$;
for $i \in V,\ j \in V,\ j \neq i$ **do**
    if $\pi_{ij} \leq \pi^{(L)}$ **then**
        $\hat{G}_{ij} \longleftarrow 1$;
        $\hat{G}_{ji} \longleftarrow 1$;

Examining more in-depth the analytical results in (2.10a), we foresee some issues that might hinder the applicability of our algorithm, yielding a small value of $\mathscr{P}_{j\to i}$, even though a strong tie connecting i to j exists. In particular, this can occur in two cases. First, if both degrees d_i and d_j are large, such that the two nodes have a large degree centrality in the backbone network. Second, if both activities a_i and a_j are small. In the following, we present detailed numerical simulations with different parameter choices to demonstrate the accuracy of the algorithm.

2.4 Numerical validation

We validate our backbone detection algorithm on several synthetic datasets, to illustrate its applicability in real-world scenarios and identify potential limitations. These synthetic datasets consist of benchmark networks with $n = 200$ nodes, generated according to the RADN paradigm described in Section 2.2.1. We consider different distributions for the nodes' activities and

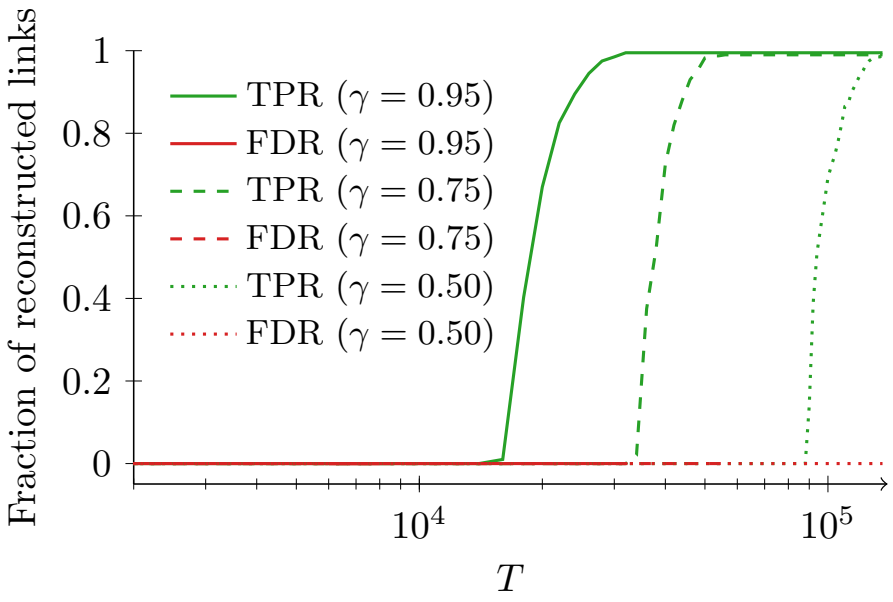


Fig. 2.3 Fraction of strong ties identified by our algorithm in the scenario with both homogeneous activity distribution and backbone degrees, for different values of the parameter γ. The backbone is a 4-regular network with 200 nodes. The other parameters are $\lambda = 0.9$, $\mu = 0.1$, and $a_i = 0.3$, for all the nodes.

backbone degrees. Specifically, the latter follows a configuration model [195]. The epidemic process is simulated using the SIS model illustrated in Section 2.2.2 with $\lambda = 0.9$ and $\mu = 0.1$. Unless otherwise specified, we set the significance level of the statistical test to $\alpha = 0.05$ and the parameter $\gamma = 0.95$.

2.4.1 Homogeneous activity distribution and homogeneous backbone

We first examine the possibility of identifying regular networks of strong ties against weak ties generated using a common activity value for all the nodes. In this scenario, the backbone is chosen to be a 4-regular random network and the activity is equal to $a_i = 0.3$, for all $i \in V$.

In Fig. 2.3, we plot the true positive rate (TPR), which is the fraction of links that the algorithm is able to correctly predict (green); and the false discovery rate (FDR), which is the ratio between the number of times it fails to properly identify a link and the number of links in the backbone (red). Perfect reconstruction is attained when the number of true positives is equal to the total number of positives (TPR$=1$) and the number of false positives is equal to zero (FDR$=0$). The computations are carried out for different values of T, such that larger

values of T imply access to a longer time-window for the estimation of the probabilities of transitions in the algorithm.

For sufficiently large values of T, our algorithm is successful in exactly reconstructing the topology of the backbone, for any choice of the parameter γ. As suggested by the analytical expression in (2.10a) where γ appears as a multiplicative coefficient, the smaller γ, the larger values of T are required by our algorithm. Choosing small values of T may hamper the correct identification of links, but it rarely results in the identification of false positives (for instance, only four false positives are overall identified for $\gamma = 0.95$). Thus, increasing T, we progressively improve the detection of strong ties, attributing a very small quantity of wrong links to the backbone. This is an important feature of the algorithm, whereby all the links it discovers can be relied upon with extremely high confidence. When little data is available, that is, for small T, the output of our algorithm could be poor. A possible strategy to circumvent the issue of limited data could be to not perform the multiple comparison correction, which, however, could beget a larger number of erroneous identifications.

2.4.2 Heterogeneous activity distribution and homogeneous backbone

To better proxy a real-world setting, we release the assumption that all the nodes have the same activity. As a stepping stone, we consider the case in which nodes are randomly divided into two activity classes with 100 nodes each: low-activity nodes ($a_i = 0.2$) and high-activity nodes ($a_i = 0.8$). Similar to the previous analysis, the backbone is a 4-regular random network. To help tease out the role of heterogeneity, we also simulate the scenarios in which all the nodes are either in the low- or high-activity classes.

Again, we examine the effect of T on true and false positives, with respect to the number of positives. Results in Fig. 2.4 confirm those from Fig. 2.3, whereby the fraction of correctly identified links increases with T, and the fraction of misclassified links is always negligible. Comparing the three scenarios, we observe that large values of the activity have a negative effect on the performance of the algorithm. In fact, an increased observation window is required to detect strong ties in the homogeneous case with high activity, with respect to the scenario with low activity.

Heterogeneity further reduces performance, hampering the detection of strong ties between low-activity nodes. Even though networks with a heterogeneous activity distribution require a longer window to correctly detect all the strong ties, we observe that, for sufficiently large T, our algorithm is able to correctly reconstruct the backbone, with a negligible fraction of

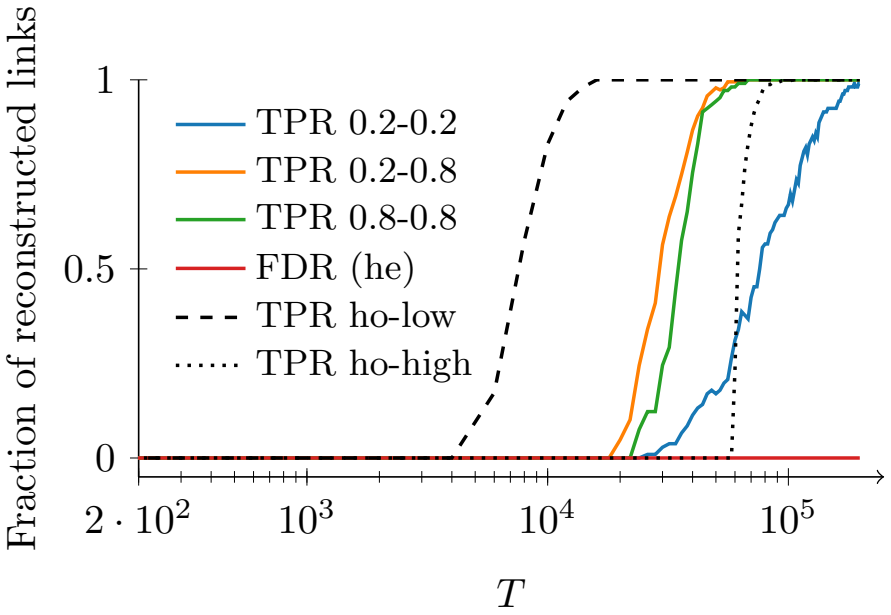


Fig. 2.4 Fraction of strong ties correctly identified by our algorithm for both heterogeneous and homogeneous activity distributions, and for homogeneous degree in the backbone. The backbone is a 4-regular network with $n = 200$ nodes. The other parameters are $\gamma = 0.95$, $\lambda = 0.9$, and $\mu = 0.1$. Three cases for the activity distribution are examined: all the nodes have the same activity $a_i = 0.2$ (ho-low, dashed), $a_i = 0.8$ (ho-high, dotted), and half the nodes have $a_i = 0.2$ and half have $a_i = 0.8$ (he, colored). For the last case of heterogeneous activities, the TPR curve is plotted with respect to links between nodes with low activity (blue), links between nodes of different activity (orange), and links between nodes with high activity (green). Only one FDR curve is plotted for all the cases since they are practically indistinguishable (he, red).

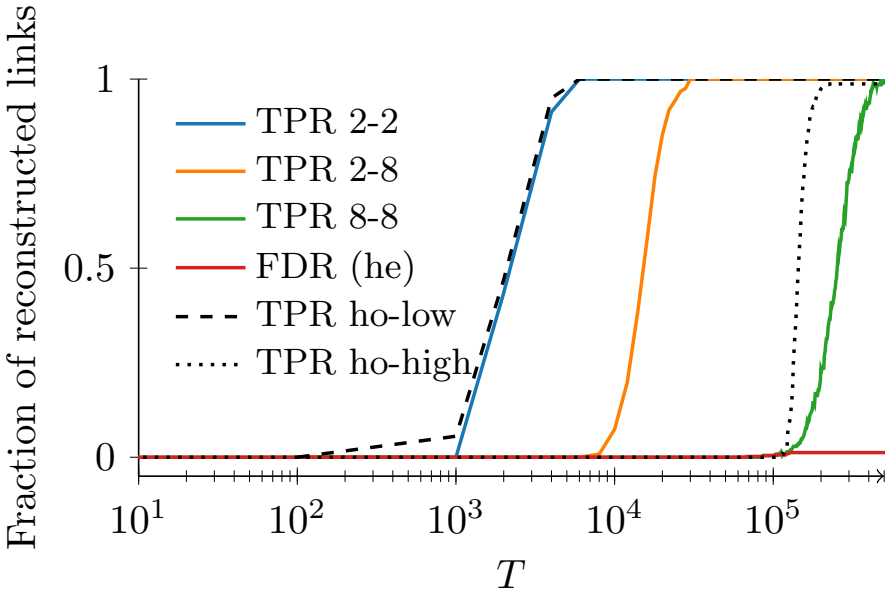


Fig. 2.5 Fraction of strong ties correctly identified by our algorithm for both heterogeneous and homogeneous backbones, and homogeneous activities $a_i = 0.3$, for all the nodes. The other parameters are $n = 200$, $\gamma = 0.95$, $\lambda = 0.9$, and $\mu = 0.1$. Three cases for the backbone are examined: all the nodes have the same low-degree $d_i = 2$ (ho-low, dashed); all the nodes have the same high-degree $d_i = 8$ (ho-high, dotted); and half the nodes have $d_i = 2$ and half have $d_i = 8$ (he, colored). For the last case of heterogeneous degrees, the TPR curve is plotted with respect to links between nodes with low degree (blue), links between nodes of different degree (orange), and links between nodes with high degree (green). Only one FDR curve is plotted for all the cases since they are practically indistinguishable (he, red).

erroneous identifications. Overall, these results are in agreement with the theoretical analysis in section 2.3, whereby decreasing the activities causes a reduction in the probability difference in (2.10a).

2.4.3 Homogeneous activity distribution and heterogeneous backbone

Next, we examine a backbone where the degree of the nodes is not held constant throughout the network. Specifically, we consider a network in which nodes are partitioned into two classes of 100 nodes each with low- ($d_i = 2$) or high-degree ($d_i = 8$). To avoid confounding, we maintain the activity at a common value of $a_i = 0.3$, similar to the results in Fig. 3. Once again, to facilitate the assessment of the effect of a heterogeneous degree distribution on the algorithm performance, we analyze two control cases in which all the nodes have the same low- or high-degree.

Figure 2.5 illustrates the fraction of links predicted as a function of T for three considered settings. Consistently with our previous results, we observe that increasing the length of the observation steadily benefits the algorithm's precision in inferring strong ties, as shown in Fig. 2.5. The number of false positives is always negligible, even for small values of T, confirming that the algorithm can be reliably utilized for backbone inference.

Comparing the two homogeneous cases of low- and high-degree distributions, we register an expected decrease in performance when dealing with higher degrees. In this case, the value of added knowledge regarding the state of health of one node is diluted by the presence of many other neighbors that could have triggered the infection. Analytical results in the section 2.3 provide a theoretical basis for this explanation, whereby increasing the values of the degree causes a reduction in the probability difference in (2.10a).

As one might expect, the performance of the algorithm toward the inference of the heterogeneous network is in between the two cases of homogeneous networks. To gain further insight into the relationship between topological features and successful reconstruction, we can isolate the specific links that are first detected by the algorithm for small values of T. In agreement with our analytical result in (2.10a), the links that require shorter observations are incident to low-degree nodes. These links encompass both strong ties between low-degree nodes and strong ties between nodes with high and low degrees that might exemplify disassortative structures of real networks [230, 231]. Longer time-windows are required for detecting links that connect pairs of high-degree nodes.

2.4.4 Highly-heterogeneous activity distribution and backbone

To offer insight on the performance of our algorithm over a wider class of RADNs, we systematically examine a two-dimensional grid of salient parameters. We assume that both the activity and the degree distributions follow a power-law with exponents β_a and β_d, respectively. We vary each parameter from -5 to -2, which are representative of real-world scenarios [232]. Parameters are varied in 11 steps with cutoffs at 0.1 and 1 for the activity, and at 1 and $n-1$ for the degree.

We observe that smaller values of the exponent of power-law yield distributions with a larger dispersion, in which most of the nodes have small activity (degree) and few have an extremely high activity (degree). Two different realizations are examined, one with $T = 10,000$ and $T = 30,000$, respectively. The weight γ is reduced to 0.5 to guarantee the spread of the epidemic diseases for all the choices of parameters investigated and the network size is

increased to $n = 300$ to ensure the presence of high-degree (activity) nodes in the power-law distributions. The epidemic parameters are set as $\lambda = 0.9$ and $\mu = 0.1$, similar to the simulations in Section 2.4.

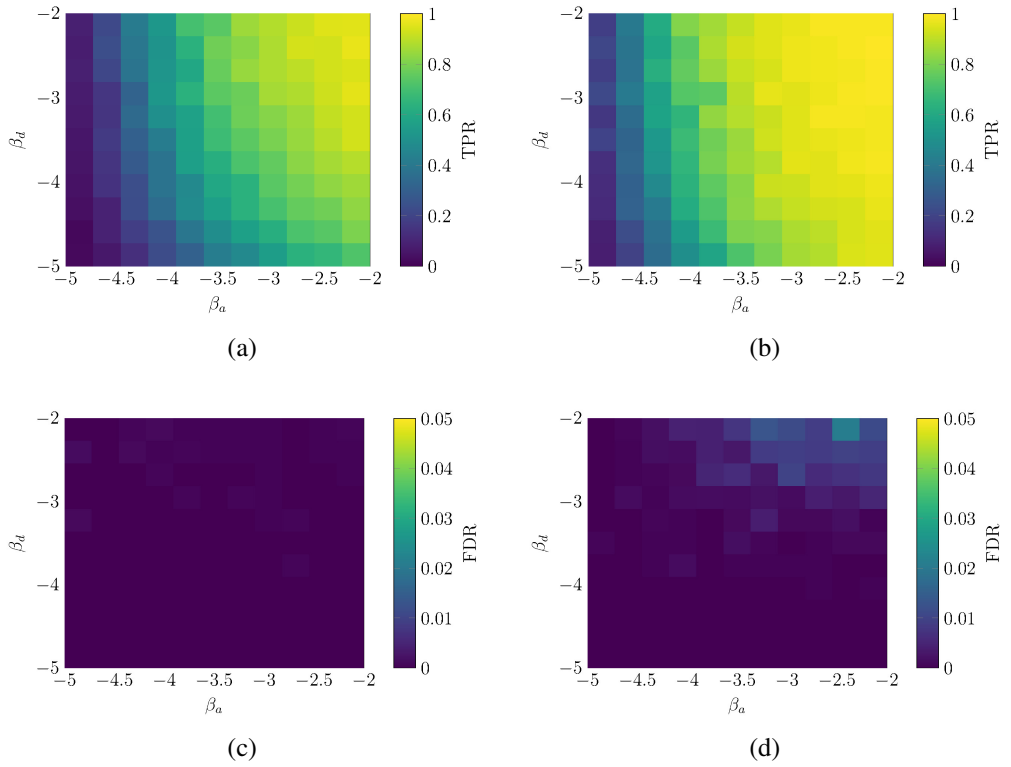


Fig. 2.6 TPR (a,b) and FDR (c,d) of our algorithm implemented on a network of $n = 300$ nodes with heterogeneity in both activity distribution and backbone degree, for an observation window of $T = 10,000$ time-steps (a,c) or $T = 30,000$ time-steps (a,c). Both activities and backbone degrees follow power-law distributions with exponents β_a and β_d, respectively. Other parameters are set to $\lambda = 0.9$, $\mu = 0.1$, and $\gamma = 0.5$. Each point is an average of ten independent simulations.

From Fig. 2.6, we recognize a marked effect of the parameters on the performance of our algorithm. For lower values of both parameters, β_a and β_d, our algorithm fails to identify the backbone, under-predicting the number of strong ties. This is in agreement with Figs. 2.4 and 2.5, which indicate that longer observation windows are required to infer the backbone when the RADN is dominated by high-degree and high-activity nodes. The best performance is attained for higher values of the two parameters. In this case, the algorithm correctly detects all the strong ties, with a very small quantity of false positives.

Comparing the results for $T = 10,000$ and $T = 30,000$, interestingly, β_a seems to have a stronger effect on performance than β_d, whereby at $T = 30,000$, the algorithm is able to detect most of the strong ties for small values of β_d but its performance is strained when examining

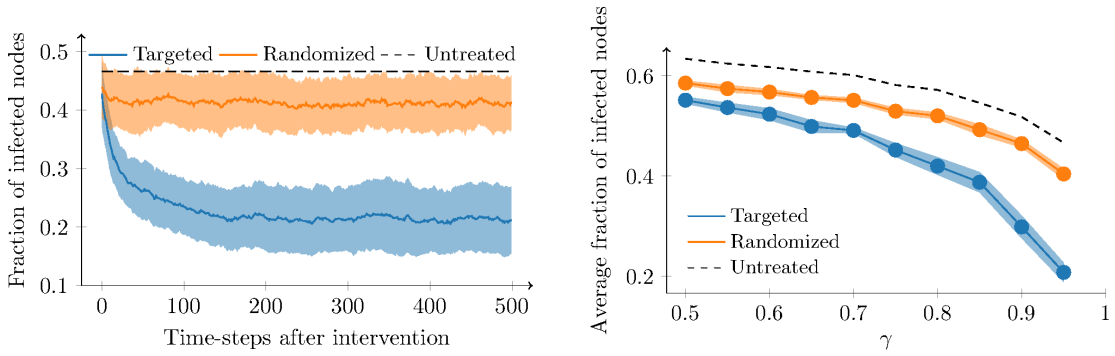
small values of β_a. This confirms our preliminary observation from Fig. 2.4 that heterogeneity in the activity distribution hampers the detection of strong ties.

2.5 Application to targeted immunization

In epidemiology, knowledge about the backbone network might offer valuable information about how diseases spread and which is the role played by individuals [233]. In this vein, we conclude this chapter by presenting an application of our algorithm to design a targeted immunization protocol. Our control strategy observes the disease spreading for a finite time-window to identify the backbone network, and then utilizes such an inference to prioritize immunization of nodes in the network according to a centrality criterion. Specifically, we immunize nodes according to decreasing values of their PageRank centrality [234]. By means of Monte Carlo numerical simulations, we evaluate the performance of the approach against a randomized immunization, where no information regarding the backbone is utilized.

Similar to the analysis in Section 2.4.4, we examine a benchmark network with $n = 300$ nodes. The backbone is generated using a configuration model with power-law degree distribution of power $\beta_d = -3$ and cutoffs at 1 and $n-1$. Activities are also drawn from a power-law distribution with exponent $\beta_a = -3$ and a lower cutoff at 0.1. We consider an SIS epidemic with $\lambda = 0.9$ and $\mu = 0.1$. We run the model over a window of $50,000$ time-steps implementing our algorithm to identify the backbone. At this time, we execute two control strategies (targeted and randomized), with the number of interventions limited to 5% of the total number of nodes. We perform Monte Carlo simulations by averaging over 100 independent runs of the two control strategies.

The results of these simulations are summarized in Fig. 2.7. In Fig. 2.7a, we compare the performance of the two immunization strategies for $\gamma = 0.95$, as in the numerical analysis in Section 2.4. While randomized immunization decreases the portion of infected nodes by 13%, targeted intervention decreases it by 55%, on average. The difference between these two strategies is statistically significant (p-value $\ll 0.0001$, according to a two-sample z-test) comparing the average fraction of infected individuals after the implementation of the immunization strategy, for 100 independent runs. In Fig. 2.7b, instead, the comparison between the two techniques is conducted for different values of the parameter γ, spanning from 0.5 to 0.95 in steps of 0.05. Therein, we report the average fraction of infected nodes in the 500 time-steps that follow the application of the control strategy. Predictably, the larger the parameter γ, the stronger the improvement of the targeted immunization with respect to



(a) (b)

Fig. 2.7 Monte Carlo estimation over 100 runs of the effect of randomized (orange) and targeted (blue) immunization on the fraction of infected nodes. Dotted lines indicate the fraction of infected nodes in the absence of any immunization technique. In (a), we show the entire realizations for $\gamma = 0.95$. The solid line is the average, while the light band is one standard deviation. In (b), we compare the average fraction of infected nodes for different values of γ. Bands identify 95% confidence intervals. Other parameters are $n = 300$, $\beta_d = \beta_a = -3$, $\lambda = 0.9$, and $\mu = 0.1$.

the randomized one. In fact, for small values of γ, the backbone has a marginal role on the link formation process, reducing the effect of targeted immunization exploiting the centrality measures in the backbone. However, the difference between the two strategies is statistically significant in all the performed simulations.

Encouraged by these promising results, we apply our targeted immunization technique to real-world face-to-face interactions measured through proximity sensors in a high school [194], available at [38]. The dataset comprises 188,508 temporal links, generated over $T = 7,375$ time-steps among $n = 327$ nodes. We run an SIS epidemic model for half of the available dataset, starting from a fraction of one-third of infected nodes, selected uniformly at random. Then, 5% of the nodes is immunized following either the randomized or the targeted strategy. By performing an extensive Monte Carlo simulation with 1,000 runs, we compare the two strategies for different values of the epidemic parameters λ and μ. Figure 2.8 demonstrates that our immunization technique should always be preferred to randomized immunization, whereby, for most parameter choices, it outperforms randomized immunization.

2.6 Conclusions

In this chapter, we propose an algorithm to unveil the backbone of strong ties in a temporal network from empirical data of a spreading process unfolding on the network nodes. Building

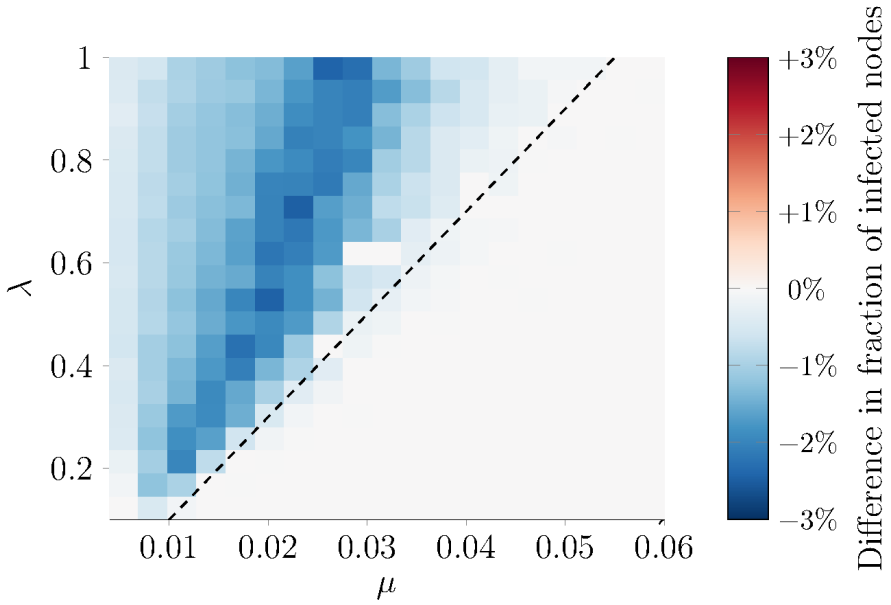


Fig. 2.8 Difference in the fraction of infected nodes after the immunization phase, between the randomized and the targeted strategy (color coded) in the high-school case study. The dashed line represents the epidemic threshold, below which none of the nodes is infected at the onset of the immunization strategy. Darker blue areas identify parameter regions where targeted immunization has a superior outcome. Each point is an average of $1,000$ independent simulations.

on analytical insight regarding the role of strong ties in the process, we have put forward a statistically-principled approach to discover strong ties from empirical data. Extensive simulations are performed to assess the effectiveness of the proposed technique, which has proved to be reliable in a variety of scenarios. Finally, we examined the integration of the proposed algorithm in the solution of an important challenge in epidemiology, namely, targeted immunization during an outbreak. Therefore the main innovations proposed in this chapter are: *i*) the analytical computation of the effect of strong ties on the infection probability for a susceptible–infected–susceptible epidemic model on routed activity driven networks; *ii*) the design of a backbone detection algorithm and its numerical validation; and *iii*) the implementation of a targeted immunization technique.

The promising preliminary results of our numerical analysis pave the way for several avenues of future research. Future efforts should focus on the development of accurate methods to deal with limited data, without increasing the number of erroneous identifications. In the analytical derivation of our bounds, we specialize the computations to the SIS epidemic model. However, the generality of our proving argument suggests that similar bounds could be established for other models, as well, provided that they do not admit permanently attractive states. These achievements would be key to providing a theoretical basis for the generalization

of our algorithm to deal with other dynamics, including richer epidemic processes or opinion diffusion. Such an extension will be part of our future research. In most real-world scenarios, it is not tenable to have access to the entire node set, thereby calling for methods to discover missing nodes, beyond links, as we will explore in chapter 3.

The ability to reconstruct the structure of the backbone of a complex system from the observation of an unfolding spreading process finds application in disparate fields of investigation. Our study on targeted immunization has demonstrated how information about the backbone can be leveraged to design effective control techniques that could steer the behavior of dynamical systems. Extending the framework to other disease models and mathematically proving performance bounds is the objective of future research.

Chapter 3

Hidden nodes in activity driven networks

In this chapter, we approach the problem of recollecting the number of nodes in a network given only partial information about it. Once we were able to find a viable approach to tackle the problem of reconstructing the nature of links in activity-driven networks, we questioned the feasibility of identifying hidden nodes. Our curiosity stems from the consideration that in almost all data-driven approaches a perfect sampling of the network under study is impossible, and dealing with missing data may greatly mislead researchers while inferring spreading processes in different contexts. In our approach, we set to only rely on the data originating from a diffusive process, in order to be able to assess how many nodes were hidden from our knowledge in the network under study. We propose an analytic framework capable of assessing the imprint left by hidden nodes on the dynamics of a network. Focusing on a variant of the voter model, and using statistical approaches based on combinatorial analysis, a descriptive tool is developed to assess the variation caused by inaccessible nodes in activity-driven networks. Our technique proves robust to noise and parameters variation, bringing a new perspective on node identification and network reconstruction.

3.1 Background

Network science, since its beginnings, always had two entangled souls, a theoretical one and an experimental one [6, 64, 195, 235–237]. Relying on data and observation has proven essential to discover counter-intuitive phenomena and emerging behaviors. Since the advent of social networks and electronic databases, the methodologies utilized to study human interactions have changed [238–243]. Data integrity has become an issue, sometimes due to the sheer amount of

information that needs to be cleaned of noise, and sometimes due to an unreliable collection technique that leaves researchers in the dark about missing data [194, 244–246].

New techniques are needed to assess the quality and completeness of both obtained and observable data. Unrecognized connections and interactions among individuals, and the inability to retrieve all the parties involved in a diffusion process, may be extremely detrimental to the process of assessing and evaluating the characteristic of the phenomena under study.

In this regard we focused our attention on activity-driven networks, which emerged as a powerful paradigm to describe diffusion processes on networks [154]. Thanks to the co-evolution, at the same timescale of the diffusion process and the network itself, ADN can capture events that happen at a fast pace [156] and may leave a feeble imprint on the network dynamic. This property is often essential to properly studying a phenomenon, such as an epidemic diffusion [12, 23, 147, 155, 227], information diffusion[157, 247] or opinion formation [15, 160, 248].

Different techniques have been proposed to infer the size and the nodes involved in a network system, with different uses of Bayesian models [83, 249–251], spectral clustering algorithms [252, 253], optimization processes [254] or multiple techniques combined [76, 255]. More exact and reliable results may be limited to some epidemic parameters [256] or near a specific stable state of the dynamic [20]. The process of reconstructing the real size of a network has been faced in data-driven approaches, using compressing sensing algorithms [257], maximum likelihood approaches [258], or scale-up methods [259, 260]. A more in-depth focus is needed on the subject, in order to develop reliable and robust tools available to any researcher working with data sets.

In this study, we build on previous works that have focused on retrieving fundamental structures in a network [27, 208, 209, 213] obtained thanks to the observable effects these structures produce. Our technique focuses on the ability to discern if there are nodes hidden in an observable system from the external observer. If this is the case, we aim to assess the number of hidden nodes participating in the unraveling diffusion process.

Our main goal is to analytically assess the effect caused by hidden nodes in a diffusion process on Activity Drive Network. Once the diffusion process is established, all the states of the visible nodes are recorded, computing the probability for each to change state given the total number of visible infected nodes in the system. A statistical analysis, rooted in a combinatorial and probabilistic calculation, allows estimating the shift caused by one or more hidden nodes in this probability of changing state. The robustness of the analysis is tested through an extensive

simulation campaign. A simple optimization technique is then proposed to infer the correct size of the system leveraging the main results of our work.

The main contribution of this chapter are: 1) defining a technique to identify the presence of hidden nodes in a system; 2) computing the exact effect that hidden nodes have on visible nodes; 3) validating the robustness and efficiency of the technique, and finally 4) the definition of a simple technique to infer the exact number of hidden nodes in the system is proposed.

The rest of the chapter introduces in sec.3.2 the model and analyzes the dynamics of the diffusion process under study, and in sec.3.3 a technique to classify if a model has hidden nodes is defined. In sec.3.4 a thorough analysis of the visible effects of hidden nodes left in the diffusion process is carried out, and a simple application using optimization techniques is proposed for real applications. In sec.3.5 numerous simulations are carried out to assess the robustness and sensibility of the analysis.

3.2 Dynamics

For our study, we choose a modified version of the voter model for its straightforward analytical description and the symmetrical properties in the state change of each node [261]. The constituent elements of our model are the interacting nodes of the system. Each node i is characterized by two elements: a Boolean state x_t^i that can vary at each time-step, and an activity $a_i \in [0,1]$, fixed for the whole process, representing the propensity to create a link with another node at each time-step. The complete system is composed of N nodes and following the ADN paradigm [154], the population size does not change during the diffusion process. Different sets of edges $\mathscr{E}_t$ among the nodes are created at the beginning of each time-step and removed at the end of it. Parameter μ is the same for all the nodes in the system at all time-steps, representing noise affecting the propagation process.

At the beginning of a time-step, each node has a probability $\mu \in (0,1)$ of changing its Boolean state from 0 to 1 or vice-versa. if such state-change does not happen, node i creates an edge $(i,j) \in \mathscr{E}_t$, with another node j, with probability a_i, independent form all the others. Node j is chosen with uniform probability among $\mathscr{N} \setminus \{i\}$. When the edge (i,j) is created by node i, it copies the status of node j: $x_{t+1}^i = x_t^j$, node j is not affected otherwise by such interaction. At the end of the time-step, all edges are removed, and the state of each node is updated. To

clarify this process, we show in eq.(3.1)

$$P(x_{t+1}^i = 1 | x_t^i = 0) = \mu + (1-\mu) \cdot a_i \cdot \frac{I_t}{N-1} \tag{3.1}$$

the probability of changing state for a given node i, at time-step t.

The set of nodes with state equal to 1 will have size I_t, and we will commonly refer to the nodes in this set as *infected*. We want to compute the probability of changing state, for a node i, conditioned to the number of infected nodes in the system, as it will vary for different values of $I_t = k$.

$$P(x_{t+1}^i = 1 | x_t^i = 0, I_t = k) = \mu + (1-\mu) \cdot a_i \cdot \frac{k}{N-1} \tag{3.2}$$

From eq.(3.2) we can derive all the probabilities of changing from one state to another for each node i:

$$p_{10}^i(k) := P(x_{t+1}^i = 1 | x_t^i = 0, I_t = k) = \mu + (1-\mu) \cdot a_i \cdot \frac{k}{N-1} \tag{3.3}$$

$$p_{01}^i(k) := P(x_{t+1}^i = 0 | x_t^i = 1, I_t = k) = \mu + (1-\mu) \cdot a_i \cdot \frac{N-k}{N-1} \tag{3.4}$$

$$p_{00}^i(k) := P(x_{t+1}^i = 0 | x_t^i = 0, I_t = k) = 1 - p_{10}^i(k) \tag{3.5}$$

$$p_{11}^i(k) := P(x_{t+1}^i = 1 | x_t^i = 1, I_t = k) = 1 - p_{01}^i(k) \tag{3.6}$$

The value of $k \in [0, N-1]$ for eq.(3.3,3.5), having those equation the assumption of $x_t^i = 0$, and $k \in [1, N]$ for eq.(3.4,3.6), for the assumption of $x_t^i = 1$.

The value $I = \lim_{t\to\infty} I_t$ will be the number of infected nodes in the system at the steady state of the dynamic, following a probability distribution $\Pi_N(k) = \lim_{t\to\infty} P(I_t = k)$, better explored in sec.3.4. Must be noted that, if parameter $\mu \neq 0$, the process has no absorbing state. If $\mu = 0$, the states $I = 0$ and $I = N$ are absorbing, since $p_{10}^i(0) = a_i \cdot \frac{0}{N-1} = 0$ implying that no node can ever be infected again, and $p_{01}^i(N) = a_i \cdot \frac{0}{N-1} = 0$ implying that no node can ever be non-infected.

3.3 System identification

In many real systems, the extent of the number of subjects involved in a diffusion process is unknown, and the analysis is carried out on the visible population. Following that line of thought, we assume of being able to observe only a subset of nodes in our system, the set $\mathcal{N}'$ of size N'. A subset of $\mathcal{Q}$ nodes, of size Q, is fully participating in the dynamic but inaccessible to an external observer. The size of the observed system will then be $N' = N - Q$. Even if the nodes in the subset $\mathcal{Q}$ are not possible to observe, they are fully participating in the network dynamics, getting infected and infecting nodes over time. In the model here under study, we have that the probability of changing state is directly proportional to the number of infected nodes in the system I_t. Given that some nodes are hidden, the number of visible infected, denoted by I'_t, will not always be coincident with the real number of infected I_t. The probabilities in eq. (3.3,3.4,3.5,3.6) will be denoted as $p^i_{10}(k')$ when the observed number of infected in the system is k' and the real number of infected k is unknown.

3.3.1 Homogeneous activities

Let's start our analysis assuming that all nodes have the same activity $a_i = a \; \forall i \in \mathcal{N}$. Thanks to this assumption, we can drop the index i from all the probabilities in eq.(3.3,3.4,3.5,3.6), because all values will be equivalent for each node. Any hidden node in the system will have the same probability of participating in the diffusion process as any other. Therefore, the influence on any hidden given node $h \in \mathcal{Q}$ will be the same. Discrepancies of the computed probabilities of eq.(3.3,3.4,3.5,3.6) with the observed probabilities from a simulation or a data-set can be assessed using a z-test for each different value of k in the system.

We illustrate the methodology by comparing the analytic prediction to a preliminary simulation. After computing the probability $p_{10}(k)$ as in eq.(3.3), we compare it against the measured probability obtained from a simulation with $N = 4$, aggregated over 10 runs each of time-steps $= 1.000$, $\mu = 0.5, a = 0.5$, and no hidden nodes. For each node we computed the fraction of times a transition from state 0 to state 1 happens given the total number of infected nodes in the system. Since all probabilities are equivalent, the data points are aggregated for each value of k. The adherence between the predicted values and the observed is assessed with a null hypothesis that in our system no hidden node is present by performing a z-test between the predicted value and the measured probabilities, for each value of k, with an acceptance of $\alpha = 0.05$. For this situation, we compute that all pvalues above the threshold α, therefore all

z-test are accepted and our null hypothesis of no hidden nodes participating in the system is confirmed.

We then perform the same procedure on a system with not all nodes visible. After computing $p_{10}(k')$ as in eq.(3.3), we compare the computation with the measured probability obtained in a simulation with $N' = 4$, aggregated over 10 run each of time-steps $= 1.000$,$\mu = 0.5, a = 0.5$, and 1 hidden node. The real size of the system will then be $N = 5$. Testing this scenario for the null hypothesis that we are observing a system with no hidden nodes, we observe all pvalues below the acceptance threshold α, thus correctly rejecting the assumption that there are no hidden nodes.

3.3.2 Heterogeneous activities

Let's now explore systems in which the activity for each node is different, normally distributed around a mean value $\langle a \rangle$, with a small variance. The activity of each node i will be $a_i = \langle a \rangle + \sigma \cdot s_i$, where the term s_i is normally distributed around 0 with unit standard deviation, σ represent the standard deviation of the activities. In this framework, each node will be subject to different probability $p^i_{10}(k)$ of changing state, and will not be possible to fully understand the dynamics of the system unless all activities are known.

A simple method that can address this issue is not to consider the probability of changing state of each node but consider the average of all the probabilities of changing state.

$$\begin{gathered} \frac{1}{N}\sum_{i\in\mathcal{N}} p^i_{10}(k) = \frac{1}{N}\sum_{i\in\mathcal{N}} \mu + (1-\mu)\cdot a_i \cdot \frac{k}{N-1} = \\ \mu + (1-\mu)\cdot\frac{k}{N-1}\cdot\frac{1}{N}\sum_{i\in\mathcal{N}} a_i = \mu + (1-\mu)\cdot\langle a\rangle\cdot\frac{k}{N-1} \end{gathered} \tag{3.7}$$

Thanks to the result of eq.(3.7), we can correctly predict the average probability of changing state for all nodes in the system, knowing only the average activity. Similarly to the previous scenario of homogeneous activities, we will preliminarly compare our predicted value against the aggregated set of $p^i_{10}(k)\ \forall i \in \mathcal{N}$ against a simulation aggregated over 10 runs each of time-steps $= 1.000$. We illustrate the methodology by comparing the analytic prediction to a simulation, our acceptance threshold will be $\alpha = 0.05$ as in the previous examples for a null hypothesis of no hidden nodes in the system. For our test we start with a system composed of $N = 4$ and no hidden nodes, with parameters $\mu = 0.5,\ \langle a \rangle = 0.5,\ \sigma = 0.01$. In this scenario,

all the pvalues for the different values of k are above the threshold, so we correctly accept the null hypothesis that no hidden nodes are present in the system.

As before, we move on to explore the situation in which hidden nodes are indeed present. To test our claim we compare the analytical results with a simulation aggregated over 10 runs each of time-steps $= 1.000$. Testing in this scenario for the null hypothesis that we are observing a system with no hidden nodes, we obtain all pvalues below the specified threshold, therefore correctly rejecting the assumption that there are no hidden nodes.

3.4 Node Reconstruction

Once a suitable method to classify if a system has or not hidden nodes, it is of certain interest to be able to estimate the number of hidden nodes in the system under study. The problem involves computing all the possible combinations of states of visible and hidden nodes, in order to be able to assess the effect of each node on the probability of changing state for all the other nodes in the system.

3.4.1 Homogeneous activity

As before, we start from the assumption of homogeneous activity, $a_i = a \; \forall i \in \mathcal{N}$. We will describe a system having Q hidden nodes, and $N' = N - Q$ visible nodes. A generic hidden node will be h, and its state at a generic time-step x_t^h. While we try to compute the probability $p_{10}^i(k)$ of changing state for a visible node, we are affected by not knowing if the state of the hidden node is $x_t^h = 0$ or $x_t^h = 1$. For example, while we observe the system having $I' = k'$ infected nodes, it could be having $I = k = k'$ or $I = k = k' + 1$ infected nodes. The limit probability $p_{10}^i(k') = \lim_{t\to\infty} P(x_{t+1}^i = 1 | x_t^i = 0, I_t = k')$, in a system with N' visible nodes and k' visible infected, must be re-written to account for hidden nodes. Equation (3.15) describes the corrected probability for a node in the system to change state from 0 to 1 when k' infected are visible in the system, and Q hidden are present.

The probability of changing state from 0 to 1 for a given node i in the presence of Q hidden nodes is computed as follows. The probability of changing state, for a node i, directly depends on the number of infected nodes in the system. Specifically, from (3.2), we obtain the, for

$I_t = k$, it holds

$$p_{10}(k) := P(x^i_{t+1} = 1 | x^i_t = 0, I_t = k) = \mu + (1-\mu)\cdot a \cdot \frac{k}{N-1} \tag{3.8}$$

If hidden nodes are present, the quantity I_t is not observable, hence the probability in (3.8) cannot directly be computed. Let us assume that the number of visible nodes is equal to $N' = N - Q$ and the number of visible infected nodes is equal to $I'_t = k'$. We want to compute

$$\tilde{p}_{10}(k') := P(x^i_{t+1} = 1 | x^i_t = 0, I'_t = k'). \tag{3.9}$$

Using (3.8) and the law of total probability, we write

$$\tilde{p}^i_{10}(k') = \sum_{h=0}^{Q} p_{10}(k'+h) P(I_t = k \,|\, I'_t = k'). \tag{3.10}$$

Let us consider the process I_t. For homogeneous networks of agents I_t is an ergodic Markov chain, whose transition probability matrix $\mathbb{P} \in [0,1]^{(N+1)\times(N+1)}$ can be computed following (3.17). Let $\Pi_N \in [0,1]^{N+1}$ be its invariant distribution (computed as the left eigenvector of $\mathbb{P}$ associated with eigenvalue 1), so that $\Pi_N(k)$ is the probability of having k infected nodes in the steady state.

When I_t is in the steady state, we can use Bayes' theorem to compute

$$P(I_t = k'+h \,|\, I'_t = k') = \frac{P(I'_t = k' \,|\, I_t = k'+h) P(I_t = k'+h)}{P(I'_t = k')} \tag{3.11}$$

where, using a combinatorial argument, we write

$$P(I'_t = k' \,|\, I_t = k) = \binom{k'+h}{k'} \cdot \binom{N-k-h}{N'-k'} \tag{3.12}$$

Hence, in the steady state, we can write (3.11) using the expression computed in (3.12) as

$$P(I_t = k'+h \,|\, I'_t = k') = \frac{\binom{k'+h}{k'} \cdot \binom{N-k-h}{N'-k'} \Pi_N(k'+h)}{\Theta} \tag{3.13}$$

where the denominator is computed using the law of total probability as

$$\begin{aligned}\Theta &= \sum_{h=0}^{Q} P(I'_t = k' \,|\, I_t = k'+h) P(I_t = k'+h) \\ &= \sum_{h=0}^{Q} \binom{k'+h}{k'} \cdot \binom{N-k-h}{N'-k'} \cdot \Pi_N(k'+h).\end{aligned} \tag{3.14}$$

Hence, substituting (3.13) in (3.10), (3.9) reads

$$\begin{aligned}\tilde{p}_{10}(k') = \mu + (1-\mu) a \frac{1}{N-1} \frac{1}{\Theta} \cdot \\ \sum_{h=0}^{Q} (k'+h) \binom{k'+h}{k'} \binom{N-k-h}{N'-k'} \Pi_N(k'+h).\end{aligned} \tag{3.15}$$

The combinatorial terms are needed to consider the correct realizations of extracting indistinguishable infected nodes, given the N' visible and N total in the system.

We indicate with the notation $\Pi_N(k) := \lim_{t\to\infty} P(I_t = k)$ the probability for system composed of N nodes to being in a state with k infected nodes. Therefore, when the system it's in the steady state, $\Pi_N(k)$ is the probability distribution of having k infected nodes. To solve (3.15) we must find the value for the stationary distribution $\Pi_N(k)$ of the associated discrete time Markov Chain process defined by

$$P(I_t = k) = \sum_{f=0}^{N} P(I_t = k \mid I_{t-1} = f) \cdot P(I_{t-1} = f) \tag{3.16}$$

If parameters μ, a are known, the steady state $\Pi_N(k)$ of the Markov chain is easily computed from the transition matrix $\mathbb{P}(N) \in [0,1]^{(N+1)\times(N+1)}$ encompassing all the transition probabilities from one state to the other.

Still under the assumption that $a_i = a, \forall i \in \mathcal{N}$, we compute the stationary distribution $\Pi_N(k)$ of transitioning from $I_t = f$ to $I_{t+1} = k$ infected in a system of size N.

The diffusion process is re-formulated as a discrete-time Markov Chain, in which each state is the number of infected in the system. Since the future state of each node only depends on the current number of infected the process is memoryless. The transition probability from a state with f infected to one with k infected will be $P(I_t = k \mid I_{t-1} = f)$, and the space of the chain will be $S = \{0, \ldots N\}$. The stationary state of the Markov Chain to be found is defined in eq.(3.16).

Transition probabilities from one state to another can be computed as described in eq.(3.17).

$$
\begin{aligned}
&P(I_t = k \mid I_{t-1} = f) := \\
&if\ k > f: \\
&\sum_{p=0}^{\min\{f,N-k\}} \binom{f}{p} [p_{01}(f)]^p \cdot [p_{11}(f)]^{f-p} \cdot \binom{N-f}{k-f+p} [p_{10}(f)]^{k-f+p} \cdot [p_{00}(f)]^{N-k-p} \\
&if\ k < f: \\
&\sum_{p=0}^{\min\{k,N-f\}} \binom{N-f}{p} [p_{10}(f)]^{f-k+p} \cdot [p_{00}(f)]^{N-f-p} \cdot \binom{f}{f-k+p} [p_{01}(f)]^p \cdot [p_{11}(f)]^{k-p}
\end{aligned}
\tag{3.17}
$$

If parameters μ, a are known, the steady state $\Pi_N(k)$ of the Markov chain is easily computed from the transition matrix $\mathbb{P}(N) \in [0,1]^{(N+1)\times(N+1)}$ encompassing all the transition probabilities from one state to the other.

The result in eq.(3.17) is obtained through the combination of all the possible state transitions realized by indistinguishable nodes. Here we show the derivation concerning $k > f$ shown in eq.(3.18), all considerations will also be valid for $k < f$.

The initial state at time t has $I_t = f$ infected nodes, each one can either change status with probability p_{01} or not with probability p_{11}. Conversely, the non-infected nodes initially are $N - f$, with probability p_{10} of changing state and p_{00} of not changing state. Since $k > f$ the number of transitions from non-infected to infected must be higher than the opposite and at minimum $k - f$.

These $k - f$ transitions must be chosen among all the initially non-infected nodes, $N - f$. The event of all the transitions happening is Bernoulli trials that involve an exact number of infected f and $N - f$ non-infected. If no infected node perform a state transition, $p = 0$, the transition from state $I_t = f \longrightarrow I_{t+1} = k$ implies extracting exactly $k - f$ new infected in the population of non-infected available

$$
\binom{N-f}{k-f} [p_{10}(f)]^{k-f} \cdot [p_{00}(f)]^{N-k} \tag{3.18}
$$

The number of infected nodes at time t that can perform a transition is fixed, and besides the minimum number of $k - f$, transitions from infected to non-infected are possible as long as

an equal number of transitions from non-infected to infected balance them out. If $p \neq 0$, these transitions will be extracted among all the infected f, following a Bernoulli trial.

$$\binom{f}{p} [p_{01}(f)]^p \cdot [p_{11}(f)]^{f-p} \tag{3.19}$$

The number of transitions p is bounded either from the number of infected initially, f, or the number of non-infected in the final state, $N-k$ Since multiple values of p, from 0 to $min(f, N-k)$, may be permitted to reach the same final state with k infected, it is necessary to sum over all the possible values of the index p. Therefore, combining the two Bernoulli trials of eq.(3.18),(3.19) we obtain all the possible realizations for state $I_{t+1} = k$, starting form state $I_t = f$ in eq.(3.20).

$$P(I_t = k \mid I_{t-1} = f) = \sum_{p=0}^{\min\{f, N-k\}} \binom{f}{p} [p_{01}(f)]^p \cdot [p_{11}(f)]^{f-p} \cdot \binom{N-f}{k-f+p} [p_{10}(f)]^{k-f+p} \cdot [p_{00}(f)]^{N-k-p} \tag{3.20}$$

An example is given in eq.(3.21) of the matrix for a system with homogeneous activity and $N = 2$.

$$\mathbb{P}(N=2) = \begin{bmatrix} (1-\mu)^2 & (\mu + (1-\mu)\cdot a) - (\mu + (1-\mu)\cdot a)^2 & \mu^2 \\ 2(\mu - \mu^2) & 1 - 2(\mu + (1-\mu)\cdot a) + 2(\mu + (1-\mu)\cdot a)^2 & 2(\mu - \mu^2) \\ \mu^2 & (\mu + (1-\mu)\cdot a) - (\mu + (1-\mu)\cdot a)^2 & (1-\mu)^2 \end{bmatrix} \tag{3.21}$$

Let's now explicitly compute the influence caused by a single hidden node, in a system of size N, to the probability of changing state of all visible nodes. The hidden node h will be infected, or not, following eq.(3.22), when k' infected are visible. This proportion will be crucial to weigh the probability for all visible nodes to change status.

$$\lim_{t\to\infty} \frac{\Pi_N(k)\cdot P(x_t^h=1|I_t=k,|\mathscr{N}|=N)}{\Pi_N(k)\cdot P(x_t^h=1|I_t=k,|\mathscr{N}|=N)+\Pi_N(k')\cdot P(x_t^h=0|I_t=k',|\mathscr{N}|=N)} = P_{h1}(N',k') = \frac{\Pi_N(k)\cdot\frac{k}{N}}{\Pi_N(k)\cdot\frac{k}{N}+\Pi_N(k')\cdot\frac{k'}{N}} \tag{3.22}$$

$$p_{10}(k') = \mu + (1-\mu)\cdot a\cdot\frac{1}{N-1}\cdot\left\{(k'+1)\cdot P_{h1}(N',k') + k'\cdot\left[1-P_{h1}(N',k')\right]\right\} \tag{3.23}$$

Thanks to all the results obtained to this point, we can solve eq.(3.22) and provide a corrected version of eq.(3.3) for any system with one hidden node. Equation (3.23) describes the corrected probability for a visible node in the system to change state from 0 to 1 when k' infected are visible in the system.

3.4.2 Heterogeneous activity

In order to approach the reconstruction of the number of nodes in the system considering heterogeneous activities, we will follow a process like the one used in sec.3.3. The activity $a_i = \langle a\rangle + \sigma s_i$ of each node i will be distributed around a mean value$\langle a\rangle$, the term s_i is defined such that $\sum_i s_i = 0$ with normal distribution and unit standard deviation, and σ will be the standard deviation of the activities.

As we see eq.(3.16,3.22) still stand, but eq.(3.17) must be reformulated. Since we cannot assume now that all nodes are equal, we must approach the problem in terms of distinguishable nodes. Let's start by unfolding the probability of changing state for a single node:

$$p_{10}^i = \mu + (1-\mu)\cdot(\langle a\rangle + \sigma\cdot s_i)\cdot\frac{k}{N-1} \tag{3.24}$$

If in eq.(3.24) the term σ is small enough, an exponential term of p_{10}^i can be expanded as follows:

$$
\begin{aligned}
[p_{10}^{i}]^{\alpha} = \left[\mu + (1-\mu)\cdot(\langle a\rangle + \sigma s_i)\cdot\frac{k}{N-1}\right]^{\alpha} = \\
\left[\mu + (1-\mu)\cdot\langle a\rangle\cdot\frac{k}{N-1}\right]^{\alpha} + \\
(1-\mu)\cdot\sigma\cdot s_i\cdot\alpha\cdot\frac{k}{N-1}\cdot\left[\mu + (1-\mu)\cdot\langle a\rangle\cdot\frac{k}{N-1}\right]^{\alpha-1} + \\
O\left((1-\mu)\cdot\sigma\cdot s_i\cdot\alpha\cdot\frac{k}{N-1}\right)^{2}
\end{aligned}
\tag{3.25}
$$

And if we were to multiply two probabilities of different nodes i and j, we would obtain:

$$
\begin{aligned}
p_{10}^{i}\cdot p_{10}^{j} = \left[\mu + (1-\mu)\cdot\langle a\rangle\cdot\frac{k}{N-1}\right]^{2} + \\
\left[(1-\mu)\cdot\sigma\cdot\frac{k}{N-1}\right]\cdot\left(\mu + (1-\mu)\cdot\langle a\rangle\cdot\frac{k}{N-1}\right)\cdot(s_i+s_j) + \\
\left[(1-\mu)\cdot\sigma\cdot\frac{k}{N-q}\right]\cdot s_i\cdot s_j
\end{aligned}
\tag{3.26}
$$

Therefore, deriving from the previous eq.(3.25), we can express the product of multiplying N different probability terms:

$$
\begin{aligned}
\prod_{i=0}^{N} p_{10}^{i} = \left(\mu + (1-\mu)\cdot\langle a\rangle\cdot\frac{k}{N-1}\right)^{N} + \\
\left(\mu + (1-\mu)\cdot\langle a\rangle\cdot\frac{k}{N-1}\right)^{N-1}\cdot(1-\mu)\cdot\sigma\cdot\frac{k}{N-1}\cdot\sum_{i=0}^{N} s_i + \\
O\left((1-\mu)\cdot\sigma\cdot\sum_{i=0}^{N} s_i\cdot\frac{k}{N-1} + (1-\mu)\cdot\sigma\cdot\frac{k}{N-1}\cdot\prod_{i=0}^{N} s_i\right)^{2} = \\
\left(\mu + (1-\mu)\cdot\langle a\rangle\cdot\frac{k}{N-1}\right)^{N} + O\left((1-\mu)\cdot\frac{k}{N-1}\cdot\sigma\cdot\prod_{i=0}^{N} s_i\right)^{2}
\end{aligned}
\tag{3.27}
$$

The condition will grant us that for heterogeneous activities, equation for $P(I_t = k|I_{t-1} = f)$ that contains $N!$ distinct elements, can be approximated with eq.(3.17). Therefore, the influence

of a single hidden node for the heterogeneous activity system will approximate that of a homogeneous system, under the condition that σ is small or N is large enough.

3.4.3 Optimization

We propose a simple method, based on optimization techniques, to assess the number of hidden nodes in a system. Since eq.(3.15) gives the exact estimate for the probability of changing state, we propose to asses the exact number of hidden nodes minimizing the Least-Square sum between the average over all observed values and the estimate obtained for a different number of hidden nodes.

First, we define the function (3.28) of the parameters (k',q):

$$f(k',q) := \mu + (1-\mu)\cdot a \cdot \frac{1}{N'+q-1}\cdot\frac{1}{\theta(q)}\cdot \sum_{h=0}^{q}(k'+h)\cdot\binom{k'+h}{k'}\cdot\binom{N'+q-k'-h}{N'-k'}\cdot\Pi_{N'+q}(k'+h) \tag{3.28}$$

Where the normalization term is defined as

$$\theta(q) := \sum_{h=0}^{q}\binom{k'+h}{k'}\cdot\binom{N'+q-k'-h}{N'-k'}\cdot\Pi_{N'+q}(k'+h) \tag{3.29}$$

The values computed using function (3.28) is compared to the average of the observed probability of changing state $p^i_{10}(k')$ for each node i, as explained in sec.3.4.2, assuming a Gaussian distribution around the average.

$$LS(q) := \sum_{k'=0}^{N'}\left(f(k',q) - \frac{1}{N'}\sum_{i\in\mathcal{N}'}p^i_{10}(k')\right)^2 \tag{3.30}$$

The parameter q that minimizes the value for $LS(q)$ will be the best estimate for the number of hidden nodes in the system.

3.5 Numerical validation

Once the analytic framework has been developed, we want to explore the effectiveness of our model with respect to different parameters, to asses where it performs better.

We will use the pvalue to estimate if the predicted probability of changing state, for each k, is well described by our model. In particular, we will compare the aggregated set of $p^i_{10}(k) \forall i \in \mathcal{N}'$ with the predicted values from eq.(3.3), the Null-hypothesis, and eq.(3.15), the Correct-hypothesis. We will accept a model to be descriptive of the data if all pvalues for each k' in the system are above the threshold of $\alpha = 0.05$.

To mitigate false rejections of a model due to stochastic noise, we will average the number of successful acceptances of a model over $r = 50$ repetitions. Wherever the Null-hypothesis will be rejected more than the Correct-hypothesis, we can assess that our analysis to infer hidden nodes is correct.

First, in fig. 3.1 we run a simulation with $N = 10$ nodes and one hidden node. All nodes having the same activity $a = 0.5$ and the number of time-steps fixed at $time - steps = 10.000$. What we obtain is the portion of times the Null-hypothesis of no hidden nodes in the system is accepted.

We notice that for higher values of activity and lower values of μ we never accept the Null-hypothesis, as explored in sec.3.3 and expressed by eq.(3.3), this is correct since one hidden node is present in the system. For increasing values of the parameter μ, we see that the probability of erroneously accepting the Null-hypothesis increase, to the point that this test alone is not useful anymore. This is due to the increased noise in the diffusion process, hiding the effects of nodes interactions.

Conversely, if we apply the analysis discussed in sec.3.4 and expressed by eq.(3.15), our Correct-hypothesis will be that we are observing a system with hidden nodes. In fig.3.2 we notice that over all the parameter-space the performances of our analysis are consistent, and we correctly accept the hypothesis of one hidden node with a higher ratio than the Null-hypothesis.

We want to explore how different parameters affect the ability to identify if one or more nodes are hidden and if the technique is robust to heterogeneous activities. In fig.3.3 we display the difference in erroneously accepting the Null-hypothesis and correctly accepting the Correct-hypothesis. By doing so, we highlight the contribution given by our exact analytic solution in a system with heterogeneous activities. Here we see a system with $N = 10$ nodes and one hidden, with average activity $\langle a \rangle = 0.5$, for different values of standard deviation and

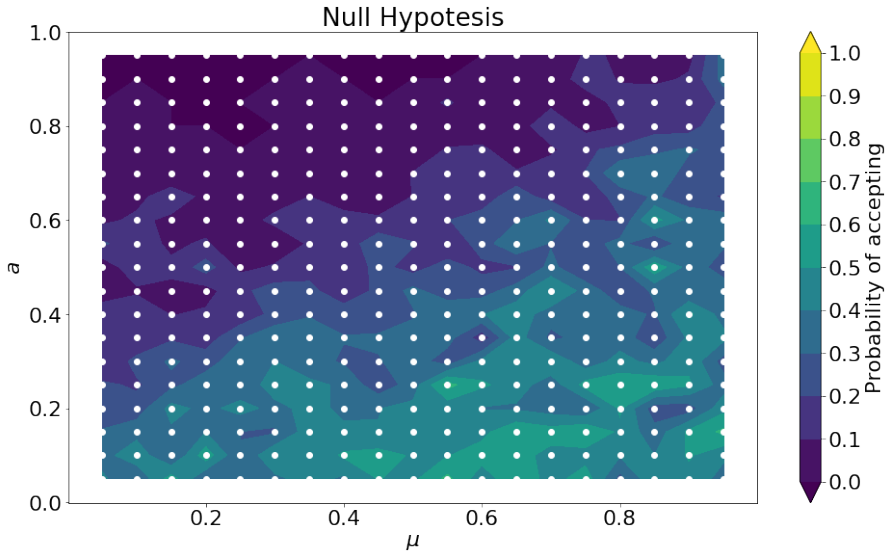


Fig. 3.1 Averaged probability of accepting the Null-hypothesis of no hidden node, time-steps = 10.000, r=50 repetitions.

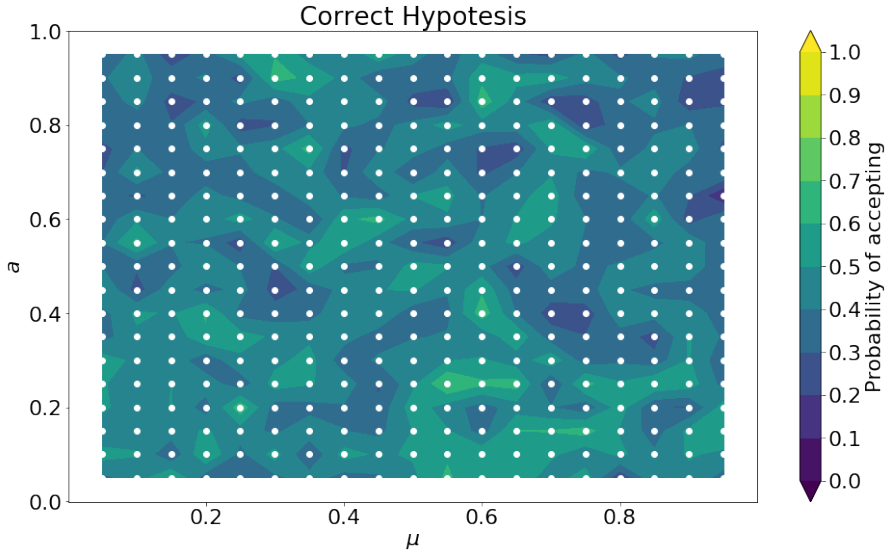


Fig. 3.2 Averaged probability of accepting the Correct-hypothesis of one hidden node, time-steps = 10.000, r=50 repetitions.

time-steps. The major difference between our Correct-hypothesis against the Null-hypothesis, and by extension the more exact predictions, are obtained for long time-series with low standard deviation in the activity distribution.

A critical aspect of a detection technique is its sensibility. To explore this characteristic, we simulated different systems of increasing size all having a single hidden node. We notice

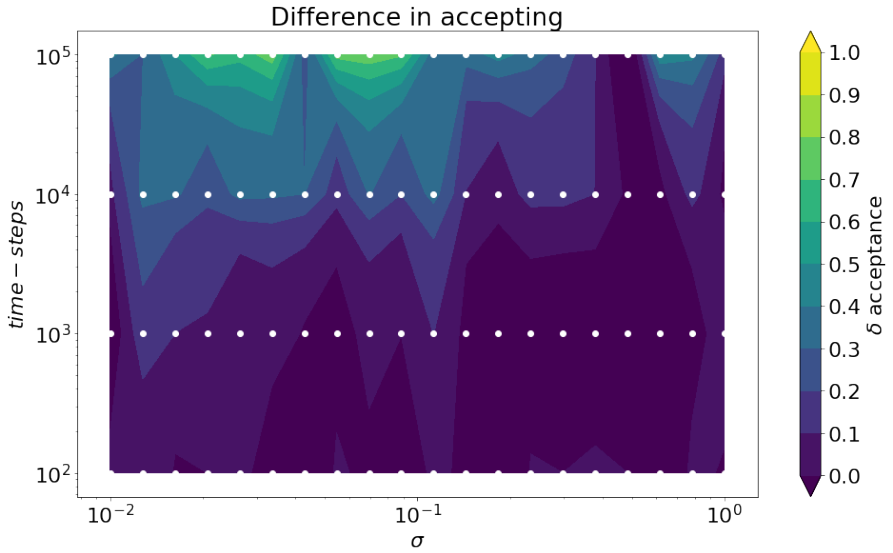


Fig. 3.3 Difference probability of accepting the Null or Correct-hypothesis, $\mu = 0.5, < a >= 0.5$, one hidden node, over $r = 50$ repetitions.

that isolating the contribution of the hidden node in bigger systems requires an increasing number of time-steps. This highlight that the contribution given by the hidden node is ever-less impacting networks of increasing size. In fig.3.4 we display the difference in the probability of erroneously accepting the Null-hypothesis and correctly accepting the Correct-hypothesis. Here we see a system with average activity $\langle a \rangle = 0.5$ and variance $\sigma = 0.01$, with exactly one hidden node. The best identification occurs for systems with small sizes and longer time steps, while the sensibility decreases as the system gets larger.

Furthermore, we test the robustness of our technique to identify a varying ratio of hidden nodes. This task is computationally expensive because, as we have previously seen, a high number of time-steps is required to have a good reconstruction. In fig.3.5 we see that our system is capable of identifying a high ratio of hidden nodes, even with an elevated heterogeneity in activities. Such capability is proof of the robustness ad accuracy of our analysis, at the cost of a longer time-series needed.

Finally, as an example, in fig.3.6 we show a realization of the process to identify the number of hidden nodes through minimization of eq.(3.30). A system with $N = 20$ nodes, of which $Q = 4$ hidden, and heterogeneous activity with $\sigma = 0.01$ is simulated. We see how the minimum value is coincident with the real number of hidden nodes, therefore allowing the observer to correctly asses the size of the system.

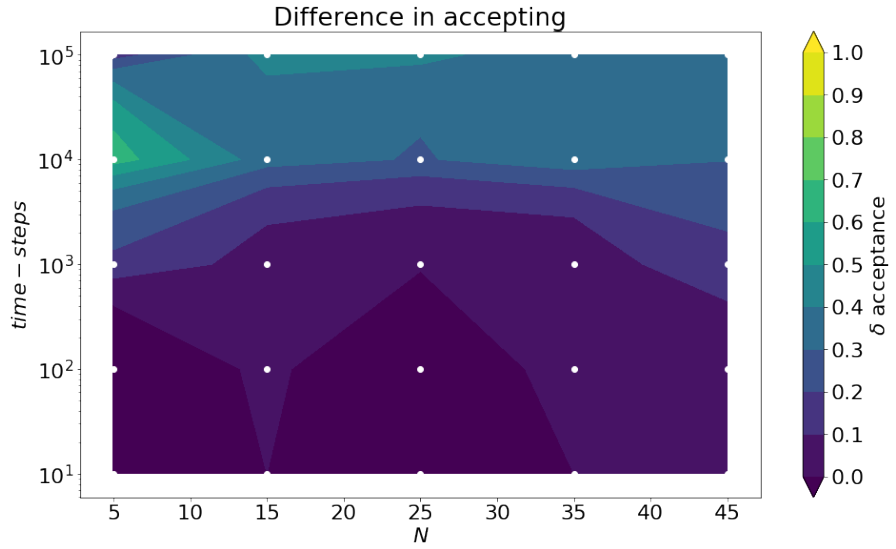


Fig. 3.4 Difference probability of accepting the null or correct hypothesis, $\mu = 0.5, < a >= 0.5, \sigma = 0.01$, over $r = 50$ repetitions.

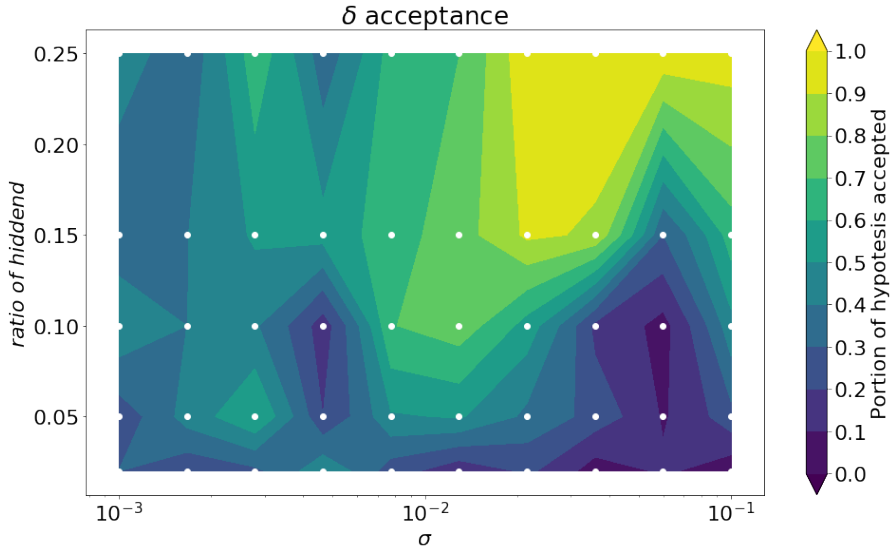


Fig. 3.5 Difference probability of accepting the null or correct hypothesis, $\mu = 0.5, < a >= 0.5$, over $r = 50$ repetitions and $time - steps = 100.000$.

3.6 Discussion

We explored a derivation of the voter model on activity driven networks. In sec.3.2 we described the dynamics of the system, and the interest it held in our work, delineating the main equations to describe it.

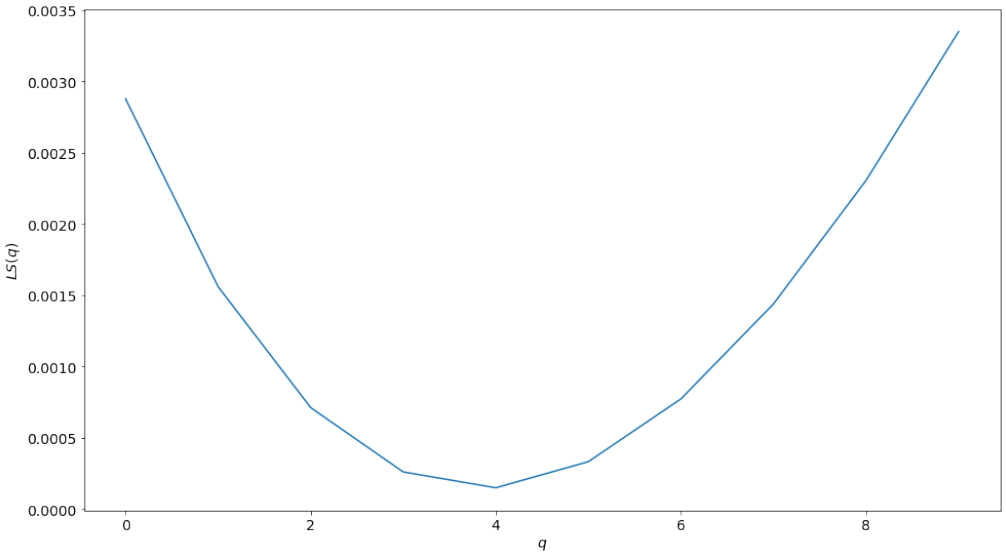


Fig. 3.6 Least Square for different values of q, $\mu = 1/N$, $< a >= 0.75$, $\sigma = 0.01$ over $time - steps = 100.000$.

Based on the described dynamic, in sec.3.3 we proposed a method to identify if a system has hidden nodes participating in the dynamics or not. Leveraging the probability of changing state of the visible nodes, we developed a statistical test to detect the presence of hidden nodes, for systems with homogeneous and heterogeneous nodes' activity.

After, in sec.3.4 we computed the exact influence that hidden nodes have on a system. Analyzing scenarios with homogeneous and heterogeneous nodes' activity, exact results are computed to infer the probability of changing state for each node.

Finally, we compared our results with different simulations in sec.3.5, in order to assess the reliability and the sensibility of our technique with respect to varying levels of noise, the number of time-steps, and variability in the activities of the nodes. A simple optimization process is proposed taking advantage of the work's results, to estimate the correct size of the system under study.

Our technique seems to be bounded by the high number of time-steps needed to correctly infer the number of hidden nodes. This is linked to the statistical nature of the techniques used to infer the probability of changing state of each node involved. Such limitation may be accentuated in large systems, where pairwise interactions among nodes carry less information, hindering the detection of hidden nodes. These limitations should be better explored in light of the big-data available for research, as we will try to do in chapter 4.

Future work should build on the present results to develop higher sensibility techniques requiring fewer time-steps to correctly infer the presence of hidden nodes. Furthermore, the presence of a backbone in the system should be studied, since stronger dyadic interaction could carry more information for an external observer on the structure under study.

Part II

A data-scientific approach

Chapter 4

Analysis of lockdown perception in the United States during the COVID-19 pandemic

In this chapter, we explore the online diffusion of sentiments and the perception of sensible topics. To contain the diffusion of the Sars-CoV-2 virus across the United States, central and local governments tried to assess the best course of action to stop the spread of this pathogen. Among the strongest actions taken, there were different forms of limitations imposed onto citizens regarding mobility and access to public spaces, leading to heterogeneous epidemiological, social, and economic effects. These actions were commonly referred to as *lockdown* measures.

We present a spatio-temporal analysis of a Twitter dataset comprising 1.3 million geo-localized Tweets about lockdown, from January to May 2020. Through sentiment analysis, we classified tweets as expressing positive or negative emotions about lockdown, demonstrating a change in perception during the course of the pandemic modulated by socio-economic factors. A transfer entropy analysis of the time-series of tweets unveiled that the emotions in different parts of the country did not evolve independently. Rather, they were mediated by spatial interactions, which were also related to socio-ecomomic factors and, arguably, to political orientations. This study constitutes a first, necessary step toward isolating the mechanisms underlying the acceptance of public health interventions from highly resolved online datasets.

4.1 Background

The word *lockdown* originated in the context of criminal justice in the middle of the 20th century [262], indicating an emergency measure in which people are temporarily prevented from entering or leaving a restricted area. Since the first wave of the SARS-CoV-2 outbreak in 2019, this word has been utilized to broadly define the measures adopted by governments and local administrations to curb the diffusion of the epidemic, by reducing individuals' mobility and in-person interactions. These measures include restricted access to shops, workplaces, and other public spaces, along with travel limitations. With their high population densities and productive and economic fabric, cities have been dramatically affected by the pandemic and its containment measures [263]. Lockdowns have had a broad and strong impact on the life of individuals and communities [264–266], who have experienced different psychological responses that evolved over time. While such measures are undoubtedly beneficial from an epidemiological point of view, their economic, social, and psychological costs cannot be denied.

The adoption of lockdown measures to curb the diffusion of COVID-19 has impacted social interactions, accelerating massive use of online platforms at a rate even faster than the spread of the epidemic [267–269]. Among social media, Twitter is one of the preferred platforms for users to express their reactions to the ongoing epidemics and related policies [270, 271]. Twitter is a micro-blogging platform, where users can write posts of up to 280 characters, including images and URLs. Users interact through re-Tweets, by forwarding the text of others on their own post stream; mentions, where users explicitly refer to others in their tweets; and follows, where users decide to permanently incorporate others' Tweets in their stream.

Twitter has been studied by researchers to investigate public opinion on a variety of topics. Notably, Twitter was extensively used to understand how the political debate evolved and was perceived [272–276], to investigate how rumors and opinions spread [277, 278], and to test the validity of models of complex social behavior [225, 279, 280]. Other efforts aimed at understanding the spread of contagious diseases that would be otherwise hard to track with traditional medical testing [281], such as influenza [241–243, 282–284], Ebola virus disease [285–287], and, more recently, COVID-19 [288].

The availability of data about COVID-19 diffusion and the access to Twitter data enabled different studies on the perception and reaction to the pandemic [289]. Typically, these studies rely on sentiment analysis, also known as opinion mining [290]. These tools are statistical techniques that explore and extract emotions conveyed by selected texts [291–296], in terms of a discrete classification or a continuous score. Twitter data on COVID-19 pandemic has

been used to study reactions to the outbreak in different countries [297–299], benchmark and validate new models for natural language processing [300–302], perform sentiment analysis about the pandemic [303–305], and to perform analysis surrounding a specific event [263].

Of the entire body of knowledge on the topic, only the study by Rahman et al. [263] frames the sentiment analysis within a socio-economic perspective, although relying on a relatively small dataset. Other authors have used Twitter to study real-time events [71], mostly relying on a limited number of interactions [306] or tackling the analysis mainly from a theoretical point of view [307]. To the best of our knowledge, sentiment analysis on a big dataset collected over long periods of time remains elusive, especially in the context of a disruptive event, such as the COVID-19 pandemic.

In this vein, the present study explores temporal variations in the emotions expressed online by Twitter users regarding lockdown measures in the United States (U.S.), starting from what is commonly referred to as the first wave of the virus (January–May, 2020). To identify the drivers of sentiment dynamics, we consider spatio-temporal variations in the severity of the pandemic, along with social, economical, and political aspects. Within an information-theoretic approach, we use the notion of transfer entropy [114] to discover causal relationships that underlie the spread of emotional content among different geographical regions in the U.S. Toward the identification of salient factors, we then proceed to a dimensionality reduction using principal component analysis. In light of the granularity and extent of the available data, we are successful in spatially correlating emotional shifts to the epidemic prevalence and socio-economic factors.

4.2 Methods

We examined the sentiment expressed in the online debate surrounding the containment policies in the U.S. between January 21st and May 31st 2020. The data we processed comprise about 55 million Tweets in English [289], as defined by Twitter's metadata. The data was subsequently filtered to retain only those originating from one of the 50 U.S. states and the District of Columbia. We performed a polar sentiment analysis [308] on all Tweets containing the word "lockdown," categorizing them as expressions of positive, negative, or neutral emotions. For each U.S. state and the District of Columbia, we recorded the daily portion of positive and negative Tweets. Alongside these data, we collected the number of daily infections in the U.S. from the publicly available dataset of the New York Times [309], and several socio-economic indicators from the Census Bureau website [310].

4.2.1 Data, pre-processing, and post-processing

Our analysis is based on the ongoing collection of data curated by Chen et al.[289], which started on January 21st, 2020, and which included more than 123 million Tweets in several languages when this project started. Complying with the Twitter privacy policy, the database contains only Tweets IDs. We used the software Hydrator [311] to retrieve the Tweets' text and metadata. Specifically, metadata are used to select only Tweets written in English. Re-tweets are not distinguished from ordinary Tweets, under the premise that a re-Twitting user expresses a form of endorsement [293].

We filtered the data set by restricting the search to Tweets containing the keywords established by the data set curator before February 16th 2020. Specifically, we used the following keywords: "Coronavirus", "Corona", "CDC", "Ncov", "Wuhan", "Outbreak", "China", "Koronavirus", "Wuhancoronavirus", "Wuhanlockdown", "N95", "Kungflu", "Epidemic", "Sinophobia", and "Covid-19". Starting from such a filtered data set, we restricted our field of analysis to those Tweets containing the term "lockdown," either as Tweet text or as a hashtag, regardless of any capitalization. Only Tweets that originated in the U.S. have been retained, through a geo-localization procedure detailed in what follows. Eventually, the data set contained about 1.3 million Tweets, monthly distributed as follows: January, 56,920; February, 40,030; March, 322,877; April, 857,612; and May, 32,865.

Multiple metadata are associated with Tweets, thereby allowing for inferring the position of the user at the time of content creation or their home and workplace. The largest portion of Tweets monthly, ranging in $(99.69\% - 99.92\%)$ have a user-defined location. This is likely connected to users' home or workplace [312], although it may not reflect their exact position and, sometimes, does not contain meaningful information (referring, for example, to imaginary places, or to whole countries [312]). A much smaller portion of Tweets is associated with platform-generated locations, based on the Tweet content $(0.11\% - 0.26\%)$. An even smaller portion of Tweets contains a GPS location $(0.02\% - 0.08\%)$.

To associate specific coordinates to each Tweet we relied on the geoparsing software CLIFF-CLAVIN [313]. Upon retrieval of a geographical entity in the Tweet, we used the open data provided by OpenStreetMap Contributors© to determine the country of origin. If the Tweet originated in the U.S., we sought to narrow the origin to any of the 50 states or the District of Columbia. In case of conflicting information regarding the state of origin, we discarded the Tweet.

We studied polarization and changes in sentiment in the online debate about the topic of lockdown using a classification of emotions aroused by text, in positive, neutral, or negative. Such an analysis was performed using VADER [308], a valence-aware sentiment analysis tool. For each Tweet, VADER assigns a composite score that is used for classification. Specifically, following [308], we selected three thresholds to assign an emotional quality to each Tweet. Composite scores below -0.050 were classified as carrying negative emotions; between -0.050 and 0.050 as neutral; and beyond 0.050 as carrying positive emotions.

By performing sentiment analysis on the geo-localized Tweets, we created two local time-series for each region (all the U.S. states and the District of Columbia), namely, daily fractions of positive Tweets $\rho_P(t)$ and negative Tweets, $\rho_N(t)$. In total, we collected 102 local time-series, with the resolution of one day, each one with a length of 132 days.

To acknowledge country-wise changes in the perception of the pandemic, we partitioned each time-series (from the fifty states and the District of Columbia) into three sections: before the onset of the pandemic (the first day in which the incidence of $5/10,000,000$ daily cases in the population of the corresponding region was registered), from the onset of the pandemic to the first peak of the infection incidence (evaluated using a moving weekly average), and from the peak to the end of May 2020.

For each region, we studied the time-series of the portion of positive and negative Tweets over the total number of Tweets, $\rho_P(t)$ and $\rho_N(t)$. From each of these time-series, we computed the average values over the three sections, ρ_P^i and ρ_N^i, and the standard deviations, σ_P^i and σ_N^i, with $i = \{1,2,3\}$. To ascertain time variations in the positive and negative sentiments across the three sections, we used Welch's t-test with a significance level of 0.050.

4.2.2 Socio-economic factors

We considered education and wealth indicators from the 2018 data of the U.S. Census Bureau [310]. For each region (U.S. state and the District of Columbia), we collected the corresponding data for Population (POP), Median Household Income (MHI), and the following rates: Poverty (PR), Employment (ER), Uninsured (UR), High School Diploma (or higher level, HSD), Bachelor Degree (BD), and Professional or Doctoral Degree (PDD).

To consolidate the number of explanatory variables into interpretable indicators [314], we performed a principal component analysis on these socio-economic factors [315]. We retained three main components, accounting for 73% of the total variance and all having a corresponding

eigenvalue above 0.995. We excluded variables contributing to a principal component with an absolute loading lower than 0.500. The first principal component, accounting for 37% of the variance, is interpreted as "Wealth" and is mainly associated with the poverty rate (principal component loading equal to -0.958), the employment rate (0.816), rate of Bachelor Degree (0.768), and median household income (0.673). The second principal component, accounting for 27% of the variance, is interpreted as "Education" and is mainly associated with the rate of Professional or Doctoral Degree (loading equal to 0.940), the Median Household Income (0.599), the rate of Bachelor Degree (0.557), and the rate of High School Diplomas (-0.523). Finally, the third principal component, accounting for 10% of the variance, is interpreted as "Social Exclusion" and is mainly associated with the rate of high school degree (-0.562) and the rate of uninsured (loading equal to 0.553).

The obtained principal component scores were used as dependent variables in a Kendall correlation test [316] with combinations of sentiment analysis parameters. The null-hypothesis of independence was tested with a two-sided test with $p < 0.050$.

4.2.3 Spatial interactions

Given the massive use of Twitter throughout the country, it is tenable to expect that local sentiment does not evolve in silos, but is the result of a spatial influence process. Hence, we studied the influence of sentiments among regions. We pursued this analysis through an information-theoretic approach based on the notion of transfer entropy. Transfer entropy is designed to unveil cause-and-effect relationships in a Wiener-Granger sense. Specifically, a process X is said to cause another process Y if knowledge of the present state of X improves the prediction of the future of Y from its present [114].

We separately studied spatial interactions associated with positive and negative Tweets. For each type of Tweet, we computed transfer entropy between any pair of local time-series, totaling $51 \times 50 = 2,550$ values of transfer entropy. To control for common-driver effects in the evolution of time-series (for example, one state simultaneously influencing two other states that would otherwise be independent), we conditioned over the average of positive or negative Tweets across the entire country. Specifically, given a source process X (local time-series of positive or negative Tweets), a target process Y (local time-series of positive or negative Tweets), and the conditioning process Z (national average of time-series of positive or negative

Tweets), we computed conditional transfer entropy as

$$TE_{X\to Y|Z} = H(Y(t+1)|Y(t),Z(t)) - H(Y(t+1)|Y(t),X(t),Z(t)), \tag{4.1}$$

where $H(\cdot)$ is the Shannon entropy.

In the computation of transfer entropy, we used a symbolic representation with a binary alphabet to ensure the accuracy of the estimation of the probability mass functions in the Shannon entropy, similar to our previous work [317]. Specifically, we first detrended the local time-series of positive and negative Tweets by subtracting at each instant of time the average value of the corresponding time section (before the onset of the pandemic, from the onset of the pandemic to the incidence peak, from the incidence peak to the end of May 2020); we verified the stationarity of the time-series using a Dickey-Fuller test [318]. Then, we symbolized the time-series into a sequence of binary symbols: $\uparrow$ and $\downarrow$, associated with daily values above or below the median, respectively. This transformation is performed separately for both the time-series of positive and negative Tweets, obtaining a total of 102 symbolic time-series.

Statistical testing was performed by following the approach presented in [319]. To test whether transfer entropy in Eq. (4.1) was different from chance, we created a surrogate distribution by shuffling the values of the source process, while preserving the associations between the target and conditional processes. A total of $10,000$ permutations were executed for each statistical test and a significance level of 0.050 was considered.

Hence, for every pair of candidate target and source processes, we rejected (or failed to reject) the null hypothesis that their directional interaction from positive or negative Tweets was due to chance. Through this analysis, we determined two directed networks, one from spatial influences inferred from positive Tweets, and the other from negative Tweets, in which a link signifies rejection of the null hypothesis. To highlight the strongest patterns of spatial influence, we studied the normalized in-degree centrality, $K_{(N,P),in}$ and the normalized out-degree centrality, $K_{(N,P),out}$ [320] of the obtained networks.

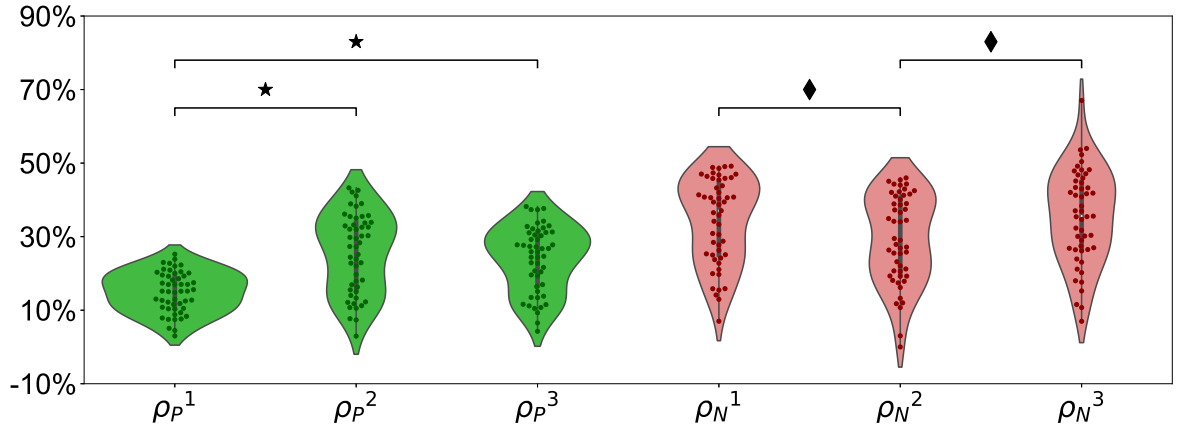
Using the directed networks and the centralities described above, we investigated potential associations between socio-economic factors and spatial influence patterns through Kendall-τ correlation tests using a two-sided significance threshold of $p < 0.050$. In addition, we sought to connect these patterns to political ideology, as defined by Berry et al. [321] and using updated 2018 data from professor R.C. Fording [322]. To this aim, we assigned to each region a label, either “liberal" or “conservative", and then we counted in any of the two networks the number of links connecting nodes with the same or different ideology.

4.3 Results

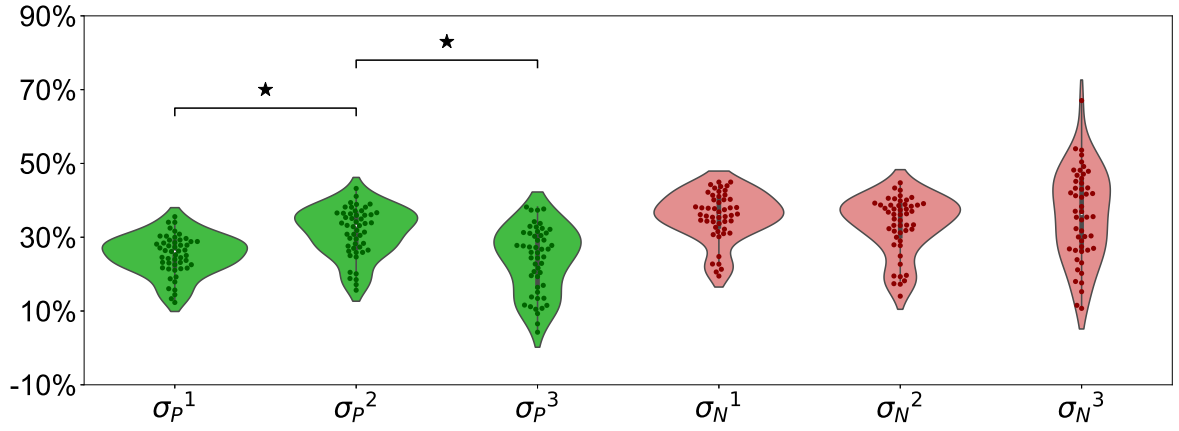
Across time, we registered a variation in both the means of the positive and negative Tweets (Fig. 4.1a). Specifically, the portion of positive Tweets before the onset of the pandemic was lower than the section between the onset of the pandemic and the incidence peak ($t_{74.33} = 6.24$, $p < 0.001$) and then the section from the incidence peak to the end of May 2020 ($t_{83.12} = 6.12$, $p < 0.001$). We did not register a difference between the portion of positive Tweets from the central section to the last section ($t_{96.56} = 0.82$, $p = 0.416$). Likewise, we determined a temporal variation in the portion of negative Tweets, whereby the central section was higher than the initial one ($t_{99.88} = 2.04$, $p = 0.045$) and the last section was higher than the central section ($t_{99.67} = 2.44$, $p = 0.016$). However, such differences did not reverberate into a significant change from the first to the last section ($t_{99.18} = 0.50$, $p = 0.620$).

Differences in the mean portion of positive Tweets in time were accompanied by changes in their variability (Fig. 4.1b). Specifically, the standard deviation showed an inverted U-shape, by increasing from the first to the second section ($t_{96.16} = 5.26$, $p < 0.001$) and decreasing from the second to the third ($t_{91.78} = 4.81$, $p < 0.001$); no difference was registered when comparing the first with the last section ($t_{81.86} = 0.84$, $p = 0.405$). On the other hand, the variability of the portion of negative Tweets was indistinguishable in time (first versus second: $t_{95.06} = 1.66$, $p = 0.101$; second versus third: $t_{83.18} = 1.27$, $p = 0.207$; and first versus third: $t_{74.82} = 0.11$, $p = 0.910$).

We further investigated the correlation between socio-economic factors and the shift in sentiment across the three time sections (Table 4.1). The variation in the portion of positive Tweets before the onset of the pandemic and between the onset of the pandemic and the incidence peak correlates with all the identified socio-economic factors: negatively with Wealth ($\tau = -0.442$, $p < 0.001$), and positively with Education and Social Exclusion ($\tau = 0.500$, $p < 0.001$; $\tau = 0.487$, $p < 0.001$; respectively). We did not observe a correlation when examining the variation in the portion of positive Tweets between the onset and the peak and after the peak with neither Wealth ($\tau = 0.183$, $p = 0.058$) nor Social Exclusion ($\tau = -0.228$, $p = 0.270$). On the other hand, we recorded a correlation with Education ($\tau = -0.235$, $p = 0.015$). Exploring the correlation between socio-economic factors and the variation in the portion of negative Tweets, we did not find a correlation between the variation from the first to the second time sections and Wealth ($\tau = -0.112$, $p = 0.245$), Education ($\tau = -0.079$, $p = 0.412$) or Social Exclusion ($\tau = 0.082$, $p = 0.393$). Likewise, we did not register a correlation between the variation in negative Tweets between the second and the third time



(a) Average portion of daily Tweets in each period



(b) Standard deviation in time of the portion of daily Tweets in each period

Fig. 4.1 Green and red violin plots represent Tweets corresponding to positive and negative sentiments, respectively. Each point represent the value for any of the state or the District of Columbia. Stars indicate significant comparisons at $p < 0.001$ and diamonds at $p < 0.050$.

sections and Wealth ($\tau = 0.106$, $p = 0.273$), Education ($\tau = -0.101$, $p = 0.295$), or Social Exclusion ($\tau = -0.107$, $p = 0.266$).

Not only were socio-economic factors associated with the averages of the portions of Tweets, but also were they related to the standard deviations in time of the portions of Tweets (Table 4.1). Across the first and second time sections, we did not register a correlation of the change of the standard deviation of positive Tweets with Wealth ($\tau = 0.082$, $p = 0.394$), Education ($\tau = -0.049$, $p = 0.609$) or Social Exclusion ($\tau = -0.059$, $p = 0.542$). Differently, such a correlation for the same data is observed between the second and the third time sections, namely, negatively with Wealth ($\tau = -0.536$, $p < 0.001$) and positively with both Education ($\tau = 0.550$, $p < 0.001$) and Social Exclusion ($\tau = 0.540$, $p < 0.001$). The variation in standard deviation of the portion of negative Tweets between the first and the second time sections did not correlate with Wealth ($\tau = 0.061$, $p = 0.530$), Education ($\tau = -0.086$, $p = 0.380$), or Social exclusion ($\tau = -0.086$, $p = 0.380$). With respect to the standard deviation in the portion of negative Tweets between the second and the third time sections, we registered a negative correlation with Wealth ($\tau = -0.528$, $p < 0.001$), and a positive correlation with both Education ($\tau = 0.556$, $p < 0.001$) and Social Exclusion ($\tau = 0.543$, $p < 0.001$).

In Fig. 4.2, we illustrate a cartographic map obtained from the transfer entropy analysis. Therein, each state is colored based on the in-degree (top images) and out-degree (bottom images) centrality as computed from the time-series of positive (green) and negative (red) Tweets: the higher the out-degree (in-degree) the higher the influence exerted (experienced) by a node on (from) the rest of the network. In total, the network of positive Tweets has 249 directed edges, whereas the network of negative Tweets is composed of 146 directed edges.

Table 4.1 Kendall-τ coefficients for the correlation between socio-economic factors and changes in the averages and standard deviations of the portions of positive and negative Tweets. Numbers in parentheses report the p-value from the correlation; a bold value indicates $p < 0.050$.

Kendall-τ	Wealth	Education	Social Exclusion
$\rho_P^2 - \rho_P^1$	**-0.442** ($p < 0.001$)	**0.500** ($p < 0.001$)	**0.487** ($p < 0.001$)
$\rho_P^3 - \rho_P^2$	0.183 ($p = 0.058$)	**-0.235** ($p = 0.015$)	-0.228 ($p = 0.270$)
$\rho_N^2 - \rho_N^1$	-0.112 ($p = 0.245$)	0.079 ($p = 0.412$)	0.082 ($p = 0.393$)
$\rho_N^3 - \rho_N^2$	0.106 ($p = 0.273$)	-0.101 ($p = 0.295$)	-0.107 ($p = 0.266$)
$\sigma_P^2 - \sigma_P^1$	0.082 ($p = 0.394$)	-0.049 ($p = 0.609$)	-0.059 ($p = 0.542$)
$\sigma_P^3 - \sigma_P^2$	**-0.536** ($p < 0.001$)	**0.550** ($p < 0.001$)	**0.540** ($p < 0.001$)
$\sigma_N^2 - \sigma_N^1$	0.061 ($p = 0.530$)	-0.086 ($p = 0.380$)	-0.086 ($p = 0.380$)
$\sigma_N^3 - \sigma_N^2$	**-0.528** ($p < 0.001$)	**0.556** ($p < 0.001$)	**0.543** ($p < 0.001$)

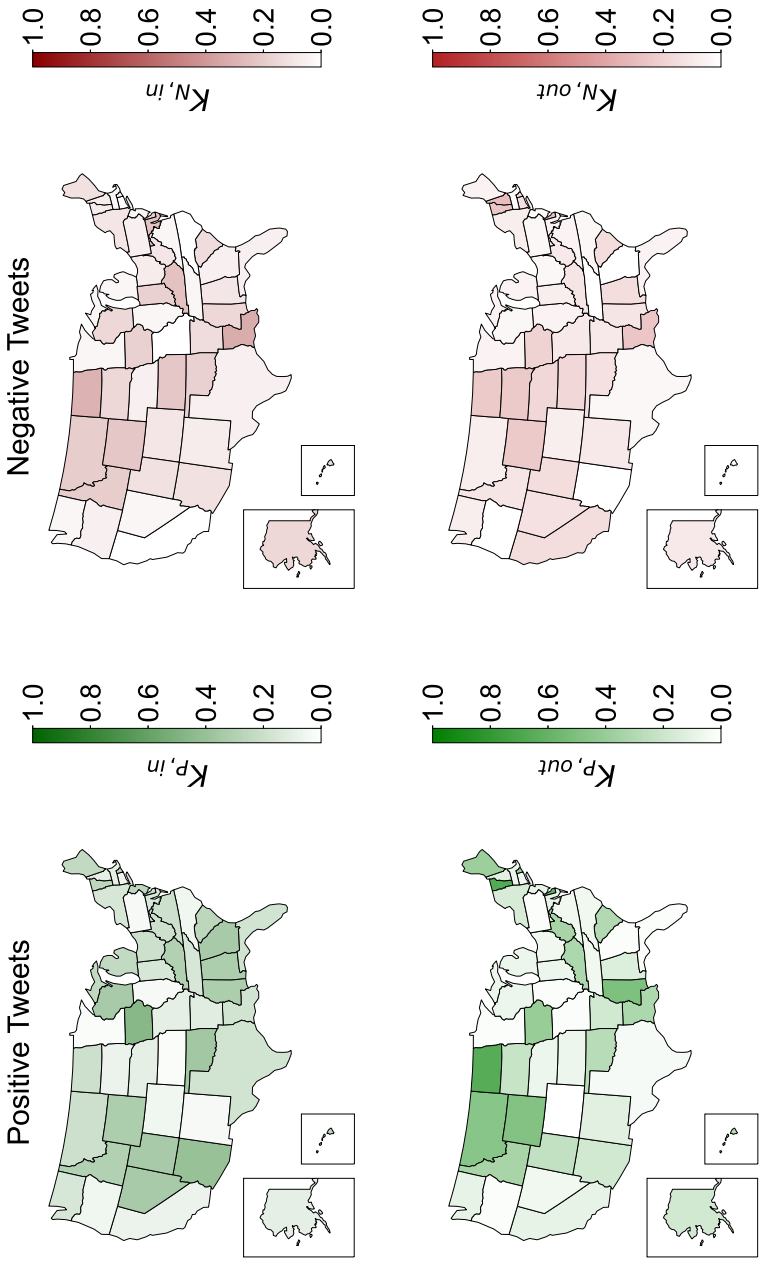


Fig. 4.2 Maps of the U.S. showing the in-degree (top) and out-degree (bottom) distributions associated with the networks for positive (green) and negative (red) Tweets.

Table 4.2 Kendall-τ coefficients between socio-economic factors and either in- or out-degrees from the portions of positive and negative Tweets. Numbers in parentheses report the p-value from the correlation; a bold value indicates $p < 0.050$.

Kendall-τ	Wealth	Education	Social Exclusion
$K_{P,out}$	**0.468** ($p < 0.001$)	**-0.525** ($p < 0.001$)	**-0.517** ($p < 0.001$)
$K_{P,in}$	0.087 ($p = 0.383$)	-0.099 ($p = 0.322$)	-0.102 ($p = 0.306$)
$K_{N,out}$	**0.428** ($p < 0.001$)	**-0.437** ($p < 0.001$)	**-0.440** ($p < 0.001$)
$K_{N,in}$	**0.313** ($p = 0.002$)	**-0.341** ($p < 0.001$)	**-0.342** ($p < 0.001$)

The in-degrees of each region, computed from the network of positive Tweets, correlate negatively with Wealth ($\tau = -0.442, p < 0.001$) and positively with Education and Social Exclusion ($\tau = -0.442, p < 0.001$; $\tau = -0.442, p < 0.001$; respectively, Table 4.2). On the other hand, the out-degrees computed from the same network do not correlate with any of the socio-economic factors, let them be Wealth ($\tau = 0.087, p = 0.383$), Education ($\tau = -0.099, p = 0.322$), or Social Exclusion ($\tau = -0.102, p = 0.306$). The same analysis was performed on the centrality measures for the network of negative Tweets. Here, we recorded a positive correlation between the out-degree and Wealth ($\tau = 0.428, p < 0.001$), and a negative correlation with Education and Social Exclusion ($\tau = -0.437, p < 0.001$; $\tau = -0.440, p < 0.001$; respectively). A similar pattern was noted for the in-degree, which also entailed a positive correlation with Wealth ($\tau = 0.313, p = 0.002$) and a negative correlation with Education and Social Exclusion ($\tau = -0.341, p < 0.001$; $\tau = -0.342, p < 0.001$; respectively).

Finally, we performed a cluster analysis on the networks based on the liberal or conservative ideologies of the corresponding nodes. For the network associated with the positive Tweets, out of the existing 249 edges, we determined 87 (34.9%) links from conservative to liberal, 76 (30.5%) from conservative to conservative, 46 (18.5%) from liberal to conservative, and 40 (16.1%) from liberal to liberal. For the network related to negative Tweets, out of the 146

edges, 34 (23.3%) were from conservative to liberal nodes, 48 (32.9%) from conservative to conservative, 37 (25.3%) from liberal to conservative, and 27 (18.5%) from liberal to liberal.

4.4 Discussion

The first wave of SARS-CoV-2 has impacted the health, wealth, and the life of millions of people all over the country. Information about the pandemic has spread over the globe, creating waves of polarized emotions and, at times, influencing actions in response to the ongoing crisis. A controversial debate has emerged about the application of strict containment policies, such as severe lockdowns and travel bans. Opinions have been extremely heterogeneous across geographical regions and social strata [323].

Here, we analyzed online sentiment on Twitter from January 21st to May 31th, 2020 in the U.S. about lockdown measures. Beyond qualitatively describing the opinion throughout the country, we sought to dissect potential explanations and causal mechanisms. In this vein, we pursued a principal component analysis of socio-economic factors to consolidate variations across the country in a few salient explanatory variables (Wealth, Education, and Social Exclusion). Alongside, we conducted a transfer entropy study to unveil spatial interactions among different regions of the country (states and the District of Columbia).

In agreement with our expectations, we registered a time variation of public opinion regarding lockdown measures. People expressed support for lockdown measures in the early stage of the pandemic, whereby the portion of positive Tweets increased and the portion of negative Tweets decreased. It is likely that risk perception regarding the spreading of the infection caused fear in the population, spurring emotional changes toward containment measures that were evident from our Twitter dataset. As the pandemic progressed, the portion of positive Tweets remained leveled and the negative Tweets raised, suggesting that pandemic fatigue, stress, and isolation started taking a toll [324] in how people felt about lockdowns.

Interestingly, the U.S. did not react uniformly, so that different parts of the country responded differently to the pandemic as a function of socio-economic factors. In the initial stage of the pandemic, lower Wealth and higher Education and Social Exclusion contributed to the raise in positive emotions around lockdown policies. Educated individuals, but also those fearing for their health due to poverty and lack of social safety nets, were more favorable to containment measures.

As the pandemic progressed and people changed their views regarding lockdowns, these correlations were lost and, sometimes, even reversed. In particular, neither Wealth nor Social Exclusion were explanatory of the changes in positive emotions regarding lockdown. Education became negatively correlated with the sentiment change so that people living in more affluent regions with a higher portion of college graduates were those who reduced the most their support to lockdown measures. Perhaps, this reflected some sort of cheering for the end of restrictions or the final acceptance of the new normalcy by those individuals who kept abreast of advancements in the combat against the pandemic. We warn care when interpreting this claim, whereby its statistical significance was drastically lower than any other of the observed associations and higher education was also positively correlated with changes in the temporal variability of positive sentiments, registered in our Twitter dataset and echoed by online debates [325]. As a result, claims drawn on changes in the mean values may not be indicative of a true change in sentiment.

It is tenable that the complex response of the U.S. to lockdown was mediated by spatial interactions supporting the spread of opinions across state borders. Our transfer entropy analysis offers evidence in this direction, whereby we detected close to four hundred dyadic interactions in relation to positive and negative Tweets. In agreement with one's expectation, the distribution of these links was not random, but rather it was informed by socio-economic factors. People living in regions with a higher Wealth tended to have a higher influence on how the rest of the country perceived lockdowns, whether through positive or negative emotions. Such an influence was, instead, moderated by Education and Social Exclusion, which may exacerbate political and cultural polarization, as well as differences in the very use of Twitter [326, 327].

Interestingly, we discovered that these associations would also underlie the tendency of a region to be influenced by, rather than influence, others with respect to negative emotions. Negative emotions are likely to resonate more in wealthier parts of the country, which could have been more worried about the downturn caused by the pandemic [328]. Such a worry was indeed mitigated by higher levels of education and the presence of social safety nets. Perhaps, political orientations could play a role in these spatial interactions, but present evidence is not conclusive. We speculate that the positions on lockdowns taken by the two major parties were partly responsible for the observed spatial interactions, with conservative states playing a more influential role in opinion spreading.

Our approach is not free of limitations. First, we acknowledge that the Twitter database could be excessively widespread [329], thereby challenging the retrieval of pertinent information from selected keywords, especially when dealing with a new topic. Second, sentiment

analysis may not allow for a deeper understanding of nuances or sarcasm [308], thereby confounding the classification of some of the Tweets in a database. Third, the use of aggregated socio-economic data only allows for the study of macroscopic phenomena without capturing fine details of human behavior.

There are several routes for future inquiry from this effort. In principle, our analysis could be expanded to encompass different sentiment analysis of Tweets than a simple positive/negative classification, at the cost of a more intricate interpretation of results. Likewise, our correlation studies could be undertaken without the use of a principal component analysis on socio-economic factors, thereby allowing for a more detailed assessment of potential drivers. Further work could also address a finer resolution of time effects, rather than the coarse three-section representation proposed in this chapter. The use of a finer resolution may help elucidate sentiment dynamics in the online debate, potentially assisting in the inference of key attributes of Tweets that become viral. Although our focus was the ongoing pandemic, the approach presented in this chapter could be beneficial to policy-makers when dealing with unpopular, yet timely, interventions in general [323]

Part III

An experimental approach

Chapter 5

An open source framework to study face-to-face interactions

In this chapter, we present an open-source framework that allows for realizing ad-hoc experimental settings in order to gather data on face-to-face and proximity interactions between individuals, with a high temporal and spatial resolution. This tool, composed of multiple sub-systems, was created over the course of three years to fill the need for a ready-to-use solution for researchers to quickly, easily, and inexpensively perform social experiments on human interactions. Therefore we developed two smartphone applications for the main operating systems in the market, that leverage Bluetooth®'s received signal strength indication (RSSI) to assess the approximate distance between devices, and therefore participants. The applications do not impact the normal operation of smartphones, allowing anonymized and GDPR-compliant [330] data to be sent to a server of the researchers' choice. Here we present the main features and functions of the system, as well as the results of preliminary experiments and data analysis, which provide an assessment of the framework's capability to be used in order to research human interactions in real-life scenarios.

5.1 Background

Two of the main fields of network science have been the theoretical one and the experimental one [6, 64, 195, 235–237]. Leveraging data and quantitative observations has been essential to discover counter-intuitive phenomena and emerging behaviors. Sparked from observation, analytic descriptions of networks and dynamical systems shape our field of research.

The *Zachary Karate Club* [331] is obtained through oral interviews in the seventies, leaving to the researcher the decision on what to define as an interpersonal link. Since the advent of social networks and of electronic databases, the methodologies used to study human interactions have changed [238–243]. Particularly in recent years, new technologies have been put to use of understanding how humans interact and behave. Ad-hoc devices or re-purposed ones can be leveraged to infer coarse positions of individuals [259, 332], or in-depth day-to-day routines [333]. With adequate resources, extensive experiments and analysis can be performed, where participants are equipped with dedicated devices capable of multiple measures for extended periods of time [334–337]. Yet these solutions are out of reach for most researchers, and only selected institutions may afford to embark in such projects.

Recently, more cost-effective devices have shown great potential to perform social experiments of in-person interactions. Still needing to be provided to each participant, and with reduced capabilities, single-function low-cost devices have allowed for the study of face-to-face contact patterns [84, 194, 338, 339]. Wearable sensors have achieved good results in recording and quantifying meaningful interactions when compared to participant interviews [87, 244]. But this technology still suffers from a range of drawbacks: the usable range is limited, the devices can break or malfunction, the devices have to be worn by participants and need to be recharged frequently, and only one experiment at a time can be performed since the number of wearable sensors is limited. Furthermore, a special infrastructure and logistics must be set up for the data collection to work properly.

5.2 Development

5.2.1 Functioning principle

The core technology of the proposed framework is the Bluetooth® wireless technology. Bluetooth® has been proposed and used [340–343] extensively for indoor positioning solutions, both in research and industrial applications. This technology has been widely adopted on the vast majority of consumer-level handheld devices, particularly smartphones, where it is used to interface the device with accessories. Bluetooth® technology is nowadays mainly used to transfer data, such as audio streams, between two or more devices. The maximum distance that such a wireless connection can cover is usually between 10 and 100 meters, depending on the protocol version and hardware specifications [344].

Bluetooth® technology can be leveraged for indoor positioning because the protocol also encompasses the reading of received signal strength indication (RSSI), a simple measure in decibels of the attenuation of the received signal. Being the power of the transmitter known, and set within a range by the nature of the protocol itself, a logarithmic regression can be used to infer the distance between two Bluetooth® enabled devices. Sadly, different hardware producers have different tolerances in upholding the standard, and usually, a direct comparison between signals from different hardware needs to be compensated. Furthermore, the gain and the orientation of the antenna can differ from one handheld device to another, and that coupled with the unknown orientation of the device and the surrounding effects of the environment, make this technology usable only for moderate precision, in the order of magnitude of one meter [345].

5.2.2 Smartphone applications

Right now the market for smartphone's operating systems is dominated by two platforms [346]: Android™ developed by Google Limited Liability Company; and iOS, developed by Apple Incorporated. Both operative systems originate from the UNIX® operative system, and both are designed for ARM® chip architecture. This makes for a good degree of interoperability and communication between the two platforms. Both platforms are capable of running a compatible Bluetooth® module to transmit and receive data. We decided to develop two smartphone applications, one for Android® and one for iOS operating systems, capable of exchanging data via Bluetooth® in order to exploit the previously mentioned capability to infer the relative position of the devices participating in an experiment. The two applications are made freely available on the respective online stores:

- *https://play.google.com/store/apps/details?id=com.polito.humantohuman*
- *https://apps.apple.com/it/app/human2human/id1585798094*

In order to develop an effective tool for measuring participants' proximity and interactions, we aimed at some fundamental goals in the development of our smartphone application:

- The application shall not cause abnormal use of energy, avoiding draining the device's battery too fast
- Minimal impact should be felt on the processing power and memory capacity of the device

- No disruption of the normal Bluetooth® radio functionality should be felt in the use of the device
- It must be possible for the user to limit or stop data transmission from and to the server, as it is in some countries very expensive to use cellular connection for data transmission

From Bluetooth® version 4.0 the feature Bluetooth Low Energy (BLE) has been introduced in the protocol, which limits the energy consumption of the Bluetooth® hardware. Leveraging this version's new modulation protocol and features, we adopted BLE for our project. It has proven a reliable and stable solution, compatible with a multitude of devices.

Special attention should be put on granting anonymity to all participants involved. In order to do so, we assign to each device joining an experiment via the developed applications a unique identifier, which is shared with a predetermined backend server. The association between the identifier and the device is known only to the user, who does not have to disclose it to join an experiment. Other devices in proximity of a device enrolled in an experiment will not be registered, due to the inherent implementation of the unique identifier into the overflow area of the Bluetooth® protocol.

In order to achieve the goals listed before, we used in an innovative way the overflow area of the Bluetooth® protocol. This area is designed to give some basic information to a device before a connection is established, therefore it is not shown to the user, as it is part of the communication that happens in the background between devices. Each Bluetooth® device that is available for a connection advertises some of these features, like being a sound accessory, a vehicle multimedia system, or a printer, to other devices by providing a 128bit identification code, the UUID. Usually part of this code is unused, and therefore can be manipulated without repercussion on the device functionality. We exploited an 8bit section of the UUID to set a unique identifier linked to the device in our experiment. In this way, even if the normal operation of our application is interrupted by the operating system, the UUID will remain set and visible for other devices to be picked-up, allowing for the device to be still visible to others running the application.

The application's user-interface is very simple, as the interaction with the user is minimal. Here we will focus on the application developed for Android™ operative system to explain the general functions as seen by the user. The application developed for the iOS operative system has been designed to have similar functionalities and interaction, with minor layout and options differences.

The landing page of the application has general information about its purpose and the researchers that developed it, see fig.5.1a or 5.2a. As the user is not enrolled in any experiment yet, no other option is available but the 'SETTINGS' button. Pressing the button lands the user on the second page of the application, see fig.5.1b or 5.2b, in which a box is available to set the server address chosen for the experiment to join. The aforementioned server must have been implemented using the complimentary backend package, freely and openly available on our GitHub repository [347, 122], and better described in section 5.2.3. Once the server address is set up, the user is automatically prompted to the consent form, a text downloaded from the server. The consent form is customizable and uploaded from the server side by researchers, as it must be prepared to fit the privacy requirements for the experiment being conducted, eg. the GDPR laws in Europe. If the user does not accept the consent form, the whole application is reset to the initial state. This is done in order to avoid any data being collected or transmitted unwillingly from the user. Once the consent form is accepted, the application creates a random ID for the device, which is sent to the server, and the overflow Bluetooth® area is set accordingly. Must be noted that no connection can be inferred by researchers looking at the data on the server between a user and the generated ID, as no details concerning the user of the device are recorded.

When the device is properly enrolled in the experiment on the chosen server, a set of new options are made available. First of all in the settings page, in figure 5.1b or 5.2b, the user can review the consent form by clicking the button 'CONSENT FORM', which opens again the consent form page. If the user decides to withdraw the consent by un-ticking the selection box, a command is automatically sent to the server associated with the experiment: all data connected with the ID communicated by the device will be permanently erased. This may seem like a radical solution, but it ensures perfect data protection from the users' side, as they can remove all data collected autonomously. This measure does not act on backups or copies that may have been done outside the scope of the provided server backend package. The second button available on this page will be 'LEAVE EXPERIMENT'. By selecting this command, the application un-enrolls from the experiment, without erasing data from the server. In fact, it just resets the application to the initial stage, and stops all actions of recording and sending data to the server.

When the device is enrolled in an experiment, see figure 5.1a and 5.2b, the user will have two new controls available: 'COLLECT DATA' and 'ONLY WIFI' for the Android® system, and 'COLLECT DATA' for the iOS system. The first command 'COLLECT DATA' allows the users to select when their device is going to collect data to be sent to the server, and

when to advertise the device presence via Bluetooth® . The first time this action is triggered, the operative system may request the user to authorize the application to access Bluetooth® capabilities, per the creator of the operative system policy. When this command gets turned off, the device stops the recording of Bluetooth® data but completes the sending of registered data to the server. The command 'ONLY WIFI' could be implemented only for the Android® version of the application, due to the system's restrictions in iOS. This option makes the application send the collected information to the server only if the device's WiFi connection is active. This option is made available to the users as sending data through the cellular network may be expensive in some countries. When the option is not selected, the device will use the system default to send data to the server.

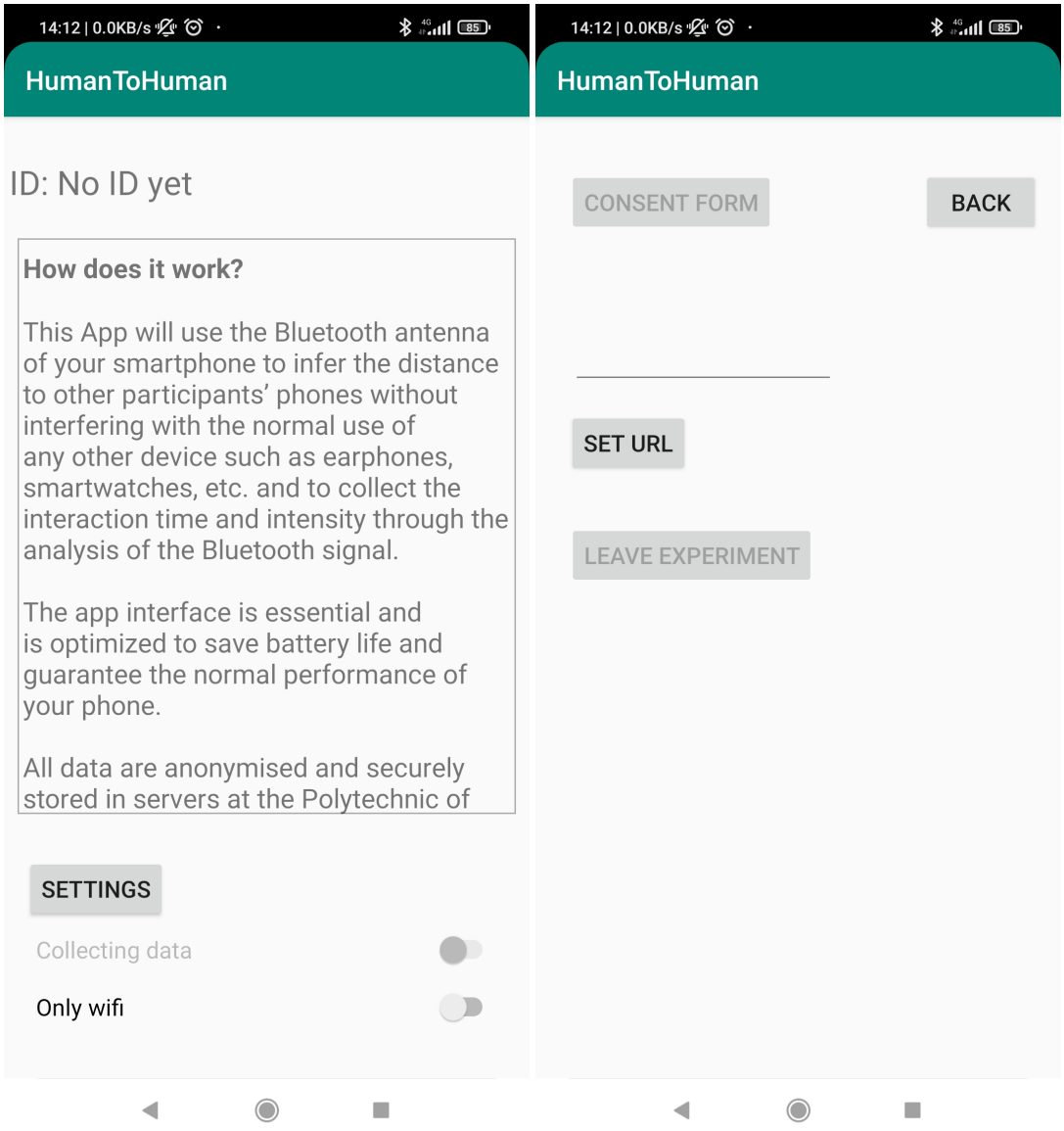
5.2.3 Server backend

The server backend is made available to complement the smartphone applications we are releasing. Our goal was to implement software that could allow researchers from different backgrounds and skills to use it effectively:

- The system must be easy to set up
- The software must run on low-power and low specs machines
- Data extraction must be simple
- A level of customization must be available
- Data protection must be integrated

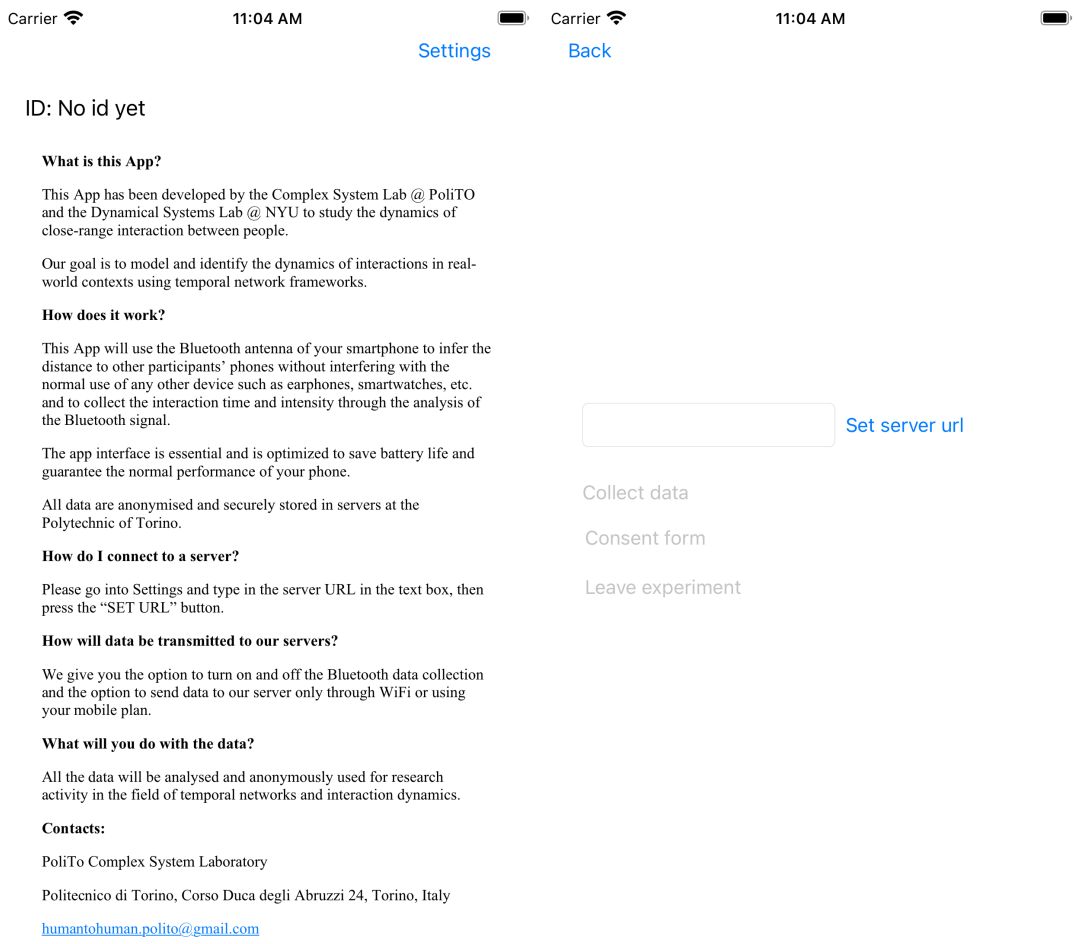
To achieve these goals, we used Python and GO programming languages to develop a REST API to interface with the mobile application and exchange data using JSONL format, subsequently all data is stored in a PostgreSQL database. The backend server is designed to run on Linux-based operating systems, and can be freely downloaded from our GitHub repository [347, 122]. Computers can be costly to acquire and to run, especially for long periods of time, when energy consumption and maintenance can weight on the cost of the experiment. As per our tests, our software is capable of running on system-on-a-chip machines with modest performances, such as the Raspberry Pi™[1] model $3B+$. This device mounts an ARM® quad-core processing unit running at 1.4GHz, with 1Gb of RAM. The memory

[1] Raspberry Pi is a trademark of Raspberry Pi Ltd



(a) Landing page (b) Address setting

Fig. 5.1 User interface developed for the Android™ operative system



(a) Landing page (b) Address setting

Fig. 5.2 User interface developed for the iOS operative system

needed to store the data connected to an experiment is widely dependent on the amount of data collected and exchanged, but the JSONL format for data transfer and the lean architecture of our system suggests that hundreds of megabytes of memory should suffice for tens of devices running an experiment for multiple weeks. An available internet connection is needed in order to have the server receive data from the devices, HTTP Secure connection is required to ensure full protection of the data exchanged and to be compliant with the latest iOS requirements.

Furthermore, the server is designed to run multiple experiments at the same time, without the devices involved interfering with one another. Each device enrolls in an experiment by setting in the application the URL for the server compounded with the unique identifier of the experiment defined by the researchers, such as:

$$https://serveraddress/humantohuman/experiment/identifier$$

The first time the device enrolls, it sends to the server the generated ID, and the server records to which experiment it must be assigned. Cross-control is performed each time data is received from a device: if the advertising and scanning devices are not involved in the same experiment, or not enrolled at all, the data is discarded.

As we mentioned in section 5.2.2, the server can receive from each device the command to remove a device from the list of enrolled devices in an experiment, or delete all data sent and concerning a device. This behavior ensures that users have full control of their data, and can remove it at any time without the need for a specific request to an operator.

As we mentioned before, the server is responsible for enrolling a new device in the experiment, receiving data from devices, and collecting such data. The server also provides researchers with a user interface to create new experiments, delete them, or download the collected data. For all these operations, researchers provide a user interface, as shown in figure 5.3.

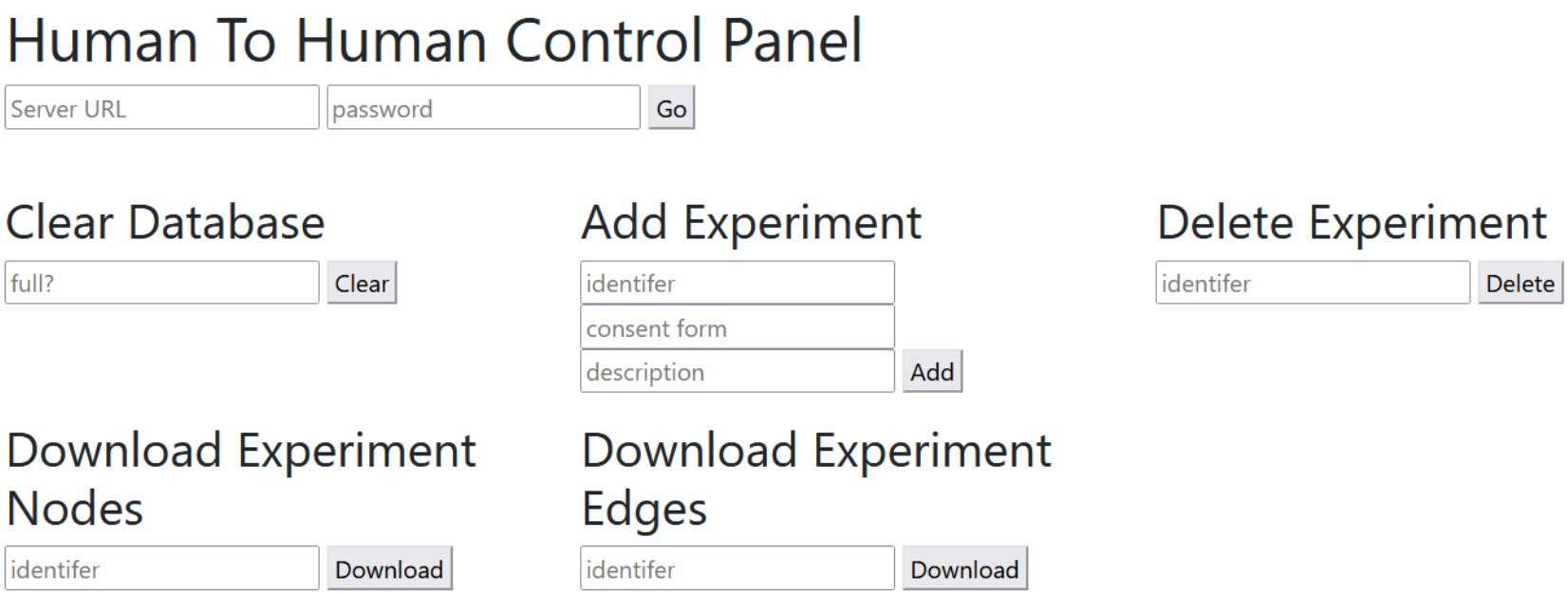


Fig. 5.3 User interface developed for the server backend

The interface can be accessed via a webpage and is protected by a password that needs to be set when preparing the server. From this webpage a researcher can:

- Clear the whole database of all experiments, by typing 'yes' in the 'Clear Database' section.
- Create a new experiment for users to join, by defining a unique identifier and uploading the proper description and consent form in the 'Add Experiment' section.
- Download a list of all the IDs for the devices enrolled in a specific experiment in the section 'Download Experiment Nodes'. This action will provide a comma-separated values file.
- Delete an experiment, by filling in the unique identifier in the section 'Delete Experiment'
- Download all the data collected from an experiment, in the section 'Download Experiment Edges'. This command will produce a comma-separated values file heaving for each event the data of time, scanned ID, advertiser ID, RSSI, and Power Level.

The user interface has been developed to be easy to use and intuitive, not needing specialized knowledge to create, delete or download an experiment from the server. Furthermore, we decided to provide comma-separated values files in the downloads as this file format is particularly common, and can be processed with a variety of software while maintaining a modest size in the memory.

5.3 Preliminary tests

The first tests performed were those connected with the strict functionality of the two applications. After passing the requirements of Google LLC and Apple Inc. to be released in the respective online stores and be downloadable, we checked that the two systems were in fact interoperable and that the communication with the server was as expected. Once the system was fully set and functional, we wanted to characterize the quality of data that could be extracted and used for our goal, inferring the relative distance of participants. As mentioned before, the chipset used for Bluetooth® signal transmission varies from manufacturer to manufacturer, so our experiment was focused on establishing a baseline for the sampling rate and the accuracy in distinguishing the distance between two devices. All tests were performed within 10 meters, as

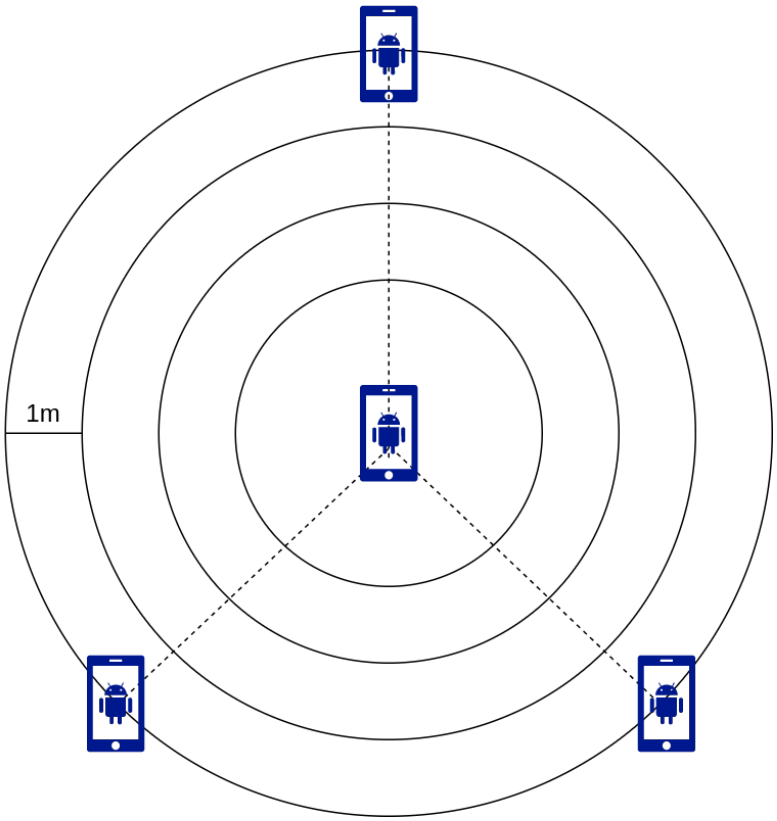


Fig. 5.4 Diagram of the experimental setting

such distance is the upper limit for the Bluetooth® communication protocol, and no reasonable human interaction is expected to happen farther between participants.

5.3.1 Experimental setup

The experiment took place in an open field, outside a building, to safeguard the health of participants and respect the social-distancing rules dictated by the alert for COVID-19. The experimental setup was done using four devices of different manufacturers all running Android™ operative system. Although the interoperability with iOS products was thoroughly tested, we could not have at hand such a device for the whole duration of the experimental phase.

As shown in figure 5.4, one device was sitting on the ground at the center of concentric circles, each circle being $1mt$ in radius bigger than the other, up to $3mt$ from the center. Three participants roughly positioned at 120° from each other, had their device on a necklace, standing in a circle and facing the center. This configuration allows for a direct line of sight of every device, reducing the effect of absorption of radio frequencies from human bodies. Once the

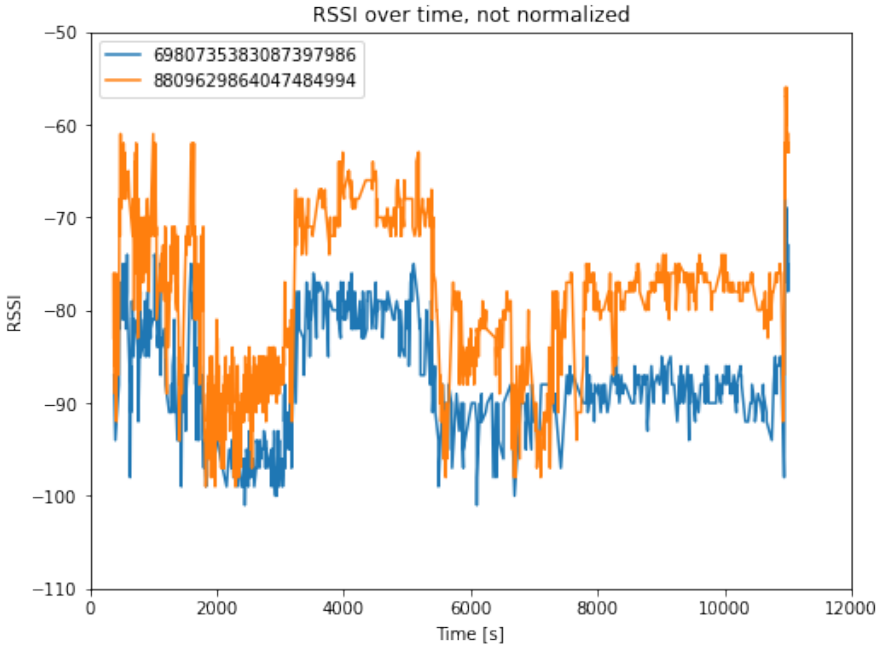


Fig. 5.5 Raw signal strength versus distance as seen by one device held by a participant and the central device and vice-versa, for the duration of the experiment.

experiment is set, the participants start by standing for one minute at a distance of 1*mt* from the central device, then each minute they move outward by 1*mt* on pre-marked spots.

Our objective is to be able to assess if two devices are within 1*mt* of distance or further. Being able to reliably assess if two people are close to each other is essential to characterize face-to-face interactions, and to unveil a number of other dynamics such as the diffusion of pathogens [348, 349].

5.3.2 Spatial and temporal resolution

The collected data are RSSI values for each participant, plus the central device. As noted before all manufacturers respect the same Bluetooth® specifications, but by implementing different technical solutions. So the first step of our analysis was to match the reading of each pair of devices to establish a common scaling factor, in order to be able to compare one device to another. An example of non-scaled measurements obtained from two devices, one looking at the other, is shown in figure 5.5.

Our approach to the rescaling process was centered on the central device by the design of the experiment, nevertheless, the choice of the device that will be used to rescale all other

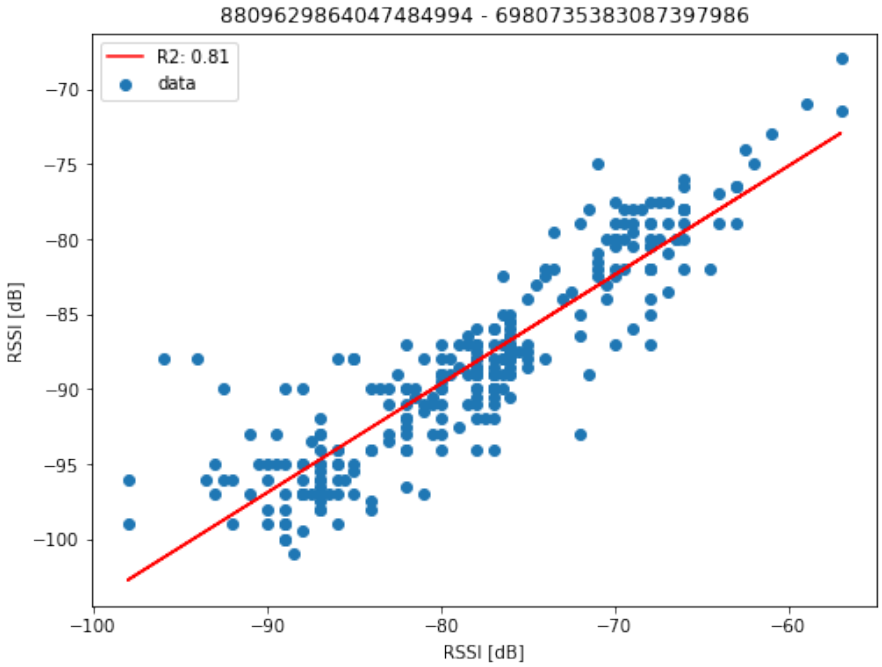


Fig. 5.6 Signal strength versus distance for one device held by a participant and the central device.

signals can be chosen using other factors or procedures. First, we compared the signal reading over the course of the experiment for each participant's device of the central device, with that obtained by the central device of the specific participant's device. Then we performed a linear regression among the two, of which an example is shown in figure 5.6. We obtained a total of six linear regressions, of which the average $R^2 = 0.65$, with the lowest value being $R^2 = 0.54$. Given the proof-of-concept nature of this experiment, we accepted these regressions as always explaining more than 50% of the variance in the data. Each regression gives us the parameters needed to rescale the data collected by each device to the power levels recorded by the chosen central device.

In figure 5.7 we show the result of the normalization process applied to the raw data of figure 5.5. Having applied this technique, we have all the data from all the devices scaled and normalized in a way to be comparable one to another. Such a process can be done iteratively also for devices not in direct line of sight, given that those share at least a common contact. In other terms, if the aggregated network of contacts over time has a unique connected component, this normalizing process can be iteratively applied to have all the signals recorded, scaled, and normalized in the same way.

Once we completed the normalization process, we explored the data first by aggregating each record of each device in time-windows of 1 second, for which the average value of signal

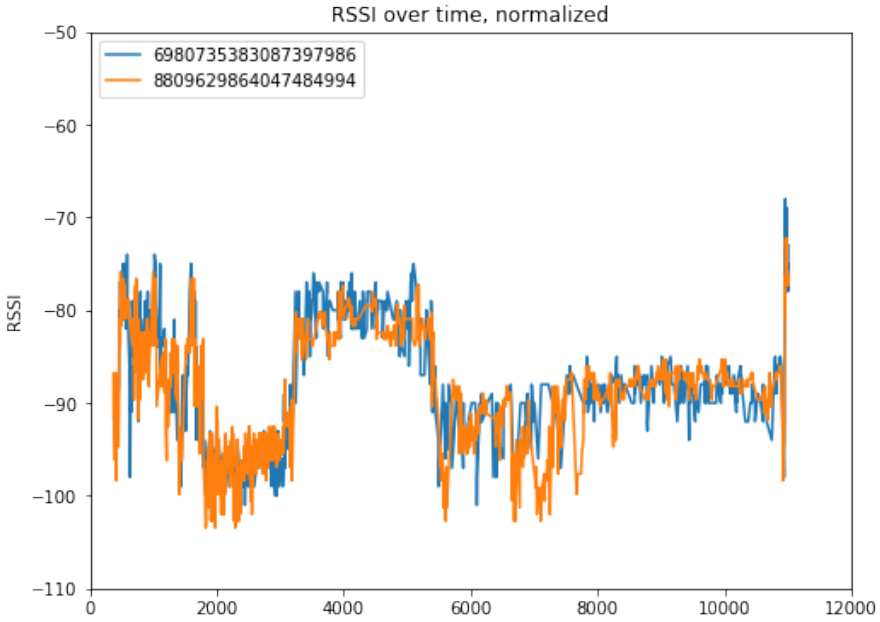


Fig. 5.7 Normalized signal strength versus distance as seen by one device held by a participant and the central device and vice-versa, for the duration of the experiment.

strength recorded over that time-window is associated. By doing so we explored the sampling rate for each device, and noted a slight variation in sampling rates, with values from $3.2Hz$ to $9.3Hz$. These sampling rates are much higher than other solutions available to researchers [38], and allow for a more precise data collection.

Finally, we recall that the participants stood at the same distance from the central device for 1 minute before moving outwards. We show in figure 5.7 the data aggregated over 1 minute for the signal recorded by a participant's device as an example of such collected and normalized data. At first glance, we see that signal strength is decreasing with the distance.

Our goal was not to model the RSSI of a device as a function of distance, but instead to evaluate the feasibility of distinguishing if two devices are close to each other or not. To do so, we proceeded by establishing if the collected signals are distinguishable at different distances. We aggregate all the data concerning the participant's devices seeing the central device for each distance, from $1mt$ to $3mt$. This is done because we are interested in assessing if the signal at distances of interest can be discerned, independently from the device used.

We perform a Welch t-test between all the data recorded at $1mt$ and all the data recorded at $2mt$, under the null hypothesis of the data coming from the same population. We reject the null hypothesis with a pvalue < 0.05. Then we perform a Welch t-test between all data recorded at $1mt$ and the aggregated data recorded at $2mt$ and $3mt$, with the null hypothesis of all the data

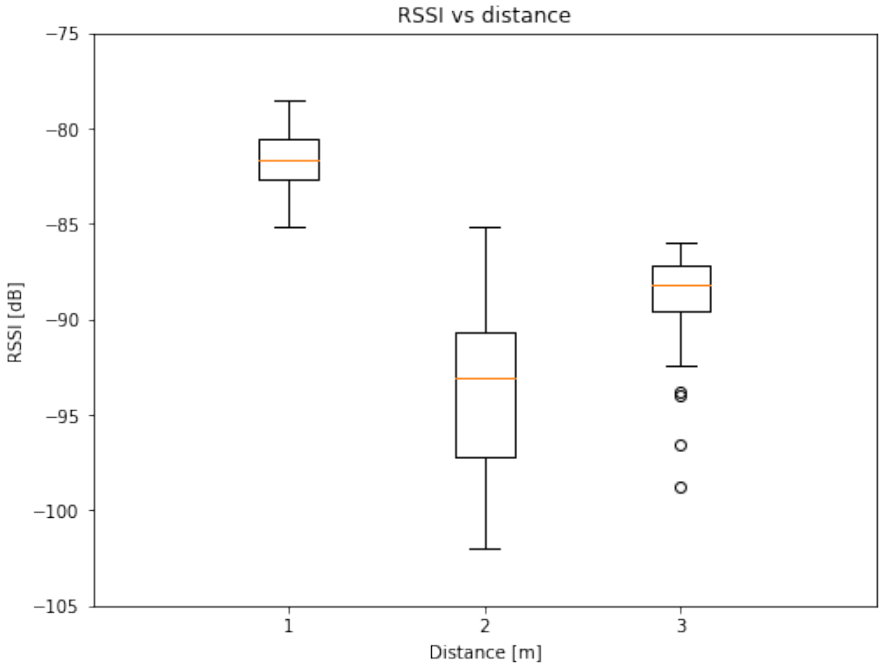


Fig. 5.8 Signal strength versus distance for one device held by a participant and the central device.

coming from the same population. We reject the null hypothesis with a pvalue < 0.05. These results confirm that the signals of devices at different distances can be indeed recognized with adequate data analysis.

5.4 Conclusions

In this chapter, we presented two smartphone applications and a server backend software aimed at providing a set of simple and cost-effective tools to study face-to-face interactions. Driven by the need to better analyze face-to-face interactions, we recognized a lack of tools for this kind of experiment, and set ourselves to develop a novel framework that could be openly used by the whole academic community. We decided to build upon the large diffusion of smartphones to allow for powerful yet economical hardware, for which we developed adequate software. The applications developed for Android™ and iOS operative systems are easy to set up and friendly to use. Alongside these applications, we also released software designed to run on simple Linux-based machines that acts at the same time as a data collection server for the data received by smartphones and as a control panel to create, run and remove experiments, download, and manage data.

We performed standardized quality tests on the applications to be able to release them on the aforementioned smartphone's operating systems. Thorough tests are done to ensure interoperability and the correct functions of the system as a whole, meaning the correct collection of signals between different devices and the correct reception of collected data by the server. Furthermore, we performed a simple yet informative experiment in which we tested the usefulness of the collected data to assess if two participants are close to one another or further apart. The results we obtained are very promising and show the usefulness of our approach.

We believe that more experiments have to be conducted to fully assess the potential and limits of our system. More refined techniques, maybe incorporating machine learning and automated classification can be developed to better analyze the obtained data and provide more insight to researchers. Some of the techniques exposed in previous chapters could be leveraged with the data obtained from this system to assess the integrity of collected data, as shown in chapter 3, and fully reconstruct the network of interaction among participants, as seen in chapter 2.

Finally, by making this set of tools fully and freely available online, on the Android™online store, iOS online store, and on GitHub [347, 350], we think that new and exciting work will spark on face-to-face interactions, both in our research community and in research fields that previously considered this kind of experimental campaign too costly or complex to set up.

Part IV

Conclusions

Chapter 6

Conclusions and future directions

In this Ph.D. dissertation, we have explored different aspects of the problem of network reconstruction.

First, we approached and explored the theoretical challenges of reconstructing a network. Starting with the binary data collected by observing the epidemiological states of agents in an activity driven network, we developed techniques to reconstruct the underlying backbone of the network.

Working on the model of routed activity driven networks, we developed a technique to discriminate between the interactions originated by the backbone network of the system and random connections. The technique is based on the comparison of the probability distribution of being infected, given the knowledge of the past states for each node. Once the technique is developed, it is implemented and tested first on Monte-Carlo simulations to evaluate its accuracy and precision, then on real data of face-to-face interactions to assess the viability of using this technique to implement immunization strategies in real scenarios.

Another side of network reconstruction is that of accurately knowing the size of the network under study. Knowing the boolean state of a node in a modified version of the voter model, we developed an exact solution of the probability distribution for the state of a node. We then extended those computations for an arbitrary size of the network the node is into, allowing for statistical testing to take place, thereby assessing the presence or absence of one or more hidden nodes in the system. A simple method based on optimization is proposed to analyze bigger networks, and extensive numerical simulations are run to evaluate the accuracy of the technique.

Although very effective and precise, we found the developed techniques to be bound by the availability of sufficient data. Identifying elements in presence of noise or other signals is a difficult task, and although innovative, our solutions were not perfect. The most pressing issue is probably having to rely on too many datapoints to reach a high level of accuracy in the reconstruction. In our numerical validation, the length of the associated time-series in the Monte-Carlo simulations had to be significant with the increasing size of the network. Furthermore, we relied on boolean dynamics to explore the diffusive processes on our networks. This is not necessarily a limitation, but more rich dynamics could have an unexpected effect on the information available for the reconstruction.

In the second section of this dissertation, we approached challenges connected with the use of data for network reconstruction.

Our attention was directed at developing proper analysis for data extracted by the social network Twitter. During the COVID-19 pandemic that started in 2019, the public opinion on the different policies enacted to stop the spread were applied in different manners across the states and territories of the United States. Subsequently, we analyzed over 55 million tweets at the beginning of 2020 to explore how the public discourse was unraveling around the topic of lockdown policies. After extracting sentiments from each post in our dataset, we were able to reconstruct the network of sentiment influence among U.S. states by means of applying transfer entropy in an innovative way. Our findings could also be linked to socio-economic factors, which we explored.

Our data-based approach to network reconstruction was undoubtedly insightful, yet like other data-based approaches, bounded by the source and domain of the data. First of all, we recognize that other dimensionality-reduction techniques are available, and a more widespread analysis must be conducted to assess the best tools for network analysis. Working with social media data can be difficult, as data extraction is domain- and platform-specific. One of the challenges we faced was choosing and applying the appropriate filters, such as keywords, location, and text interpretation. Moreover, our sentiment analysis was effective yet limited in the range of classified emotions. This latter technique can be extended and improved, but at the cost of not being able to apply some of the techniques that rely on entropy analysis which we implemented, and that led to uncovering of correlations and causality effects among the online debates in different U.S. states.

In the third section of this dissertation, we explored how to improve data collection of real-life interactions.

After recognizing the lack of suitable solutions in order to run experiments aimed at face-to-face interactions, we set to develop one ourselves. First, we defined the requirements and decided to leverage omnipresent smartphones as the main hardware solution. Those devices are capable, thanks to Bluetooth® technology, of being used to infer rough distances among them. We developed two smartphone applications, for the two main operating systems on the market, and thoroughly implemented and tested their interoperability. Alongside a server backend software was developed to allow for smooth data collection from researchers. Once the solution was ready, preliminary tests were run to ensure that distances among participants can be reconstructed, and interactions between people studied. All the software has been made open-source, for the whole academic community to use.

The developed apps are a leap forward in experimental tools for human interactions, their ease of use however comes at the cost of precision and control. From our preliminary tests, we understood that fine positioning with Bluetooth® is yet to be easy to implement. Coarse distances, with meter or sub-meter accuracy are possible, but to increase the precision more control over the hardware is required. Furthermore, extended experimental campaigns are needed to asses how environmental factors impact data collection, such as crowds, buildings, and signal interference.

Many sides of network reconstruction have been tackled and explored in this dissertation, and a connecting thread is laid out to design an integrated procedure to perform experiments, collect and process data, and from the processed data infer and reconstruct the underlying network of people's interactions.

Comprehensive studies should be conducted to explore the full potential of integrating these techniques and the possible benefits that could emerge. Increasingly, there is a need to organically integrate the different phases of research, especially in network theory. In this field, huge theoretical advances have been made, alongside complex projects of data-analysis and various efforts in experimental campaigns. The interconnections among those areas are not always well developed, and at times theoretical models and results are not fully validated and compared with available data analyses or experimental results. Cross-contamination of techniques and approaches will improve each area while making the whole field more coherent and robust.

Activity driven networks have proved to be a reliable model to describe human interactions, and more work is needed to properly tune this model on real data, and fully translate our theoretical findings into workable solutions and policy propositions. In the aim of network reconstruction, higher precision in the number of links or nodes must be sought when smaller

databases are available. Different epidemiological and diffusion processes should be analyzed for the collection of information when reconstructing networks, to expand the use cases that can benefit from network reconstruction.

Information-theoretic approaches, joined with data-analysis techniques have proven essential to understand complex opinion dynamic processes. More refined approaches in natural language processing could extract even more information from online data, enriching the dataset at hand for network analysis. Applying machine-learning techniques seems the natural next step in this area of research, due to the sheer amount of available data and the complex nonlinear processes that we expect in online interactions. Deepening the knowledge and application of entropy-based analysis could prove interesting in unveiling not just correlations, but causality effects among the elements of the system under study, allowing researchers to develop better models to explain human behavior.

Extensive campaigns exploring face-to-face interactions are now more than ever needed, as deepening our knowledge of how people behave is pivotal to facing societal challenges, from epidemics to fake news diffusion. Leveraging the open-source software developed in this dissertation, new experiments should be conducted by increasing the number of participants to uncover more complex behaviors in people's interactions. Achievable precision in distance measurement has to be explored, and innovative analysis techniques should be implemented to increase it. New techniques, leveraging machine-learning and signal processing, should be developed and implemented in the system, alongside visualization tools, to allow for a wider pool of researchers to run face-to-face experiments on human interactions.

References

[1] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8:128–140, 1741.

[2] Mark S Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[3] John Graham White, Eileen Southgate, J.N. Thomson Brenner, and Sydney Brenner. The structure of the nervous system of the nematode Caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 314(1165):1–340, nov 1986.

[4] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, apr 1998.

[5] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, jun 1998.

[6] Albert László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, oct 1999.

[7] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, feb 2010.

[8] Tiago P. Peixoto. Network Reconstruction and Community Detection from Dynamics. *Physical Review Letters*, 123(12):128301, sep 2019.

[9] Jianxi Gao, Baruch Barzel, and Albert László Barabási. Universal resilience patterns in complex networks. *Nature*, 530(7590):307–312, 2016.

[10] Xin Yuan, Yang Dai, H. Eugene Stanley, and Shlomo Havlin. K -core percolation on complex networks: Comparing random, localized, and targeted attacks. *Physical Review E*, 93(6):1–10, 2016.

[11] Yanjun Lei, Xin Jiang, Quantong Guo, Yifang Ma, Meng Li, and Zhiming Zheng. Contagion processes on the static and activity-driven coupling networks. *Physical Review E*, 93(3):032308, mar 2016.

[12] Suyu Liu, Nicola Perra, Márton Karsai, and Alessandro Vespignani. Controlling contagion processes in activity driven networks. *Physical Review Letters*, 112(11):1–5, 2014.

[13] Inder M. Verma. Editorial expression of concern: Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(29):10779, 2014.

[14] Xi Xiong, Yuanyuan Li, Shaojie Qiao, Nan Han, Yue Wu, Jing Peng, and Binyong Li. An emotional contagion model for heterogeneous social media with multiple behaviors. *Physica A: Statistical Mechanics and its Applications*, 490:185–202, 2018.

[15] A. Li, S. P. Cornelius, Y. Y. Liu, L. Wang, and A. L. Barabási. The fundamental advantages of temporal networks. *Science*, 358(6366):1042–1046, 2017.

[16] Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, oct 2012.

[17] Giulia Cencetti, Federico Battiston, Bruno Lepri, and Márton Karsai. Temporal properties of higher-order interactions in social networks. *Scientific Reports*, 11(1):1–10, 2021.

[18] Ingo Scholtes, Nicolas Wider, and Antonios Garas. Higher-order aggregate networks in the analysis of temporal networks: path structures and centralities. *European Physical Journal B*, 89(3):1–15, 2016.

[19] Bastian Prasse and Piet Van Mieghem. Exact Network Reconstruction from Complete SIS Nodal State Infection Information Seems Infeasible. *IEEE Transactions on Network Science and Engineering*, 6(4):748–759, 2019.

[20] Ye Yuan, Guy-Bart Bart Stan, Sean Warnick, and Jorge Goncalves. Robust dynamical network structure reconstruction. *Automatica*, 47(6):1230–1235, jun 2011.

[21] Christof Van Mol and Joris Michielsen. The Reconstruction of a Social Network Abroad. An Analysis of the Interaction Patterns of Erasmus Students. *Mobilities*, 10(3):423–444, may 2015.

[22] Duygu Balcan, Hao Hu, Bruno Goncalves, Paolo Bajardi, Chiara Poletto, Jose J Ramasco, Daniela Paolotti, Nicola Perra, Michele Tizzoni, Wouter Van den Broeck, Vittoria Colizza, and Alessandro Vespignani. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC Medicine*, 7(1):45, dec 2009.

[23] Alessandro Rizzo, Biagio Pedalino, and Maurizio Porfiri. A network model for Ebola spreading. *Journal of Theoretical Biology*, 394:212–222, apr 2016.

[24] Adam J Kucharski, Timothy W Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M Eggo, Fiona Sun, Mark Jit, James D Munday, Nicholas Davies, Amy Gimma, Kevin van Zandvoort, Hamish Gibbs, Joel Hellewell, Christopher I Jarvis, Sam Clifford, Billy J Quilty, Nikos I Bosse, Sam Abbott, Petra Klepac, and Stefan Flasche. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(5):553–558, may 2020.

[25] Moritz U. G. Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M. Pigott, Louis du Plessis, Nuno R. Faria, Ruoran Li, William P. Hanage, John S. Brownstein, Maylis Layan, Alessandro Vespignani, Huaiyu Tian, Christopher Dye, Oliver G. Pybus, and Samuel V. Scarpino. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, 368(6490):493–497, may 2020.

[26] Francesco Parino, Lorenzo Zino, Giuseppe C. Calafiore, and Alessandro Rizzo. A model predictive control approach to optimally devise a two-dose vaccination rollout: A case study on COVID-19 in Italy. *International Journal of Robust and Nonlinear Control*, page rnc.5728, aug 2021.

[27] Francesco Vincenzo Surano, Christian Bongiorno, Lorenzo Zino, Maurizio Porfiri, and Alessandro Rizzo. Backbone reconstruction in temporal networks from epidemic data. *Physical Review E*, 100(4):1–11, 2019.

[28] Matt J. Keeling and Pejman Rohani. Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecology Letters*, 5(1):20–29, jan 2002.

[29] Bastian Prasse and Piet Van Mieghem. Maximum-Likelihood Network Reconstruction for SIS Processes is NP-Hard. *arXiv preprint*, jul 2018.

[30] Alfredo Braunstein, Alessandro Ingrosso, and Anna Paola Muntoni. Network reconstruction from infection cascades. *Journal of The Royal Society Interface*, 16(151):20180844, feb 2019.

[31] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528, New York, NY, USA, apr 2012. ACM.

[32] Hongwu Ma, Anatoly Sorokin, Alexander Mazein, Alex Selkov, Evgeni Selkov, Oleg Demin, and Igor Goryanin. The Edinburgh human metabolic network reconstruction and its functional analysis. *Molecular Systems Biology*, 3(1):135, jan 2007.

[33] Markus J Herrgård, Neil Swainston, Paul Dobson, Warwick B Dunn, K Yalçin Arga, Mikko Arvas, Nils Blüthgen, Simon Borger, Roeland Costenoble, Matthias Heinemann, Michael Hucka, Nicolas Le Novère, Peter Li, Wolfram Liebermeister, Monica L Mo, Ana Paula Oliveira, Dina Petranovic, Stephen Pettifer, Evangelos Simeonidis, Kieran Smallbone, Irena Spasié, Dieter Weichart, Roger Brent, David S Broomhead, Hans V Westerhoff, Betül Kürdar, Merja Penttilä, Edda Klipp, Bernhard Ø Palsson, Uwe Sauer, Stephen G Oliver, Pedro Mendes, Jens Nielsen, and Douglas B Kell. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature Biotechnology*, 26(10):1155–1160, oct 2008.

[34] Michael Breakspear. Dynamic models of large-scale brain activity. *Nature Neuroscience*, 20(3):340–352, mar 2017.

[35] Ri-Qi Su, Wen-Xu Wang, and Ying-Cheng Lai. Detecting hidden nodes in complex networks from time series. *Physical Review E*, 85(6):065201, jun 2012.

[36] Ri-Qi Su, Ying-Cheng Lai, Xiao Wang, and Younghae Do. Uncovering hidden nodes in complex networks in the presence of noise. *Scientific Reports*, 4(1):3944, may 2015.

[37] Suma V. Community Based Network Reconstruction for an Evolutionary Algorithm Framework. *Journal of Artificial Intelligence and Capsule Networks*, 3(1):53–61, mar 2021.

[38] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, and Philippe Vanhems. High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School. *PLoS ONE*, 6(8):e23176, aug 2011.

[39] Ciro Cattuto, Wouter van den Broeck, Alain Barrat, Vittoria Colizza, Jean François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE*, 5(7):1–9, 2010.

[40] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security*, 10(4):1–26, jan 2008.

[41] Lijuan Chen and Jitao Sun. Optimal vaccination and treatment of an epidemic network model. *Physics Letters A*, 378(41):3028–3036, aug 2014.

[42] Nguyen Manh Hung, Quoc van Tran, Ji Liu, and Hyo-Sung Ahn. Resource Allocation for Epidemic Network under Complications. In *2019 19th International Conference on Control, Automation and Systems (ICCAS)*, pages 1512–1516. IEEE, oct 2019.

[43] Giulia Berlusconi, Francesco Calderoni, Nicola Parolini, Marco Verani, and Carlo Piccardi. Link Prediction in Criminal Networks: A Tool for Criminal Intelligence Analysis. *PLOS ONE*, 11(4):e0154244, apr 2016.

[44] Francesco Calderoni, Domenico Brunetto, and Carlo Piccardi. Communities in criminal networks: A case study. *Social Networks*, 48:116–125, jan 2017.

[45] Francis Crick and Edward Jones. Backwardness of human neuroanatomy. *Nature*, 361(6408):109–110, jan 1993.

[46] Christopher S. von Bartheld, Jami Bahney, and Suzana Herculano-Houzel. The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. *Journal of Comparative Neurology*, 524(18):3865–3895, dec 2016.

[47] Patric Hagmann, Leila Cammoun, Xavier Gigandet, Stephan Gerhard, P. Ellen Grant, Van Wedeen, Reto Meuli, Jean-Philippe Thiran, Christopher J. Honey, and Olaf Sporns. MR connectomics: Principles and challenges. *Journal of Neuroscience Methods*, 194(1):34–45, dec 2010.

[48] D.C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T.E.J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S.W. Curtiss, S. Della Penna, D. Feinberg, M.F. Glasser, N. Harel, A.C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S.E. Petersen, F. Prior, B.L. Schlaggar, S.M. Smith, A.Z. Snyder, J. Xu, and E. Yacoub. The Human Connectome Project: A data acquisition perspective. *NeuroImage*, 62(4):2222–2231, oct 2012.

[49] Alex Fornito, Andrew Zalesky, and Michael Breakspear. The connectomics of brain disorders. *Nature Reviews Neuroscience*, 16(3):159–172, mar 2015.

[50] Seyed Hani Hojjati, Ata Ebrahimzadeh, Ali Khazaee, and Abbas Babajani-Feremi. Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM. *Journal of Neuroscience Methods*, 282:69–80, apr 2017.

[51] Van J. Wedeen, Patric Hagmann, Wen Yih Isaac Tseng, Timothy G. Reese, and Robert M. Weisskoff. Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging. *Magnetic Resonance in Medicine*, 54(6):1377–1386, 2005.

[52] Olaf Sporns, Giulio Tononi, and Rolf Kötter. The human connectome: A structural description of the human brain, 2005.

[53] Karl J. Friston. Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, 2(1-2):56–78, 1994.

[54] Karl J. Friston. Functional and Effective Connectivity: A Review. *Brain Connectivity*, 1(1):13–36, jan 2011.

[55] Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, sep 2010.

[56] Olaf Sporns. Network attributes for segregation and integration in the human brain. *Current Opinion in Neurobiology*, 23(2):162–171, apr 2013.

[57] M. Madan Babu. Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Research*, 31(4):1234–1244, feb 2003.

[58] Lian En Chai, Swee Kuan Loh, Swee Thing Low, Mohd Saberi Mohamad, Safaai Deris, and Zalmiyah Zakaria. A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine*, 48:55–65, may 2014.

[59] M. G. Kann. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Briefings in Bioinformatics*, 11(1):96–110, jan 2010.

[60] T GARDNER and J FAITH. Reverse-engineering transcription control networks. *Physics of Life Reviews*, 2(1):65–88, mar 2005.

[61] C. Alfarano. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research*, 33(Database issue):D418–D424, dec 2004.

[62] G. Joshi-Tope. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Database issue):D428–D432, dec 2004.

[63] Lude Franke, Harm van Bakel, Like Fokkens, Edwin D. de Jong, Michael Egmont-Petersen, and Cisca Wijmenga. Reconstruction of a Functional Human Gene Network, with an Application for Prioritizing Positional Candidate Genes. *The American Journal of Human Genetics*, 78(6):1011–1025, jun 2006.

[64] Albert-László László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, feb 2004.

[65] EP van Someren, LFA Wessels, E Backer, and MJT Reinders. Genetic network modeling. *Pharmacogenomics*, 3(4):507–525, jul 2002.

[66] Peng Wang, Bao Wen Xu, Yu Rong Wu, and Xiao Yu Zhou. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, 58(1):1–38, 2015.

[67] Pin Luarn, Jen-Chieh Yang, and Yu-Ping Chiu. The network effect on information dissemination on social network sites. *Computers in Human Behavior*, 37:1–8, aug 2014.

[68] Zsolt Katona, Peter Pal Zubcsek, and Miklos Sarvary. Network Effects and Personal Influences: The Diffusion of an Online Social Network. *Journal of Marketing Research*, 48(3):425–443, jun 2011.

[69] June Ahn. The effect of social network sites on adolescents' social and academic development: Current theories and controversies. *Journal of the American Society for Information Science and Technology*, 62(8):1435–1445, aug 2011.

[70] F. A. Binti Hamzah, C. H. Lau, H. Nazri, D. C. Ligot, G. Lee, C. L. Tan, and Et al. CoronaTracker: World-wide Covid-19 outbreak data analysis and prediction. *Bulletin of the World Health Organization*, (March):Submitted, 2020.

[71] Sepandar D. Kamvar and Jonathan Harris. We feel fine and searching the emotional web. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, number July, page 117, New York, New York, USA, 2011. ACM Press.

[72] Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web*, 6(2):1–33, may 2012.

[73] Sen Wu, Jimeng Sun, and Jie Tang. Patent partner recommendation in enterprise social networks. In *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, page 43, New York, New York, USA, 2013. ACM Press.

[74] Milen Pavlov and Ryutaro Ichise. Finding experts by link prediction in co-authorship networks. *CEUR Workshop Proceedings*, 290:42–55, 2007.

[75] Junichiro Mori, Yuya Kajikawa, and Hisashi Kashima. Finding business partners and building reciprocal relationships - A machine learning approach. In *First International Technology Management Conference*, pages 1069–1073. IEEE, jun 2011.

[76] Myunghwan Kim and Jure Leskovec. The network completion problem: Inferring missing nodes and edges in networks. *Proceedings of the 11th SIAM International Conference on Data Mining, SDM 2011*, pages 47–58, 2011.

[77] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, 2002.

[78] Krzysztof Juszczyszyn, Katarzyna Musial, and Marcin Budka. Link Prediction Based on Subgraph Evolution in Dynamic Social Networks. In *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, pages 27–34. IEEE, oct 2011.

[79] Bjoern Bringmann, Michele Berlingerio, Francesco Bonchi, and Arisitdes Gionis. Learning and Predicting the Evolution of Social Networks. *IEEE Intelligent Systems*, 25(4):26–35, jul 2010.

[80] Haifeng Liu, Zheng Hu, Hamed Haddadi, and Hui Tian. Hidden link prediction based on node centrality and weak ties. *EPL (Europhysics Letters)*, 101(1):18004, jan 2013.

[81] Iftikhar Ahmad, Muhammad Usman Akhtar, Salma Noor, and Ambreen Shahnaz. Missing Link Prediction using Common Neighbor and Centrality based Parameterized Algorithm. *Scientific Reports*, 10(1):364, dec 2020.

[82] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[83] Darko Hric, Tiago P. Peixoto, and Santo Fortunato. Network structure, metadata, and the prediction of missing nodes and annotations. *Physical Review X*, 6(3):031038, sep 2016.

[84] A. Barrat, C. Cattuto, A. E. Tozzi, P. Vanhems, and N. Voirin. Measuring contact patterns with wearable sensors: Methods, data characteristics and applications to data-driven simulations of infectious diseases. *Clinical Microbiology and Infection*, 20(1):10–16, jan 2014.

[85] Wouter Van den Broeck, Ciro Cattuto, Alain Barrat, Martin Szomszor, Gianluca Correndo, and Harith Alani. The Live Social Semantics application: a platform for integrating face-to-face presence with on-line social networking. In *2010 8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 226–231. IEEE, mar 2010.

[86] Alain Barrat, Ciro Cattuto, Martin Szomszor, Wouter Van den Broeck, and Harith Alani. Social Dynamics in Conferences: Analyses of Data from the Live Social Semantics Application. In *Lecture Notes in Computer Science*, pages 17–33. Springer, Berling, Heidelberg, 2010.

[87] Rossana Mastrandrea, Julie Fournet, and Alain Barrat. Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS ONE*, 10(9):1–26, 2015.

[88] Marc Timme. Revealing Network Connectivity from Response Dynamics. *Physical Review Letters*, 98(22):224101, may 2007.

[89] Marc Timme and Jose Casadiego. Revealing networks from dynamics: an introduction. *Journal of Physics A: Mathematical and Theoretical*, 47(34):343001, aug 2014.

[90] Emily S. C. Ching, Pik-Yin Lai, and C. Y. Leung. Reconstructing weighted networks from dynamics. *Physical Review E*, 91(3):030801, mar 2015.

[91] Mor Nitzan, Jose Casadiego, and Marc Timme. Revealing physical interaction networks from statistics of collective dynamics. *Science Advances*, 3(2), feb 2017.

[92] Jingwen Li, Zhesi Shen, Wen-Xu Xu Wang, Celso Grebogi, and Ying-Cheng Cheng Lai. Universal data-based method for reconstructing complex networks with binary-state dynamics. *Physical Review E*, 95(3):032303, mar 2017.

[93] Srinivas Gorur Shandilya and Marc Timme. Inferring network topology from complex dynamics. *New Journal of Physics*, 13(1):013004, jan 2011.

[94] Marco Tulio Angulo, Jaime A. Moreno, Gabor Lippner, Albert-László László Barabási, and Yang-Yu Yu Liu. Fundamental limitations of network reconstruction from temporal data. *Journal of The Royal Society Interface*, 14(127):20160966, feb 2017.

[95] Dongchuan Yu, Marco Righero, and Ljupco Kocarev. Estimating Topology of Networks. *Physical Review Letters*, 97(18):188701, nov 2006.

[96] Dongchuan Yu and Ulrich Parlitz. Inferring Network Connectivity by Delayed Feedback Control. *PLoS ONE*, 6(9):e24333, sep 2011.

[97] Wen-Xu Wang, Ying-Cheng Lai, Celso Grebogi, and Jieping Ye. Network Reconstruction Based on Evolutionary-Game Data via Compressive Sensing. *Physical Review X*, 1(2):021021, dec 2011.

[98] William Ogilvy Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, aug 1927.

[99] Chris Milling, Constantine Caramanis, Shie Mannor, and Sanjay Shakkottai. On identifying the causative network of an epidemic. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 909–914. IEEE, oct 2012.

[100] T. M. L. Wigley, K. R. Briffa, and P. D. Jones. On the Average Value of Correlated Time Series, with Applications in Dendroclimatology and Hydrometeorology. *Journal of Climate and Applied Meteorology*, 23(2):201–213, feb 1984.

[101] P Gopikrishnan, V Plerou, Y Liu, L.A.N Amaral, X Gabaix, and H.E Stanley. Scaling and correlation in financial time series. *Physica A: Statistical Mechanics and its Applications*, 287(3-4):362–373, dec 2000.

[102] Boris Podobnik and H. Eugene Stanley. Detrended Cross-Correlation Analysis: A New Method for Analyzing Two Nonstationary Time Series. *Physical Review Letters*, 100(8):084102, feb 2008.

[103] J. Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, jul 2018.

[104] Chuang Ma, Han-Shuang Chen, Ying-Cheng Lai, and Hai-Feng Zhang. Statistical inference approach to structural reconstruction of complex networks from binary time series. *Physical Review E*, 97(2):022301, feb 2018.

[105] Katerina Hlaváčková-Schindler, Milan Paluš, Martin Vejmelka, and Joydeep Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, 2007.

[106] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, jun 1983.

[107] Tiago P. Peixoto. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. *Physical Review X*, 4(1):011047, mar 2014.

[108] Tiago P. Peixoto. Reconstructing Networks with Unknown and Heterogeneous Errors. *Physical Review X*, 8(4):41011, 2018.

[109] Roger Guimerà and Marta Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, dec 2009.

[110] Jie Sun, Dane Taylor, and Erik M. Bollt. Causal Network Inference by Optimal Causation Entropy. *SIAM Journal on Applied Dynamical Systems*, 14(1):73–106, jan 2015.

[111] C. W.J. Granger. Testing for causality. A personal viewpoint. *Journal of Economic Dynamics and Control*, 2(C):329–352, 1980.

[112] Lionel Barnett, Adam B. Barrett, and Anil K. Seth. Granger causality and transfer entropy Are equivalent for gaussian variables. *Physical Review Letters*, 103(23):238701, dec 2009.

[113] Kuo-Ching Liang and Xiaodong Wang. Gene Regulatory Network Reconstruction Using Conditional Mutual Information. *EURASIP Journal on Bioinformatics and Systems Biology*, 2008:1–14, 2008.

[114] Terry Bossomaier, Lionel Barnett, Michael Harré, and Joseph T. Lizier. *An Introduction to Transfer Entropy*. Springer International Publishing, Cham, 2016.

[115] Raul Vicente, Michael Wibral, Michael Lindner, and Gordon Pipa. Transfer entropy-a model-free measure of effective connectivity for the neurosciences. *Journal of Computational Neuroscience*, 30(1):45–67, 2011.

[116] Federica Parisi, Guido Caldarelli, and Tiziano Squartini. Entropy-based approach to missing-links prediction. *Applied Network Science*, 3(1), 2018.

[117] Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000.

[118] Maurizio Porfiri and Manuel Ruiz Marín. An information-theoretic approach to study spatial dependencies in small datasets: Spatial dependencies in small datasets. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2242):1–15, 2020.

[119] Christian Bongiorno, Alessandro Rizzo, and Maurizio Porfiri. An information-theoretic approach to study activity driven networks. *Proceedings - IEEE International Symposium on Circuits and Systems*, 2018-May:1–5, 2018.

[120] Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz-Marí, Egbert H. van Nes, Jonas Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1):1–13, 2019.

[121] Jakub Kořenek and Jaroslav Hlinka. Causal network discovery by iterative conditioning: Comparison of algorithms. *Chaos*, 30(1), 2020.

[122] Francesco Vincenzo Surano, Maurizio Porfiri, and Alessandro Rizzo. Analysis of lockdown perception in the United States during the COVID-19 pandemic. *The European Physical Journal Special Topics*, 231(9):1625–1633, jul 2022.

[123] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, apr 2006.

[124] Emmanuel Candes, Justin Romberg, and Terence Tao. Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. *arXiv preprint*, sep 2004.

[125] A. C. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. Strauss. Near-optimal sparse fourier representations via sampling. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing - STOC '02*, page 152, New York, New York, USA, 2002. ACM Press.

[126] C.E. Shannon. Communication in the Presence of Noise. *Proceedings of the IRE*, 37(1):10–21, jan 1949.

[127] A.J. Jerri. The Shannon sampling theorem—Its various extensions and applications: A tutorial review. *Proceedings of the IEEE*, 65(11):1565–1596, 1977.

[128] Long Ma, Xiao Han, Zhesi Shen, Wen-Xu Wang, and Zengru Di. Efficient Reconstruction of Heterogeneous Networks from Time Series via Compressed Sensing. *PLOS ONE*, 10(11):e0142837, nov 2015.

[129] Zhesi Shen, Wen-Xu Xu Wang, Ying Fan, Zengru Di, and Ying-Cheng Cheng Lai. Reconstructing propagation networks with natural diversity and identifying hidden sources. *Nature Communications*, 5(1):4323, sep 2014.

[130] Xiao Han, Zhesi Shen, Wen Xu Wang, and Zengru Di. Robust reconstruction of complex networks from sparse data. *Physical Review Letters*, 114(2):028701, jan 2015.

[131] Marton Posfai and Albert-Laszlo Barabasi. *Network Science*. Cambridge University Press, 2016.

[132] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, jan 2007.

[133] P Erdös and A Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[134] Jesús Gómez-Gardeñes and Yamir Moreno. From scale-free to Erdos-Rényi networks. *Physical Review E*, 73(5):056124, may 2006.

[135] C. Seshadhri, Tamara G. Kolda, and Ali Pinar. Community structure and scale-free collections of Erdős-Rényi graphs. *Physical Review E*, 85(5):056109, may 2012.

[136] Eytan Katzav, Ofer Biham, and Alexander K. Hartmann. Distribution of shortest path lengths in subcritical Erdős-Rényi networks. *Physical Review E*, 98(1):012301, jul 2018.

[137] Qawi K. Telesford, Karen E. Joyce, Satoru Hayasaka, Jonathan H. Burdette, and Paul J. Laurienti. The Ubiquity of Small-World Networks. *Brain Connectivity*, 1(5):367–375, dec 2011.

[138] Derek J. de Solla Price. Networks of Scientific Papers. *Science*, 149(3683):510–515, jul 1965.

[139] Aaron Clauset, Cosma Rohilla Shalizi, and M. E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, nov 2009.

[140] Ivan Voitalov, Pim van der Hoorn, Remco van der Hofstad, and Dmitri Krioukov. Scale-free networks well done. *Physical Review Research*, 1(3):033034, oct 2019.

[141] R. Dean Malmgren, Daniel B. Stouffer, Adilson E. Motter, and Luís A.N. Amaral. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences of the United States of America*, 105(47):18153–18158, 2008.

[142] Ayan Bhattacharya, Bohan Chen, Remco van der Hofstad, and Bert Zwart. Consistency of the PLFit estimator for power-law data. *arXiv preprint*, pages 1–38, feb 2020.

[143] Luis E. C. Rocha, Fredrik Liljeros, and Petter Holme. Information dynamics shape the sexual networks of Internet-mediated prostitution. *Proceedings of the National Academy of Sciences*, 107(13):5706–5711, mar 2010.

[144] Petter Holme. Network reachability of real-world contact sequences. *Physical Review E*, 71(4):046119, apr 2005.

[145] C. S. Riolo. Methods and Measures for the Description of Epidemiologic Contact Networks. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, 78(3):446–457, sep 2001.

[146] P Holme. Network dynamics of ongoing social relationships. *Europhysics Letters (EPL)*, 64(3):427–433, nov 2003.

[147] Giovanni Petri and Alain Barrat. Simplicial Activity Driven Model. *Physical Review Letters*, 121(22):228301, nov 2018.

[148] Billings Jacob, Saggar Manish, Hlinka Jaroslav, Keilholz Shella, and Giovanni Petri. Simplicial and Topological Descriptions of Human Brain Dynamics. *bioRxiv*, 2021.

[149] Martin Rosvall and Carl T. Bergstrom. Mapping Change in Large Networks. *PLoS ONE*, 5(1):e8694, jan 2010.

[150] Giovanna Miritello, Esteban Moro, and Rubén Lara. Dynamical strength of social ties in information spreading. *Physical Review E*, 83(4):045102, apr 2011.

[151] Mladen Kolar, Le Song, Amr Ahmed, and Eric P. Xing. Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1), mar 2010.

[152] Steve Hanneke and Eric P. Xing. Discrete Temporal Models of Social Networks. In *Statistical Network Analysis: Models, Issues, and New Directions*, pages 115–125. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[153] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173–191, may 2007.

[154] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani. Activity driven modeling of time varying networks. *Scientific Reports*, 2(1):469, dec 2012.

[155] Alessandro Rizzo, Mattia Frasca, and Maurizio Porfiri. Effect of individual behavior on epidemic spreading in activity-driven networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 90(4):042801, oct 2014.

[156] Lorenzo Zino, Alessandro Rizzo, and Maurizio Porfiri. Modeling memory effects in activity-driven networks. *SIAM Journal on Applied Dynamical Systems*, 17(4):2830–2854, 2018.

[157] Alessandro Rizzo and Maurizio Porfiri. Innovation diffusion on time-varying activity driven networks. *European Physical Journal B*, 89(1):1–8, 2016.

[158] Iacopo Pozzana, Kaiyuan Sun, and Nicola Perra. Epidemic spreading on activity-driven networks with attractiveness. *Physical Review E*, 96(4), 2017.

[159] Dandan Li, Dun Han, Jing Ma, Mei Sun, Lixin Tian, Timothy Khouw, and H. Eugene Stanley. Opinion dynamics in activity-driven networks. *Epl*, 120(2), 2017.

[160] Jalil Hasanyan, Lorenzo Zino, Daniel Alberto Burbano Lombana, Alessandro Rizzo, and Maurizio Porfiri. Leader–follower consensus on activity-driven networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2233):20190485, jan 2020.

[161] Wei Wang, Quan Hui Liu, Shi Min Cai, Ming Tang, Lidia A. Braunstein, and H. Eugene Stanley. Suppressing disease spreading by using information diffusion on multiplex networks. *Scientific Reports*, 6(7600):29259, 2016.

[162] Lucas Lacasa, Inés P. Mariño, Joaquin Miguez, Vincenzo Nicosia, Édgar Roldán, Ana Lisica, Stephan W. Grill, and Jesús Gómez-Gardeñes. Multiplex Decomposition of Non-Markovian Dynamics and the Hidden Layer Reconstruction Problem. *Physical Review X*, 8(3):1–36, 2018.

[163] Iacopo Iacopini, Giovanni Petri, Alain Barrat, and Vito Latora. Simplicial models of social contagion. *Nature Communications*, 10(1):1–5, 2019.

[164] Christian Bongiorno, András London, Salvatore Miccichè, and Rosario N. Mantegna. Core of communities in bipartite networks. *Physical Review E*, 96(2):1–10, 2017.

[165] Chu-Xu Zhang, Zi-Ke Zhang, and Chuang Liu. An evolving model of online bipartite networks. *Physica A: Statistical Mechanics and its Applications*, 392(23):6100–6106, dec 2013.

[166] Yiming Li, Jingzhi Fang, Yuxiang Zeng, Balz Maag, Yongxin Tong, and Lingyu Zhang. Two-sided online bipartite matching in spatial data: experiments and analysis. *GeoInformatica*, 24(1):175–198, jan 2020.

[167] Jean-Loup Guillaume and Matthieu Latapy. Bipartite graphs as models of complex networks. *Physica A: Statistical Mechanics and its Applications*, 371(2):795–813, nov 2006.

[168] John Scott. Social Network Analysis. *Sociology*, 22(1):109–127, feb 1988.

[169] P.V. Bindu, P. Santhi Thilagam, and Deepesh Ahuja. Discovering suspicious behavior in multilayer social networks. *Computers in Human Behavior*, 73:568–582, aug 2017.

[170] M. Kivela, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, sep 2014.

[171] S. Boccaletti, G. Bianconi, R. Criado, C.I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, nov 2014.

[172] Manlio De Domenico, Albert Solé-Ribalta, Emanuele Cozzo, Mikko Kivelä, Yamir Moreno, Mason A. Porter, Sergio Gómez, and Alex Arenas. Mathematical Formulation of Multilayer Networks. *Physical Review X*, 3(4):041022, dec 2013.

[173] Marvin E. Shaw. Group Structure and the Behavior of Individuals in Small Groups. *The Journal of Psychology*, 38(1):139–149, jul 1954.

[174] Di Zhou, H. Eugene Stanley, Gregorio D'Agostino, and Antonio Scala. Assortativity decreases the robustness of interdependent networks. *Physical Review E*, 86(6):066103, dec 2012.

[175] Rogier Noldus and Piet Van Mieghem. Assortativity in complex networks. *Journal of Complex Networks*, 3(4):507–542, dec 2015.

[176] R. Xulvi-Brunet and I. M. Sokolov. Reshuffling scale-free networks: From random to assortative. *Physical Review E*, 70(6):066102, dec 2004.

[177] S.L. Hakimi. On the degrees of the vertices of a directed graph. *Journal of the Franklin Institute*, 279(4):290–308, apr 1965.

[178] Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, dec 1966.

[179] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, dec 1959.

[180] Frank Schulz, Dorothea Wagner, and Karsten Weihe. Dijkstra's algorithm on-line. *ACM Journal of Experimental Algorithmics*, 5:12, dec 2000.

[181] DongKai Fan and Ping Shi. Improvement of Dijkstra's algorithm and its application in route planning. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, pages 1901–1904. IEEE, aug 2010.

[182] Linton C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35, mar 1977.

[183] Ulrik Brandes. A faster algorithm for betweenness centrality*. *The Journal of Mathematical Sociology*, 25(2):163–177, jun 2001.

[184] Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, may 2008.

[185] Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2(1):113–120, jan 1972.

[186] Phillip Bonacich. Some unique properties of eigenvector centrality. *Social Networks*, 29(4):555–564, oct 2007.

[187] Anand Bihari and Manoj Kumar Pandia. Eigenvector centrality and its application in research professionals' relationship network. In *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, pages 510–514. IEEE, feb 2015.

[188] Kam-Fung Cheung, Michael G.H. Bell, Jing-Jing Pan, and Supun Perera. An eigenvector centrality analysis of world container shipping network connectivity. *Transportation Research Part E: Logistics and Transportation Review*, 140:101991, aug 2020.

[189] Maksim Kitsak, Lazaros K. Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hernán A. Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, nov 2010.

[190] Jian-Guo Liu, Zhuo-Ming Ren, and Qiang Guo. Ranking the spreading influence in complex networks. *Physica A: Statistical Mechanics and its Applications*, 392(18):4154–4159, sep 2013.

[191] Antonios Garas, Frank Schweitzer, and Shlomo Havlin. A k -shell decomposition method for weighted networks. *New Journal of Physics*, 14(8):083030, aug 2012.

[192] Per Hage and Frank Harary. Eccentricity and centrality in networks. *Social Networks*, 17(1):57–63, jan 1995.

[193] Frank Takes and Walter Kosters. Computing the Eccentricity Distribution of Large Graphs. *Algorithms*, 6(1):100–118, feb 2013.

[194] Julie Fournet and Alain Barrat. Contact patterns among high school students. *PLoS ONE*, 9(9):e107878, sep 2014.

[195] M. E.J. Newman. The structure and function of complex networks, 2003.

[196] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3):925–979, aug 2015.

[197] Andrea Montanaria and Amin Saberi. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 107(47):20196–20201, 2010.

[198] Ping Ping Li, Da Fang Zheng, and P. M. Hui. Dynamics of opinion formation in a small-world network. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 73(5):56128, may 2006.

[199] Petter Holme. Modern temporal network theory: a colloquium. *European Physical Journal B*, 88(9):234, sep 2015.

[200] Noah E. Friedkin. Information flow through strong and weak ties in intraorganizational social networks. *Social Networks*, 3(4):273–285, jan 1982.

[201] Valerio Gemmetto, Alessio Cardillo, and Diego Garlaschelli. Irreducible network backbones: unbiased graph filtering via maximum entropy. *arXiv preprint*, jun 2017.

[202] J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A. L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18):7332–7336, 2007.

[203] Panpan Shu, Ming Tang, Kai Gong, and Ying Liu. Effects of weak ties on epidemic predictability on community networks. *Chaos*, 22(4):43124, 2012.

[204] Márton Karsai, Nicola Perra, and Alessandro Vespignani. Time varying networks and the weakness of strong ties. *Scientific Reports*, 4, 2014.

[205] Kaiyuan Sun, Andrea Baronchelli, and Nicola Perra. Contrasting effects of strong ties on SIR and SIS processes in temporal networks. *European Physical Journal B*, 88(12):1–8, 2015.

[206] Lorenzo Zino, Alessandro Rizzo, and Maurizio Porfiri. Continuous-Time Discrete-Distribution Theory for Activity-Driven Networks. *Physical Review Letters*, 117(22):228302, nov 2016.

[207] Michele Starnini and Romualdo Pastor-Satorras. Temporal percolation in activity-driven networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 89(3):1–7, 2014.

[208] Christian Bongiorno, Lorenzo Zino, and Alessandro Rizzo. On Unveiling the Community Structure of Temporal Networks. In *2018 IEEE Conference on Decision and Control (CDC)*, volume 2018-Decem, pages 6210–6215. IEEE, dec 2018.

[209] Christian Bongiorno, Lorenzo Zino, and Alessandro Rizzo. A novel framework for community modeling and characterization in directed temporal networks. *Applied Network Science*, 4(1):1–20, 2019.

[210] Matthieu Nadini, Alessandro Rizzo, and Maurizio Porfiri. Epidemic Spreading in Temporal and Adaptive Networks with Static Backbone. *IEEE Transactions on Network Science and Engineering*, 7(1):549–561, 2020.

[211] Hyewon Kim, Meesoon Ha, and Hawoong Jeong. Scaling properties in time-varying networks with memory. *European Physical Journal B*, 88(12):1–8, dec 2015.

[212] Hyewon Kim, Meesoon Ha, and Hawoong Jeong. Dynamic topologies of activity-driven temporal networks with memory. *Physical Review E*, 97(6), 2018.

[213] Matthieu Nadini, Christian Bongiorno, Alessandro Rizzo, and Maurizio Porfiri. Detecting network backbones against time variations in node properties. *Nonlinear Dynamics*, 99(1):855–878, jan 2020.

[214] Linyuan Lü, Ci Hang Jin, and Tao Zhou. Similarity index based on local paths for link prediction of complex networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 80(4):46122, 2009.

[215] L. L. Linyuan and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, mar 2011.

[216] Hao Liao and An Zeng. Reconstructing propagation networks with temporal similarity. *Scientific Reports*, 5, 2015.

[217] Maurizio Porfiri and Manuel Ruiz Marin. Information Flow in a Model of Policy Diffusion: An Analytical Study. *IEEE Transactions on Network Science and Engineering*, 5(1):42–54, 2018.

[218] Chuang Ma, Hai Feng Zhang, and Ying Cheng Lai. Reconstructing complex networks without time series. *Physical Review E*, 96(2):22320, aug 2017.

[219] Yu Zhong Chen and Ying Cheng Lai. Sparse dynamical Boltzmann machine for reconstructing complex networks with binary dynamics. *Physical Review E*, 97(3):32317, 2018.

[220] Wen-Xu Wang, Ying-Cheng Lai, and Celso Grebogi. Data based identification and prediction of nonlinear and complex dynamical systems. *Physics Reports*, 644:1–76, jul 2016.

[221] Michele Tumminello, Salvatore Miccichè, Fabrizio Lillo, Jyrki Piilo, and Rosario N. Mantegna. Statistically Validated Networks in Bipartite Complex Systems. *PLoS ONE*, 6(3):e17994, mar 2011.

[222] J. Meier, X. Zhou, A. Hillebrand, P. Tewarie, C. J. Stam, and P. Van Mieghem. The epidemic spreading model and the direction of information flow in brain networks. *NeuroImage*, 152:639–646, 2017.

[223] William Hedley Thompson, Per Brantefors, and Peter Fransson. From static to temporal network theory: Applications to functional brain connectivity. *Network Neuroscience*, 1(2):69–99, 2017.

[224] Olaf Sporns. Graph theory methods: Applications in brain networks. *Dialogues in Clinical Neuroscience*, 20(2):111–120, 2018.

[225] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on twitter networks: Validation of Dunbar's number. *PLoS ONE*, 6(8), 2011.

[226] P. Giles and Norman T. J. Bailey. *The Mathematical Theory of Infectious Diseases and Its Applications*, volume 28. Griffin, London, UK, 2 edition, 1977.

[227] Lorenzo Zino, Alessandro Rizzo, and Maurizio Porfiri. An analytical framework for the study of epidemic models on activity driven networks. *Journal of Complex Networks*, 5(6):924–952, 2017.

[228] Valerii V Kozlov. Weighted averages, uniform distribution, and strict ergodicity. *Russian Mathematical Surveys*, 60(6):1121–1146, 2005.

[229] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, jan 1995.

[230] M. E.J. Newman. Mixing patterns in networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 67(2):13, 2003.

[231] Romualdo Pastor-Satorras, Alexei Vázquez, and Alessandro Vespignani. Dynamical and correlation properties of the internet. *Physical Review Letters*, 87(25):258701–1–258701–4, nov 2001.

[232] William Aiello, Fan Chung, and Linyuan Lu. A random graph model for massive graphs. *The Structure and Dynamics of Networks*, 9781400841(1):259–268, 2011.

[233] A. Ganesh, L. Massoulié, and D. Towsley. The effect of network topology on the spread of epidemics. In *Proceedings - IEEE INFOCOM*, volume 2, pages 1455–1466, 2005.

[234] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The Anatomy of the Facebook Social Graph. Technical Report 1999-66, Stanford InfoLab, 2011.

[235] Réka Albert and Albert László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[236] M. E.J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):404–409, 2001.

[237] M. Girvan and M. E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, 2002.

[238] Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30(4):330–342, oct 2008.

[239] Atif Nazir, Saqib Raza, and Chen Nee Chuah. Unveiling facebook: A measurement study of social network based applications. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, pages 43–56, New York, New York, USA, 2008. ACM Press.

[240] Salvatore A. Catanese, Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. Crawling Facebook for social network analysis purposes. In *ACM International Conference Proceeding Series*, 2011.

[241] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1568–1576, 2011.

[242] Aron Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*, pages 115–122, 2010.

[243] Vasileios Lampos, Tijl De Bie, and Nello Cristianini. Flu detector-tracking epidemics on twitter. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6323 LNAI, pages 599–602, 2010.

[244] Timo Smieszek, Stefanie Castell, Alain Barrat, Ciro Cattuto, Peter J. White, and Gérard Krause. Contact diaries versus wearable proximity sensors in measuring contact patterns at a conference: Method comparison and participants' attitudes. *BMC Infectious Diseases*, 16(1):1–14, 2016.

[245] Mathieu Génois and Alain Barrat. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science*, 7(1), 2018.

[246] Mathieu Génois, Christian L. Vestergaard, Julie Fournet, André Panisson, Isabelle Bonmarin, and Alain Barrat. Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science*, 3(3):326–347, 2015.

[247] Lorenzo Zino, Giacomo Como, and Fabio Fagnani. On imitation dynamics in potential population games. In *2017 IEEE 56th Annual Conference on Decision and Control, CDC 2017*, volume 2018-Janua, pages 757–762, 2018.

[248] Lorenzo Zino, Lorenzo Zino, Lorenzo Zino, Alessandro Rizzo, Alessandro Rizzo, and Maurizio Porfiri. Consensus over activity-driven networks. *IEEE Transactions on Control of Network Systems*, 7(2):866–877, 2020.

[249] Matthew J. Beal, Francesco Falciani, Zoubin Ghahramani, Claudia Rangel, and David L. Wild. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3):349–356, 2005.

[250] Juliane Schäfer and Korbinian Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

[251] Tiago P. Peixoto and Stefan Bornholdt. Evolution of robust network topologies: Emergence of central backbones. *Physical Review Letters*, 109(11):1–5, 2012.

[252] Francesc Comellas and Jordi Diaz-Lopez. Spectral reconstruction of complex networks. *Physica A: Statistical Mechanics and its Applications*, 387(25):6436–6442, 2008.

[253] Ron Eyal, Avi Rosenfeld, Sigal Sina, and Sarit Kraus. Predicting and identifying missing node information in social networks. *ACM Transactions on Knowledge Discovery from Data*, 8(3), 2014.

[254] Shashidhar Sundareisan, Jilles Vreeken, and B. Aditya Prakash. Hidden hazards: Finding missing nodes in large graph epidemics. *SIAM International Conference on Data Mining 2015, SDM 2015*, pages 415–423, 2015.

[255] Rundong Shi, Weinuo Jiang, and Shihong Wang. Detecting network structures from measurable data produced by dynamics with hidden variables. *Chaos*, 30(1), 2020.

[256] Christian L. Vestergaard, Eugenio Valdano, Mathieu Génois, Chiara Poletto, Vittoria Colizza, and Alain Barrat. Impact of spatially constrained sampling of temporal contact networks on the evaluation of the epidemic risk. *European Journal of Applied Mathematics*, 27(6):941–957, 2016.

[257] Ri Qi Su, Wen Xu Wang, Xiao Wang, and Ying Cheng Lai. Data-based reconstruction of complex geospatial networks, nodal positioning and detection of hidden nodes. *Royal Society Open Science*, 3(1), 2016.

[258] Chuang Ma, Han Shuang Chen, Xiang Li, Ying Cheng Lai, and Hai Feng Zhang. Data based reconstruction of duplex networks. *SIAM Journal on Applied Dynamical Systems*, 19(1):124–150, jan 2020.

[259] Peter D. Killworth, Christopher McCarty, H. Russell Bernard, Gene Ann Shelley, and Eugene C. Johnsen. Estimation of seroprevalence, rape, and homelessness in the united states using asocial network approach. *Evaluation Review*, 22(2):289–307, apr 1998.

[260] Tyler H. McCormick, Matthew J. Salganick, and Tian Zheng. How many people do you know?: Efficiently estimating personal network size. *Journal of the American Statistical Association*, 105(489):59–70, 2010.

[261] Balázs Ráth and Daniel Valesin. Percolation on the stationary distributions of the voter model. *Annals of Probability*, 45(3):1899–1951, feb 2017.

[262] Dictionary Merriam-Webster. Lockdown, 2021.

[263] Md Mokhlesur Rahman, G. G.Md Nawaz Ali, Xue Jun Li, Jim Samuel, Kamal Chandra Paul, Peter H.J. Chong, and Michael Yakubov. Socioeconomic factors analysis for COVID-19 US reopening sentiment with Twitter and census data. *Heliyon*, 7(2):e06200, 2021.

[264] Federico Mucci, Nicola Mucci, and Francesca Diolaiuti. Lockdown and isolation: Psychological aspects of covid-19 pandemic in the general population. *Clinical Neuropsychiatry*, 17(2):63–64, 2020.

[265] G. James Rubin and Simon Wessely. The psychological effects of quarantining a city, 2020.

[266] Emanuele Caroppo, Pietro De Lellis, Ilaria Lega, Antonella Candelori, Daniela Pedacchia, Alida Pellegrini, Rossella Sonnino, Virginia Venturiello, Manuel Ruiz Marìn, and Maurizio Porfiri. Unequal effects of the national lockdown on mental and social health in Italy. *Annali dell'Istituto Superiore di Sanita*, 56(4):497–501, 2020.

[267] Alessandro Rovetta and Akshaya Srikanth Bhagavathula. COVID-19-related web search behaviors and infodemic attitudes in Italy: Infodemiological study. *JMIR Public Health and Surveillance*, 6(2), 2020.

[268] Anneliese Depoux, Sam Martin, Emilie Karafillakis, Raman Preet, Annelies Wilder-Smith, and Heidi Larson. The pandemic of social media panic travels faster than the COVID-19 outbreak. *Journal of Travel Medicine*, 27(3):1–2, may 2020.

[269] Cristina M. Pulido, Beatriz Villarejo-Carballido, Gisela Redondo-Sama, and Aitor Gómez. COVID-19 infodemic: More retweets for science-based information on coronavirus than for false information. *International Sociology*, 35(4):377–392, jul 2020.

[270] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The COVID-19 social media infodemic. *Scientific Reports*, 10(1):1–18, 2020.

[271] Daniel Allington, Bobby Duffy, Simon Wessely, Nayana Dhavan, and James Rubin. Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. *Psychological Medicine*, 51(10):1763–1769, 2021.

[272] Pablo Barberá and Gonzalo Rivero. Understanding the Political Representativeness of Twitter Users. *Social Science Computer Review*, 33(6):712–729, 2015.

[273] Andreas Jungherr. Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology and Politics*, 13(1):72–91, 2016.

[274] Marija Anna Bekafigo and Allan McBride. Who Tweets About Politics?: Political Participation of Twitter Users During the 2011Gubernatorial Elections. *Social Science Computer Review*, 31(5):625–643, 2013.

[275] Chang Sup Park. Does Twitter motivate involvement in politics? Tweeting, opinion leadership, and political engagement. *Computers in Human Behavior*, 29(4):1641–1648, 2013.

[276] Julian Ausserhofer and Axel Maireder. NATIONAL POLITICS ON TWITTER: Structures and topics of a networked public sphere. *Information Communication and Society*, 16(3):291–314, 2013.

[277] Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. A few hubs with many connections share with many individuals with few connections. *Communications of the ACM*, 55(6):70–75, jun 2012.

[278] Sardar Hamidian and Mona T. Diab. Rumor identification and belief investigation on Twitter. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA 2016 at the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3–8, Stroudsburg, PA, USA, 2016. Association for Computational Linguistics.

[279] Linda Lombi. La ricerca sociale al tempo dei Big Data: sfide e prospettive. *Studi di Sociologia*, 2:215–227, 2015.

[280] Itai Himelboim, Stephen Mccreery, and Marc Smith. Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter. *Journal of Computer-Mediated Communication*, 18(2):40–60, 2013.

[281] Lauren Sinnenberg, Alison M. Buttenheim, Kevin Padrez, Christina Mancheno, Lyle Ungar, and Raina M. Merchant. Twitter as a tool for health research: A systematic review. *American Journal of Public Health*, 107(1):e1–e8, 2017.

[282] Michael J. Paul, Mark Dredze, and David Broniatowski. Twitter Improves Influenza Forecasting. *PLoS Currents*, pages 1–12, 2014.

[283] David A. Broniatowski, Michael J. Paul, and Mark Dredze. National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS ONE*, 8(12):1–8, 2013.

[284] Qanita Bani Baker, Farah Shatnawi, Saif Rawashdeh, Mohammad Al-Smadi, and Yaser Jararweh. Detecting epidemic diseases using sentiment analysis of arabic tweets. *Journal of Universal Computer Science*, 26(1):50–70, 2020.

[285] Liza G.G. Van Lent, Hande Sungur, Florian A. Kunneman, Bob Van De Velde, and Enny Das. Too far to care? measuring public attention and fear for ebola using twitter. *Journal of Medical Internet Research*, 19(6), 2017.

[286] Meg Carter. How Twitter may have helped Nigeria contain Ebola. *BMJ (Online)*, 349(November):1–2, 2014.

[287] Erin Hea Jin Kim, Yoo Kyung Jeong, Yuyoung Kim, Keun Young Kang, and Min Song. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. *Journal of Information Science*, 42(6):763–781, 2016.

[288] Xiaoling Yuan, Jie Xu, Sabiha Hussain, He Wang, Nan Gao, and Lanjing Zhang. Trends and Prediction in Daily New Cases and Deaths of COVID-19 in the United States: An Internet Search-Interest Based Model. *Exploratory Research and Hypothesis in Medicine*, 000(000):1–6, 2020.

[289] Emily Chen, Kristina Lerman, and Emilio Ferrara. Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. *JMIR Public Health and Surveillance*, 6(2), 2020.

[290] Bing Liu. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, may 2012.

[291] Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of Twitter sentiment analysis methods. *ACM Computing Surveys*, 49(2), 2016.

[292] Rajani Shree Manjappa and Aditya Kumar. Twitter Sentiment Analysis. *SSRN Electronic Journal*, pages 212–216, 2020.

[293] Jundong Chen, Md Shafaeat Hossain, and Huan Zhang. Analyzing the sentiment correlation between regular tweets and retweets. *Social Network Analysis and Mining*, 10(1), 2020.

[294] Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. Contextual semantics for sentiment analysis of Twitter. *Information Processing and Management*, 52(1):5–19, 2016.

[295] Laura Pollacci, Alina Sîrbu, Fosca Giannotti, Dino Pedreschi, Claudio Lucchese, and Cristina Ioana Muntean. Sentiment spreading: An epidemic model for lexicon-based sentiment analysis on twitter. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10640 LNAI:114–127, 2017.

[296] Xiao Long Deng, Ya Qi Tang, and Yi Hua Huang. Opinion mining for emergency case risk analysis in spark based distributed system. *Proceedings of the 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Emergency Management, EM-GIS 2015*, pages 3–10, 2015.

[297] Bishwo Prakash Pokharel. Twitter Sentiment Analysis During Covid-19 Outbreak in Nepal. *SSRN Electronic Journal*, (March):1–9, 2020.

[298] Klaifer Garcia and Lilian Berton. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing*, 101:107057, 2021.

[299] Cherish Kay L. Pastor. Sentiment analysis of Filipinos and effects of extreme community quarantine due to coronavirus (COVID-19) Pandemic. *Journal of Critical Reviews*, 7(7):91–95, 2020.

[300] Usman Naseem, Imran Razzak, Matloob Khushi, Peter W. Eklund, and Jinman Kim. COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. *IEEE Transactions on Computational Social Systems*, 8(4):976–988, 2021.

[301] László Nemes and Attila Kiss. Social media sentiment analysis based on COVID-19. *Journal of Information and Telecommunication*, 5(1):1–15, 2021.

[302] Jim Samuel, G. G.Md Nawaz Ali, Md Mokhlesur Rahman, Ek Esawi, and Yana Samuel. COVID-19 public sentiment insights and machine learning for tweets classification. *Information (Switzerland)*, 11(6):1–22, 2020.

[303] Sakun Boon-Itt and Yukolpat Skunkan. Public perception of the COVID-19 pandemic on twitter: Sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, 6(4):1–17, 2020.

[304] A. H. Alamoodi, B. B. Zaidan, A. A. Zaidan, O. S. Albahri, K. I. Mohammed, R. Q. Malik, E. M. Almahdi, M. A. Chyad, Z. Tareq, A. S. Albahri, Hamsa Hameed, and Musaab Alaa. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert Systems with Applications*, 167(April 2020):114155, 2021.

[305] Raina M. Merchant and Nicole Lurie. Social Media and Emergency Preparedness in Response to Novel Coronavirus. *JAMA - Journal of the American Medical Association*, 323(20):2011–2012, 2020.

[306] Younggue Bae and Hongchul Lee. Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science and Technology*, 63(12):2521–2535, dec 2012.

[307] Huizhi Liang, Umarani Ganeshbabu, and Thomas Thorne. A Dynamic Bayesian Network Approach for Analysing Topic-Sentiment Evolution. *IEEE Access*, 8:54164–54174, 2020.

[308] C. J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, pages 216–225, 2014.

[309] New York Times GitHub Database - https://github.com/nytimes/covid-19-data.

[310] Andrew Carswell. United States Census Bureau, 2020.

[311] Documenting the Now. Hydrator - https://github.com/docnow/hydrator.

[312] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from justin bieber's heart: The dynamics of the "location" field in user profiles. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 237–246, 2011.

[313] Catherine D'Ignazio, Rahul Bhargava, Ethan Zuckerman, and Luisa Beck. CLIFF-CLAVIN: Determining Geographic Focus for News Articles. In *Proceedings of the NewsKDD: Data Science for News Publishing*, 2014.

[314] J. O. Lee, R. Kosterman, T. M. Jones, T. I. Herrenkohl, I. C. Rhew, R. F. Catalano, and J. D. Hawkins. Mechanisms linking high school graduation to health disparities in young adulthood: a longitudinal analysis of the role of health behaviours, psychosocial stressors, and health insurance. *Public Health*, 139:61–69, 2016.

[315] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, aug 1987.

[316] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81, 1938.

[317] Maurizio Porfiri, Roni Barak-Ventura, and Manuel Ruiz Marín. Self-Protection versus Fear of Stricter Firearm Regulations: Examining the Drivers of Firearm Acquisitions in the Aftermath of a Mass Shooting. *Patterns*, 1(6):100082, 2020.

[318] David A. Dickey and Wayne A. Fuller. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, 74(366a):427–431, jun 1979.

[319] Maurizio Porfiri, Raghu Ram Sattanapalle, Shinnosuke Nakayama, James Macinko, and Rifat Sipahi. Media coverage and firearm acquisition in the aftermath of a mass shooting. *Nature Human Behaviour*, 3(9):913–921, 2019.

[320] Ted Lewis, Stefan Pickl, Ben Peek, and Guoliang Xu. Network science. *IEEE Network*, 24(6):4–5, nov 2010.

[321] William D. Berry, Evan J. Ringquist, Richard C. Fording, and Russell L. Hanson. Measuring Citizen and Government Ideology in the American States, 1960-93. *American Journal of Political Science*, 42(1):327, 1998.

[322] Richard C. Fording. State Ideology Data - Https://Rcfording.Wordpress.Com/State-Ideology-Data/.

[323] Ben Balmford, James D. Annan, Julia C. Hargreaves, Marina Altoè, and Ian J. Bateman. Cross-Country Comparisons of Covid-19: Policy, Politics and the Price of Life. *Environmental and Resource Economics*, 76(4):525–551, aug 2020.

[324] David Badre. How We Can Deal with Pandemic Fatigue. *2021*.

[325] Lauren Leatherby and Rich Harris. States that imposed few restrictions now have the worst outbreaks. *2020*.

[326] G. Veletsianos. Higher education scholars' participation and practices on Twitter. *Journal of Computer Assisted Learning*, 28(4):336–349, aug 2012.

[327] Charles Wankel. Educating Educators with Social Media. *Development and Learning in Organizations: An International Journal*, 26(3):dlo.2012.08126caa.012, apr 2012.

[328] Laura Montenovo, Xuan Jiang, Felipe Lozano Rojas, Ian Schmutte, Kosali Simon, Bruce Weinberg, and Coady Wing. Determinants of Disparities in Covid-19 Job Losses. Technical report, National Bureau of Economic Research, Cambridge, MA, may 2020.

[329] Aaron Smith and Joanna Brenner. Twitter use 2012. *Pew Internet & American Life Project*, 4:1–12, 2012.

[330] The European Parliament and The Council of Europe. REGULATION (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, 2016.

[331] Wayne W. Zachary. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33(4):452–473, dec 1977.

[332] Santi Phithakkitnukoon, Zbigniew Smoreda, and Patrick Olivier. Socio-geography of human mobility: A study using longitudinal mobile phone data. *PLoS ONE*, 7(6):1–9, 2012.

[333] Sune Lehmann and Document Version. *Dynamics of High-Resolution Networks Vedran Sekara*. PhD thesis, Danmarks Tekniske Universitet, 2015.

[334] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. Measuring large-scale social networks with high resolution. *PLoS ONE*, 9(4):no pagination, apr 2014.

[335] Vedran Sekara, Arkadiusz Stopczynski, and Sune Lehmann. Fundamental structures of dynamic social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 113(36):9977–9982, 2016.

[336] Arkadiusz Stopczynski, Alex Sandy Pentland, and Sune Lehmann. Physical Proximity and Spreading in Dynamic Social Networks. *arXiv preprint*, sep 2015.

[337] Arkadiusz Stopczynski, Alex 'Sandy' Pentland, and Sune Lehmann. How Physical Proximity Shapes Complex Social Networks. *Scientific Reports*, 8(1):17722, 2018.

[338] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. *PLoS ONE*, 5(7):e11596, jul 2010.

[339] Moses C. Kiti, Michele Tizzoni, Timothy M. Kinyanjui, Dorothy C. Koech, Patrick K. Munywoki, Milosch Meriac, Luca Cappa, André Panisson, Alain Barrat, Ciro Cattuto, and D. James Nokes. Quantifying social contacts in a household setting of rural Kenya using wearable proximity sensors. *EPJ Data Science*, 5(1), 2016.

[340] Jason Yipin Ye. Atlantis : Location Based Services with Bluetooth. *Ad Hoc Networks*, pages 1–8, 2005.

[341] Nathan Eagle and Alex Pentland. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[342] Jenny Röbesaat, Peilin Zhang, Mohamed Abdelaal, and Oliver Theel. An improved BLE indoor localization with Kalman-based fusion: An experimental study. *Sensors (Switzerland)*, 17(5):1–26, 2017.

[343] Arkadiusz Stopczynski, Jakob Eg Larsen, Sune Lehmann, Lukasz Dynowski, and Marcos Fuentes. Participatory bluetooth sensing: A method for acquiring spatio-temporal data about participant mobility and interactions at large scale events. *2013 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2013*, (March):242–247, 2013.

[344] A. K.M.Mahtab Hossain and Wee Seng Soh. A comprehensive study of bluetooth signal parameters for localization. *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, 2007.

[345] Song Chai, Renbo An, and Zhengzhong Du. An Indoor Positioning Algorithm using Bluetooth Low Energy RSSI. *Amsee*, (Amsee):276–278, 2016.

[346] Siniša Husnjak, Ivan Jovović, Ivan Cvitić, and Josip Štefanac. Overview: Operating Systems of Modern Terminal Devices. *Proceedings of The 5th International Virtual Research Conference In Technical Disciplines*, 6:8–13, 2018.

[347] Francesco Vincenzo Surano. Human To Human GitHub repository - https://github.com/fvsura/humanToHuman.

[348] Nicholas R Jones, Zeshan U Qureshi, Robert J Temple, Jessica P J Larwood, Trisha Greenhalgh, and Lydia Bourouiba. Two metres or one: what is the evidence for physical distancing in covid-19? *BMJ*, page m3223, 2020.

[349] Lydia Bourouiba. Turbulent Gas Clouds and Respiratory Pathogen Emissions. *JAMA*, pages 1837–1838, 2020.

[350] Francesco Vincenzo Surano. Dynamical Systems Laboratory, Human To Human GitHub repository - https://github.com/Dynamical-Systems-Laboratory/humanToHuman.