



UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

UNDERSTANDING SOCIAL INTERACTIONS VIA TEMPORAL NETWORK ANALYSIS

Antonio Longa

Advisor

Prof. Bruno Lepri

Fondazione Bruno Kessler

Co-Advisor

Prof. Andrea Passerini

Università degli Studi di Trento

Dr. Giulia Cencetti

Fondazione Bruno Kessler

March 2023

Abstract

The COVID-19 epidemic has had a significant impact on society, affecting not only physical health but also mental health, social interactions, and the economy. Measures such as lockdowns, travel restrictions, and social distancing have altered the way we live, work, and interact. Digital contact tracing is a valuable tool in managing infectious disease outbreaks and can help avoid severe lockdowns and excessive quarantines. Our initial research explored the effectiveness of contact tracing apps, but we faced the challenge of accessing actual temporal interaction networks to accurately simulate disease spread. To gain a deeper understanding of the underlying structures of temporal networks, we delved into the study of temporal networks in our second project. Our focus was on developing an effective approach for identifying temporal motifs in interaction networks, and we introduced the concept of egocentric temporal neighborhoods (ETN) and egocentric temporal motifs (ETM). Finally, we proposed a generative model for temporal networks called ETNgen, which takes into account the intrinsic temporal correlations present in real-world temporal networks. The model captures the time-evolving network structure of egocentric temporal neighborhoods (ETN), thus providing a more accurate representation of real-world networks.

Keywords

[Temporal network motifs, Temporal network generation, human interaction networks]

Contents

1	Introduction	3
1.1	Contributions	5
1.1.1	A Real-world application: the Digital Contact Tracing	5
1.1.2	Mining in temporal networks	6
1.1.3	Generating temporal networks	8
1.2	Publications	10
2	Background	13
2.1	Networks	13
2.2	Datasets of human interaction	16
3	Digital Contact Tracing	21
3.1	Methods	22
3.1.1	A modeling framework for digital contact tracing on empirical contact networks	23
3.1.2	Spreading and tracing on the real network	26
3.1.3	Aggregation and parameter estimation	27
3.2	Results	28
3.2.1	Tracing efficiency based on empirical contact data	28

3.2.2	How infectiousness depends on duration and proximity	30
3.2.3	Design of appropriate policies	32
3.2.4	Digital tracing enables containment for moderate reproductive numbers	34
3.2.5	Any effective containment comes at a cost	36
3.3	Discussion	38
3.3.1	Policies for digital contact tracing: implications and constraints . .	38
3.3.2	Digital contact tracing: insights and limitations	40
3.4	Code Availability	43
4	Egocentric Temporal Motifs Miner	45
4.1	Method	47
4.1.1	Mining egocentric temporal motifs	47
4.1.2	ETM-based graph distance	52
4.1.3	Experimental setup	53
4.2	Results	56
4.2.1	Egocentric vs non-egocentric temporal motifs	56
4.2.2	Egocentric vs non-egocentric graph distances	59
4.2.3	Sensitivity analysis	61
4.2.4	Results on distance communication and synthetic datasets	62
4.3	Discussions	64
4.4	Code availability	65
5	Generating Temporal Networks	67
5.1	Method	68
5.1.1	The neighborhood generation process	68

5.1.2	Model evaluation	71
5.1.3	Computational complexity and space complexity	72
5.1.4	Size expansion: preserving interaction density.	72
5.1.5	Alternatives approaches for generating networks.	73
5.2	Results	74
5.2.1	Temporal periodicity	74
5.2.2	Topological similarity evaluation	76
5.2.3	Dynamical similarity evaluation	79
5.2.4	Dataset expansion and extension	81
5.3	Discussion	84
5.4	Code availability	86
6	Conclusion	87
6.1	Limitations & Future Directions	88
	Bibliography	91
7	Appendix	111
7.1	Digital Contact Tracing	111
7.1.1	Characteristic parameters of the disease	111
7.1.2	Evaluation of additional containment measures and refined policies	120
7.2	Generating Temporal Networks	134
7.2.1	Execution time comparison	134
7.2.2	Scalability	134
7.2.3	Varying K	136
7.2.4	Multiple versus single probabilistic model	138

Chapter 1

Introduction

Our world is structured upon the interactions between various entities. Society arises by the connections between individuals, molecules are formed by bonding atoms together, and the internet is a network of interlinking computers, among other examples. These discrete and pairwise relationships can be represented by making use of discrete mathematical objects known as *networks*, which consist of collections of nodes (entities) that are connected by edges or links (relations). The study of networks has been the subject of considerable research in recent decades, making network science one of the major disciplines in the study of complex systems.

Many real-world interacting systems, including social relationships, are dynamic and evolve over time. These changes can be captured by temporal networks, where edges can appear and disappear in time. Schematizing the real world via such temporal architectures allows researchers to analyze and comprehend complex dynamics. The study of temporal networks has proven to be essential in many fields, including Chemistry, Physics, Biology, and Computational Social Sciences. Computational Social Science (CSS) is an interdisciplinary field that combines social science theories and methods with computer science techniques and technologies to study social phenomena. The goal of this discipline is to better understand and explain social behaviours, social systems, and social interactions. The increasing amount of data involving people and social interactions provide us with an extremely useful instrument, which has opened the perspective of this new field. The possibility to acquire real privacy-preserving data allows research to evolve and, starting with the analysis of real interactions, to get insight and formulate new theories. In my thesis, I have made use of several open data on social interactions, namely face-to-face and co-location data which have been collected in different environments.

The temporal nature of face-to-face interactions can be well represented by temporal networks. Both network science and computational social science are complex disciplines well studied in the last decades. However, physical interactions in a social context represent a complex problem and a complete characterization of them is still missing. Open problems involve how social bonds arise and evolve in time, how each individual's behaviour is affected by past interactions and by the features of its neighbours, and the effect that a single individual may have on a group in a specific social context. Moreover, if the temporal network represents the physical structure on which a dynamical process evolves, it becomes natural to wonder how the dynamics are affected by the topology of such networks.

With this in mind, we started by exploring the spreading of COVID-19 in a face-to-face interaction network. In this study we investigated how isolating and quarantining infected people affects the virus spreading. Both isolations and quarantines correspond to removing connections of the original network, thus implying a change in the topology of the temporal network. The impact of this change on spreading dynamics is very significant and it can be quantified via appropriate indicators that we define. This highlights the complex interplay between network structure and dynamics and makes us understand the importance of a deeper investigation of temporal network topology. To this end, we decompose temporal networks finding their constitutive building blocks. These correspond to small subnetworks with an extension in time and in space, that are centred on specific ego nodes: we refer to those subnetworks as *Egocentric Temporal Neighbourhood (ETN)*. We found a method to identify the most significant ones for each temporal network and we gave them the name of *Egocentric Temporal Motifs (ETM)*. It is worth mentioning that finding important structures in networks is extremely expensive. This is not the case for our *ETM* since they can be converted into binary strings, thus the mining becomes extremely fast. Decomposing a temporal network into substructures not only allows for a better understanding of the underlying dynamics but also allows for generating synthetic ones. Generating temporal networks is crucial for many applications. For instance, the main issue with our first work is the lacking of big face-to-face interaction data. In fact, we exploit the novel concept of *ETN* as a building block to generate synthetic networks similar to the original one, namely *surrogate* temporal networks. Our method, not only generates temporal networks with high temporal granularity, but it also allows for the extension (increase in the number of temporal snapshots) and expansion (increase in the number of nodes) of temporal networks, something not feasible so far. Moreover, since our generative model is based on *ETN* the procedure is extremely fast compared to state-of-the-art algorithms.

This thesis is structured as follows. In Chapter 1, the contributions developed during the doctoral program are presented. Chapter 2 introduces the mathematical backgrounds and datasets that were utilized throughout this thesis. Chapter 3 provides a real-world application of digital contact tracing and its effectiveness in reducing the spread of COVID-19. In Chapter 4, ETN and ETM are introduced, and then we demonstrate their effectiveness in calculating the distances between temporal networks collected from different social contexts. Chapter 5 presents our generative model for fine-grained temporal networks, with a particular focus on temporal extension and expansion. Finally, in the concluding chapter, several future developments are presented. In the rest of this chapter, we discuss in more detail the original contributions presented in this thesis.

1.1 Contributions

The upcoming subsections aim to provide an overview of the issues discussed in the subsequent chapters. The purpose of this introduction is to offer readers an in-depth understanding of the challenges and complexities that will be addressed in the forthcoming sections.

1.1.1 A Real-world application: the Digital Contact Tracing

As of mid-January 2021, the COVID-19 pandemic has resulted in over 85 millions detected cases worldwide (29), overwhelming the healthcare capacities of many countries and thus presenting extraordinary challenges for governments and societies (104; 114; 154). Rigorous restrictions such as lockdowns and quarantine have proven to be effective in many countries as a measure to curb the spread of COVID-19, limit contagions and reduce the effective reproductive number (63; 104; 4; 86; 79; 180; 32; 22; 38; 54). Many areas slowly started to lift the restrictions, but new outbreaks appeared again, arriving in waves as anticipated by several early models (47; 82). An effective and affordable long-term plan is required, since the fraction of the population that has been infected is still far too low to provide herd immunity (154).

Despite their efficacy, large-scale quarantine and lockdown strategies carry large costs (4). Moreover, in a situation where most of the population is not infected, population-wide lockdowns are far from optimal, and interventions at smaller scale, selectively targeting individuals at higher risk of spreading the disease, are more desirable.

While the testing and isolation of symptomatic cases is crucial, it is insufficient in the

case of COVID-19, since there is clear evidence of presymptomatic and asymptomatic transmission (117; 53; 48; 96; 141). Thus, the identification and isolation of infected cases must be coupled with a strategy for tracing their contacts and preventively quarantining them (113; 93; 71; 48). Traditional manual contact tracing, besides being slow and labor intensive (46; 83; 69), is not able to entirely reconstruct close proximity contacts (162; 120). Thus, technologies based on digital sensors have been developed to complement manual tracing. The idea is to leverage the widespread dissemination of smartphones to develop proximity-sensing apps based on the exchange of Bluetooth radio packets between them (170; 41; 48; 136; 146; 68; 115; 25), within a privacy-preserving contact tracing framework (170).

The efficacy of digital contact tracing (DCT) (85; 80; 59; 49; 118; 1; 71; 112; 93; 14; 113) has been discussed in several recent papers. We draw inspiration from the work by Fraser *et al.*(52), recently adapted to the case of COVID-19 by Ferretti *et al.*(48). This work models the pandemic evolution using recursive equations describing the number of infected individuals in a homogeneously mixed population, taking into account the evolving infectiousness of the infected individuals. The analysis is based on two effective parameters, ε_I and ε_T , to represent the ability to identify and isolate infected individuals, and to correctly trace their contacts, respectively. Assuming an exponential growth for the number of infected individuals (applicable in early phases of an uncontrolled epidemic outbreak) the authors studied how the growth rate depends on these intervention parameters.

Chapter 3 discusses a new approach to understanding the effectiveness of contact tracing in the real world. First, the mathematical framework proposed by Ferretti *et al.*(48) is restructured to apply to any epidemic growth pattern and phase, and the epidemiological aspects of the model are modified to consider asymptomatic cases and delays in isolating infected individuals. The tracing ability is quantified through simulations on real-world data sets, allowing for a more accurate evaluation of the impact of tracing on disease spread. The probability of contagion events is also considered, and policies for identifying risky contacts are developed to minimize false positives and negatives. The approach allows for the evaluation of the effect of different tracing policies on disease spread and their impact on the population.

1.1.2 Mining in temporal networks

Complex networks play a pivotal role in describing and analyzing complex systems in multiple natural and artificial scenarios, representing a fundamental tool for modeling biological, cognitive and social systems (132). Interestingly, the small substructures that

compose the complex topology of a network are sometimes recurrently emerging as essential constituents for the specific network at hand. They consist in sub-networks composed by a small number of nodes with a specific structure of connections. The substructures which are identified as the most significant take the name of *motifs* (124; 3). The significance of each specific substructure within the overall network architecture is assessed in relation to its frequency and usually referring to a null model: a structure is considered a motif if the number of its occurrences in the network is substantially higher than the number of occurrences in the null model.

The identification of specific repeated motifs offers a unique opportunity to investigate the complex and intricate dynamics of human behavior and interactions (177; 123). As a matter of fact, when analyzing social dynamics we usually need to deal with time-dependent structures (88; 87). Social interactions are indeed characterized by links which appear and disappear in time and are associated with variable duration. The appropriate topological tool to describe systems of dynamical interactions is represented by temporal networks with a fixed set of nodes connected by edges that vary over time (72). In such framework the identification of motifs becomes more challenging, since a substructure can be repeated both in time and in space.

In the last years, a number of solutions has been developed to mine motifs in temporal networks (see (78) for a survey). In this thesis, we focus on temporal networks where nodes are fixed and edges can change over time. Currently, two popular strategies have been followed to adapt graph mining approaches to deal with a changing network topology. The first strategy (7; 42; 169) consists of aggregating temporal information, i.e., building a static network containing all connections in the temporal graph regardless of the time associated with them. While this simple strategy allows using standard techniques for motifs discovery, it loses the ability to capture the temporal dynamics of the interactions between nodes. The second one consists in building a growing network, where nodes and edges can be added but never deleted (147; 100). However these approaches are not appropriate to deal with data containing social interactions which are necessarily transient.

Most methods for mining motifs of transient interactions have been developed in the field of communication networks. Kovanen *et al.* (91) define the concept of Δt - *connected* graph as the connected temporal graph containing edges within a temporal gap Δt , and search for temporal motifs inside them. Zhao *et al.* (184) extended this concept to *communication motifs*, basically requiring a number of occurrences greater than a given threshold. Later, Gurukar *et al.* (62) proposed COMMIT, an algorithm that converts con-

nected temporal subgraphs in sequences using graph invariants and then mines frequent sub-sequences as communication motifs. More recently, Kosyfaki *et al.* (90) proposed a new definition of max-flow communication motifs, in which flow refers to data (e.g., money, messages, etc.). Hulovatyy *et al.* (75) introduced *dynamic graphlets*, which extend the concept of graphlets from static networks to temporal graphs. However, they do not search for temporal motifs but rather use all dynamic graphlets (up to a given complexity) to generate vectorial representations of the network and its nodes. A related line of research aims at characterizing temporal networks in terms of dense subgraphs (89; 153; 152). Finally, Paranjape *et al.* (139) propose a mining strategy that extracts static motifs from the aggregate network (obtained collapsing all the temporal layers together and thus dropping the temporal information) and expands them into temporal motifs by considering the order of appearance of edges within a given temporal gap. Other studies investigated approximate methods for counting temporal motifs (105; 176).

None of these approaches tries to capture the temporal evolution of the interactions of a single node, which is the focus of our work. The egocentric perspective allows to extract meaningful patterns of interaction that are hard to find with non-egocentric solutions. Additionally, it allows to devise an efficient procedure to compare these types of patterns that can substantially speed up the mining process.

In Chapter 4, a new approach to analyzing temporal networks called egocentric temporal motifs (*ETM*) is presented. This approach focuses on a specific node, observing its neighbors and how its connections change over time from an "ego perspective". By comparing the egocentric temporal sub-networks using a bit vector signature, this method simplifies the motif identification procedure compared to traditional methods that require graph isomorphism. The effectiveness of this approach was evaluated by applying it to various interaction datasets, including close proximity interaction networks and distance communication networks, and comparing it to existing micro-scale, meso-scale, and global-scale alternatives. Results showed that the *ETM* approach was more effective in discriminating between different types of graphs and characterizing a wide range of interactions, although there were some limitations.

1.1.3 Generating temporal networks

Across the past decade, temporal networks have driven breakthroughs in real world systems across Biology, communications, social interactions, and mobility. The power of temporal networks resides in their ability to capture complex dynamics such as diffusion

and contagion (74; 64; 43; 31; 11; 26; 164; 35; 95; 34). In order to model realistic dynamics, it is often necessary to employ large temporal networks, including a large number of nodes and many temporal layers (163; 150; 28). Many state-of-the-art temporal datasets, however, are limited both in the number of agents and in the number of temporal layers (26; 77; 167; 2; 156). When the available data are insufficient – e.g. in long-term epidemiological simulations – datasets are extended by simply repeating the same temporal sequence multiple times, a procedure which is known to result in biases (167). An appealing solution to the problem of insufficient or privacy sensitive data is to use *surrogate temporal networks* (142). Surrogate temporal networks are synthetic datasets which mimic the real-world temporal patterns relevant for a desired use-case. They can involve the desired number of nodes and number of temporal layers, where the actual dynamics are known through smaller studies or via available small datasets. The real dynamics, that surrogate temporal networks aim to reproduce, is known to be characterized by typical patterns of interactions, different in different domains (social, biological, infrastructural, etc.), but that we can often recognize and delineate (128; 81). Moreover, in the case of sensible data, such as fine-grained records of social interactions (37), surrogate data can freely be shared. Knowledge of actual dynamics is often available because real temporal networks are known to be characterized by typical patterns of interactions, different in different domains (social, biological, infrastructural, etc.), but that we can often recognize and delineate (128; 81). For these reasons, it is clear that surrogate temporal networks are highly desirable from the perspective of a number of applications. Over the past years, a large number of successful algorithms for *static* network generation have been proposed (23; 36); however, extending these models to the dynamic regime has proven prohibitively difficult, due to the greatly increased complexity introduced by the temporal dimension.

Indeed, it has become clear that temporal networks are characterized by a highly non-trivial interplay between the network topology (adjacency, degree distribution, clustering, etc.) at a given time and the temporal activation of nodes and links – how each connection changes over time (duration of interactions, patterns by which new links appear and old ones disappear, etc.). From the ‘egocentric’ perspective of an individual node, these two dimensions imply that models must take into account (*i*) the history of what has occurred in the preceding timesteps and (*ii*) the current activations of the neighboring nodes. Current network theory has not yet been able to fully understand and model the interplay between the two dimensions. So far, it is not even clear which statistics to measure (73). In fact, the scientific literature is full of studies focusing on the spatial dimension but unable to take into account possible temporal correlations (12; 92; 39; 20; 40; 138), or – alternatively – work dedicated to model the behavior of individual nodes in time (for

example activity driven models (140; 164)) which does not reproduce realistic network topologies (55). There exist models for link prediction that try to combine temporal and topological dimensions by using small local temporal patterns (16) or building over a backbone of significant links (142). However, there is currently a dearth of models for generating surrogate networks from scratch that are able to take into account the two dimensions simultaneously. The few works that do this rely on temporal motifs, like *Dymond* (179) and *STM* (143), or on deep learning like *TagGen* (185). These three models represent the state-of-the-art. We show, however, that all these techniques generate temporal networks with massive macroscopic differences compared with the original temporal network datasets, and that – in many cases – the output-networks do not reflect the dynamic behavior of the original network.

Chapter 5 proposes a method called Egocentric Temporal Neighborhood Generator (*ETN-gen*) for generating synthetic networks that match real-world networks in terms of topological and dynamic measures. The generative algorithm uses the idea of egocentric temporal neighborhood, which includes a small number of prior time steps for a node without interactions between its neighbors. The algorithm characterizes a given real-world network in terms of neighborhoods and uses them as building blocks for a new synthetic network. A local probabilistic model suggests new temporal interactions for each node at subsequent time steps based on its behavior during the prior time steps. The algorithm is scalable, easily interpretable, and can be used for generating any type of graph. The surrogate networks generated by *ETN-gen* match original networks with high accuracy, not just in terms of local features, but also in terms of global features such as the number of interactions, the number of interacting individuals in time, and density of their connections. The method can generate large datasets without resolution limits and mitigates privacy issues.

1.2 Publications

My doctoral program has brought to the publication of the following papers, inherent to this thesis.

- Giulia Cencetti, Gabriele Santin, **Antonio Longa**, Emanuele Pigani, Alain Barrat, Ciro Cattuto, Sune Lehmann, Marcel Salathé and Bruno Lepri (2021, March) *Digital proximity tracing on empirical contact networks for pandemic control*. Nature communications. (28)
- **Antonio Longa**, Giulia Cencetti, Bruno Lepri and Andrea Passerini (2021, Novem-

ber) *An efficient procedure for mining egocentric temporal motifs*. Data Mining and Knowledge Discovery. (110)

- **Antonio Longa**, Giulia Cencetti, Sune Lehmann, Andrea Passerini and Bruno Lepri. *Neighbourhood matching creates realistic surrogate temporal networks*. Under review at Communication Physics.(109)

The program was instrumental to study other topics, which I chose not to include in this thesis, and have however produced those publications:

- Hazem Peter Samoaa, **Antonio Longa**, Mazen Mohamad, Morteza Haghiri Chehrehgani and Philipp Leitner (2022, November) *TEP-GNN: Accurate Execution Time Prediction of Functional Tests Using Graph Neural Networks*. Proceeding of Product-Focused Software Process Improvement: 23rd International Conference, PROFES 2022, Jyväskylä, Finland.(155)
- Giovanni Mauro, Massimiliano Luca, **Antonio Longa**, Bruno Lepri and Luca Pappalardo (2022, December) *Generating mobility networks with generative adversarial networks*. EPJ data science. (122)
- Steve Azzolin, **Antonio Longa**, Pietro Barbiero, Pietro Liò and Andrea Passerini (2023 March) *Global explainability of GNNs via logic combination of learned concepts*. Proceedings of International Conference on Learning Representations: 11th International Conference, ICLR 2023. (10)
- **Antonio Longa**, Steve Azzolin, Gabriele Santin, Giulia Cencetti, Pietro Liò, Bruno Lepri and Andrea Passerini. *Explaining the Explainers in Graph Neural Networks: a Comparative Study*. Under review at ACM Computing surveys.(108)
- **Antonio Longa**, Veronica Lachi, Gabriele Santin, Monica Bianchini, Bruno Lepri, Pietro Lio, Franco Scarselli and Andrea Passerini. *Graph Neural Networks for temporal graphs: State of the art, open challenges, and opportunities*. Under review at International Joint Conferences on Artificial Intelligence Organization, IJCAI 2023.(111)
- Hazem Peter Samoaa, Linus Aronsson, **Antonio Longa**, Mazen Mohamad, Morteza Haghiri Chehrehgani and Philipp Leitner. *A Unified Active Learning Framework for Annotating Graph Data with Application to Software Source Code Performance Prediction*. Under review at ECMLPKDD 2023

Chapter 2

Background

In this chapter, we delve into the concept of networks and provide a formal definition of what they are. We begin by discussing the fundamental building blocks of networks, such as nodes and edges, and how they can be used to model various types of relationships between entities. We also explore different types of networks, such as directed and undirected networks.

Additionally, we provide background information on important concepts and techniques used in network analysis. We explain how these tools can be used to extract valuable insights from network data and uncover hidden patterns and structures. Towards the end of the chapter, we showcase some of the networks used in this work, highlighting their characteristics.

2.1 Networks

At their core, both molecular and internet systems rely on the interaction and communication between different entities. In the case of molecules, the atoms are linked together by specific chemical bonds, forming a complex network of interactions that give rise to a vast array of chemical properties and behaviors. Similarly, in the internet, individual computers are connected through a complex network of cables, routers, and other devices, allowing for the seamless transfer of data and communication across the globe.

Relational data is a crucial concept in understanding these systems, as it refers to any type of data where the relationships between different entities are essential to understanding the overall system. In addition to molecules and the internet, other examples of relational

data include social networks, economic systems, biological ecosystems, and many others.

Understanding and analyzing these complex systems requires a range of tools and techniques, from computational modeling and simulation to network analysis and data visualization. By studying the relationships between different entities in these systems, researchers can gain a deeper understanding of their underlying structure, dynamics, and function, and develop new insights and innovations that can help to address a wide range of scientific and societal challenges.

A network, also known as a graph, is a mathematical construct that provides a way to represent and analyze relational data. In a network, entities are represented as vertices or nodes (denoted as V), while the relationships between them are represented as edges (denoted as E). These edges connect pairs of vertices, indicating a specific type of relationship or interaction between them.

Definition 1 (*Graph*). A graph G can be defined as a pair (V, E) , where V is a set of vertices or nodes, and E is a set of edges between the nodes, i.e., $E \subseteq \{(u, v) | u, v \in V\}$. A graph can be represented with a squared adjacency matrix A of size $|V| \times |V|$ in which the element $a_{i,j}$ is one if the graph has an edge from i to j . The graph is undirected if it does not contain self-loops and the associated adjacency matrix is symmetric, directed otherwise. At each node (or edge) of the graph can be associated features.

The provided illustration, depicted in Figure 2.1, serves as an exemplar of three types of graphs, namely an undirected graph, a directed graph, and a graph that encompasses both node and edge features. Notably, each of these graphs comprises seven vertices and nine edges.

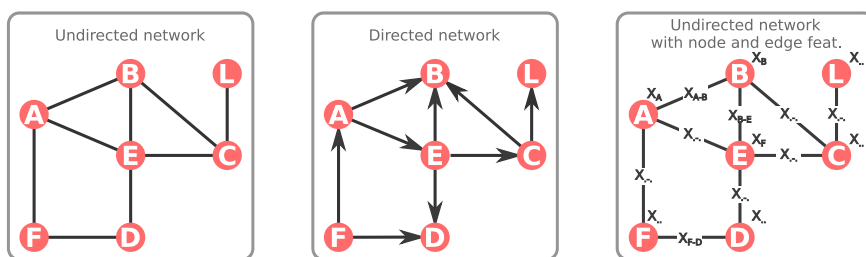


Figure 2.1: Example of undirected, directed and graph with node and edge feature

Definition 2 (*Graph isomorphism*). Two graphs $G = (V, E)$ and $G' = (V', E')$ are said to be isomorphic if and only if there exists a bijection π between their vertex sets such that for all $(u, v) \in E$ it holds that $(\pi(u), \pi(v)) \in E'$ (edge-preservation). Graph isomorphism is denoted as $G \simeq G'$.

Definition 3 (*Node neighborhood*). Given a graph $G = (V, E)$, the neighbors of a node $v \in V$ are the set of nodes adjacent to v , i.e., $\mathcal{N}(v) = \{u \in V \mid (u, v) \in E\}$. The node neighborhood is the subgraph of G containing v and its neighbors as nodes and all edges connecting them as edges.

Network motifs are patterns of connections occurring on a given network significantly more often than in random networks (124). The next definition formalizes the concept.

Definition 4 (*Network motif*). Given a graph G and a set of n random graphs G_0 , a sub-graph M of G is a network motif if and only if: (i) $\Pr(\bar{N}_{G_0} > N_G) < \alpha$ (*over-representation*); (ii) $N_G - \bar{N}_{G_0} \geq \beta \bar{N}_{G_0}$ (*minimum deviation*); (iii) $N_G \geq \gamma$ (*minimum frequency*). Here N_G is the number of occurrences of sub-graph M in G , \bar{N}_{G_0} is the average number of occurrences of sub-graph M in the random graphs (G_0) and $\alpha \in [0, 1]$, $\beta \in [0, 1]$ and $\gamma \in \mathbb{N}$ are parameters.

The *over-representation* condition requires that the probability of observing a motif in the random graphs more than in the original one is lower than a certain threshold α . *Minimum deviation* instead prevents the detection as motifs of subgraphs with a slight difference in occurrences between the graph under investigation and the random graphs. Finally, *minimum frequency* avoids detecting statistically significant but infrequent motifs.

Definition 5 (*Temporal graph*). A temporal graph $\mathcal{G} = (V, E)$ is a pair of sets where V is a set of vertices or nodes and E is a set of temporal edges, i.e., edges enriched with temporal information. Each temporal edge $e \in E$ is a quadruple $(u, v, t_{start}, t_{end})$, where u and v are nodes ($u, v \in V$) and t_{start} and t_{end} are time instants representing, respectively, the beginning and the end of the interaction between node u and node v . Given a temporal graph \mathcal{G} , its corresponding (static) aggregate graph G is obtained removing temporal information from the edges of \mathcal{G} .

Definition 6 (*Temporal graph snapshot*). Given a temporal graph $\mathcal{G} = (V, E)$ and a temporal gap Δt , a temporal graph snapshot at time t is a static graph $G_t = (V_t, E_t)$ such that $V_t = V$ and there is a static edge $(u, v) \in E_t$ if and only if the corresponding temporal interaction $(u, v, t_{start}, t_{end}) \in E$ exists within Δt , i.e. $t_{start} \in [t, t + \Delta t) \vee t_{end} \in [t, t + \Delta t)$. A temporal graph $\mathcal{G} = (V, E)$ can be represented as a sequence of temporal graph snapshots $G_{t_1}, G_{t_2}, \dots, G_{t_m}$ where t_1 is the smallest t_{start} in E , $t_{i+1} = t_i + \Delta t$ and t_m is smaller than the largest t_{end} in E .

A simple example of temporal graphs and temporal graph snapshots is given in figure 2.2

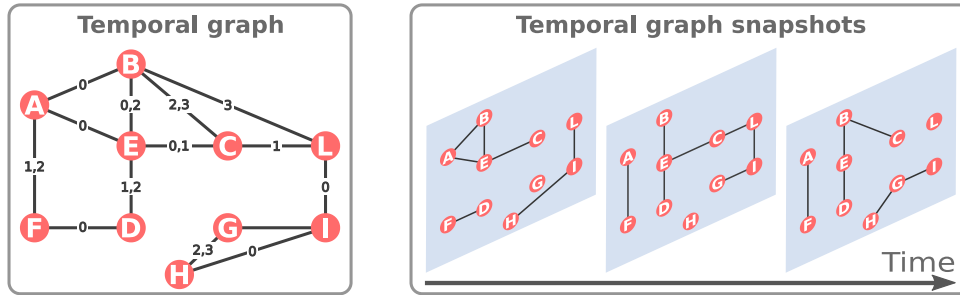


Figure 2.2: The left panel shows a temporal graph \mathcal{G} . The right panel shows three graph snapshots.

2.2 Datasets of human interaction

As mentioned in the opening section of this chapter, graphs provide an effective tool for modeling a wide range of real-world situations. Specifically, the focus of this study is on network modeling human interactions, including face-to-face interactions, proximity networks, phone call networks, SMS networks, and email networks. In the following sections, we delve into each of these categories, presenting the specific networks utilized in our analysis.

Face-to-face interactions networks: Those networks have been collected using the wearable sensors developed by the SocioPatterns¹ collaboration, equipped with radio-frequency identification devices (RFIDs) capturing face-to-face interactions. The devices record an interaction if and only if there is at least one exchanged signal within 20 seconds, so 20 seconds is the smallest time resolution.

- **HighSchool11 (50).** The dataset has been collected in 2011 in Lycée Thiers, Marseilles, France, over four days (Tuesday to Friday). It reports the interactions among 118 students and 8 teachers in three different high school classes.
- **HighSchool12 (50).** The dataset has been collected in 2012 in Lycée Thiers, Marseilles, France, over seven days (Monday to Tuesday of the following week). It reports the interactions among 180 students in five different high school classes.
- **HighSchool13 (120).** The dataset has been collected in 2013 in Lycée Thiers, Marseilles, France, over five days in December. It reports the interactions among 327 students in nine different high school classes.
- **InVS13 (58).** The dataset has been collected in 2013 at the *Institut National de*

¹<http://www.sociopatterns.org/>

Veille Sanitaire, a health research institute near Paris, over two weeks. The dataset contains 92 individuals divided in five departments: DISQ, DMCT, SFLE, DSE and SRH.

- **LH10 (172)**. The dataset has been collected in the geriatric ward of a university hospital (173) in Lyon, France, over four days in December 2010. The individuals belong to four classes: medical doctors (MED), paramedical staff (NUR), administrative staff (ADM) and patients (PAT).
- **Primary school (168)**. The dataset has been collected in a primary school in France, over two days in October 2009. The individuals belong to two classes: teachers (10 individuals) and children (232 individuals).

Proximity networks: Those networks have been Bluetooth technology to infer physical co-location. In particular, we explore the DTU dataset ((156)) representing proximity interactions among university students, collected over a month.

- **DTU (156)**. The dataset represents the interactions among university freshmen students at Copenhagen University. In particular, DTU represents the network of interactions among students collected over a month using Bluetooth technology to infer physical proximity.

Phone calls networks: Those networks represent the phone calls between people. In this work, we explore two phone calls dataset.

- **DTU calls (156)**. The dataset represents phone calls among university freshmen students in the Copenhagen University. Number of edges: 605, number of nodes: 525.
- **Friends and Family calls (2)**. The dataset represents phone calls among members of a young-family residential living community adjacent to a major research university in North America.

SMS networks: Those networks represent the SMS between people. In this work, we explore two SMS dataset.

- **DTU SMS (156)**. The dataset represents SMSs among university freshmen students in the Copenhagen University. Number of edges: 697, number of nodes: 568.

- **Friends and Family SMS (2).** The dataset represents SMSs among members of a young-family residential living community adjacent to a major research university in North America.

Email networks: Those networks represent the emails exchanged between people. In this work, we explore two email networks.

- **Email EU (139)** The dataset is a collection of emails between members of a European research institution.
- **Email DNC (151)** The dataset is a collection of leaked emails between members of the 2016 Democratic National Committee.

The key attributes of the networks being analyzed, which encompass the number of nodes, temporal edges, density, duration in days, and data collection technology, are presented in Table 2.1.

Network name	nodes	temporal edges	density	length	technology
HighSchool 11	126	28560	0.217	4 days	SocioBadge
HighSchool 12	180	45047	0.138	7 days	SocioBadge
HighSchool 13	327	188508	0.109	5 days	SocioBadge
InVS13	92	9827	0.180	10 days	SocioBadge
LH10	75	32424	0.410	5 days	SocioBadge
PrimarySchool	242	125773	0.285	2 days	SocioBadge
DTU	692	2426279	0.333	30 days	Bluetooth
DTU calls	525	3489	0.004	30 days	Calls
Friends and Family calls	95	1613	0.027	7 days	Calls
DTU SMS	405	6366	0.004	30 days	SMS
Friends and Family SMS	49	796	0.034	7 days	SMS
Email EU	986	332334	0.027	803 days	email
Email DNC	1900	37400	0.021	27 days	email

Table 2.1: The fundamental characteristics of the networks being examined

Data availability

The datasets employed in our experiments can be downloaded at:

- Sociopattern networks <http://www.sociopatterns.org>
- DTU networks <https://doi.org/10.6084/m9.figshare.7267433>
- Friends&Family networks <http://realitycommons.media.mit.edu/friendsdataset.html>

- Emails <http://snap.stanford.edu/data/email-Eu-core-temporal.html>
- Emails DNC <http://networkrepository.com/email-dnc.php>

Chapter 3

Digital Contact Tracing

Recalling the work by Ferretti *et al.*(48), in which they adapt the work by Fraser *et al.*(52), to the case of COVID-19. In particular, Ferretti *et al.*(48) model the pandemic evolution using recursive equations describing the number of infected individuals in a homogeneously mixed population, taking into account the evolving infectiousness of the infected individuals. The analysis is based on two effective parameters, ε_I and ε_T , to represent the ability to identify and isolate infected individuals, and to correctly trace their contacts, respectively. Assuming an exponential growth for the number of infected individuals (applicable in early phases of an uncontrolled epidemic outbreak) the authors studied how the growth rate depends on these intervention parameters.

Here, we discuss a new approach to understand the effectiveness of contact tracing in the real world. First, we restructure and generalize the mathematical framework. proposed by Ferretti *et al.*(48), to allow us to completely avoid assumptions regarding the functional form of the epidemic growth. This development makes the setting applicable to any possible evolution shape and any phase of the epidemic. Moreover, we modify the epidemiological aspects of the model according to recent literature on COVID-19 (66; 181; 30), to properly consider asymptomatic cases and the delay in isolating individuals after they are identified as infected. We consider different values of R_0 , reduced with respect to the one assigned to the free pandemic, to take into account the widely implemented additional containment strategies, e.g., physical distancing and masks wearing (Appendix 7.1.2). Second, we provide a realistic quantification of the tracing ability ε_T by performing simulations of contact tracing strategies on real-world data sets collected across different social settings (i.e., a university campus, a workplace, a high school) (157; 57; 121). Hence, the tracing ability ε_T , defined by Ferretti *et al.*(48) as a free parameter, becomes here an empirically estimated quantity, which directly depends on the contact network. The impact

of the tracing procedure on the spread can then be evaluated by inserting ε_T into the mathematical model. Third, we assume that the probability of a contagion event occurring during an interaction between a susceptible and an infected individual also depends on the duration and on the degree of proximity of the contact (148; 161) (along with other epidemiological variables such as the infectiousness of the individual). This can be simulated on real contact data sets, in particular on the Copenhagen Networks Study (CNS) data set (157) that provides proximity information, via the strength of Bluetooth radio packets exchanged between their smartphones. Finally, we investigate in detail the contact tracing procedure, designing appropriate policies in terms of the definition of the most risky contacts. We thus implement a system where tracing does not necessarily imply a massive preventive quarantine of the population. We define duration and proximity thresholds to discriminate between “risky” contacts and contacts that instead correspond to a low contagion probability. Note that, as contagion events are stochastic in nature, not all contacts that we consider at risk lead to infection events. This leads to “false positives”, i.e., non-infected individuals who will be quarantined. Similarly, among the contacts considered as “non-risky” by the contact tracing, some might actually have led to a contagion event (“false negatives”). Quantifying these outcomes represents crucial information to calibrate the policies for contact tracing apps. Quarantine too few and omit many potential spreaders. Quarantine too many and incur unnecessarily high social cost.

Overall, our approach allows to evaluate the effect of different contact tracing policies, not only on the disease spread but also in terms of their impact on the population, as quantified by the fraction of quarantined individuals.

3.1 Methods

In this section we introduce our model for contact tracing. The tracing procedure allows to identify individuals who are considered to be at the highest infection risk and quarantine them without necessarily isolating a large fraction of the population. This allows for devising ad-hoc strategies to control the epidemic.

3.1.1 A modeling framework for digital contact tracing on empirical contact networks

We consider a population within which a virus is spreading, and the spread is determined by the contacts between individuals. As we do not consider geography nor large-scale mobility, our modeling can be considered as referring to a limited geographical area or community, similar to previous modeling efforts (48; 93). The spreading process is designed in order to mimic the COVID-19 epidemic, thus characterized by values of R_0 , viral load and fraction of asymptomatic individuals that are typical of SARS-CoV-2. We assume that two types of non-pharmaceutical interventions are at play: isolation and contact tracing. Infected individuals are isolated when they self-report as symptomatic or if they are identified through randomized testing. Isolated individuals do not have contacts with other individuals, thus can not infect anyone else once they have been identified. In other words, they are removed from the system. Individuals who have had a potentially contagious contact with identified infected individuals are traced and can be warned through a privacy preserving app on their smartphone (170), and they quarantine preemptively.

The only difference between isolation and quarantine is that the latter is only precautionary: if quarantined individuals show symptoms before the end of quarantine they immediately become isolated and their past contacts (before quarantine) are traced, otherwise they are released at the end of the quarantine.

A natural baseline for the work we present here is the model by Fraser *et al.* (52), recently adapted to the COVID-19 case in Ferretti *et al.*(48). The mathematical model is based on recursive equations designed to quantify the number of newly infected individuals at time intervals, given a characterization of the disease in terms of infectiousness and manifestation of symptoms. The model is designed to consider the two interventions described above, whose effectiveness are quantified by two parameters $\varepsilon_I, \varepsilon_T$ varying from 0 to 1, where $\varepsilon_I = 0$ means “no isolation” and $\varepsilon_I = 1$ represents a perfectly successful identification and isolation of all infected individuals; analogously, ε_T quantifies the efficacy of contact tracing.

Here we use this model as a stepping stone in order to define a more general approach. The generalization of the equations of Fraser *et al.* (52) is derived in detail in the Appendix and resolves an important limitation. Indeed, it identifies a solution at finite time t , while the original model only shows the asymptotic behavior, for t going to infinity. The equation models the number $\Lambda(t, \tau)$ of people who are infected at time t by people that have been in turn infected for a time $\tau \leq t$. In the equation, R_0 is the reproductive number of the

disease, $\omega(\tau)$ is the infectiousness of individuals at time τ after being infected, and $s(\tau)$ is the probability of symptom onset at time τ after infection. The details of each of these quantities are discussed in Appendix 7.1.1. The equation reads

$$\Lambda(t, \tau) = R_0 \omega(\tau) (1 - \varepsilon_I s(\tau)) \int_0^{t-\tau} \left(1 - \varepsilon_T \frac{s(\rho + \tau) - s(\rho)}{1 - s(\rho)} \right) \Lambda(t - \tau, \rho) d\rho, \quad (3.1)$$

where the integration variable ρ spans the time range between 0 and $t - \tau$, meaning that the contagion at time t from people infected at time $t - \tau$ is in turn affected by contagion at time ρ before $t - \tau$.

For $\varepsilon_I = 0$ and $\varepsilon_T = 0$ we obtain a free spreading without control. The quantity of interest, which can be derived by numerically solving the above equations, is the incidence $\lambda(t) := \int_0^t \Lambda(t, \tau) d\tau$ of newly infected individuals at time t . We use the model to predict the evolution of $\lambda(t)$ up to time $t = 50$ days, which is sufficient for the numerical solutions to reach a stationary growth or decline regime (constant growth or decline rate of $\lambda(t)$), and we consider the average growth or decline in the last 10 days as an indicator of the long-term behavior of the epidemic. A negative number indicates that the epidemic is declining, while a positive one corresponds to growth (uncontained epidemic).

An important feature of the model is given by the probability $s(\tau)$. The ideal case in which all infected individuals can eventually be identified because they exhibit symptoms ($s(\tau)$ approaching 1 for large times) is reported in Fig. 3.1a: this represents the best case scenario, considered in the previous studies of this model (52; 48). Next, we assume instead that 40% of infected individuals are asymptomatic (96; 141; 48; 135; 137) and that only symptomatic individuals can be identified: no randomized testing is performed. This represents our worst case scenario. We represent the presence of asymptomatic individuals by considering that the probability of an infected individual to display symptoms is a growing function of time, which however never reaches 1. In this case, the model predicts epidemic containment for the upper half of the range of values of the parameters ε_I and ε_T (Fig. 3.1c).

In the following, we assume an alternative scenario where 50% of the asymptomatic individuals are identified by a policy of randomized testing (38). These, added to the symptomatic individuals, result in a detection of 80% of the total infected cases. We remark that this scenario is equivalent to assuming that asymptomatic individuals account for only 20% of the infected population (18; 125). Indeed, there is still no agreement in the scientific community about the fraction of asymptomatic infections for COVID-19, and different possible scenarios should be considered (38) (Appendix 7.1.1). This is our baseline for the following investigations and the resulting model predictions are plotted in Fig. 3.1b.

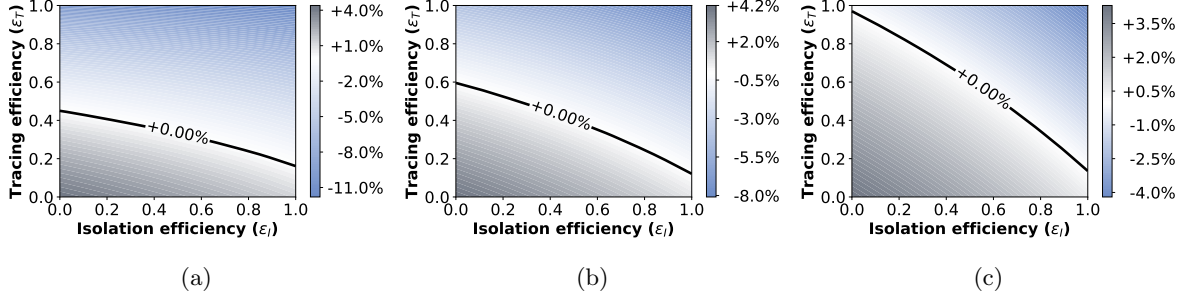


Figure 3.1: **Infection rate scenarios.** Growth or decrease rate of the number of newly infected individuals, assuming either that all the infected people can eventually be identified and isolated (Fig. 3.1a); or that only symptomatic people can be isolated with 20% of asymptomatic infected individuals (Fig. 3.1b); or that only symptomatic people can be isolated with 40% of asymptomatic infected individuals (Fig. 3.1c). Infection rates are reported as a function of the isolation efficiency ϵ_I and the tracing efficiency ϵ_T . In all three settings the cases are reported with a delay of 2 days.

Note also that we take into account in all settings a delay of 2 days between the detection of an infected individual and the time when this person is actually isolated and contact tracing is implemented. A delay of 3 days is considered in Appendix 7.1.2.

The algorithm modelling the spreading and containment of the virus is implemented on the real contact network and coupled with the mathematical model.

This simulation is used in two ways. First, it produces results which are averaged over the network and then aggregated into a quantity, ϵ_T , that can be plugged into the mathematical model. In this step, the network simulation is used as an estimator of a real-world parameter value. We remark in particular that the prediction of the outcome of the policies (epidemic containment or exponential contagion) is obtained solely from the mathematical model, informed with these real-world parameters.

On the other hand, the simulation on the network goes beyond the mathematical model in that it captures complex and non-uniform events and the heterogeneity of individual behaviors. The simulations thus give also access to several fine-grained quantities of interest that provide a complementary view on the epidemic. In particular, we can measure the number and time evolution of false and true positives, offering a quantification of the cost of the quarantine measures.

In the following we detail the implementation of the numerical simulations (Section "3.1.2") and the methods used to extract the aggregated parameters (Section "3.1.3").

3.1.2 Spreading and tracing on the real network

The contact data set is represented as a temporal sequence of undirected and weighted graphs. The nodes of the graphs are the individuals stored via their unique identifiers, and an edge connects two of them if their respective Bluetooth devices have recorded each other. The weight of each edge is the pair of the signal strength and the duration of this contact. These two values are obtained by aggregating the continuous measures of the data set on successive time windows of duration $300s$.

The simulation keeps track of the status of each node, which is updated depending on the spread of the infection (which is a stochastic phenomenon regulated by the infection probability ω_{data}) and on the enforcement of the tracing and isolation policy (which is again stochastic, and dependent on the definition of the policy's thresholds).

The simulation is parametrized by two types of inputs: disease-dependent parameters, which are discussed in Section 3.1.1 and Appendix Table 7.2, and tracing-dependent parameters, which are the isolation efficiency $\varepsilon_I \in [0, 1]$, the memory length of the contact tracing, the duration of the quarantine, and the fraction of app adopters in the population.

Once these parameters are set the algorithm works as follows:

Setup. A fraction of the nodes, extracted uniformly at random, is set to non-adopters, i.e., not using the app. They will contribute to the spread of the virus and they can be isolated, but their contacts cannot be traced and they cannot be quarantined. Observe that we make the simplifying assumption that the app influences only the quarantining of individuals, but not the isolation policy. Namely, we assume to be able to detect and thus isolate an infected individual independently of the app, while we are able to trace the contacts only between pairs of app adopters.

Initialization. A randomly extracted node from the first graph of the sequence is set to infected. It is assigned a time since infection chosen uniformly at random in $[0, 10]$ days.

Time evolution. For each temporal step the following steps are repeated:

Update contacts. The list of contacts of each app adopter node is updated by adding the contacts of other app adopters at the current time, if they fall within the policy's thresholds. Each list stores the contacts for a fixed maximum number of days (which is set to 7 days in the main simulations).

Update quarantined. The list of quarantined nodes is scanned. Nodes who completed the quarantine time (10 days in the main simulations) are just removed

from the list if healthy, or removed and added to the list of isolated if they developed symptoms.

Update infected. The list of infected nodes is scanned. Those who became symptomatic or are tested positive, depending on the probability $\text{onset_time}(\cdot)$ (see Section "3.1.1" and Appendix Table 7.2) are added to the list of infected identified by the health authority. Then, the list of identified infected is scanned, and each of its nodes is isolated with a probability ε_I . For each successfully isolated node who is an app adopter, the tracing policy is enforced on its contacts, i.e. all the nodes registered as contacts are quarantined. All the other infected nodes instead can spread the infection: each of their neighbors is infected independently with a probability modeled by ω_{data} (Appendix 7.1.1).

Check quarantined. The list of quarantined is scanned again to find symptomatic nodes. If a symptomatic node is found, it is isolated and the tracing policy is enforced on its contacts who are app adopters.

Observe that the contacts taken into account for the contact tracing are defined according to a given policy's thresholds (distance and duration), i.e. only those interactions with sufficient duration and small enough distance are stored in the contact lists. However, the spreading process can a priori occur between an infected node and any of its neighbors, the probability of a contagion event being given by ω_{data} .

Moreover, the simulation assumes that each individual that is required to quarantine is willing to do so. We consider in Appendix 7.1.2 the situation where individuals have a decreasing acceptance to comply, based on the number of times that they are asked to quarantine.

On the other hand, the compliance to isolation is already modelled by the user-defined parameter ε_I , which represents the effective fraction of identified infected who successfully isolate, where the value of this fraction may depend on the health system capacity, but also on the nodes' compliance and possibility to isolate.

3.1.3 Aggregation and parameter estimation

During the simulation, whenever the tracing and quarantine policy is enforced a quarantine error e_T is computed to score its success. This value is defined for each isolated node as the ratio between the number of its secondary infections (i.e. the nodes that it infected) that did not quarantine, and the total number of its secondary infections.

The list of values e_T (one for each isolated individual) is collected and averaged over the entire simulation to obtain a mean score $\langle e_T \rangle$. This value encodes the contributions of the chosen policy, of the adoption rate, of the duration of the memory of contacts and in general of the heterogeneity of the network dynamics.

This allows to assign to each policy a tracing efficiency ε_T observed over the simulation as a function of its inputs and of the network dynamics. We define it as the product of two independent factors modelling the efficiency of the isolation (individuals who are not isolated are automatically excluded from the contact tracing, so their contacts do not quarantine) and the effect of the quarantine error, as:

$$\varepsilon_T = \varepsilon_I (1 - \langle e_T \rangle). \quad (3.2)$$

A perfect efficiency of the tracing policy ($\varepsilon_T = 1$) is possible only under perfect isolation ($\varepsilon_I = 1$) and zero quarantine error ($\langle e_T \rangle = 0$).

Considering ε_I as a free parameter allows us to explore different scenarios, thus providing a full range of predictions. This choice accounts for the fact that in a realistic scenario the ability to identify and consequently isolate an infected individual is set by the number of tests that are implemented and by their accuracy, features whose identification is out of the scope of this work. We mention that the adoption of an app might have a positive effect on this quantity if the possibility of self-reporting when symptoms appear is implemented in the device.

3.2 Results

3.2.1 Tracing efficiency based on empirical contact data

The proposed mathematical framework makes it possible to address our main goal: characterizing the efficiency of contact tracing. This can be quantified by ε_T , which instead of being a free parameter can be estimated numerically, by observing how well the implemented policies enable to find the infected individuals. More precisely, we assume that a fraction ε_I of infected individuals is identified at each time step. Their recent contacts are then traced and, according to the nature of their interaction, as we explain in detail in the next sections, some of them will be classified as “at risk and thus possibly contagious”. Tracing is therefore strongly dependent on the ability to identify those primary infected individuals that caused the secondary infections, and we thus assume that ε_T is

proportional to ε_I . Moreover, it is influenced by the actual ability to find the secondary cases, given the primary infected. This in turn depends on multiple factors, involving the spreading model, the definition of a risky contact, the app adoption, the compliance to quarantine and clearly the quantity and nature of contacts in the population. For this reason we need a numerical model that takes into account all these factors and simulates the spreading, with isolation and tracing, in a population of individuals with realistic contacts. To this end, we make use of three different data sets of empirical contacts involving large groups of people, in a high school, in a university campus and in an office building. The variable ε_T will be computed by counting, for each primary infection, the fraction of the corresponding secondary cases that are actually quarantined according to some contact tracing strategy, see Section "3.1.3" for the details on the derivation of ε_T .

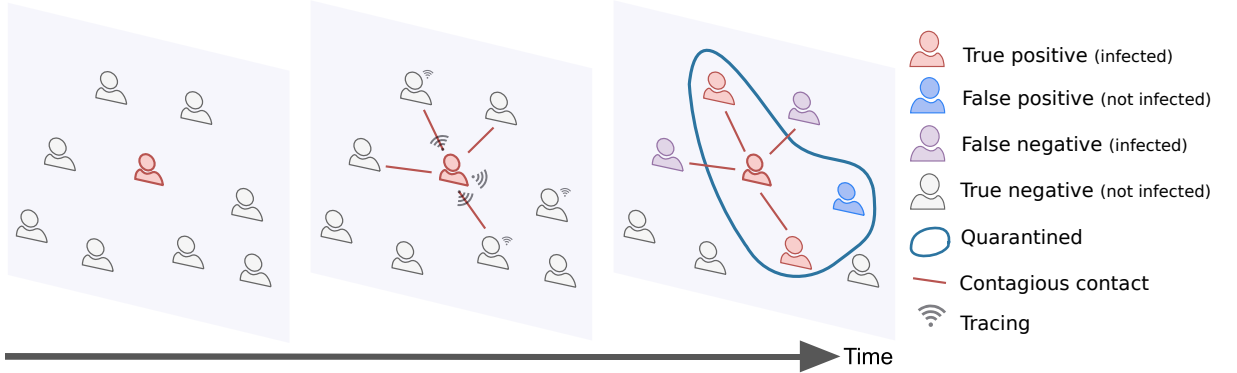


Figure 3.2: **Contagion, tracing and quarantines.** The contacts among users of the contact tracing app are registered via the app. When individuals are identified as infected they are isolated, and the tracing and quarantine policy is implemented. Depending on the policy design, the number of false positives and false negatives may vary significantly.

The data that we use have been collected using wearable devices in different populations of individuals and contain time-resolved information on their pairwise close-range proximity interactions. In each case, we simulate an epidemic spread starting from a single random individual. The epidemic propagates from person to person via their interactions and we assume that the recent contacts of each individual are stored in their mobile phones. Each infected individual has a probability of being identified equal to ε_I . When this happens, all the identified people are isolated, i.e. removed from the simulation, and their recent stored contacts are automatically traced (i.e., warned by the app). In order to avoid quarantining a large portion of the population we define specific criteria to determine which contacts are at risk, and only the corresponding individuals go into quarantine. As the definition of risky contacts is made a priori, and as infection events occur stochastically, quarantines will not only concern individuals who have been infected, but also some who have been in

contact but were not infected (false positives), while some other individuals who have been infected although their contact was not considered at risk, will not receive any warning by the app and thus remain outside quarantines (false negatives). Note also that individuals who did not adopt the app cannot be notified nor quarantined, and contribute both to the true and false negatives. This is schematically explained in Fig. 3.2.

Different policies to define the risky contacts will be delineated in Section "3.2.3" and their efficiency will be quantified by not only observing their ability in controlling the epidemic but also by their efficiency in minimizing the number of false positives, i.e. unnecessary quarantines.

In the following we will mainly rely, for the numerical evaluations of tracing, on the Copenhagen Networks Study (CNS) data set (157). These data describe the interactions of 706 students, as registered by the exchange of Bluetooth radio packets between smartphones, for a period of one month. From the complete data set we extract the proximity measures in the form of Bluetooth signal strength. We therefore have access to two important properties of contacts: their duration and the proximity of the two individuals at the time of the interaction. We are hence able to refine the spreading model by including the dependence on these variables too, as explained in the next section. Moreover, the risk assessment in the tracing procedure will be based on contacts proximity and duration thresholds, corresponding to different policies which will be discussed in Section "3.2.3".

In the Supplementary Information we also show simulations performed using two other data sets collected by the SocioPatterns collaboration in two environments: a high school (121) and an office building (57).

It is important to emphasize that these simulations are specifically used to evaluate the impact of isolation and tracing in different contexts and under different policies and to extract the resulting values of isolation and tracing efficiencies. On the other hand, the epidemic model we use to understand which policies are efficient is the theoretical one described by Equation 3.1 and is thus not restricted to any specific setting.

3.2.2 How infectiousness depends on duration and proximity

In the theoretical model (3.1), infectiousness is simply given by the curve $\omega(\tau)$ multiplied by R_0 ; on the other hand, as stated above, the numerical simulations make it possible to take into account several crucial factors, like duration and proximity of contacts.

We thus multiply $\omega(\tau)$ by two independent factors, $\omega_{\text{exposure}}(e)$ and $\omega_{\text{dist}}(s_s)$. They repre-

sent the probability for an infected individual to transmit the disease respectively given the duration e of a contact and given the signal strength s_s of a contact. Here, the Bluetooth received signal strength can be considered as a proxy for the distance between two individuals, where signal attenuations (in dBm) with smaller absolute value tend to correspond to smaller distances (160). We refer to Appendix 7.1.1. for a detailed discussion on the functional shapes of $\omega_{\text{exposure}}(e)$ and $\omega_{\text{dist}}(s_s)$. In particular, as both are parametric functions, it is possible to tune their parameters by imposing some physical constraints regarding duration, distance and R_0 . The reproductive number of COVID-19 can be extracted from the literature as being close to $R_0 = 3$ (30), while there is little evidence for the dependence on proximity and duration; we thus consider multiple possible infection curves corresponding to different combinations of $\omega_{\text{exposure}}(e)$ and $\omega_{\text{dist}}(s_s)$, keeping $R_0 = 3$ fixed. To this aim, we elaborate a procedure aimed at choosing the function parameters starting from physical constraints so as to always consider meaningful infectiousness curves. The procedure is explained in detail in Appendix 7.1.1, where we characterize three different possible curves. The constraint given by R_0 requires to find a good balance between the two functions $\omega_{\text{exposure}}(e)$ and $\omega_{\text{dist}}(s_s)$. If for instance we suppose that infectiousness is high even at long distances we should thus set ω_{exposure} such that contacts are contagious only for long durations in order not to have a huge R_0 . (e.g., the pink curves in Figure 7.1 in the Appendix). Vice versa, if ω_{dist} is adjusted such that only close proximity contacts are contagious, we should give more importance to duration and suppose that also short durations are at risk (e.g., the blue curves in Appendix Figure 7.1). In Appendix 7.1.1, we show the results of simulations in these different cases. We observe that for the controllability of the epidemics, the different types of infectiousness do not lead to significant differences. However, from the point of view of cost versus effectiveness of the restrictive measures, different curves lead to different results. We discuss this point in Appendix 7.1.1.

Here, we choose for definiteness one of the obtained pairs of curves ($\omega_{\text{exposure}}(e)$, $\omega_{\text{dist}}(s_s)$) compatible with $R_0 = 3$, and we assume in the following that infectiousness is governed by these. They correspond to an $\omega_{\text{exposure}}(e)$ which reaches 90% infectiousness after 2 hours of contact, and to an ω_{dist} such that the contagion probability drops by 50% at a distance of 2.5 meters, and by 99% at 7.0 meters.

Finally, in the numerical model we rescale the curves of infectiousness of a factor r_{R_0} , which plays a pivotal role. Indeed, the procedure described above for parameter setting is aimed at reconstructing a scenario without restrictions, where the epidemic of COVID-19 is free to spread and is characterized by a reproductive number equal to 3. However, in this work we analyze the effect of isolation and tracing in the context of reemerging epidemics

where a number of protective measures are in place, such as face masks and physical distancing. Such measures contribute to mitigate the spreading and enter in our model as an overall reduction of R_0 , in a range suggested by recent literature (33; 144; 84; 94; 60). This can be obtained by setting the reduction factor r_{R_0} to specific values, reported in Appendix Table 7.5 in the Appendix.

3.2.3 Design of appropriate policies

	Signal strength threshold T_p (dBm)	Duration threshold T_d (min)	Fraction of CNS contacts
● Policy 1	-73	30	2.2%
● Policy 2	-80	20	7.3%
● Policy 3	-83	15	13.4%
● Policy 4	-87	10	25.9%
● Policy 5	-91	5	56.7%

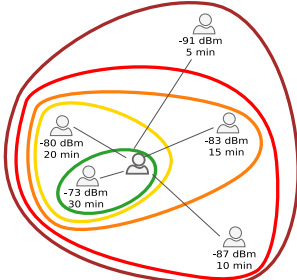


Figure 3.3: **Policies based on distance and duration.** In the left table, the signal strength threshold T_p and the duration threshold T_d defining the policies are reported. Contacts with a duration larger than T_d and signal strength larger than T_p are considered at risk. The last column gives the fraction of the total number of interactions of the CNS data set that they correspond to. A larger value of the magnitude of the signal strength tends to correspond to a larger distance, such that in the second column the thresholds go from the least to the most restrictive policy. The policies are sketched in the left figure.

As mentioned above, the empirical CNS data set provides us with the opportunity to devise policies for tracing in order to avoid a massive preventive quarantine of the population.

We can classify contacts at low and high probability of contagion on the basis of thresholds of duration and proximity: only contacts with duration above a threshold T_d and Bluetooth signal strength above a threshold T_p are considered as at risk and thus stored in the individual's devices (when both individuals in contact have adopted the app). Assuming that the dependence of infectiousness from duration and proximity is unknown, we consider several possible values for the thresholds T_d and T_p , thus defining multiple possible policies, reported in Fig. 3.3, from the least to the most restrictive. We also consider two additional policies in Appendix 7.1.2, corresponding to either close range but short exposure interactions or long range but long exposure interactions.

We remark that the policies implement distance detection directly as a measure of the Received Signal Strength Indicator (RSSI) values, since a precise and reliable conversion

to an actual distance is a notoriously difficult task (160; 130) that would only add a layer of uncertainty to our analysis, without any gain in terms of accuracy. It is in general true that weak signal strengths correspond to large distances between users and vice versa but the link between RSSI and actual distance is affected by multiple factors, from the smartphone brand to the presence of obstacles between devices, and more (160; 130).

In substance, we simulate the epidemic and at the same time implement the contact tracing, supposing that we do not know which individuals are infected. We then compare the set of quarantined individuals with the set of people who have actually been infected in the spreading simulation, and measure the performances of each tracing policy (i.e. of each definition of thresholds T_p and T_d). The performance of a policy is quantified first of all by its ability to find the infected individuals, and consequently by its ability to contain the epidemic according to our mathematical model; in addition, we will measure the efficacy of a policy in quarantining only infected individuals (i.e. in limiting the number of false positives), in order to limit the social and economic damage to society.

Figure 3.4 shows the distributions of RSSI and contact durations of the interactions contained in the CNS data set. Most contacts have short duration and low signal strength, but long lasting durations are also observed, with overall a broad distribution of contact durations as is typical for data on human interactions (27; 160). The thresholds defined by the tracing policies determine the fraction of these contacts that can be traced by the app. Even slight variations in the tracing policy thresholds may strongly influence the capacity to identify the contacts corresponding to the highest risks of infection, as shown in Fig. 3.4 by comparing the RSSI and contact duration distributions with the infectiousness curves.

In line with many privacy preserving contact tracing apps, we additionally assume that each individual device stores the anonymous IDs received from other devices only for a limited time, such that every device does not keep track of all its past contacts but only those of the last n days. This is already implemented in apps used by most countries, applying the privacy-preserving DCT model (170). We assume $n = 7$ days, and we show in the Supplementary Information (Appendix 7.1.2) alternative results for shorter and longer tracing memories.

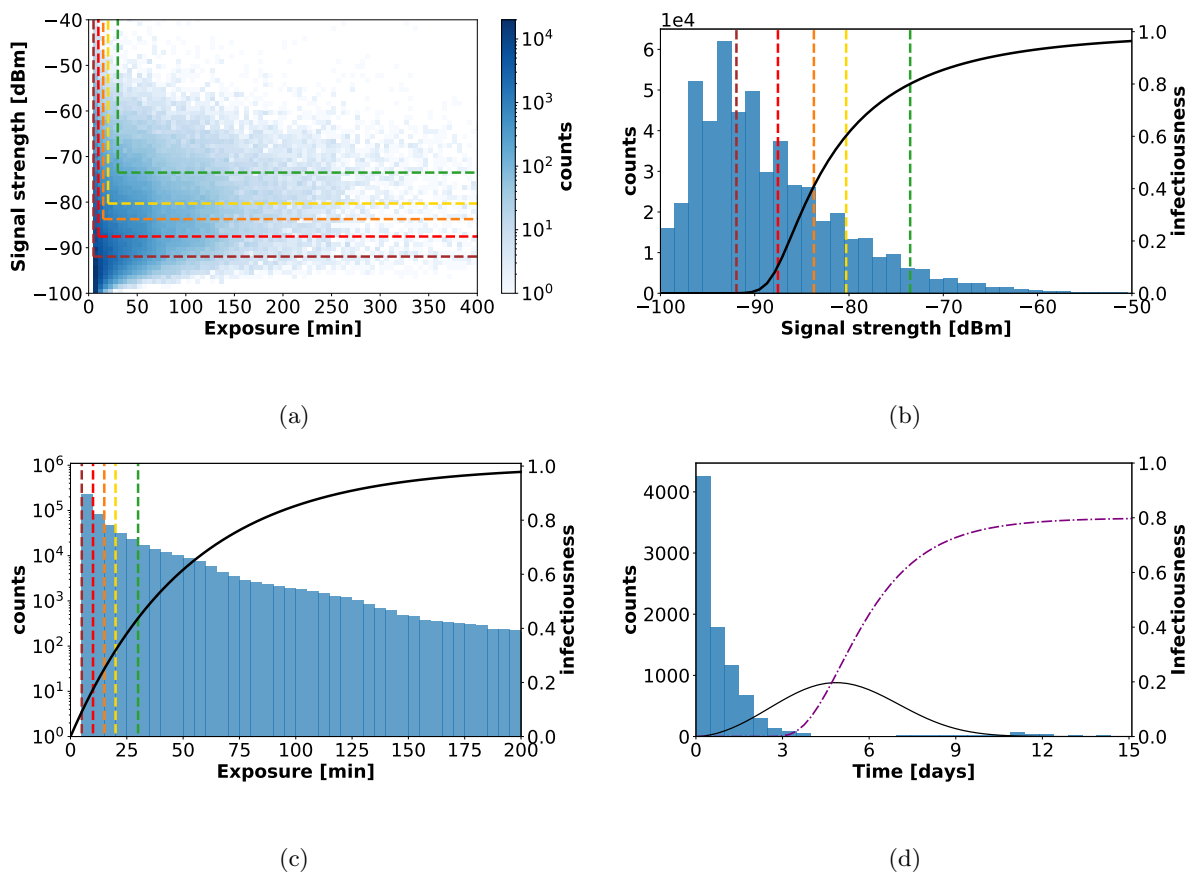


Figure 3.4: **Contacts in CNS data set: signal strength, exposure and inter-contact time.** Fig. 3.4a shows a scatterplot of signal strength vs. duration for all contact events in the CNS data set, and displays the thresholds defining the various policies (T_p for signal strength and T_d for duration): the contacts identified as “at risk” are those situated above and to the right of the dashed colored lines. Fig. 3.4b and Fig. 3.4c separately depict the distributions of signal strength and duration, together with the infectiousness functions ω_{dist} and ω_{exposure} , respectively (black curves) see Appendix 7.1.1. Fig. 3.4d shows the distribution of time elapsed between the infection of an individual and their successive contacts, obtained with $\varepsilon_I = 0.8$ and for Policy 5 in the CNS data set. The black curve shows the normalized infectiousness $\omega(\tau)$ as a function of time, and the purple dashed line is the cumulative probability $s(\tau)$ to identify an infected individual.

3.2.4 Digital tracing enables containment for moderate reproductive numbers

In this section we show the results provided by the combination of numerical simulations on empirical data and the theoretical model. The five policies described in Fig. 3.3 are tested in different scenarios corresponding to different levels of app adoption and different values of R_0 . Only individuals adopting the app participate to contact tracing;

the remaining individuals are outside the reach of the tracing and quarantining policies, but they are still isolated whenever detected because symptomatic or through random testing. We consider as possible levels of app adoption: 20%, 40%, 60%. These levels constitute realistic cases, as the fraction of population that owns a smartphone rarely reaches larger levels (64% for instance for the French population (112; 97)), and a certain level of non-compliance should be also considered (from the point of view of the app, non-compliance or non-adoption can be considered as equivalent). As of mid-October 2020, for example adopters represent 24% of the population in Germany (165), 32% in the U.K. (67), and 20% in Italy (76).

In addition, each policy is tested with the isolation efficiency values $\varepsilon_I = 0.2, 0.5, 0.8, 1$, which encode isolation capacities ranging from rather poor to perfect isolation of any symptomatic or tested positive person.

The results are shown in Fig. 3.5. We observe that if $R_0 = 2$, practically none of the policies is able to stop the spreading, even with high app adoption. However, this pessimistic scenario changes under the hypothesis of $R_0 = 1.5$ (second line of panels in Fig. 3.5), where a larger portion of the phase space implies that the spread can be controlled. An app adoption above 40% is then sufficient to obtain good results: all policies manage to contain the spread for $\varepsilon_I = 0.8$ (except Policy 1 for 40% adoption), and all of them for $\varepsilon_I = 1$. The situation is even better with $R_0 = 1.2$, as all policies are effective as soon as the isolation efficacy is at least 0.5, even in the case of an app adoption of only 20% (bottom left panel in Fig. 3.5).

We notice that the tracing efficiency ε_T varies considerably with different levels of app adoption, but does practically not depend on R_0 . Indeed, ε_T only accounts for the fraction of secondary infections that are correctly traced, independently on the spread of the virus and the amount of infected individuals in the population.

The different scenarios explored above draw a framework where R_0 is limited by implementing several primary containment measures. DCT is added on top of them and its effect is observed as a component of a broader general effort. While in the absence of DCT a value of R_0 larger than one may rapidly lead to a new exponential outbreak and thus to renewed (possibly local) lockdown measures, we have shown here the possible improvement that can be obtained thanks to the deployment of a contact tracing app. The results however highlight that DCT should be accompanied by additional measures and by a sufficient app adoption in order to be effective.

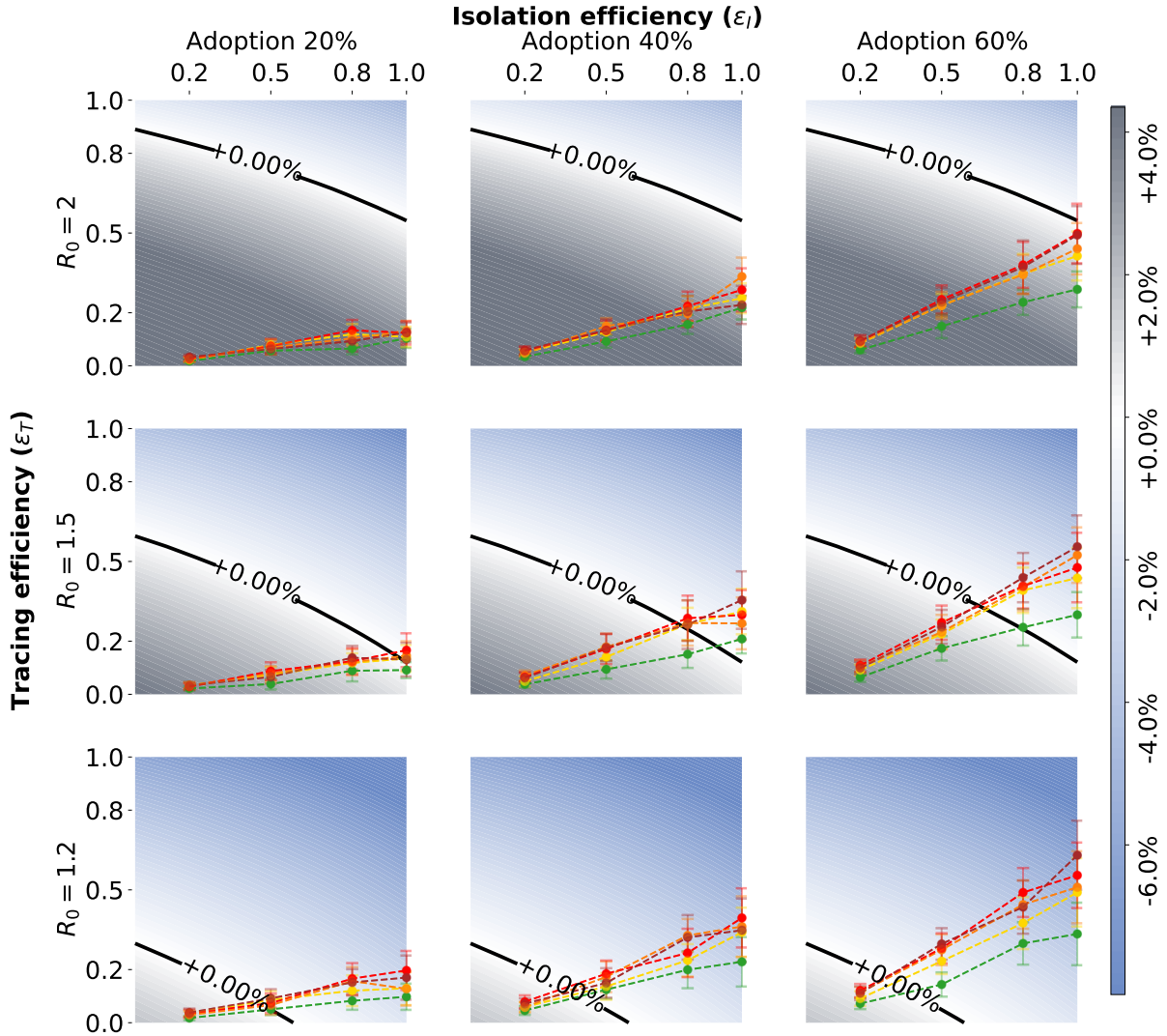


Figure 3.5: **Tracing policy efficiency.** Growth or decrease rate of the number of newly infected individuals assuming that symptomatic individuals can be isolated and that an additional 50% of asymptomatics can be identified via randomized testing. The points correspond to the parameter pairs such that the isolation efficiency ε_I is an input and the tracing efficiency ε_T an output of the simulations on CNS contact data, for the five policies. The different scenarios are defined by an app adoption level of 20%, 40%, or 60% (from left to right), and by a value of the reproductive number R_0 equal to 2, 1.5, or 1.2 (from top to bottom). All the points have been obtained as mean values over $n = 200$ simulations and the error bars represent the standard error.

3.2.5 Any effective containment comes at a cost

Behind the scenes of the results of the previous section, there is a complex dynamic deserving further investigation. Contact tracing produces in some cases the desirable

effect of containing the spread, but side effects emerge as well. Indeed, some of the “at risk” contacts do not actually correspond to a contagion event, while contacts classified as not risky might, as discussed above. It is thus important to quantify the ability of each policy to discriminate between contacts on which the disease actually propagated and the others, in terms of false positives (quarantined individuals who were not infected) and false negatives (non-quarantined infected individuals). To visualize this behavior, we focus on the setting with $R_0 = 1.5$ and $\varepsilon_I = 0.8$, with an app adoption of 40%, since it is representative of a situation in which some policies are effective in containing the spread and others are not (see Fig. 3.5, center). The corresponding time evolution of the average percentages of false negatives and of false positives over the population for each policy are shown in Fig. 3.6.

In terms of epidemic containment, the best policies are those that can rapidly reduce the number of active infected, i.e., of false negatives. In the case of Policy 1, this number remains quite high for the entire simulation time, whereas for all other policies the number of false negatives remains lower. These policies lead overall to a larger value of the tracing effectiveness ε_T (see Section “3.1”), thus leading to a better epidemic containment.

The smaller number of false negatives for the effective policies comes however at the cost of an increased number of false positives, as shown in Fig. 3.6 (top right). In other words, as a policy becomes more effective in tracing actually infected individuals, it also leads to the quarantine of individuals that have not been infected but that had a contact classified as risky by the tracing policy. The maximal number of false positives is very sensitive to the specific policy, contrarily to the number of false negatives. In particular, it appears from the analysis of Section “3.2.4” that Policies 2, 3, 4 and 5 have a similar effectiveness to contain the epidemic and Fig. 3.6 (top left) shows that they yield similar numbers of false negatives, but their undesired side costs are different, as the broader definition of risky contacts produces a larger number of false positives. This highlights once more the importance of the fine-tuning of the chosen policy. Since balancing between these two effects may be non trivial, we plot in Fig. 3.6 (bottom left) the effectiveness vs. cost for each policy, showing that Policy 2 is favorable in that it achieves an almost maximal effectivity (small number of false negatives) at a very low cost (small number of quarantines). The table reports the average percentage of the population that had to quarantine in the simulations (increasing from policy 1 to 5) and the percentage of those were actually infected (decreasing from policy 1 to 5).

To further facilitate the challenge of choosing the right policy, in Appendix 7.1.2 we test the behavior of the model under extended scenarios to precisely quantify the sensitivity of the outcomes with respect to changes of our fundamental assumptions. The model

robustness is assessed by changing the tracing memory (longer and shorter) in Appendix 7.1.2, the reporting delay in Appendix 7.1.2, the ability to trace second order contacts in Appendix 7.1.2, the fraction of asymptomatics infected in Appendix 7.1.2, the adoption of modified policy thresholds in Appendix 7.1.2, and a different response of the population to the request of multiple quarantines in Appendix 7.1.2.

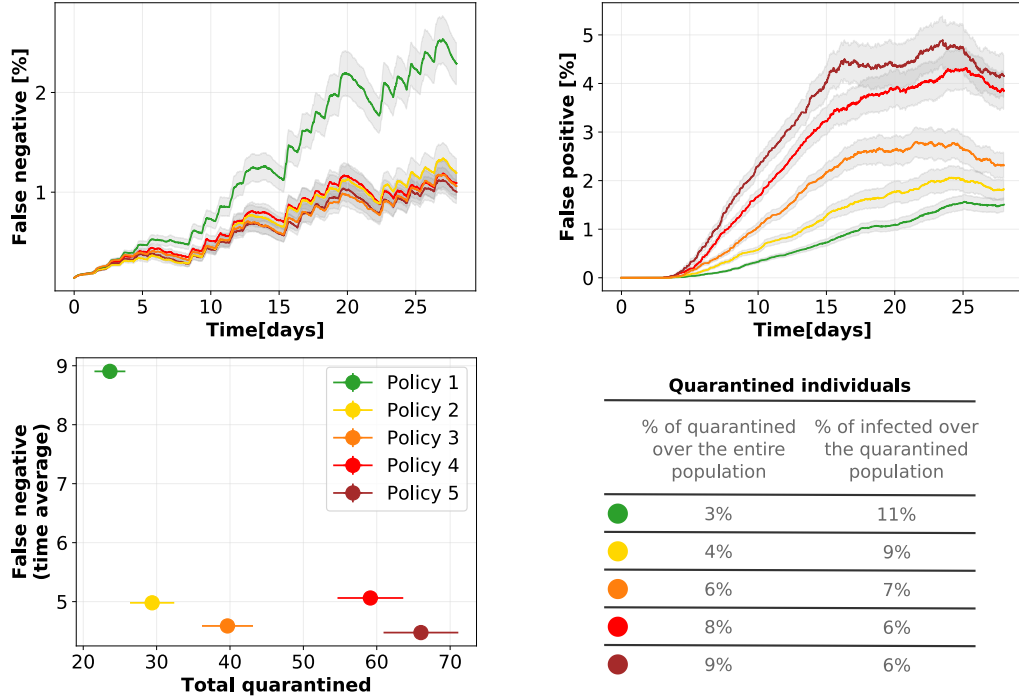


Figure 3.6: **Quarantines, false positives and negatives, with 40% app adoption and $R_0 = 1.5$.** Temporal evolution of percentages of false negatives (Fig. top left), i.e. infected individuals not quarantined, and false positives (Fig. top right), i.e. not infected individuals quarantined, over the population for the five different policies, assuming an isolation efficiency of $\varepsilon_I = 0.8$. The graphs depict the mean and standard error over 200 independent runs. Fig. in the bottom left shows the effectiveness (low number of false negatives) vs. cost (total quarantines) of the policies. The table in the bottom right reports the percentage of distinct individuals who have been quarantined over the entire population and the percentage of them who were actually infected (true positive).

3.3 Discussion

3.3.1 Policies for digital contact tracing: implications and constraints

In the modeling of contact tracing, considering several scenarios of isolation efficiency, app adoption and R_0 values is of foremost importance in order to account for the complex

and heterogeneous issues connected with concrete policy implementations.

These issues should be clear to any policy maker having to decide on containment measures, in order to understand that contact tracing is a viable containment strategy for COVID-19 only in conjunction with complementary policies, as the results of the previous sections show.

These considerations enter our modeling approach in several ways. On the one hand, some parameters are related to the healthcare system capacity and to the socioeconomic condition of the population. These include the isolation efficiency ε_I and the delay in the case reporting, that should account for potential heterogeneities in the access to tests and in the possibility of a person to isolate. This last involves in particular both the access to appropriate spaces and the economic feasibility of a temporary cessation of the working activity. Since each country has a different level of capacity to isolate individuals we considered several levels of ε_I instead of prescribing a fixed setting. The delays in turn depend on factors of different nature such as the delay in reporting, the availability and response of the call centers and of the health authorities, the app- and app-backend-related delays, etc. The analyses reported here take into account a delay of 2 days in isolating infected cases (thus in tracing and quarantining their contacts). This realistic delay does not prevent the proposed policies from keeping the epidemic under control, which is possible under some conditions. However, we observe that a larger delay, even of only one additional day, leads to a completely different scenario (reported in Appendix 7.1.2) where, assuming $R_0 = 1.5$ and 40% app adoption, none of the proposed policies proves able to contain the epidemic, even for maximal isolation efficiency, and despite the higher numbers of quarantines, false positives and false negatives.

Moreover, we have analyzed the effect of the app within epidemic scenarios of limited reproductive numbers ($R_0 = 1.2, 1.5, 2.0$), which are the result of the implementation of complementary policies in addition to DCT. Such measures include traditional manual tracing, mask use and physical distancing.

Our model also includes the level of app adoption as an explicit parameter and we consider 20%, 40%, and 60%. It should be taken into account that factors like the limited access to supported smartphones for different age and income brackets, but also the willingness to adopt the app (strongly dependent on people's trust in DCT and health system), are crucial elements that contribute to these values.

All these parameters should be set with some care. The design of our model allows us to treat them as tunable inputs and in particular no unrealistic or idealized assumption on these parameter needs to be made.

Privacy issues raised by digital tracing are also of great importance, and they have been extensively discussed (80; 131; 99; 98). For these matters we refer to the decentralized models that have been developed such as the Decentralized Privacy-Preserving Proximity Tracing (DP-3T) (170), and to the discussion therein. In particular, we adopt a tracing scheme that does not need to access the complete network of contacts at any time but is based only on decentralized exchange of anonymized keys.

3.3.2 Digital contact tracing: insights and limitations

The general model that we developed for studying the effect of isolation and contact tracing on controlling the COVID-19 epidemic is inspired from the work of Fraser *et al.* (52). The main distinctive characteristics that we have introduced are the following: (i) a general mathematical model that allows to evaluate the evolution of an epidemic in the presence of isolation and DCT at finite time; (ii) the evaluation of tracing efficiency by means of a numerical simulation on real contact data, and no more as an arbitrary parameter of the model; (iii) the dependence of infectiousness on the actual duration and physical proximity of contacts; and (iv) consequently, the design of appropriate policies.

The functional shape of the infectiousness that we devised is composed by three dependencies: the time since primary infection $\omega(\tau)$, the duration of a contact $\omega_{\text{exposure}}(e)$, and its proximity $\omega_{\text{dist}}(s_s)$. The first is originally suggested by Ferretti *et al.* (48), while the other two were introduced in this work. We have shown that the implemented model is robust to changes in all three contributions, see Appendix 7.1.1 and 7.1.1.

Our results suggest that an insufficient app adoption may render any digital tracing effort helpless on its own if the reproductive number is too high. In view of these results, bridging the gap between a realistic app adoption and the larger tracing capability required to contain the disease appears crucial. This goal can only be reached with a joint effort of policy makers and health authorities in organizing an effective manual tracing, and of individual citizens in adopting the app. We therefore tested different levels of app adoption and a range of possible values of R_0 , reduced from its original value by other restrictive measures, like masks wearing and physical distancing.

Moreover, we found that the set of parameters that allow containment of the spread is strongly influenced by the fraction of asymptomatic cases. By first assuming an ideal setting where any pair of parameters $\varepsilon_I, \varepsilon_T$ is possible, we showed (Fig. 3.1) that the area of the phase space representing the setting where it is possible to control the epidemic is reduced when considering 20% or, worst case scenario, 40% of asymptomatic individuals

in the population.

We tested five policies to define risky contacts that should be traced (Fig. 3.3), with different restriction levels. Our results highlight how isolation and tracing come at a price, and allow us to quantify this cost using real data: the policies that are able to contain the pandemic have the drawback that healthy persons are unnecessarily quarantined. In other words, achieving a rapid containment and a low number of false negatives requires accepting a high number of false positives. This stresses the importance of a fine tuning of the tracing and isolation policies, in terms of the definition of what represents a risky contact, to contain the social cost of quarantines. Let us observe that this last could be mitigated by testing the quarantined population and revealing the false negatives, thus translating the social cost in an economical burden due to swabs. Among the tested policies, those that appear to provide the best balance between effectiveness and cost are Policies 2 and 3, corresponding to considering as risky a contact longer than, respectively, 20 and 15 minutes, with distance shorter than, respectively, around 2 and around 3 meters. This is in agreement with the European guidelines for high-risk contacts (45).

We modelled the tracing procedure assuming that contacts are stored in each user's app for 7 days. Such tracing memory seems a good balance between the too short 2 days, which fails in containing the epidemic, and the too long 15 days, expensive in terms of quarantines and not leading to strong improvements in the spread containment (Appendix 7.1.2).

We also included in our model a delay of 2 days in isolating the infected individuals. This delay might however increase when the number of infected cases grows. For this reason we tested a delay of 3 days too, revealing a much worse scenario (Appendix 7.1.2). This highlights the importance of readiness in implementing the testing and isolation procedure, as increased delays might neutralize the beneficial effects of the app.

Another important result concerns the issue of privacy: we numerically tested a second order tracing, where also contacts of contacts of an infected individual are quarantined. Such procedure leads to a strongly enhanced risk in terms of privacy, but we found that it determines a useless massive quarantine while failing to bring any clear beneficial effect on controlling the epidemic (Appendix 7.1.2).

Finally, we tested the possibility that people reduce their compliance if they are notified multiple times and asked to quarantine despite not being infected. This might indeed lead to some mistrust in the DCT procedure and in the healthcare and government institutions. The results that we obtain are very similar to those found with the standard procedure, where the level of compliance is set at the beginning and does not depend on

the multiplicity of quarantines. This further confirms the robustness of our general model and of our results (Appendix 7.1.2).

Our study comes with a number of limitations. First, we have considered data corresponding to a few limited social environments (a university campus, a high school and a workplace) and we cannot provide an overall general study that includes multiple and differentiated contexts and their mutual interplay. Moreover in each data set, only people involved in the experiment have been tracked, neglecting other contacts occurring outside their school, university campus or workplace. Hence, the complete data sets only provide access to part of the interactions of the involved individuals, which is useful to analyze contact tracing in specific environments but does not provide a full picture of a society, e.g. an entire city. This limitation is due to the current lack of larger data sets involving people belonging to different environments, which would represent the general interactions within the population of a city or a larger geographical area. In addition, the implemented policies have been necessarily tailored to the specific CNS data set, depending on the available values of RSSI supported by the used smartphones. Those might differ in actual implementations of DCT apps currently in use in different countries, probably relying on a more advanced technology. Nevertheless, we emphasize that even if we used the simulations performed on these data sets to obtain a realistic quantification of the tracing ability, the controllability of the disease is itself assessed by the general mathematical framework. The results that we present are hence general, not bounded by specific data sets, but only numerically supported by real data to have a realistic implementation of tracing.

Moreover, our study is limited by the current knowledge of the contagion modalities of the SARS-CoV-2 virus, in particular concerning its dependence on physical distance among people and the duration of their contacts. The curve of infectiousness has been designed based on previous contagion studies and on reasonable assumptions (also considering a reduced transmissibility of asymptomatic people). Additional refinements of the transmission dynamic could be obtained by accounting for aerosol transmission, adding a dependence from the environment characteristics, such as being indoors or outdoors, and the presence or not of ventilation. This factor could in principle be modeled by considering information on the (co-)location of the individuals, which is available for some SocioPatterns data sets (57). Should new insights emerge in the way the virus spreads, these could be easily incorporated into our model.

Finally, we model delays in the case reporting and thus in the isolation process, but assume that the quarantine notification of the traced contacts is instantaneous. This is reasonable and it is one of the advantages of relying on DCT, but two factors may introduce a delay:

the app may check for at-risk exposures only 3-4 times a day, and the backend servers that distribute "infected" keys to the app often batches them before notification. The combination of these factors introduces an average delay of several hours (4-5 hours) and a worst-case delay of half a day.

Despite these limitations, the presented model represents an important contribution to the discussion about DCT, proposing a refined approach that allows to investigate a number of features that are unattainable with other recent models.

In conclusion, this combination of a well-established epidemic model with state-of-the-art, empirical interaction data collected via radio-based proximity-sensing methods, allows us to understand the role played by intrinsic limitations of digital tracing efforts, affording a viewpoint on the ambition of achieving containment with digital interventions. Namely, we are able to test and quantify the role that a real contact network plays both for the infectiousness of a contact and for the ability of a policy to detect it and to respond optimally.

3.4 Code Availability

We are pleased to make available the source-code accompanying this research on GitHub https://github.com/DigitalContactTracing/covid_code.

Chapter 4

Egocentric Temporal Motifs Miner

From the previous chapter, it is easy to see how isolating users (i.e. removing temporal edges) in a network directly implies a crucial change in the COVID-19 spreading. This moves our attention to a deeper study of the characterization of temporal networks, through the lens of temporal motifs. We started with the work of Paranjape *et al.* (139) in which they propose a mining strategy that extracts static motifs from the aggregate network and expands them into temporal motifs by considering the order of appearance of edges within a given temporal gap. We further extend the concept of temporal motifs going beyond the traditional point of view. The standard approach is indeed based on observing temporal networks from the outside and decomposing them into their small components. The idea of our approach is instead to jump inside the network and follow the path of a specific node, finding node-dependent spatio-temporal patterns. In particular, for each node, we observe its neighbors and how its connections to them change in a given period. We neglect the connections among neighbors of the chosen "ego" node, and we only focus on studying how the set of neighbors evolves in time, following an *ego perspective*. In social settings this allows to identify the patterns of interactions of individuals, selecting the most relevant behaviors as those which are most repeated in time by the same or different persons. We give to these ones the name of *egocentric temporal motifs (ETM)*.

The ego perspective allows to address the motif identification procedure very efficiently by comparing egocentric temporal sub-networks in terms of their *signature*, simply consisting of a bit vector. This represents a huge simplification with respect to mining standard motifs, which necessarily requires to address the graph isomorphism problem, which slows down the procedure and makes it hard to identify graph motifs with more than a handful of nodes. A graphical summary of our approach is shown in Figure 4.1.

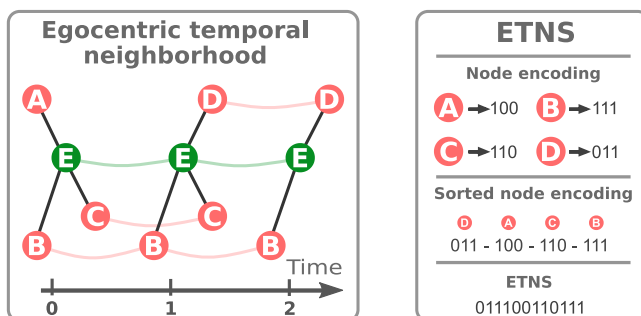


Figure 4.1: Graphical summary of the procedure for extracting egocentric temporal motifs. The left panel shows the *egocentric temporal neighborhood* of the ego node E (in green), with temporal order two and initial time instant zero. Black edges connect the central node with its neighbors (in red) at each time step, while green (resp. red) edges connect consecutive occurrences of the central (resp. a neighboring) node along the time sequence. The right panel shows how the corresponding *egocentric temporal neighborhood signature* (ETNS) is computed. Each neighboring node is encoded into a bit vector indicating the time slots when it is present. The node encodings are lexicographically sorted first and then concatenated to generate the signature.

We conducted an extensive experimental evaluation applying our mining algorithm to a number of diverse interaction datasets. First, we analyzed a set of close proximity interaction networks, including three high schools, a hospital, a research institute, a primary school and a university campus. Qualitative results indicate that, as compared to non-egocentric alternatives, egocentric temporal motifs are more intuitive and representative of the differences between these environments and the categories of the underlying egos. Quantitative results show that a metric based on egocentric temporal motifs is more effective than existing micro-scale, meso-scale and global-scale alternatives in discriminating between different types of graphs. Second, we studied the ability of egocentric temporal motifs to discriminate distance communication networks based on the technology employed (phone calls, sms or emails) and to distinguish different types of synthetic networks (i.e., temporal variants of Erdős-Rényi, scale-free and small-world networks). Results confirm the effectiveness and generality of the egocentric perspective in characterizing a wide range of interactions and highlight the conditions under which this perspective can be limiting.

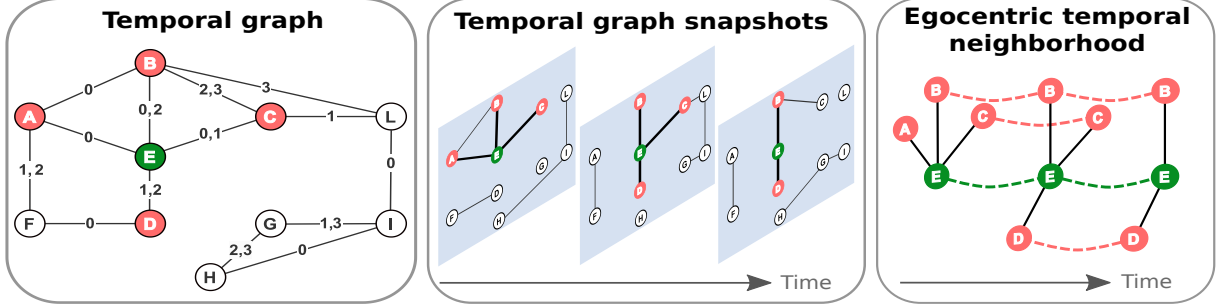


Figure 4.2: The left panel shows a temporal graph G focused on a node E . The middle panel shows three graph snapshots, and the right panel shows a $k = 2$ order ETN for E built from the sequence of its egocentric neighborhoods

4.1 Method

4.1.1 Mining egocentric temporal motifs

Let us start by introducing the notions of *egocentric neighborhood* and *egocentric temporal neighborhood*.

Definition 7 (*Egocentric neighborhood*). Given a (static) graph $G = (V, E)$ and a node $v \in V$, the egocentric neighborhood of v is the subgraph $G(v)$ obtained by taking the neighborhood of v and removing all edges not including v as one of the nodes.

Note that this simple variant of the node neighborhood focuses the attention on the central node, dropping all information not related to it. We next show how to extend this egocentric focus to the temporal aspect, by following the temporal evolution of the node neighborhood.

Definition 8 (*Egocentric temporal neighborhood – ETN*). Given a temporal graph $\mathcal{G} = (V, E)$, a temporal gap Δt , a temporal neighborhood order k and a node $v \in V$, the egocentric temporal neighborhood of v is defined as follows. Let $G_{t_1}, G_{t_2}, \dots, G_{t_m}$ be the sequence of temporal graphs' snapshots for \mathcal{G} with gap Δt . Let $G_{t_1}(v), G_{t_2}(v), \dots, G_{t_m}(v)$ be the sequence of egocentric neighborhoods of v for such temporal graph snapshots. The k -th order egocentric temporal neighborhood of v at time t_i is a graph obtained taking $G_{t_i}(v), \dots, G_{t_{i+k}}(v)$ and connecting each node to the next occurrence of the same node (if any) along the sequence. In addition, each node is labelled with its position in the sequence. We refer to this graph as $G_{t_i}^k(v)$.

Figure 4.2 shows the extraction of an ETN from a temporal graph. The structure of ETN graphs allows to efficiently compute graph isomorphism via a graph signature. To

simplify the presentation of the signature generation algorithm, we assume a function `ID` that applied to a node in an ETN returns its identifier in the original temporal graph \mathcal{G} (the letters in Figure 4.2).

Definition 9 (*Egocentric temporal neighborhood signature (ETNS)*). Given a temporal graph $\mathcal{G} = (V, E)$ and an egocentric temporal neighborhood graph $G_t^k(v)$ for node v , time t and order k , an egocentric temporal neighborhood signature $s_t^k(v)$ is a bit vector encoding $G_t^k(v)$. Two egocentric temporal neighborhoods $G_t^k(v)$ and $G_{t'}^k(v')$ have the same signature if and only if they are isomorphic.

The procedure for computing the ETNS for a given ETN graph is shown in Algorithm 1. The algorithm starts by initializing the signature s to an empty vector and collecting all nodes of the ETN graph with distinct identifiers into a set V . Here $V_{t+i}(v)$ indicates the set of nodes in the $t + i$ temporal slice of $G_t^k(v)$, and the union discards duplicates according to `ID`. For each node u , with the exception of the central node v , the algorithm then computes a bit vector encoding s_u . The encoding has length k and contains at each position i a Boolean flag stating whether the node (represented by its identifier `ID`) is present in the corresponding temporal slice, i.e., $u \in V_{t+i}(v)$. After computing this bit vector, the algorithm appends it to s . Finally, the list of neighborhood node signatures is sorted in lexicographic order and concatenated into the final signature. Figure 4.3 shows some examples of ETN and corresponding ETNS for $k = 2$.

Algorithm 1 Procedure for computing the signature of an ETN graph.

```

procedure COMPUTEETNS( $G_t^k(v)$ )
   $s \leftarrow []$ 
   $V \leftarrow \bigcup_{i=0}^k V_{t+i}(v)$ 
  for  $u \in V$  do
    if  $u \neq v$  then
       $s_u \leftarrow []$ 
      for  $i = 0, \dots, k$  do
        if  $u \in V_{t+i}(v)$  then
          APPEND( $s_u, 1$ )
        else
          APPEND( $s_u, 0$ )
      APPEND( $s, s_u$ )
   $s \leftarrow \text{SORT}(s)$ 
  return FLATTEN( $s$ )

```

Theorem 1 (*Isomorphic ETN*). Given two egocentric temporal neighborhoods $G_t^k(v)$ and $G_{t'}^k(v')$, Algorithm 1 returns the same signature if and only if they are isomorphic.

ETN and ETNS

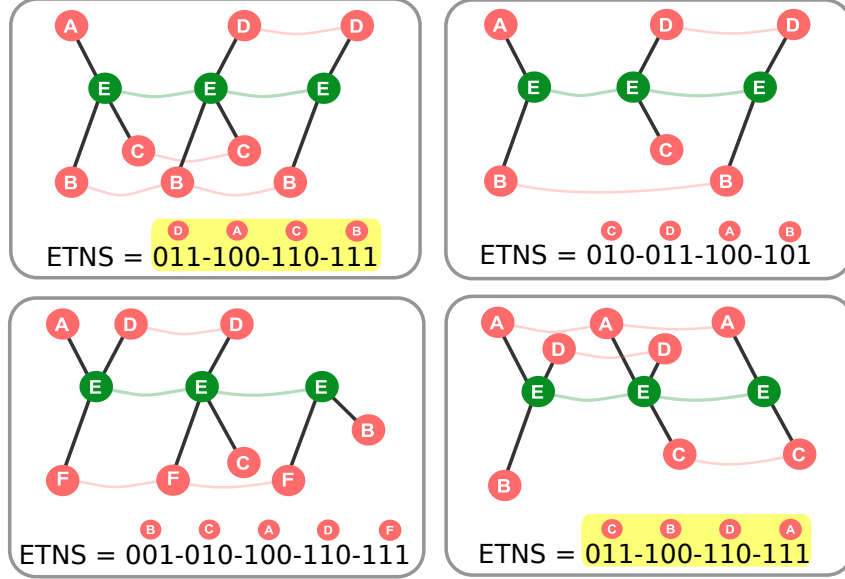


Figure 4.3: Examples of ETN and ETNS for different temporal graphs with $k = 2$. The two highlighted ETNS are identical and correspond to isomorphic ETN.

Proof. We first show that if two ETNs are isomorphic they have the same signature. Let π be a bijection for the two ETNs as from Definition 2. Note that this bijection will map central nodes to central nodes¹ (they are the only ones that can have a degree larger than one on a given temporal slice). By specifying a mapping between nodes, π also implicitly defines a mapping between node identifiers. The edge-preserving property of π implies that the mapping of identifiers is consistent (if two non-central nodes share an edge they have the same identifier). It also implies that the two paired node identifiers share the same set of edges, and thus have the same encoding. Having the same encodings for each pair of node identifiers, the resulting signatures are also the same. This concludes the first part of the proof.

We next show that if the signatures are the same the ETNs are isomorphic. We prove this by showing how to create the bijection function π . Recall that a signature is a flattened sorted list of encodings of node identifiers, and that all encodings have the same length $k + 1$. We start by pairing node identifiers in the two graphs by their positions in the respective signatures. We then map nodes with paired identifiers by matching their node labels (i.e., positions in the underlying graph sequence). Given that the node encodings of the paired identifiers are the same, the corresponding nodes appear in the same positions

¹Apart for the degenerate case consisting of a single neighbor running all along the sequence, where there is no distinction between central node and neighbor and the proof is trivial.

in the underlying sequence (thus matching by node labels produces a perfect match). We repeat the same matching for the only unpaired node identifiers, which correspond to the central node. Note that by definition of ETN the central node appears with all labels from 1 to $k + 1$. Being redundant we omit its encoding from the signature. By construction, mapped nodes share the same label, i.e., $\ell(u) = \ell(\pi(u))$ for all u . Concerning edges, by definition of ETN edges are only between the central node and the neighbors, and between consecutive instances of the same node along the sequence. The former requirement is easily satisfied as each (non-central) node is always connected to the central node having the same label. The latter is satisfied because by construction if two node identifiers have the same encoding their corresponding nodes have the same edges (recall that central nodes have the same encoding even if it is not part of the signature). This concludes the proof. \square

We are now ready to introduce the algorithm for extracting statistics on ETNs from a temporal graph. The pseudocode of the algorithm is shown in Algorithm 2.

Algorithm 2 Procedure for extracting counts of ETN graphs from a temporal graph.

```

procedure COUNTETN( $\mathcal{G}, \Delta t, k$ )
   $\mathcal{S} \leftarrow \emptyset$ 
   $G_{t_1, \dots, t_m} \leftarrow \text{EXTRACTSNAPSHOTS}(\mathcal{G}, \Delta t)$ 
  for  $i = 1, \dots, m - k$  do
    for  $v \in V_{t_i}$  do
       $G_{t_i}^k(v) \leftarrow \text{BUILDETN}(G_{t_i}(v), \dots, G_{t_{i+k}}(v))$ 
       $s_{t_i}^k(v) \leftarrow \text{COMPUTEETNS}(G_{t_i}^k(v))$ 
      if  $s_{t_i}^k(v) \in \mathcal{S}$  then
         $\mathcal{S}[s_{t_i}^k(v)] \leftarrow \mathcal{S}[s_{t_i}^k(v)] + 1$ 
      else
         $\mathcal{S}[s_{t_i}^k(v)] \leftarrow 1$ 
  return  $\mathcal{S}$ 

```

The algorithm takes as input a temporal graph \mathcal{G} , a temporal gap Δt and a temporal neighborhood order k and returns a dictionary of counts \mathcal{S} mapping ETNs to the number of occurrences of the corresponding ETN in \mathcal{G} . It starts by initializing \mathcal{S} to the empty set and extracting the sequence of temporal graph snapshots of \mathcal{G} for gap Δt . For each time t_i and node v (V_{t_i} is the set of nodes of graph G_{t_i}) it builds the corresponding ETN and computes its associated signature using Algorithm 1. The signature is finally used to update the ETN counts in \mathcal{S} . Note that this update step is extremely efficient thanks to the fact that ETNs are bit vectors.

Theorem 2 (*Complexity of COMPUTEETNS*). *The worst-case complexity of COMPUTEETNS*

is $\mathcal{O}(d^{(k)} \log d^{(k)})$, where $d^{(k)}$ is the maximal degree of the network when considering edges within a $k \cdot \Delta t$ temporal range.

Proof. Building the signature requires to create an encoding of length $k + 1$ for each of the nodes in $G_{t_i}^k(v)$ with distinct identifier, which are $|V|$. The complexity is thus $\mathcal{O}(|V|)$. Sorting the signature requires sorting each of these encodings, which costs $\mathcal{O}(|V| \cdot \log |V|)$. The worst case complexity can be obtained setting $|V| = d^{(k)}$, giving $\mathcal{O}(d^{(k)} \log d^{(k)})$. \square

Theorem 3 (*Complexity of COUNTETN*). *The worst-case complexity of COUNTETN is $\mathcal{O}(n \cdot m \cdot d^{(k)} \log d^{(k)})$, where n is the number of nodes in the network, m is the overall number of temporal snapshots, and k and $d^{(k)}$ are as in Theorem 2. The number of temporal snapshots is computed as $m = (T_{end} - T_{start})/\Delta t$, where T_{start} and T_{end} are the smallest t_{start} and the largest t_{end} in the network respectively and Δt is the temporal gap.*

Proof. Note first that the procedure EXTRACTSNAPSHOTS is introduced to simplify the explanation, but the underlying algorithm never explicitly materializes the sequence of temporal graph snapshots for the whole network but directly extracts the ETN using BUILDETN. This latter procedure costs $|G_t^k(v)|$, i.e., the number of nodes in the resulting ETN, which is upper bounded by $d^{(k)} \cdot k$. The procedure is repeated $n \cdot (m - k)$ times. Computing all ETNs thus costs $\mathcal{O}(n \cdot m \cdot d^{(k)} \cdot k)$, and converting them to ETNs costs $\mathcal{O}(n \cdot m \cdot d^{(k)} \log d^{(k)})$. The count update can be done in constant time thanks to the fact that ETNs are bit vectors, so the overall worst-case complexity is $\mathcal{O}(n \cdot m \cdot d^{(k)} \log d^{(k)})$. \square

Note that for reasonable values of k and Δt , $d^{(k)}$ is independent of the size of the network, so that the overall complexity is $\mathcal{O}(n \cdot m)$.

To extract statistically significant ETN from a temporal graph \mathcal{G} , we rely on the support of a null model $\bar{\mathcal{G}}$, defined as follows (72; 73; 78):

Definition 10 (*Temporal Graph Null Model*). *Given a temporal graph \mathcal{G} , consider the temporal graph snapshot $G_{t_1}, G_{t_2}, \dots, G_{t_m}$ (Definition 6) representation of \mathcal{G} . The null model $\bar{\mathcal{G}}$ of \mathcal{G} is obtained by randomly shuffling the snapshots $G_{t_1}, G_{t_2}, \dots, G_{t_m}$.*

Hence a null model $\bar{\mathcal{G}}$ is a temporal graph with the same number of nodes, the same number of snapshots and the same number of connections between each couple of nodes but without any temporal correlation. The procedure can be repeated an arbitrary number of times to produce a set of null models that the original temporal graph can be compared with.

As will be shown in the experimental evaluation, this allows to identify non-trivial temporal structures in a much more selective way with respect to alternative non-egocentric mining approaches.

Finally, we define the *Egocentric Temporal Motifs (ETM)* as follows:

Definition 11 (*Egocentric Temporal Motifs (ETM)*). *Given a temporal graph \mathcal{G} , n null models $\bar{\mathcal{G}}$, and the parameters α (over-representation), β (minimum deviation) and γ (minimum frequency) appearing in Definition 4, the set of ETMs for \mathcal{G} is obtained applying Definition 4 to \mathcal{G} where sub-graphs are represented by the set of its ETNs found according to Definition 8 for each of its nodes.*

We name the algorithm extracting ETM from a temporal graph ETMM, standing for Egocentric Temporal Motif Miner.

4.1.2 ETM-based graph distance

To show the importance of the egocentric perspective in networks of social interactions, we introduce a simple metric that measures the distance between graphs in terms of their respective ETM. To do this, we first define the ETN-based embedding of a temporal graph.

Definition 12 (*ETN-based embedding*). *Given a temporal graph \mathcal{G} and a list M of ETNs, we define $EMB_M(\mathcal{G})$ as the embedding of \mathcal{G} in a vector of cardinality $|M|$, in which the i^{th} element of $EMB_M(\mathcal{G})$ represents the number of occurrences of $M[i]$ in \mathcal{G} .*

Given a list of ETN, the distance between two temporal graphs is then defined as the distance between their respective ETN-based embeddings.

Definition 13 (*ETN-based distance*). *Given two temporal graphs $\mathcal{G}_1, \mathcal{G}_2$ and a list M of ETNs, we define $dist_M(\mathcal{G}_1, \mathcal{G}_2)$ as the cosine distance between the ETN-based embeddings of \mathcal{G}_1 and \mathcal{G}_2 :*

$$dist_M(\mathcal{G}_1, \mathcal{G}_2) = 1 - \frac{EMB_M(\mathcal{G}_1) \cdot EMB_M(\mathcal{G}_2)}{\|EMB_M(\mathcal{G}_1)\| \|EMB_M(\mathcal{G}_2)\|} \quad (4.1)$$

where \cdot is the dot product and $\|\cdot\|$ is the Euclidean norm.

The distance between two temporal graphs can now be computed by first extracting their respective lists of ETM, finding the set of ETM shared by the two graphs and computing their ETN-based distance using this set.

Definition 14 (*ETM-based distance*). Given two temporal graphs $\mathcal{G}_1, \mathcal{G}_2$, two corresponding sets of n null models $\bar{\mathcal{G}}_1$ and $\bar{\mathcal{G}}_2$ and three parameters α, β and γ , we define $\text{dist}(\mathcal{G}_1, \mathcal{G}_2)$ as:

$$\text{dist}(\mathcal{G}_1, \mathcal{G}_2) = \text{dist}_{M_{1,2}}(\mathcal{G}_1, \mathcal{G}_2) \quad (4.2)$$

where $M_{1,2} = M_1 \cap M_2$ and M_1 (resp. M_2) is the list of ETM obtained applying Definition 11 to \mathcal{G}_1 (resp. \mathcal{G}_2).

4.1.3 Experimental setup

In the following, we describe the different groups of network datasets we employed in our experiments and the non-egocentric miners and graph distances that we used as competitors.

Close proximity interaction datasets

The first group of datasets focuses on close proximity interactions and contains three *high school* datasets, a *workplace*, a *hospital*, a *primary school* and a university campus (*DTU*). For more details on those networks see chapter Background 2.

Distance communication datasets

The second group of datasets contains distance interactions with different communication technologies, namely phone calls, SMSs and emails. The idea is to check whether ETMs are capable of distinguishing graphs according to the underlying communication technology. The datasets are explained in the Background chapter 2.

Synthetic datasets

The last group of datasets consists in synthetic temporal networks, and aims at checking whether ETMs retain information concerning (temporal variants of) popular network topologies. Each network is built as a temporal graph where the first timestamp is a static synthetic network suitably generated, while the following temporal layers are recursively generated imposing a fixed correlation with the previous ones. In details, the timestamp $n+1$ is obtained by randomly swapping a fixed fraction f of couples of edges present in the network at timestamp n . In this way the temporal network that we obtain is characterized by a realistic temporal correlation between timestamps and each static network has the

same degree distribution. We chose $f = 0.3$ and we used as initial static networks six different graphs: two Erdős-Rényi (44) networks (with $p = 0.01$ and $p = 0.001$), two scale-free networks (12) (with the same parameters $\alpha = 0.41$, $\beta = 0.54$, $\gamma = 0.05$, $\delta_{in} = 0.2$, $\delta_{out} = 0$ of the algorithm described in (24) but with two different random seeds), and two small-world networks (178) (with $p = 2$ and $p = 8$, but $k = 3$ for both).

Table 4.1 shows the parameters of the generated graphs.

Name	# nodes	# edges	Time stamps
Erdos Renyi (p=0.01)	64	1610	301
Erdos Renyi (p=0.001)	13	78	301
Scale Free (G1)	100	2748	301
Scale Free (G2)	100	3524	301
Small World (p=2)	100	4581	301
Small World (p=8)	100	4340	301

Table 4.1: Synthetic temporal graphs parameters

Non-egocentric miners

As previously stated, no alternative approaches exist that focus on mining egocentric temporal motifs. However, to provide some comparative evaluation for the results of our mining algorithm, we also ran the state of the art non-egocentric temporal motif mining algorithm by Paranjape *et al.* (139). Note however that the motifs found by this method are prototypical of what any non-egocentric mining approach can produce. As mentioned in the related work, the method can be described as follows: (i) obtain the aggregate graph of the input temporal graph (see Definition 5); (ii) extract (static) n -node l -edges motifs, where n is the number of nodes in the motif and l is the number of edges (parameters of the algorithm), using standard approaches for determining motifs (where the null models have the same aggregate degree distribution of the input graph); and (iii) for each static motif count its isomorphic sub-graphs on the temporal network, i.e. with edges possibly appearing at different times. If the maximum distance in time among the different edges is less than a given time δ , the sub-graph is denoted as a temporal motif. In the following, we refer to the Paranjape *et al.* (139) method as TMM.

Non-egocentric graph distances

In this subsection, we present four distances based on micro-scale, meso-scale and global features of the temporal graph.

NetSimile: Berlingerio *et al.* (17) developed NetSimile, a tool for network distance. This method relies on a set of seven features of the network’s nodes. Such features are: degree of the nodes, clustering coefficient, average number of nodes in two-hop neighborhood, average clustering coefficients of the neighbors of a node, number of edges in the node egonet (induced sub-graph of node and neighbors), number of outgoing edges and number of neighbors of the ego. First, the median, mean, standard deviation, skewness, and kurtosis are computed for each feature, producing a graph embedding of $7 \times 5 = 35$ elements. Then, the distance among graphs is computed as the Canberra distance between their respective embeddings.

To apply such method to the aforementioned datasets, we compute the aggregated network, that is, the network obtained by removing the temporal dimension in the input data and the duplicated edges.

Modified NetSimile: NetSimile is not originally conceived for temporal graphs. We thus considered a variant of the method that includes the number of temporal interactions of a node as an additional feature over which to compute the statistics, thus producing an embedding of dimension 40.

Weighted Laplacian: While previous distances rely on local features of the input graph, the Weighted Laplacian leverages global features. First of all, a weighted aggregated static graph is created, in which the weights on an edge represent the number of interactions (over time) that the edge has had. Then the Laplacian matrix is defined as $L = D - W$, where D is the degree matrix and W is the matrix of edge weights.

To compute the distance among two temporal graphs \mathcal{G}_1 and \mathcal{G}_2 , we calculate the Laplacian matrices L_1 and L_2 , then we set k equal to the minimum number of nodes between \mathcal{G}_1 and \mathcal{G}_2 , and finally we compute the Euclidean distance between the first k eigenvalues of L_1 and L_2 .

Temporal motifs: To compute the distance between networks using meso-scale features, we considered a distance induced by (non-egocentric) temporal motifs. This is achieved by applying a variant of Definition 14 that uses temporal motifs as discussed in Section 4.1.3 in place of ETM.

4.2 Results

We start by showing qualitative results in which we compare egocentric and non-egocentric motifs, and then report a quantitative analysis of the effectiveness of our ETM-based graph distance as compared to alternative non-egocentric graph distance measures.

4.2.1 Egocentric vs non-egocentric temporal motifs

We compare motifs found by our ETMM with those generated by TMM. We set the number of temporal steps $k = 2$ for ETMM, while for TMM we consider 3-nodes and 3-edges structures. These values allow to generate non-trivial motifs and to find a significant amount of them in each dataset. As will be clear in the next, the difference between the methods is evident and does not depend on the specific choice for these parameters. Note that ETMM does not require to set the number of nodes and edges and it can in principle extract motifs with an arbitrary number of neighbors. Following Milo *et al.* (124) we set the number of null models $n = 100$, with parameters $\alpha = 0.01$, $\beta = 0.1$ and $\gamma = 5$.

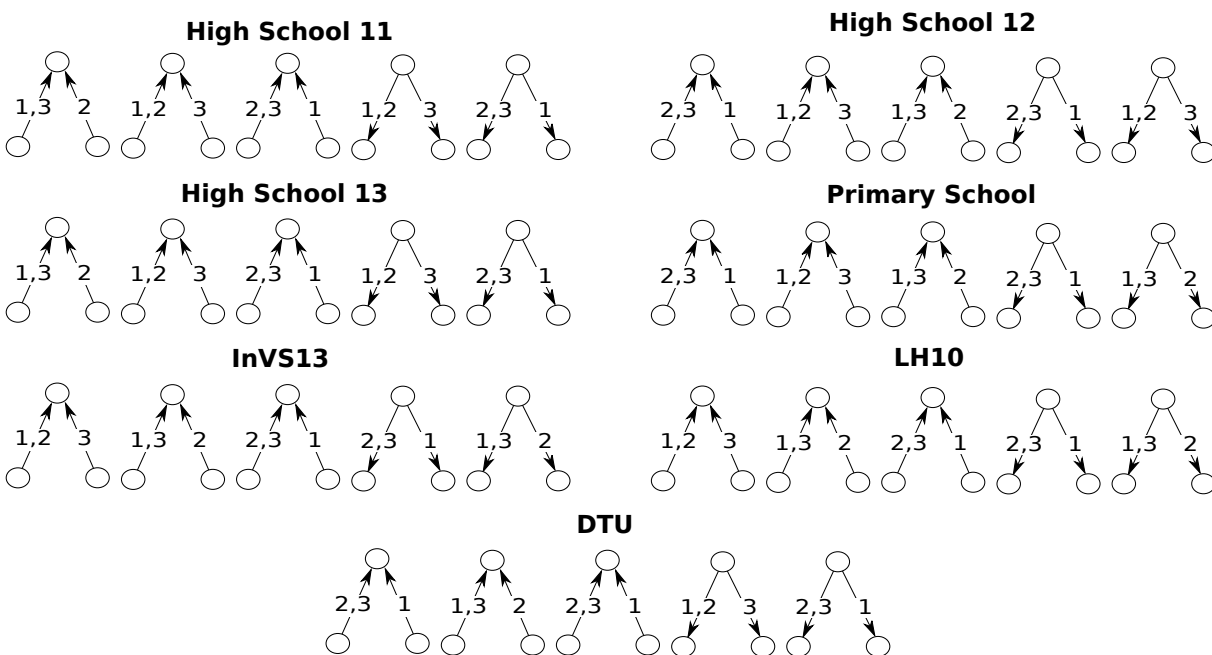


Figure 4.4: Most frequent temporal motifs discovered by TMM on the seven networks for $\Delta t = 300$.

To give an insight of the main differences between egocentric and non-egocentric motifs and highlight the usefulness of the former in discovering patterns of social interaction, we report the five most frequent motifs found by the different methods. We focus on a

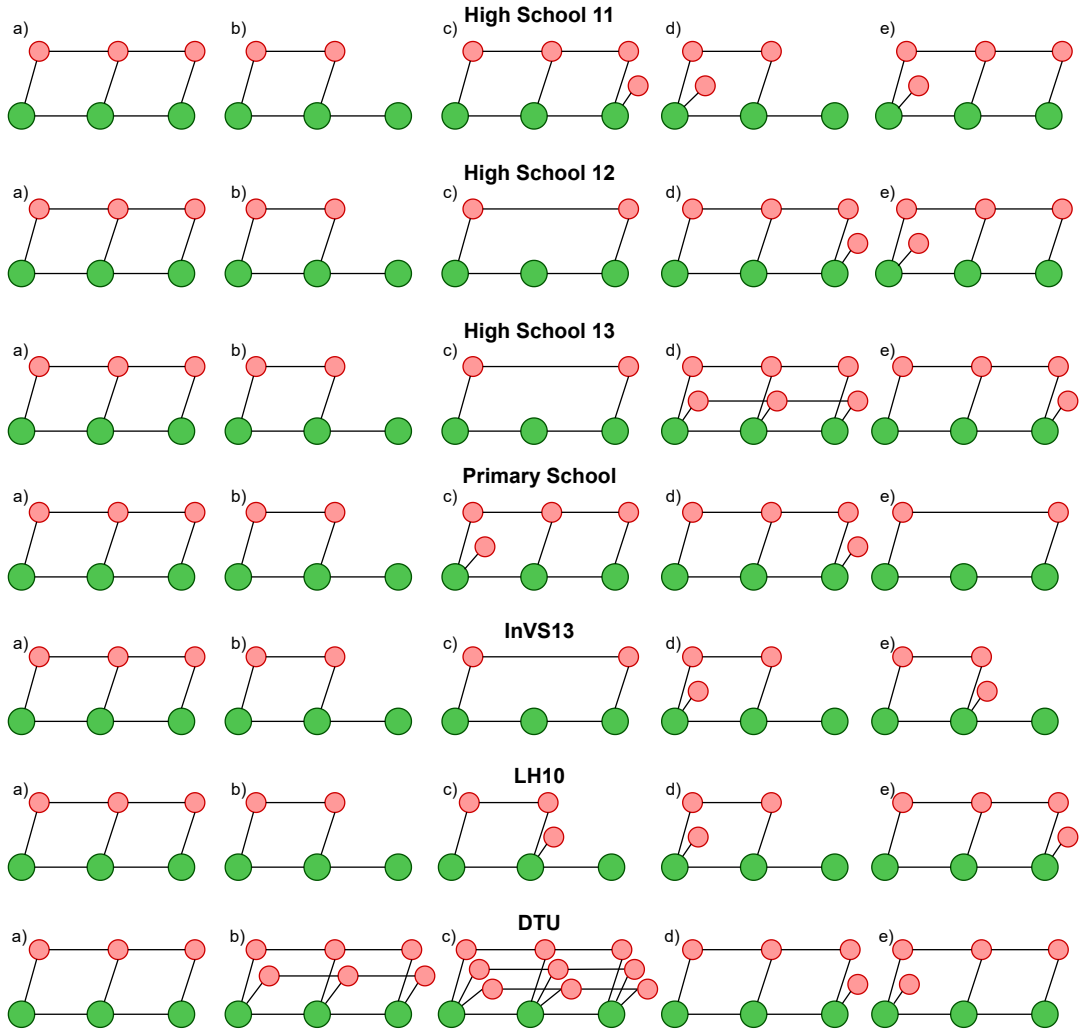
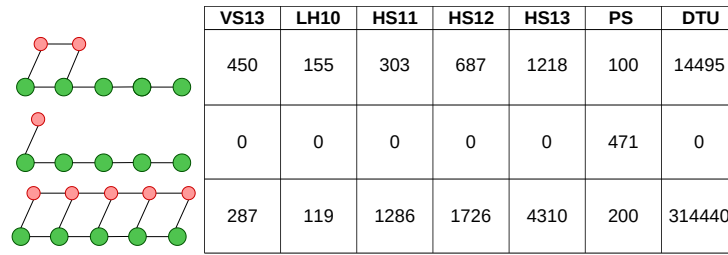


Figure 4.5: Most frequent egocentric temporal motifs discovered by ETMM on the three datasets for $\Delta t = 300$.

temporal gap $\Delta t = 300$ seconds, but results are quite similar for different temporal gaps.

Figure 4.4 shows the first five motifs found by TMM on the different datasets, ordered by frequency. These motifs show some dynamics in the interaction, but it is difficult to interpret them in terms of social interaction patterns or to identify some clear features that distinguish the various datasets. Moreover, Figure 4.4 shows that the five most frequent motifs are the same for all the datasets, with the only exception of the fifth motifs of *InVS13* and *LH10*.

The five most frequent motifs discovered by our method are reported in Figure 4.5. Note that the egocentric focus allows to generate motifs which are quite interpretable in terms of social interactions of the person under investigation (the ego). For instance, for *High-*



	VS13	LH10	HS11	HS12	HS13	PS	DTU
(a)	450	155	303	687	1218	100	14495
(c)	0	0	0	0	0	471	0
(e)	287	119	1286	1726	4310	200	314440

Figure 4.6: The figure shows the frequencies of three ETM for each dataset. The ETMs are those with the maximum variance among datasets.

School11 (first line) we identify a continuous interaction with another person (a) (c) (e), possibly combined with a third person joining at the beginning (e) or at the end (c) of the interaction.

Concerning the other datasets, even if the first two ETMs are the same (except for the *DTU* dataset), our approach does identify some differences that can be related to the different type of networks under investigation. For example, our method is able to identify motifs characterized by rich and dynamic interactions among students in high school and university, and by sparse and short interactions among employees in the research institute. The last line of Figure 4.5 shows the motifs found by ETMM on the *DTU* dataset, and it is easy to see that the structures of the discovered motifs are quite different and more complex with respect to the structures of the motifs found in the other datasets. This may also depend on the fact that the *DTU* dataset, collected using Bluetooth technology, captures co-location and not face-to-face interactions.

To provide further insights on the relationship between motifs and types of networks, we looked for the set of temporal motifs that most discriminates among different datasets. We selected the three egocentric temporal motifs with maximum variance of occurrence among the datasets, and report their frequencies in Figure 4.6. The difference between the primary school and the other datasets is striking. The former contains a motif that is totally missing in the other networks, namely the case where an individual briefly interacts with another one (for less than 5 minutes) and has no more interactions in the following 20 minutes. This small set of motifs may seem a poor description of the analyzed social settings. However, it is surprisingly accurate in catching differences and similarities among datasets, as we will see in next section.

NetSimile							
	VS13	LH10	HS11	HS12	HS13	PS	DTU
InVS13	0	15.68	11.84	14.04	15.45	19.75	26.6
LH10		0	9.8	10.15	14.12	13.72	23.19
HighSchool11			0	9.61	14.32	14.85	23.09
HighSchool12				0	12.9	15.33	23.53
HighSchool13					0	16.32	23.7
primary school						0	19.74
DTU blue							0

Modified NetSimile							
	VS13	LH10	HS11	HS12	HS13	PS	DTU
InVS13	0	17.68	14.22	15.29	18.45	23.49	30.78
LH10		0	12.29	11.41	15.89	16.4	26.95
HighSchool11			0	11.61	16.79	16.71	27.36
HighSchool12				0	15.08	18.35	27.67
HighSchool13					0	18.33	26.81
primary school						0	23.92
DTU blue							0

Weighted Laplacian							
	VS13	LH10	HS11	HS12	HS13	PS	DTU
InVS13	0	588.4	220.8	608.22	1474	849	38201
LH10		0	615	707	1584	845	37808
HighSchool11			0	570	1401	733	38319
HighSchool12				0	1158	567	37938
HighSchool13					0	1354	37006
primary school						0	37822
DTU blue							0

Temporal motifs							
	VS13	LH10	HS11	HS12	HS13	PS	DTU
InVS13	0	0.018	0.02	0.053	0.016	0.04	0.744
LH10		0	0.012	0.014	0.001	0.005	0.707
HighSchool11			0	0.049	0.017	0.019	0.678
HighSchool12				0	0.012	0.1	0.696
HighSchool13					0	0.007	0.695
primary school						0	0.651
DTU blue							0

ETMM-DIST							
	VS13	LH10	HS11	HS12	HS13	PS	DTU
InVS13	0	0.07	0.29	0.22	0.29	0.67	0.47
LH10		0	0.29	0.22	0.30	0.66	0.45
HighSchool11			0	0.04	0.04	0.59	0.06
HighSchool12				0	0.02	0.61	0.13
HighSchool13					0	0.62	0.08
primary school						0	0.62
DTU blue							0

3ETMM-DIST							
	VS13	LH10	HS11	HS12	HS13	PS	DTU
InVS13	0	0.00	0.28	0.19	0.25	0.63	0.42
LH10		0	0.23	0.14	0.20	0.61	0.36
HighSchool11			0	0.01	0.00	0.58	0.02
HighSchool12				0	0.01	0.57	0.05
HighSchool13					0	0.58	0.03
primary school						0	0.61
DTU blue							0

Figure 4.7: Distances among networks using five different methods, namely: NetSimile, Modified NetSimile, Weighted Laplacian, Temporal motifs and ETM-based distance. Each element of the table is colored in a color scale from green (minimum distance) to red (maximum distance). The last tables show the distances obtained using ETM-based distance using $\Delta t = 300$ and $k = 4$, with all motifs (left) and the three most discriminative ones (right)

4.2.2 Egocentric vs non-egocentric graph distances

To give some quantitative estimate of the descriptiveness of the motifs found by our method, we study their effectiveness in measuring the distance among the networks described in Section 4.1.3. In particular, we show the importance of the egocentric perspective in identifying similar social contexts by means of network distances (Definition 14).

Figure 4.7 shows the distances computed with the four non-egocentric methods reported in Section 4.1.3 (first two rows) and with our ETM-based distance (last row). Each table reports the pairwise distances between networks and each element is colored with a color scale starting from green (minimum distance) to red (maximum distance). The figure clearly shows that all non-egocentric methods have serious problems in producing meaningful distances between interaction networks. First, all of them consider *DTU* to be the farthest away from all other networks. However, we expect that *DTU* network, which collects the co-location behaviors of university students, should show some similar-

ities with the ones capturing the face-to-face interactions of high school students, namely *HighSchool11*, *HighSchool12* and *HighSchool13*. These similarities seem not adequately detected by these methods. As previously anticipated, the fact that *DTU* results appear so different from those obtained with the other datasets may depend on the fact that different technologies (RFID vs Bluetooth) have been used to collect the data, suggesting that non-egocentric approaches fail in revealing social patterns when different technologies are at place. We also notice that both NetSimile and Modified NetSimile (first row) detect *hospital (LH10)* as the closest network to *primary school*; this appears as an unexpected result, considering the differences between these two social contexts. Moreover, the Weighted Laplacian method (first table second row) fails in identifying similar environments, since we observe that *InVS13* is very close to *HighSchool11* but quite distant from *HighSchool13*.

Finally, according to (non-egocentric) temporal motifs (second table second row) the network *LH10* is very close to almost all datasets, being almost identical to *HighSchool13*.

The last row of Figure 4.7 shows the results of our ETM-based distance (for $\Delta t = 300$ and $k = 4$), using all ETMs (left) and only the three most discriminative ones (right), i.e., those maximizing the variance of ETM frequencies among datasets (shown in Figure 4.6).

The reported network distances provide a more satisfactory description of the similarity between the underlying datasets. First of all, the three high school networks are very close to each other, with distances around 0, while presenting larger distances with all other networks. Moreover, among the other networks, the closest one is represented by the one capturing the co-location behavior of university students (*DTU*), which are expected to share some behavioral routines with high school students (e.g., class attendance). This shows that ETM is capable of finding similar social interaction patterns despite the use of different data collection technologies, something alternative non-egocentric measures completely fail to achieve. The network which is farthest away from all the others is the *primary school* network: this may be explained by the fact that primary school children seem to experience interaction dynamics which are significantly different from the ones characterizing the social settings of young adults and adults. Finally, we observe that another sensible niche is represented by the two working places, namely the hospital and the research institute, quite similar between each other and quite distinct from all other settings. Interestingly, limiting the set of ETMs to the three most discriminative ones produces results which are very similar to those obtained with the full set of motifs (around six thousands). This is a surprising result and a confirmation of the effectiveness of the egocentric perspective in characterizing different types of social interaction settings.

4.2.3 Sensitivity analysis

In the following we provide a sensitivity analysis showing how the choice of the parameters, namely the temporal gap Δt and the temporal neighborhood order k , affect the ETM-based distance. In Figure 4.8 we report the ETM-based distance among datasets using Δt equal to 300 and 900 seconds², and k ranging from 3 to 5. We observe that results are quite stable. For intermediate values of the parameters, results are very similar to those presented for $\Delta t = 300$ and $k = 4$, with distances that tend to increase for increasing values of Δt and k . Intuitively, small values of *both* Δt and k (i.e., $\Delta t = 300$ and $k = 3$, top left matrix) produce small motifs, leading to a partial reduction in discriminative capacity, with the primary school becoming (more) similar to workplaces and high schools. On the other side, large values of *both* Δt and k (i.e., $\Delta t = 900$ and $k = 5$, bottom right matrix) determine a slight decrease in the capacity of detecting similarities among related datasets (namely between different high schools). This is again not surprising, as jointly increasing Δt and k substantially increases the required length for a temporal fragment to match a motif, making it more complex for the method to mine relevant motifs.

	$\Delta t = 300$		$k = 3$	ETMM-DIST				
	VS13	LH10	HS11	HS12	HS13	PS	DTU	
InVS13	0.00	0.09	0.20	0.13	0.18	0.19	0.40	
LH10		0.00	0.22	0.14	0.18	0.09	0.38	
HighSchool11			0.00	0.04	0.03	0.18	0.07	
HighSchool12				0.00	0.03	0.12	0.15	
HighSchool13					0.00	0.12	0.09	
primary school						0.00	0.27	
DTU blue							0.00	

	$\Delta t = 900$		$k = 3$	ETMM-DIST				
	VS13	LH10	HS11	HS12	HS13	PS	DTU	
InVS13	0.00	0.15	0.32	0.27	0.38	0.65	0.59	
LH10		0.00	0.57	0.48	0.61	0.77	0.82	
HighSchool11			0.00	0.03	0.05	0.21	0.10	
HighSchool12				0.00	0.03	0.23	0.13	
HighSchool13					0.00	0.25	0.10	
primary school						0.00	0.18	
DTU blue							0.00	

	$\Delta t = 300$		$k = 4$	ETMM-DIST				
	VS13	LH10	HS11	HS12	HS13	PS	DTU	
InVS13	0.00	0.08	0.29	0.19	0.26	0.61	0.48	
LH10		0.00	0.30	0.20	0.26	0.58	0.46	
HighSchool11			0.00	0.04	0.03	0.57	0.06	
HighSchool12				0.00	0.02	0.54	0.13	
HighSchool13					0.00	0.55	0.08	
primary school						0.00	0.60	
DTU blue							0.00	

	$\Delta t = 900$		$k = 4$	ETMM-DIST				
	VS13	LH10	HS11	HS12	HS13	PS	DTU	
InVS13	0.00	0.15	0.36	0.30	0.42	0.93	0.68	
LH10		0.00	0.61	0.54	0.68	0.96	0.40	
HighSchool11			0.00	0.04	0.07	0.82	0.13	
HighSchool12				0.00	0.05	0.83	0.17	
HighSchool13					0.00	0.81	0.14	
primary school						0.00	0.82	
DTU blue							0.00	

	$\Delta t = 300$		$k = 5$	ETMM-DIST				
	VS13	LH10	HS11	HS12	HS13	PS	DTU	
InVS13	0.00	0.06	0.37	0.24	0.33	0.65	0.57	
LH10		0.00	0.37	0.26	0.35	0.64	0.56	
HighSchool11			0.00	0.04	0.03	0.54	0.06	
HighSchool12				0.00	0.02	0.54	0.14	
HighSchool13					0.00	0.54	0.08	
primary school						0.00	0.57	
DTU blue							0.00	

	$\Delta t = 900$		$k = 5$	ETMM-DIST				
	VS13	LH10	HS11	HS12	HS13	PS	DTU	
InVS13	0.00	0.16	0.45	0.31	0.54	0.95	0.75	
LH10		0.00	0.61	0.49	0.70	0.95	0.94	
HighSchool11			0.00	0.17	0.22	0.75	0.29	
HighSchool12				0.00	0.25	0.89	0.22	
HighSchool13					0.00	0.93	0.35	
primary school						0.00	0.90	
DTU blue							0.00	

Figure 4.8: ETM-based distances obtained using $\Delta t = 300, 900$ and $k = 3, 4$ and 5 .

²A value of $\Delta t < 300$ generates a too sparse network for the DTU dataset that relies on Bluetooth to detect interactions, preventing the discovery of non-trivial motifs by any method. Results for the other datasets are similar for values of Δt as small as 60.

4.2.4 Results on distance communication and synthetic datasets

In this section we evaluate the ability of the ETM-based distance to characterize networks beyond close proximity interaction data. First of all we consider other typologies of social data representing distance communication interactions (Fig. 4.9), then we explore the algorithm performance on synthetic temporal graphs (Fig. 4.10).

		NetSimile					
		Calls		SMS		Email	
		DTU C	Frien C	Dtu S	Frien S	Email	DNC
Calls	DTU C	0.00	22.87	6.27	15.47	24.11	20.39
	Frien C		0.00	23.86	21.57	18.53	23.37
SMS	Dtu S			0.00	13.90	25.60	20.57
	Frien S				0.00	26.19	22.89
Email	Email					0.00	20.79
	Email DNC						0.00

		Modified NetSimile					
		Calls		SMS		Email	
		DTU C	Frien C	Dtu S	Frien S	Email	DNC
Calls	DTU C	0.00	27.35	9.87	18.90	27.12	23.46
	Frien C		0.00	27.13	22.78	21.04	27.89
SMS	Dtu S			0.00	17.74	28.81	22.97
	Frien S				0.00	28.58	27.22
Email	Email					0.00	24.82
	Email DNC						0.00

		Weighted Laplacian					
		Calls		SMS		Email	
		DTU C	Frien C	Dtu S	Frien S	Email	DNC
Calls	DTU C	0	10922	3803	10691	6596	915
	Frien C		0	3874	9054	8319	10405
SMS	Dtu S			0	10414	6303	4863
	Frien S				0	5808	10040
Email	Email					0	8378
	Email DNC						0

		Temporal motifs					
		Calls		SMS		Email	
		DTU C	Frien C	Dtu S	Frien S	Email	DNC
Calls	DTU C	0.00	0.96	0.29	0.96	0.66	0.73
	Frien C		0.00	0.98	0.02	0.97	0.89
SMS	Dtu S			0.00	0.98	0.81	0.87
	Frien S				0.00	0.97	0.89
Email	Email					0.00	0.07
	Email DNC						0.00

		ETMM-DIST					
		Calls		SMS		Email	
		DTU C	Frien C	Dtu S	Frien S	Email	DNC
Calls	DTU C	0.00	0.37	0.28	0.26	0.61	0.66
	Frien C		0.00	0.34	0.33	0.59	0.58
SMS	Dtu S			0.00	0.06	0.65	0.64
	Frien S				0.00	0.65	0.64
Email	Email					0.00	0.38
	Email DNC						0.00

Figure 4.9: Distances among different communication networks using five different methods, namely: NetSimile, Modified NetSimile, Weighted Laplacian, Temporal motifs and ETM-based distance. Each element of the table is colored in a color scale from green (minimum distance) to red (maximum distance). The last table shows the distances obtained using ETM-based distance with $\Delta t = 3600$ and $k = 4$.

The non-physical interaction datasets that we consider employ different communication technologies (phone calls, SMSs and emails, see Section 4.1.3). For this experiment, we choose $k = 4$ and $\Delta t = 3600$, a temporal gap for which the six temporal networks are characterized by a similar average degree (equal to 0.052 for phone calls, 0.049 for SMSs and 0.051 for emails). Results are shown in Figure 4.9. Non-egocentric methods manage to capture the similarity among some of the networks using the same technology (e.g., SMSs for NetSimile and Modified NetSimile, emails for temporal motifs), but they badly fail in most cases. On the other hand, the ETM-based distance is quite consistent in capturing the similarity between networks employing the same communication technology. Moreover, networks based on SMSs and phone calls are more similar to each other than

networks based on emails, as expected. This is a further proof of the versatility of ETM patterns to characterize temporal behaviors.

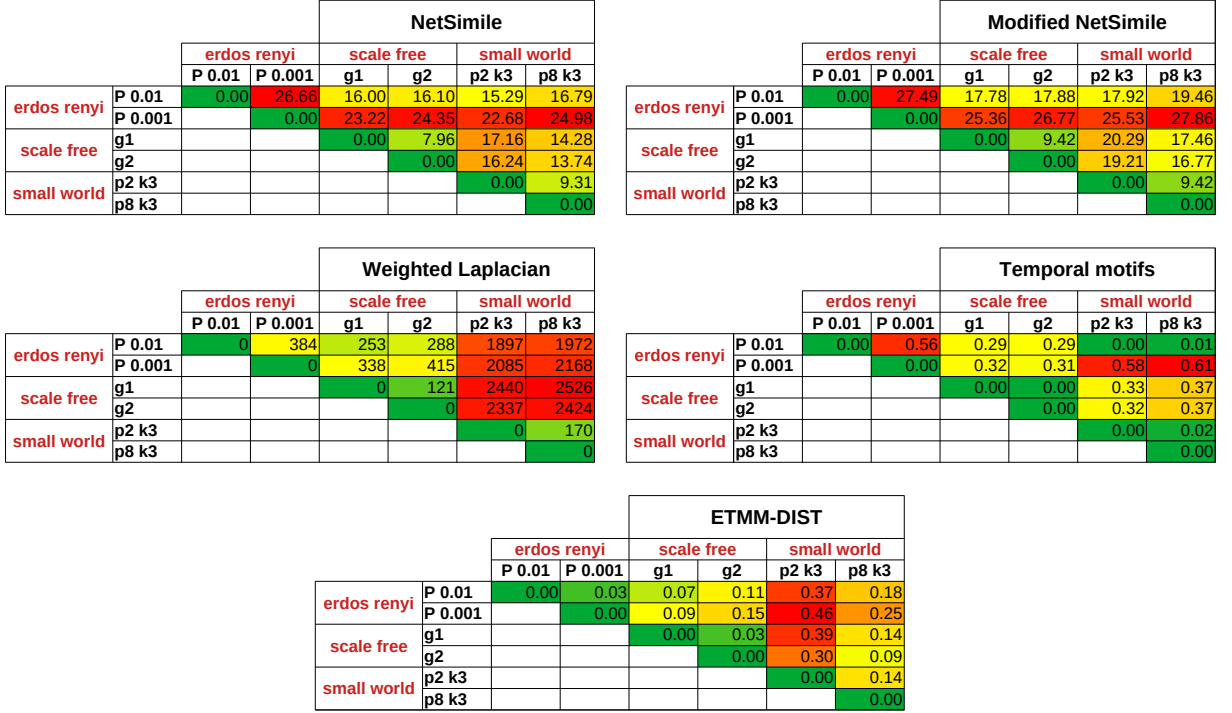


Figure 4.10: Distances among different synthetic networks using five different methods, namely: NetSimile, Modified NetSimile, Weighted Laplacian, Temporal motifs and ETM-based distance. Each element of the table is colored in a color scale from green (minimum distance) to red (maximum distance). The last table shows the distances obtained using ETM-based distance with $k = 4$. No temporal gap is needed in this case, as the networks are not extracted from time series but synthetically generated according to the procedure described in Section 4.1.3.

Results for the synthetic datasets are shown in Figure 4.10. Our ETM-based distance is clearly capable of detecting similarities among Erdős-Rényi graphs, outperforming all competitors, and among scale-free ones, which are however modelled reasonably well by all methods. On the other hand, the ETM-based distance lags behind all competitors in detecting similarities between small-world networks. This result sheds some light on the limitations of the egocentric perspective of our method. Indeed, ETMs deliberately discard the information on connections among neighbors of the ego node (we only consider the existence of neighbors and not their mutual behavior), thus neglecting the clustering structure of the network. This explains why the synthetic small-world networks, characterized by high values of clustering coefficient, are less effectively described by our method.

4.3 Discussions

In this chapter, we proposed a novel approach for mining temporal motifs based on an ego perspective. Each motif represents the evolution, during few time steps, of the set of neighbors of a specific network node. Egocentric temporal motifs present some essential characteristics that distinguish them from standard temporal motifs.

First of all, egocentric temporal motifs are simpler, at a topological level, with respect to standard temporal motifs, since they only take into account the neighboring nodes of the ego, ignoring the connections among them. This allows both to account for larger neighborhoods and to explore more in detail the temporal aspect, including duration of contacts and contemporary interactions, usually neglected in standard procedures for temporal motif mining. This is a necessary requirement when analyzing social domains like physical human interactions, where each individual can interact with multiple people at a time, with various durations.

Second, the egocentric view has substantial advantages from a computational perspective. Traditional techniques for motif mining rely on an isomorphism test for assessing if two sub-networks are equivalent or not, and this limits their applicability to mine motifs containing a handful of nodes. The focus on an ego node allows us to sidestep this problem. We show how an egocentric temporal neighborhood, which is the sub-structure representing a candidate motif, can be encoded into a bit vector in a way such that two neighborhoods have the same encoding if and only if they are isomorphic.

We made use of seven different datasets representing social interactions and applied our egocentric temporal motif miner, comparing the results with a state of the art non-egocentric temporal motif miner. Our method is shown to be more effective in terms of selectivity and quality of the extracted motifs. By visually inspecting the most frequent motifs found in each dataset, it is apparent that our method succeeds in grasping some of the peculiarities of each dataset: more rich and dynamical interactions among students in high school and university, sparser and shorter interactions for the research institute, a combination of the two in the hospital, and a different behavior at the primary school. Importantly, differences and similarities between datasets are quantified by defining a correlation measure between egocentric signatures. The results that we obtain fully reflect the social context represented by the network, especially if compared with standard non-egocentric approaches to measure temporal networks' distance. Later, we show how the egocentric perspective is crucial for the discrimination among different communication technologies, like phone calls, SMSs and emails, and how it also allows to characterize

temporal variants of popular network topologies, like Erdős-Rényi and scale-free.

The egocentric perspective surely represents an important limitation too, since we are neglecting all the second order interactions, i.e., the interactions between neighbors of an ego node. This is especially limiting in networks which are characterized by a high clustering coefficient, as shown by the suboptimal results that we achieve on small-world networks. On the other hand, this is a necessary requirement for the bit vector encoding and hence for the extreme velocity of our method (which scales linearly with the number of nodes and the timesteps of the temporal network). This allows to mine motifs covering larger structures and longer time sequences with respect to alternative solutions. Our extensive experimental results show that, even renouncing to represent second-order interactions, the proposed method is able to recognize different social settings, substantially outperforming existing alternatives.

In conclusion, we are proposing a novel efficient method to obtain temporal motifs from the node point of view. This method is not conceived to completely replace existing temporal motif mining methods, but rather to complement them in revealing a different kind of motifs. As shown in our experimental evaluation, this can be particularly useful to study social interaction networks, which could not be properly analyzed with existing approaches.

4.4 Code availability

The code used for the mining process is publicly available on GitHub: ETMM <https://github.com/AntonioLonga/Egocentric-Temporal-Motifs-Miner-ETMM>.

Chapter 5

Generating Temporal Networks

The ability to decompose temporal networks in *Egocentric temporal neighborhood*, not only gives the possibility to a better characterization of temporal networks but also allows to development of a novel generative model for fine-grained temporal networks. In this chapter, we propose a method able to generate high temporal resolution surrogate networks that are able to match real networks in terms of a wide range of topological and dynamic measures. Our generative algorithm is based on the idea of the *egocentric temporal neighborhood* (110) $G_{\{t-k, \dots, t\}}^k(n)$ for node n at time t , including a small number k of prior time steps. Here we assume that the network is represented in discrete time with each time step corresponding to a static graph, also referred to as a ‘layer’ of the network. Crucially, $G_{\{t-k, \dots, t\}}^k(n)$ does not include interactions between the neighbors of n . To avoid excessive notation in the following, we simply use the term ‘neighborhood’ to describe the egocentric temporal neighborhood when there is no risk of confusion.

Conceptually our algorithm does the following. We first characterize a given real-world network in terms of neighborhoods, and then use those neighborhoods as building blocks for a new synthetic network. When we match up neighborhoods, conflicts among the egocentric perspectives of different nodes are globally solved by combining overlapping sub-networks so as to preserve as much as possible each node’s desired neighborhood.

In order to extend the network into subsequent time steps, we build a local probabilistic model for suggesting new temporal interactions at time $t + 1$ for each node, given the behavior during $\{t - k, \dots, t\}$. We can further increase realism corresponding to activity modulation, such as day/night and week/weekend by building distinct probabilistic models for different times of the day or days of the week. In the Appendix (section 7.2.4) we show how a single probabilistic model fails to grasp temporal periodicity patterns.

A major advantage of the egocentric perspective (that ignores connections among neighbors of an ego node) is that it allows us to linearize the concept of node neighborhood sidestepping the subgraph isomorphism problem (61), making the generation process fast and scalable both in terms of the number of nodes and the number of temporal snapshots. Speed turns out to be a fundamental feature, because the other existing methods rely on algorithms of considerably higher complexity that prevent those methods from scaling to even moderately-sized networks.

We test the method, named *Egocentric Temporal Neighborhood Generator (ETN-gen)*, on a range of different temporal networks. In our testing, we mainly use social interactions datasets, because of the richness and availability of these datasets, but the method is general and can be used to generate any kind of graph. The simplicity of our algorithm makes it easily interpretable, extendable and algorithmically scalable. As we show below, the surrogate networks that we generate match original networks with a high degree of accuracy, not just in terms of local features, as one might anticipate from the local generating mechanism, but with respect to global features, such as the number of interactions, the number of interacting individuals in time and density of their connections. The ability to generate surrogate temporal graphs that reproduce real behaviors allows us to obtain large as desired data, without resolution limits, while mitigating key privacy issues.

5.1 Method

We first present the temporal graph generation process. Then, we use our method to generate temporal graphs which reproduce the temporal interaction patterns of a diverse set of face-to-face interaction networks, including a hospital (172), a workplace (58), and a high school (50).

5.1.1 The neighborhood generation process

Figure 5.1 shows a graphical representation of the Egocentric Temporal Neighborhood generation process for a small temporal network with three timesteps (see *Methods* for details). Panel A shows the egocentric temporal network(110) $G_{\{t-k, \dots, t\}}^k(n)$ – or simply ‘neighborhood’ – of a node n . We extract this neighborhood for each node in a graph. Specifically, for a given time horizon k ($k = 2$ in the figure) and a given egocentric node n ($n = E$ in the figure), $G_{\{t-k, \dots, t\}}^k(n)$ is defined as the network fragment which contains n and its neighbors at each of $k+1$ consecutive timestamps, discarding connections between

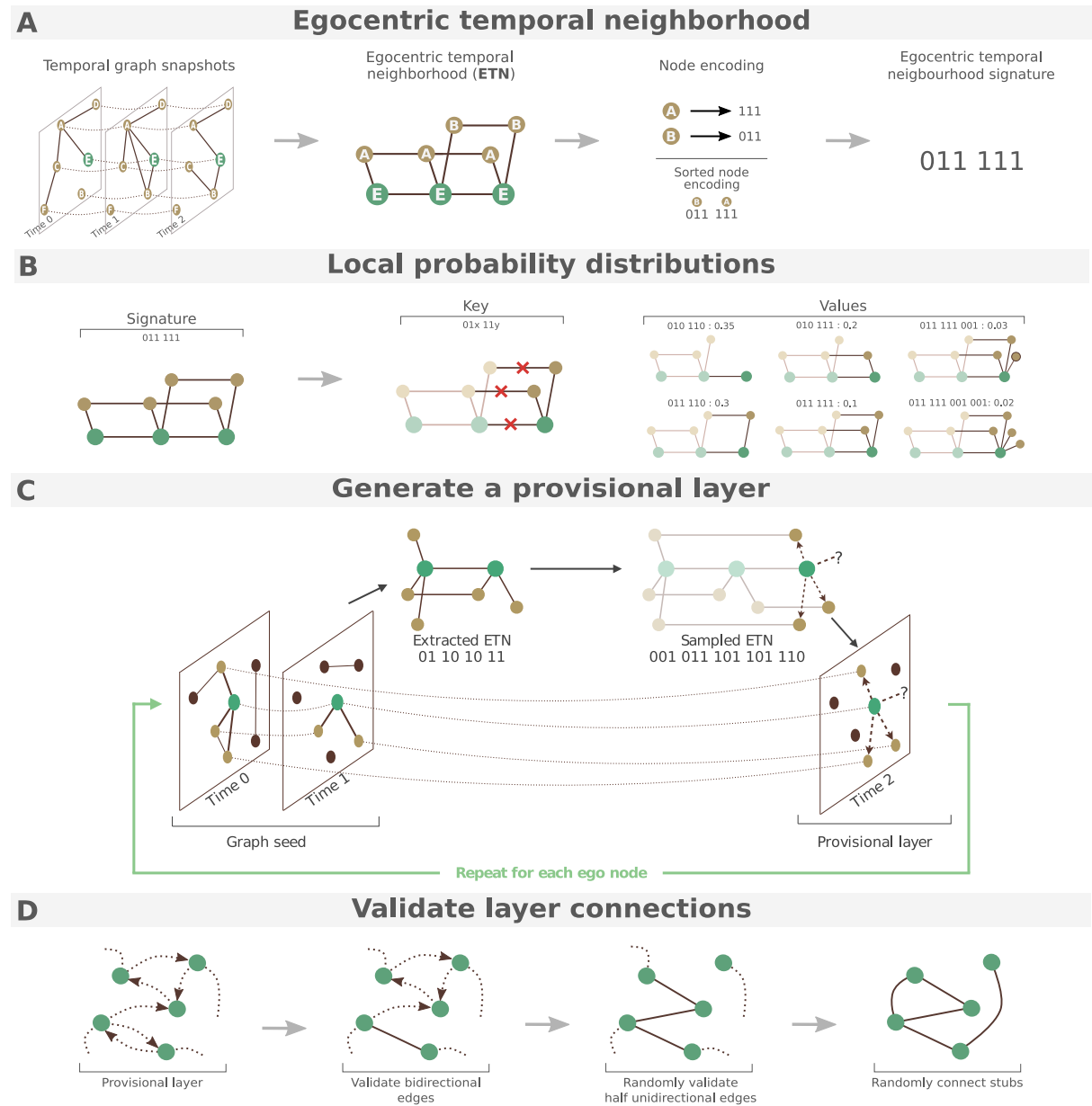


Figure 5.1: *ETN-gen*. The top panel shows how egocentric temporal neighbourhood signatures are extracted and computed. Panel **B** shows how to build the probability distribution of neighborhoods, necessary to generate a provisional layer. Panel **C** shows how to generate a provisional layer, while panel **D** explains how to convert the provisional layer into a definitive one.

the neighbors of n , and adding (temporal) connections among instances of the same node at different timestamps. Having discarded links between neighbors, $G_{\{t-k, \dots, t\}}^k(n)$ can be encoded as a binary string, where for each neighbor node and timestamp 1 (resp. 0) indicates the presence (resp. absence) of a link connecting to the node at that timestamp.

Such neighborhoods are extracted for all nodes and all timestamps by using a sliding window over time.

Second (panel B), we build a local probability distribution designed to enable simulation of activity in future time steps. This distribution to extend the graph into future time steps is based on past neighborhood activity. Specifically the local distribution maps neighborhoods of length $k - 1$ (i.e. temporal neighborhoods involving k steps) to the set of all possible extensions into the future (i.e. neighborhoods of temporal depth k , involving $k + 1$ steps), with associated probabilities estimated by Maximum Likelihood (i.e., normalized counts) over the entire original temporal network. Third (panel C), for each node in the network we generate a provisional temporal extension by sampling from the probability distribution described above. We thus obtain a provisional temporal layer of the network. Last (panel D), this provisional layer is finalized by combining provisional temporal extensions of all nodes, resolving conflicts and dangling links. To connect neighborhoods, we consider a connection from node i to node j in the provisional layer a ‘request’ of i to be connected to j . If this request is reciprocal, the link is validated and added to the new temporal layer (see the second step in panel D). All remaining one-directional links are validated with probability $\alpha = 1/2$ (third step), to preserve the overall number of connections (an $i - j$ connection can be requested by i or by j). The procedure is repeated as many times as the desired length of the final temporal graph, always considering the last k timestamps as seeds and generating an additional one.

Above, we have described the simplest possible strategy for extending a layer into the future, but note that all random choices in the link validation process could become preferential choices in order to optimize a specific characteristic of the final network (see Section *Topological similarity evaluation*). Further – which we explore below – the temporal extension can include novel nodes that are not present in the current temporal neighborhood of the ego node, and thus their identity is not known (the question mark in panel C). These nodes are connected by a ‘stub’ rather than a real connection, representing a link-to-be, and stubs are pairwise matched up at random (last step in panel D).

With the basic mechanisms in place, we take a step back and explain how to initialize the process, i.e. how to obtain the first k layers of the graph. The graph at the first timestamp is generated using a configuration model (129; 134) reproducing the degree distribution of the first layer of the original graph. The following layers up to k are generated by applying the procedure in Figure 5.1 to the first layer with $k' = 1$, to the first two layers with $k' = 2$ and so on until $k' = k$.

Finally, temporal networks are often characterized by an intrinsic periodicity (74). In social interactions data for instance this can be due to the day-night cycle or to the difference between week days and weekends, and the organization of our societies. This is accounted for in our generation procedure by using distinct local probability distributions to extend the graph during different days of the week or times of the day. In the experiments in this paper we use distinct week/weekends or daily local probability distributions, depending on the length and variability of the input network.

Neighbourhood generation process: parameters. The gap between two consecutive temporal snapshots has been set to 5 minutes for face-to-face interaction networks. The time horizon k defining the egocentric temporal neighbourhood has been set to $k = 2$ in all experiments, which is the minimal horizon that preserves some temporal correlation. Local probability models have a granularity of 1 hour and a periodicity of 1 day (i.e., between 8 and 9 am in each day we use the same probability model, and the same holds for all 1 hour slots in the day), for all networks but the ones including weekends, namely Workplace and High school 2, for which the periodicity is set to 1 week.

5.1.2 Model evaluation

To evaluate the quality of the generated networks based on interaction statistics, we compare the networks to empirical data as well as networks generated by a suite of state-of-the-art temporal network generation methods described below. We evaluate performance in terms of individual layer topology as well as temporal behaviour. The key new feature of our network generation procedure is the ability to approximately reproduce the interaction statistics of real-world data, something the existing alternatives cannot do.

It is important to underscore that these network generation methods have not necessarily been developed with the aim of generating large temporal networks with low computational cost (see sub section 5.1.3). This means that, for example, they require much more training data, need denser temporal snapshots, and therefore cannot generate high temporal resolution networks. In this regard *ETN-gen*, thanks to the linearization due to the egocentric perspective, is the first method that allows researchers to scale to arbitrarily sized temporal networks. In the rest of the paper, in order to evaluate our work relative to the other methods, we will report experiments only on the three smallest face-to-face interactions datasets, collected in the hospital (172), in the workplace (58), and in one of the high schools (50) respectively. Results applying *ETN-gen* to larger datasets are reported in the Appendix.

5.1.3 Computational complexity and space complexity

In this section, we report the time and space complexity of our model.

Time complexity: as depicted in figure 5.1, the method can be decomposed in four steps: 1) mine Egocentric Temporal Neighborhoods 2) build a local probability distribution 3) generate a provisional layer for each timestamp and 4) validate layer connections. Longa *et al.*(110) proved that the computational cost to count all Egocentric Temporal Neighborhoods in a graph is equal to $\mathcal{O}(n \cdot m_{orig} \cdot d^k \cdot \log d^k)$, where n is the number of nodes, m_{orig} is the number of timestamps in the original network, d is the maximal degree of the network and k is the length of the temporal neighborhood. The second step can be done in linear time with respect to the size of the mined Egocentric Temporal Neighbors. In the third step, we query the local probabilistic model in constant time for each node for each timestamp of the generated network (m_{gen}), thus the complexity is $\mathcal{O}(n \cdot m_{gen})$. Finally, in the validation step, for each node and each timestamp we have to go through each edge (there are at most d of them), with a complexity of $\mathcal{O}(n \cdot m_{gen} \cdot d)$. The overall complexity is thus $\mathcal{O}(n \cdot m_{orig} \cdot d^k \cdot \log d^k + n \cdot m_{gen} \cdot d)$. Note that for reasonable values of k , d^k is independent of the size of the network, so that the overall complexity is $\mathcal{O}(n \cdot m_{gen})$, assuming that $m_{gen} \gg m_{orig}$.

Space complexity: the space complexity of the method is dominated by the storage of local probabilistic models. Storing a single Egocentric Temporal Neighborhood Signature of length k costs $\mathcal{O}((k+1) \cdot d)$, where d is the maximal degree of the network. The number of Egocentric Temporal Neighborhoods is the number of all possible ordered sequences of k – bit strings of length d , which corresponds to $\binom{d+2^k-1}{d}$ and is loosely upper bounded by 2^{kd} . The overall space complexity is thus $\mathcal{O}(2^{(k+1)d})$. As discussed in the case of time complexity, for reasonable values of k , d is independent of the size of the network, so that the space complexity does not depend on it.

5.1.4 Size expansion: preserving interaction density.

The seed graphs for the size expansion experiment are generated by artificially reducing the original dataset (so that the original graph can be used as ground-truth). In this reduction process, whenever a node is dropped all its connections are dropped too. As a consequence, the resulting seed graph has a reduced mean degree with respect to the original one, and the expanded graph generated from it would inherit this reduced mean

degree. This problem can be avoided by adjusting the α parameter of the generation process (the probability to confirm the unidirectional links in each provisional layer, set to 1/2 by default). In particular, we would need to set $\alpha = 1 - \frac{1}{2} \frac{\hat{L}}{L}$, where \hat{L} is the average number of links in the seed graph and L the desired number of links in the generated graph. However, L is unknown and needs to be estimated.

Something that we know, and that we want in this case to preserve, is the density, defined as $d = \frac{\hat{L}}{\hat{N} \cdot (\hat{N} - 1) / 2}$ i.e. the fraction between the number of links in the seed graph and all possible links (\hat{N} is the number of nodes in the seed graph). If we assume a linear growth with respect to the number of all possible edges in the network, we also have: $d = \frac{L}{N \cdot (N - 1) / 2}$, with N as the number of nodes of the generated graph (that we can choose). Combining these two equations we obtain an estimate for L , from which we obtain: $\alpha = 1 - \frac{\hat{N} \cdot (\hat{N} - 1)}{N \cdot (N - 1)} \cdot \frac{1}{2}$. Hence, when we consider a seed with only 30% of the nodes of the high school dataset (so $N = 126$ and $\hat{N} = 38$) we should use $\alpha = 0.96$ to reproduce the same density. While if we start with 50% and 70% of the nodes (i.e. $\hat{N} = 63$ and $\hat{N} = 88$) in the seed we should use respectively $\alpha = 0.88$ and 0.76 .

5.1.5 Alternatives approaches for generating networks.

The state-of-the-art methods we consider are: *Dymond* (179), a model which uses the distribution of 3-nodes structures in the original graph (triads with one, two or three connections) as building blocks to generate a new temporal network; *STM* (143), a generative model based on the distribution of small temporal motifs; and *TagGen* (185), based on deep learning, which uses a generative adversarial network to generate temporal walks that are then combined into a temporal graph. *Dymond* and *STM* only consider local information, while *TagGen* is more global.

Dymond (179) builds a temporal network considering (i) the dynamics of temporal motifs in the graph and (ii) the roles nodes play in motifs (e.g., in a wedge – two links connecting three nodes – one node plays the hub, while the remaining two act as spokes). The method has no parameters to be set. *Structural Temporal Modeling (STM)* (143) extracts counts for a predefined library of (non-egocentric) temporal motifs from the original network, and turns them into generation probabilities from which to create the temporal network. This methods has no tunable parameters. *TagGen* (185) is a neural-network based approach that extracts temporal random walks from the original graph and feeds them to an assembling module for generating temporal networks. *TagGen* has been trained with the parameters used in the original paper, namely 30 epochs with a batch size of 64 and

stochastic gradient descent with a learning rate of 0.001.

5.2 Results

We evaluate the quality of the generated networks in terms of interaction statistics, considering both static and temporal network properties, highlighting the advantage of our proposed method relative to the state-of-the-art. Finally, we show how the approach can be used to expand existing temporal networks, both in time and in number of nodes, something which is not possible using competing methods for temporal network generation.

5.2.1 Temporal periodicity

Figure 5.2 reports the total number of interactions for each temporal snapshot (left) and the average number of nodes (right) in the original network, *ETN-gen* and the three competitors. The first clear finding from this figure is that *ETN-gen* (orange curves) results in time-series that are remarkably similar to those appearing in the original datasets (black curves). This is true, not just in terms of generating a number of interactions which is of the same order of magnitude as the original data (notice that different datasets have different scales on the y -axis), but also in terms of temporal patterns which are preserved with considerable accuracy, including daily and weekly periodicity.

This result, even if outstanding relative to *Dymond*, *STM*, and *TagGen* should not come as a surprise, as it is a direct consequence of our network generation procedure. The local probabilistic models store the probability distributions of the neighborhoods appearing in the original graph and this indirectly contains the key information about how nodes degree evolves in time. Further, our seed-network has the same degree distribution as the original graph, which allows us to statistically preserve the overall average number of interactions of the original graph. Moreover, we manually input periodicity via different local probabilistic models for different times and days of the week. We highlight, however, that while using only a single local probabilistic model would remove our ability to model periodic changes in graph over time, we would still be able to model the average number of interactions, as these are automatically reproduced by the rest of the algorithm. A detailed analysis is reported in the Appendix (section 7.2.4).

In contrast to *ETN-gen*, the results that we obtain from the current methods are significantly different from the empirical ones. To start, the curves representing *TagGen*, only



Figure 5.2: **Number of interactions in the generated network vs. competitors.** Number of interactions at each timestamp, each color represents a different generation algorithm, while the original graph is depicted in black. *TagGen* produces a number of interactions ten times the order of the original network and it only appears in the insets for visibility.

appear in the insets, which report the same data with different y -axes. The number of interactions generated by *TagGen* is orders of magnitude larger than that of the other methods and of the original network. *TagGen*, however, does manage to capture day-night periodicity, which is completely lost by *Dymond* and *STM*, both of which produce a number of interactions that is stable over time. Overall, *ETN-gen* is the only method capable of accurately reproducing both aggregated and temporal interaction statistics. The histograms on the right part of the figure show mean and standard deviation of the number of nodes in the networks generated by the different methods, with the horizontal dashed line representing the number of nodes in the original network. These histograms show another important result: only our method and *TagGen* always generate networks

with the same number of nodes of the input graph, with the difference that, as shown by the insets on the left, *TagGen* generates orders of magnitude more interactions. *Dymond* and *STM*, on the other hand, respectively under and over represent the number of nodes, so that only *ETN-gen* manages to reproduce *both* the number of nodes and the number of interactions in the original network.

5.2.2 Topological similarity evaluation

Having studied the temporal development, we now turn to structural similarity between the surrogate data and the original networks. We consider ten metrics for structural similarity: number of interactions, density (179), interacting individuals (163), new conversations (163), S-metric (101), duration of contacts (163), edge strength in the projected weighted network (163), global clustering coefficient (116; 177), assortativity (133), and average shortest path length (74).

The topological metrics can be divided into eight global metrics, which are computed for each temporal layer as it was a static network, and for which we report distributions over temporal layers:

- **Number of interactions.** The number of edges.
- **Density.** The ratio of edges in the graph versus the number of edges if it was a complete graph (179).
- **Interacting individuals.** The number of individuals that are interacting (163).
- **New conversations.** The number of conversations starting at this specific timestamp (163).
- **S-metric.** A measure of the extent to which a graph has a hub-like core, maximized when high-degree nodes are connected to other high-degree nodes. (101).
- **Global clustering coefficient.** The ratio of the number of closed triplets to the total number of open and closed triplets (116; 177).
- **Assortativity.** The degree-degree correlation of nodes that are connected (133).
- **Average shortest path length.** The average shortest path length for all possible pairs of nodes of the largest connected component for each temporal layer (74).

And two local metrics (distributions over edges):

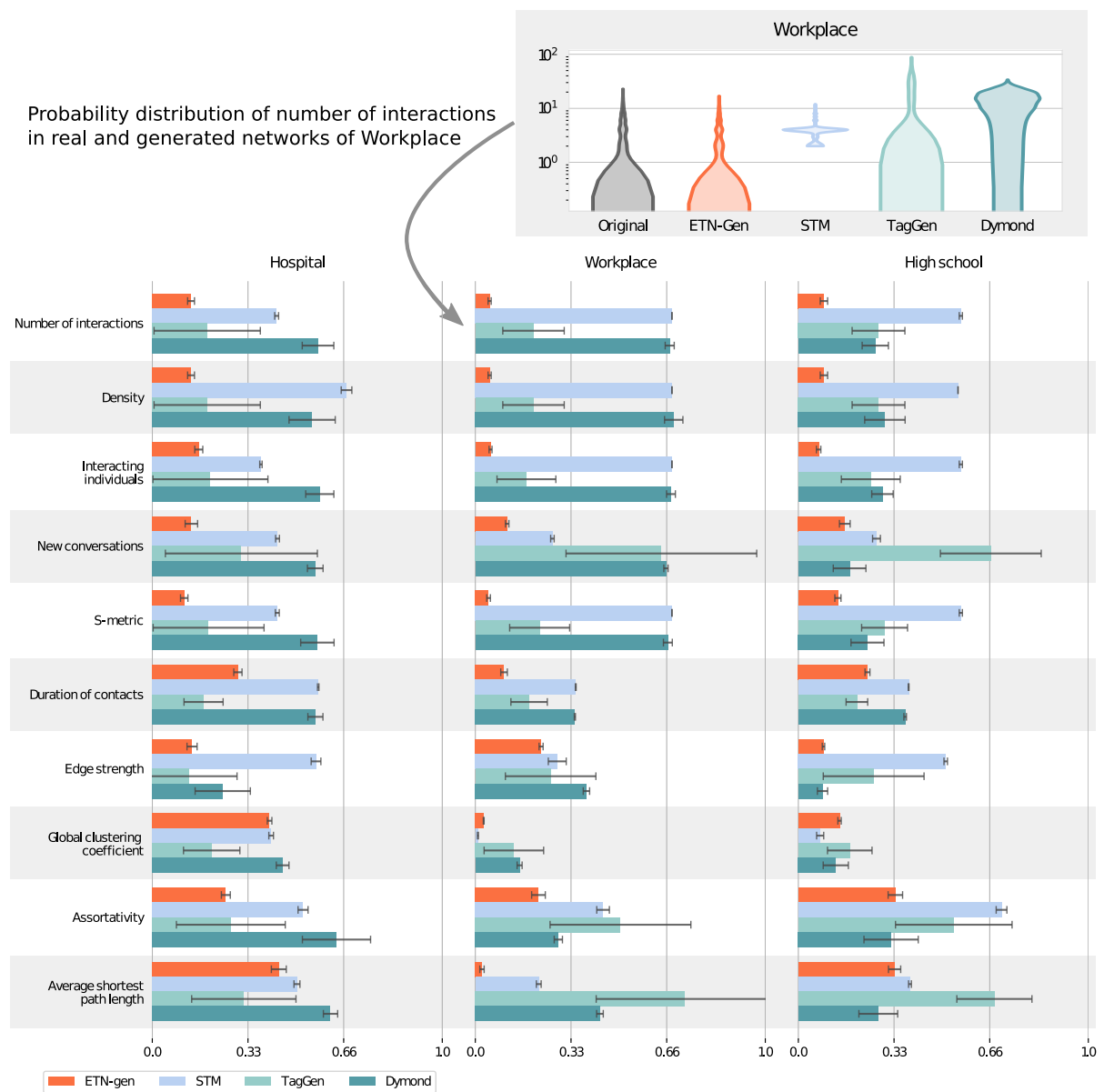


Figure 5.3: **Topological similarity.** Similarity of the original network with those generated by *ETN-gen*, *STM*, *TagGen* and *Dymond*. Each bar reports the Kolmogorov-Smirnov distance between the two distributions (original and generated) for a specific structural metric. The shorter is a bar the more similar are the distributions. Standard deviations are obtained over 10 stochastic realizations of each network. In the top inset we report the distributions of the number of interactions in real and in one instance of generated networks for the Workplace dataset.

- **Duration of contacts.** The mean duration (in timestamps) of interactions between each couple of nodes (163).

- **Edge strength in the projected weighted network.** The total number of interactions in time between each couple of nodes (163).

To compare distributions, we rely, inspired by Zeno *et al.* (179), on the Kolmogorov-Smirnov distance (119) to compare generated and original graphs. Distances between distributions are reported in Figure 5.3, where we compare graphs obtained with *ETN-gen* with those from the three alternative approaches.

ETN-gen (orange bars) clearly generates surrogate networks that are closest to the empirical networks for almost all measures, regardless of the dataset considered. Moreover, our method is substantially more stable than the competitors, as shown by the error bars which were obtained over 10 stochastic realizations of each network.

The measures for which *ETN-gen* performs best are those that, together with the number of interactions (see Figure 5.2), are preserved by construction: the density and the number of interacting individuals in time. Here, the similarity originates from the neighborhood probability distributions, which ensure that from a statistical viewpoint, the surrogate network has the same number of interactions and the same number of individuals involved in an interaction. The same holds for the number of times that a new link appears, as these statistics are also stored in the neighborhood probability distributions. Another characteristic that is entirely captured by the egocentric temporal neighborhoods is the hub-like structure that we can find in each static layer, which is measured by the S-metric (101).

Going beyond these ‘trivial’ consequences of the mechanics of the generating mechanisms, the method does well at preserving interaction durations. The k -steps memory makes various duration possible, even long durations (because of the sliding window) unlike the case of independent layers. Moreover, the ETN distributions also encode the number of times that an interaction ends, so interactions tend not to be extremely long.

Another interesting property is the distribution of edge strengths in the projected graph. Edge strength is simply the number of times that each edge has appeared over the duration of the graph. Here, we would not necessarily expect *ETN-gen* to do well as the method will tend to create networks with quite homogeneous distributions of strength. This is because it can only rely on a memory of order k for edge repetitions, and does not have a long-term memory. Hence all the heterogeneous behaviors that we can find for instance in social datasets, where individuals tend to establish relationships with specific nodes and have repeated (but not necessarily consecutive) interactions with them, are not preserved by *ETN-gen*. Nevertheless, we find that for the considered datasets *ETN-gen* remains

competitive with the other methods.

If edge strength is partially affected by the absence of long memory, the most important limitations of the egocentric perspective are highlighted by clustering, degree assortativity and average shortest path length, which are related to second-order interactions. This is the cost we pay for having a computationally efficient model applicable to arbitrary networks. Notice that while this is a problem in theory, it seems not to affect the workplace dataset, at least for clustering coefficient and average shortest path length. This is explained by the fact that the dataset is substantially sparser, hence characterized by low clustering and short paths. More importantly, it must be stressed that the other approaches also are not able to reproduce these metrics, thus our method is still the most competitive on average.

At this point we note that the limitations with respect to second-order measures could be mitigated during the last step of our new temporal layer generation. In the current version we go from a prospective layer to the actual new layer by matching up nodes with a one-way suggested connection at random. At this step, we could however apply a preferential attachment devoted to maximize or minimize a specific variable. For instance, to maximize clustering we could prefer to keep edges whose nodes have one or more common neighbors, to maximize (minimize) assortativity we could connect stubs with similar (dissimilar) degree.

5.2.3 Dynamical similarity evaluation

Having tested our method from the structural point of view, we now test the usefulness of the surrogate networks in terms of dynamical processes unfolding upon them. We study two dynamical models: random walk and a spreading model.

Random walk

We simulate a temporal random walk (163; 73) on the original and generated networks. The random walk starts in a randomly chosen node and proceeds by moving to a neighbor chosen uniformly at random.

We compute two metrics: coverage and mean first passage time (MFPT), and compare distributions over different realizations between the input and the generated temporal network using again the Kolmogorov-Smirnov distance.

- **Coverage.** The number of (distinct) visited nodes starting from a random node at an initial timestamp (163). The simulation is repeated 1000 times using a random initial node and the initial time is set equal to the first timestamp.
- **Mean First Passage Time (MFPT).** The average time taken by the random walker to arrive for the first time at a specific node i , starting from a random initial position j in the network (163). We consider each couple of nodes (i, j) in the network and repeat the simulation five times for each of them.

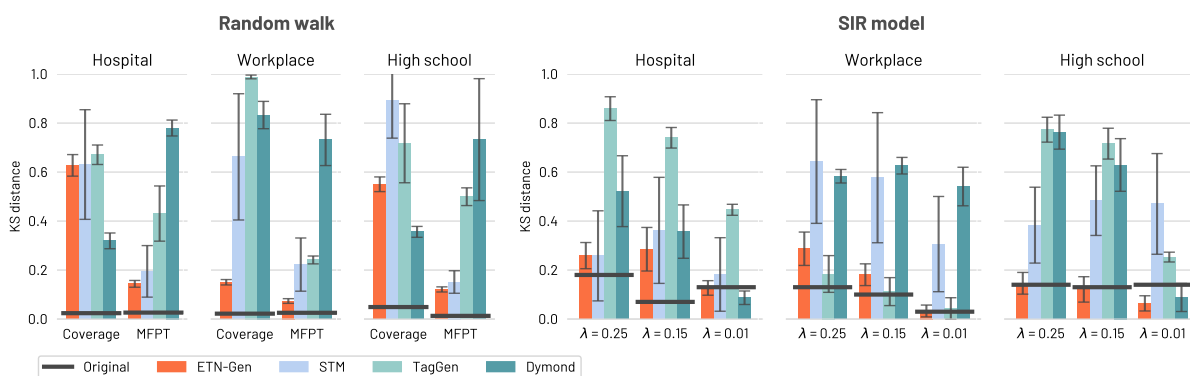


Figure 5.4: **Dynamic similarity.** The figure shows the Kolmogorov-Smirnov distance for coverage and mean first passage time in the random walk model (left panel) and the distance for distribution of R_0 values on a SIR model simulation (right panel) in each generated network. Our method is represented in orange, while the solid black line shows the stability (i.e., the same simulation on the original network).

In Figure 5.4 (left) we report the Kolmogorov-Smirnov distance for coverage and MFPT with our method and the competitors with respect to the original network. The horizontal dashed line shows the stability of each measure on the original network. The black line is obtained comparing different performances (average over 1000 simulations of random walk for coverage, and 5 times each couple of nodes for mean first passage time) by means of the Kolmogorov-Smirnov distance. We observe that in terms of the mean first passage time *ETN-gen* performs better than the competitors, while in terms of coverage performance depends on the datasets: for the most dense networks (hospital and high school) the highest similarity with the original network is achieved by *Dymond* (but *ETN-gen* is ranked second), while for the workplace *ETN-gen* again largely outperforms all the competitors, that produce very different distributions from the original one. In general we can say that the random walk process on the *ETN-gen*'s surrogate networks is quite similar to the random walk on the original graph, especially if compared with surrogate data from the other methods.

Spreading model

We simulate a Susceptible-Infectious-Recovered (SIR) model (5), with three possible values for the probability of disease transmission ($\lambda \in \{0.25, 0.13, 0.01\}$), and the recovery rate fixed at $\mu = 0.055$. In each simulation, the initial infected node is randomly selected among non-isolated nodes.

We compute the reproduction value R_0 defined as follow:

- **Reproduction value R_0 .** The average number of individuals infected by the first one, with a single random node infected as seed.

Each experiment was repeated 100 times and the distribution of R_0 obtained on the original network is, again, compared with those obtained on synthetic networks by means of the Kolmogorov-Smirnov distance. Results are shown in Figure 5.4 (right), where again a horizontal black line shows the stability of each measure on the original network (computed averaging over 100 simulations). We observe that the results obtained with *ETN-gen* are highly similar to those of the original graph and show a large degree of stability with respect to this similarity. The other methods produce networks where the dynamical behavior similarity is sometimes high and sometimes severely low, being quite sensitive to parameters and datasets. In conclusion, the method that we propose almost always produces the most similar networks to the original ones, and certainly the most stable results across datasets, across dynamical systems and across parameters.

5.2.4 Dataset expansion and extension

In the previous sections we have argued that *ETN-gen* creates realistic surrogate temporal networks that mimic real social dynamics (both in terms of structure and in reproducing dynamical systems) – and that our method outperforms alternative solutions.

Now we ask the question: How can this tool be useful in practice? A relevant application is represented by the possibility of enlarging a given temporal dataset, both in time and in size. It is indeed common that a specific analysis, in order to yield reliable results, requires a larger population or a longer time than those characterizing collected real data. In those cases we deal with the long-standing problem of data augmentation, for which we now argue that *ETN-gen* represents a promising solution. In the following we show how our method can be used for augmenting a temporal dataset, by adding temporal layers (temporal extension), but also by increasing the size of the network in terms of number

of nodes (size expansion).

Temporal extension

The temporal extension of a dataset is straight forward: once we have calculated the neighborhood probability distributions summarizing the original graph, we can repeat the process of temporal layer addition as long as needed. At the top of Figure 5.5 we show an example of temporal extension of the workplace network. We have selected this dataset to show that *ETN-gen* is capable of capturing weekends (with no interactions) as well. To evaluate the quality of the extension, we use only the the first week of the original two-week dataset (from the beginning to the vertical line) to estimate the neighborhood probability distributions. We now generate an ensemble of 10 two week networks based on that first week. The mean and standard deviation of the number of interactions in the generated graph are reported in orange. The number of interactions of the original graph are reported in black dashed curves for the first week (the “train” dataset), and in black solid curves for the following week (the “original” dataset). Results show how the generated networks accurately recreate the original behavior beyond the timespan that was used to estimate the local probability distributions.

Size expansion

Here we explore the fidelity of surrogate networks with an increased number of nodes. As discussed above, it is possible to increase the size beyond that of the original network within the *ETN-gen* framework because the number of nodes is simply a parameter to set for the method. That said, however, the concept of size expansion requires more attention than time extension. Because, as we change the number of nodes in a network we should also consider how the density of the graph and the mean degree should change accordingly.

In the following we describe an experiment of data augmentation, assuming that we only have access to incomplete data. Incomplete data are obtained by randomly removing part of the nodes from the original network. We use the high school dataset which, with its 126 nodes, is the largest among our datasets, and we consider two reduced versions, with 30% and 70% of the nodes respectively. When removing part of the nodes from a network, we naturally remove also part of the links (all those which were before connecting the eliminated nodes to the remaining ones), we hence reduce the mean degree. We should consider that an incomplete dataset has in general a reduced mean degree with respect to

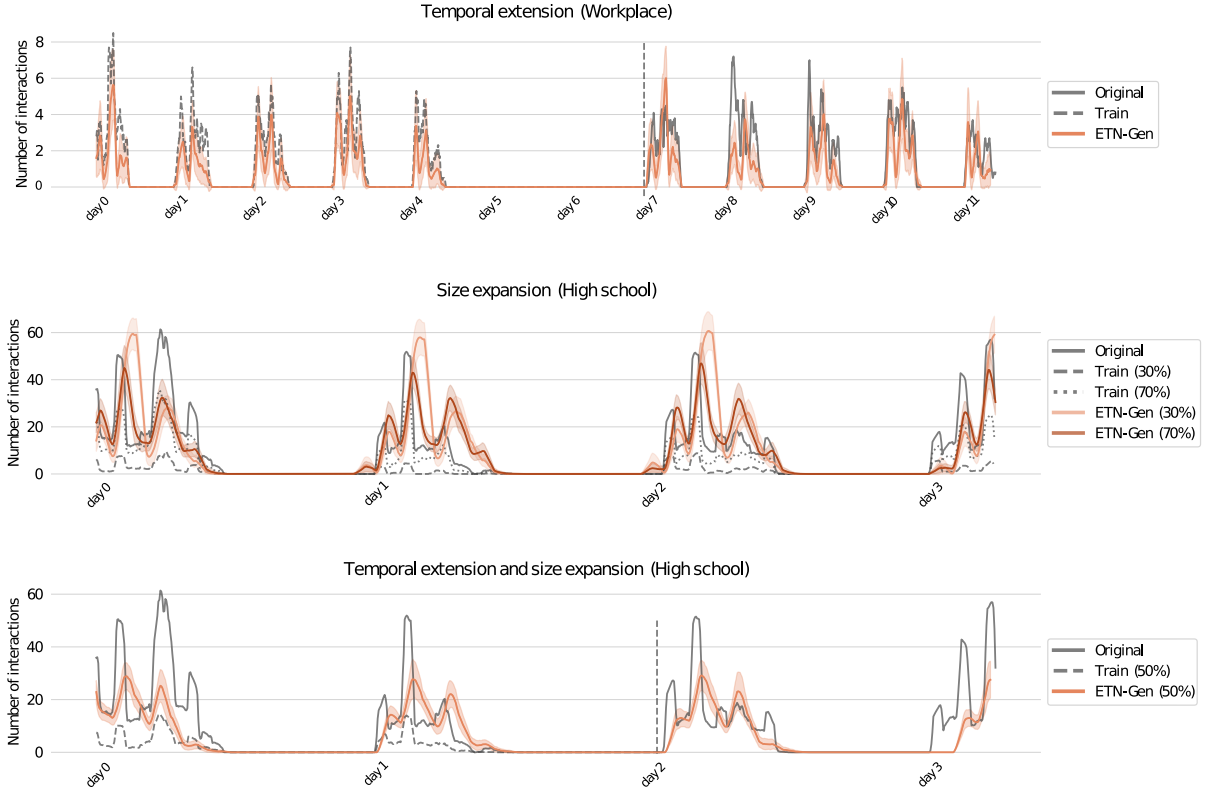


Figure 5.5: **Temporal extension and node expansion.** The mean and standard deviation of our method are shown in orange (and brown). Black dashed (and dotted) lines show the original data used to train our model, while black solid lines show the original data used to evaluate the quality of the generated network. In experiments involving temporal expansion, a vertical bar separates the temporal range used to collect training data from the one where expansion is performed. See the main text for the details.

the real-world network, and that when we try to reconstruct the original network via data augmentation we should increase the mean degree too. See Methods for a quantification of the needed increase.

Anyway, once the desired connectivity has been chosen, *ETN-gen* allows us to generate a surrogate network with the desired number of nodes and the desired degree, while maintaining the pattern of egocentric interactions of the original dataset.

The results of the experiment on the high school dataset are shown in the middle panel of Figure 5.5. For each of the two reduced temporal networks we generate a temporal network with 126 nodes to try to reconstruct the original graph. We generate the initial snapshot using the configuration model based on the degree distribution of the first snapshot of the original (not reduced) graph. Then we build local probability distributions only using

information from the reduced networks and use these local probability distributions to generate surrogate expanded networks from them. The expanded networks have the same number of nodes of the original one (126), enabling direct comparison. The expedient that we use to augment the mean degree from the reduced seed graph is to increase the parameter α of the generation process, which is the probability to confirm the unidirectional directed links in each provisional layer (set to 1/2 by default). See Methods for the details on how to compute the correct value of α given the original number of links and the desired density of the generated graph.

In the middle panel of Figure 5.5 the black solid curve represents the number of interactions in the original network, the black dashed curve those in the “train” network with 30% of the nodes and the black dotted curve those in the one with 70% of the nodes. The corresponding values for the generated networks with their standard deviations are reported in orange and brown respectively. Again, we observe the ability of our method to correctly replicate the pattern of interaction in the original network, even if fed with a small percentage of nodes from the original graph as seed.

Temporal extension and size expansion

We can also combine the two techniques above to simultaneously increase the number of nodes and the temporal snapshots. The results are shown in bottom panel of Figure 5.5 for the high school network, where the synthetic graph has been obtained by only using 50% of the nodes and the first two days of the original dataset (from the beginning to the vertical line), see the black dashed curve. Also in this case, our method is able to extend an input graph in both the temporal and the node size dimensions with remarkable accuracy.

5.3 Discussion

Here, we have proposed a model to generate surrogate temporal networks, i.e. synthetic networks that realistically capture the properties of real-world datasets, only making use of the information contained in egocentric temporal neighborhoods. Specifically, we generate temporal networks which accurately reproduce structural characteristics like density, number of interacting individuals, number of interactions in time, number of new conversations, and the possible presence of hubs.

The fidelity of our surrogate networks suggest that the egocentric temporal neighborhoods

are fundamental building blocks; building blocks which are sufficient in terms of reconstructing temporal networks, which preserve the essential characteristics of the original graph. In this sense our work illustrates the importance of the egocentric perspective in temporal networks (158; 70) opening a new direction in generating these networks.

The usefulness of surrogate networks can be evaluated by simulating dynamical systems on them, such as random walks and a SIR model. We observe that in both topological and dynamical tests, the networks generated by our model are generally closer to the original graph than those generated by different literature models. The comparison with competitors mostly highlights the fact that those models tend to neglect fundamental features that our approach is able to preserve. Indeed, even in the few cases where competing methods reproduce a single measure with slightly higher accuracy, they all have at least one measure exhibiting an extreme difference with the original graph (also including basic features like, e.g. the number of nodes).

Moreover, our approach is able to generate temporal network that have different sizes than the original one. This property can be used to increase the number of nodes and extend the network in time, providing a powerful tool for data augmentation.

The real strength of the method, however, is its simplicity. As noted above, this simplicity reveals something about the minimal fundamental building blocks of a temporal network. The same simplicity, moreover, has allowed us to formulate a fast and scalable algorithm, able to first process and then generate very large networks, with high temporal resolution, something which the existing alternatives cannot do.

The other side of the coin is that this simplicity does not capture certain topological features. This is the main limitation of the model. For instance, disregarding second-order interactions translates to a reduced ability to preserving clustering, degree correlations and average shortest path length. Similarly, the absence of long-term memory means that the model currently does not capture recurrent interactions between pairs of nodes. These features are instead well captured by more theoretical models of network generation that include aging (126; 127), edge reinforcement (56; 166), or in general some mechanism for memory such that contact durations and inter-event times are heterogeneous and depend on the past interactions (149; 174). Memory can also be used to generate a synthetic temporal network that is organized in groups, i.e. subsets of nodes highly connected among them and less connected with the other ones (183; 182). This is a characteristic often occurring in real networks, especially social networks, and it cannot be captured by small local subnetworks like egocentric temporal neighborhoods. However, long-term memory appears in literature only in theoretical models for temporal network generation,

for which the goal is to obtain realistic networks by recovering some particular characteristics of the observed dynamics in real networks, but usually do not aim at reconstructing specific real networks or environments. A model which instead is built to obtain surrogate networks with an alternative approach is the one proposed by Presigny *et al.* (142). This model does not generate a new network from scratch, it instead individuates a backbone of a real temporal network, defined as the global subnetwork composed of the most significant edges, and then reconstructs the missing links. This is based on a conceptually different idea, assuming that the important information concerns the global structure of the network, while the method that we are proposing focuses on how nodes behave given their interactions in last time steps. By recalling two different long-standing traditions in network science, a socio-centric versus an ego-centric perspective (177), we can assert that if the first one is covered, for what concerns surrogate temporal networks, by the model of Presigny *et al.*, our model places itself in the remaining gap, filling the unexplored case of the ego-centric perspective.

Finally, the sequential nature of the method that we are proposing allows us to easily extend it in many directions. For instance using a preferential attachment in the edge validation step of the procedure. Hence many additional features could be included in future developments of the model. Possible future applications may also include the possibility to share sensitive data while preserving privacy and also the possibility of merging data from different environments, simply building multiple local probability distributions.

5.4 Code availability

The codes used for the generation of temporal network are publicly available at

- *ETN-gen* <https://github.com/AntonioLonga/ETNgen>
- *STM* <https://github.com/temporal-graphs/STM>
- *TagGen* <https://github.com/davidchouzdw/TagGen>
- *Dymond* <https://github.com/zeno129/DYMOND>

Chapter 6

Conclusion

In this thesis, we showed how modifying the topology of a temporal network directly affects the dynamic in the network. In the second chapter we showed how modifying the topology of a temporal network directly affects the spreading dynamics taking place on it. The modifications we introduced in the network consisted in quarantining and isolating individuals, practically cutting their connections with the rest of the population. We tested several quarantine policies for COVID-19 containment without pharmaceutical interventions. We explored the trade-off between quarantining the minimum amount of people and containing the infection spreading.

The strong relation between topology and dynamics led us to a deep investigation of the temporal substructure of a given temporal network, proposing the novel notion of *egocentric temporal neighborhood (ETN)* and *egocentric temporal motifs (ETM)*. We proved the benefit of *ETMs* in distinguishing among different social contexts, being able to identify similar networks, for instance: high school students tend to interact with each other in a way that is closer to how university students interact, as opposed to how individuals engage with each other in a professional setting like a workplace or a hospital. Later, being able to decompose temporal networks into ETN, we developed a novel generative algorithm. We compared our algorithm with three state of the art algorithms on 10 different temporal networks, showing a higher ability to generate networks similar to the original ones. We further showed that beyond generating realistic interaction patterns, our method is able to capture the intrinsic temporal periodicity of temporal networks, all with an execution time lower than competing methods by multiple orders of magnitude. Our algorithm is able not only to extend (increase the number of temporal snapshots) but also to expand (increase the number of nodes) a given temporal network. Being able to build bigger realistic temporal networks permits us to tackle an important issue in both

Network Science and CSS. In fact, even in our first work, we had to simulate the epidemic in a relatively small network, thus we tried naive models to generate bigger temporal networks. Those experiments pointed out the difficulty of creating large realistic synthetic temporal networks.

6.1 Limitations & Future Directions

A crucial concept of this thesis relies on the notion of temporal motifs, which need a null model to be defined. However, the choice of null models directly affects the statistical significance of substructures(159). An interesting future direction would be the one to identify statistically significant substructures without relying on null models, bypassing the bias introduced in the definition of the null model. Another crucial issue of temporal motifs consists in their application(106), in fact so far they are counted or used in the generation of surrogate networks. One may explore the usefulness of temporal motifs in understanding how the type of agents interacts with other types of agents, e.g. are the structures created by doctors similar to those created by nurses in the hospital dataset? if not, can we explore these structures for a better understanding of agents in different social contexts? Another promising direction to investigate consists in the possibility to relax the definition of temporal motifs by considering similar motifs as the same. For instance, if the ego node interacts with a group of nodes, the important piece of information may be just the existence of the group, while the exact number of neighbors may not be significant. This possibility is not included in *Egocentric Temporal Motifs* but it could be admitted in a mitigated definition of motifs, like inexact motifs based on similarities. The strength of our method relies on the ability to construct a signature, which allows computing exact matching in polynomial time. The fast computation is due to the fact that second-order connections are not considered. This, however, leads to the leak of triad closure, fundamental in social interactions. To cope with this issue, one may investigate an extension of *Egocentric Temporal Neighborhood Signature* able to handle higher-order networks(19; 15).

We have clearly proven the superiority of our network generation model, not only in terms of similarity with the input network but also in terms of computation time with respect to state of the art algorithms. However, *ETN-gen* is not able to generate specific node identities, i.e. doctors, nurses, patients, etc. This issue could be easily solved by substituting ones and zeros into the *ETNS* with fixed-length identity encoding. This simple solution not only would give the possibility to generate labelled temporal networks but also would open the doors to the challenging task of policy optimization, i.e. it would

be interesting to study for instance how an epidemic spreading is affected by network modifications like addition or removal of specific nodes with specific roles, e.g. teachers in schools or medical doctors in hospitals. Moreover, being able to generate labelled networks would allow generating temporal networks that combine social contexts, e.g. different kinds of interactions during working hours and during free time. This kind of face-to-face interaction data containing multiple individuals and multiple social contexts would be fundamental for epidemic and infodemic studies.

Bibliography

- [1] Matthew Abueg, Robert Hinch, Neo Wu, Luyang Liu, William JM Probert, Austin Wu, Paul Eastham, Yusef Shafi, Matt Rosencrantz, Michael Dikovsky, et al. Modeling the combined effect of digital exposure notification and non-pharmaceutical interventions on the covid-19 epidemic in washington state. *medRxiv*, 2020.
- [2] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659, 2011.
- [3] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.
- [4] Roy M Anderson, Hans Heesterbeek, Don Klinkenberg, and T Déirdre Hollingsworth. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *The Lancet*, 395(10228):931–934, 2020.
- [5] Roy M. Anderson and Robert M. May. *Infectious diseases of humans: dynamics and control*. Oxford Science Publications, 1991.
- [6] James Andreoni and John H Miller. Rational cooperation in the finitely repeated prisoner’s dilemma: Experimental evidence. *The economic journal*, 103(418):570–585, 1993.
- [7] Miguel Araujo, Spiros Papadimitriou, Stephan Günnemann, Christos Faloutsos, Prithwish Basu, Ananthram Swami, Evangelos E Papalexakis, and Danai Koutra. Com2: fast automatic discovery of temporal (‘comet’) communities. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 271–283. Springer, 2014.
- [8] Peter Ashcroft, Jana S Huisman, Sonja Lehtinen, Judith A Bouman, Christian L Althaus, Roland R Regoes, and Sebastian Bonhoeffer. Covid-19 infectivity profile correction. *arXiv preprint arXiv:2007.06602*, 2020.

- [9] Robert Axelrod. Effective choice in the prisoner’s dilemma. *Journal of conflict resolution*, 24(1):3–25, 1980.
- [10] Steve Azzolin, Antonio Longa, Pietro Barbiero, Pietro Liò, and Andrea Passerini. Global explainability of gnns via logic combination of learned concepts. *arXiv preprint arXiv:2210.07147*, 2022.
- [11] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J. Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.
- [12] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.
- [13] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.
- [14] Alain Barrat, Ciro Cattuto, Mikko Kivela, Sune Lehmann, and Jari Saramaki. Effect of manual and digital contact tracing on covid-19 outbreaks: a study on empirical contact data. *medRxiv*, 2020.
- [15] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. Networks beyond pairwise interactions: structure and dynamics. *Physics Reports*, 874:1–92, 2020.
- [16] Michele Berlingerio, Francesco Bonchi, Björn Bringmann, and Aristides Gionis. Mining graph evolution rules. In *Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 115–130. Springer, 2009.
- [17] Michele Berlingerio, Danai Koutra, Tina Eliassi-Rad, and Christos Faloutsos. Net-simile: A scalable approach to size-independent network similarity. *arXiv preprint arXiv:1209.2684*, 2012.
- [18] Qifang Bi, Yongsheng Wu, Shujiang Mei, Chenfei Ye, Xuan Zou, Zhen Zhang, Xiaojian Liu, Lan Wei, Shaun A Truelove, Tong Zhang, et al. Epidemiology and transmission of covid-19 in 391 cases and 1286 of their close contacts in shenzhen, china: a retrospective cohort study. *The Lancet Infectious Diseases*, 20(8):911–919, 2020.
- [19] Ginestra Bianconi. *Higher-order networks*. Cambridge University Press, 2021.
- [20] Ginestra Bianconi and A-L Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54(4):436, 2001.

- [21] Robert A Blair, Benjamin S Morse, and Lily L Tsai. Public health and public trust: Survey evidence from the ebola virus disease epidemic in liberia. *Social Science & Medicine*, 172:89–97, 2017.
- [22] Per Block, Marion Hoffman, Isabel J Raabe, Jennifer Beam Dowd, Charles Rahal, Ridhi Kashyap, and Melinda C Mills. Social network-based distancing strategies to flatten the covid-19 curve in a post-lockdown world. *Nature Human Behaviour*, pages 1–9, 2020.
- [23] Frederic Y Bois and Ghislaine Gayraud. Probabilistic generation of random networks taking into account information on motifs occurrence. *Journal of Computational Biology*, 22(1):25–36, 2015.
- [24] Béla Bollobás, Christian Borgs, Jennifer T Chayes, and Oliver Riordan. Directed scale-free graphs. In *SODA*, volume 3, pages 132–139, 2003.
- [25] William J Bradshaw, Ethan C Alley, Jonathan H Huggins, Alun L Lloyd, and Kevin M Esvelt. Bidirectional contact tracing could dramatically improve covid-19 control. *Nature communications*, 12(1):1–9, 2021.
- [26] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-Francois Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS One*, 5(7):e11596, 2010.
- [27] 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, 07 2010.
- [28] Giulia Cencetti, Gabriele Santin, Antonio Longa, Emanuele Pigani, Alain Barrat, Ciro Cattuto, Sune Lehmann, Marcel Salathe, and Bruno Lepri. Digital proximity tracing on empirical contact networks for pandemic control. *Nature Communications*, 12(1):1–12, 2021.
- [29] Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. COVID-19 data repository. <https://github.com/CSSEGISandData/COVID-19>.
- [30] Diletta Cereda, Marcello Tirani, Francesca Rovida, Vittorio Demicheli, Marco Ajelli, Piero Poletti, Frédéric Trentini, Giorgio Guzzetta, Valentina Marziano, Angelica Barone, et al. The early phase of the covid-19 outbreak in lombardy, italy. *arXiv preprint arXiv:2003.09320*, 2020.
- [31] Gal Chechik, Eugene Oh, Oliver Rando, Jonathan Weissman, Aviv Regev, and Daphne Koller. Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. *Nature Biotechnology*, 26:1251–1259, 2008.

- [32] Hao-Yuan Cheng, Shu-Wan Jian, Ding-Ping Liu, Ta-Chou Ng, Wan-Ting Huang, and Hsien-Ho Lin. Contact tracing assessment of COVID-19 transmission dynamics in Taiwan and risk at different exposure periods before and after symptom onset. *JAMA Internal Medicine*, 2020.
- [33] Derek K Chu, Elie A Akl, Stephanie Duda, Karla Solo, Sally Yaacoub, Holger J Schünemann, Amena El-harakeh, Antonio Bognanni, Tamara Lotfi, Mark Loeb, et al. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of sars-cov-2 and covid-19: a systematic review and meta-analysis. *The Lancet*, 2020.
- [34] Martino Ciaperoni, Edoardo Galimberti, Francesco Bonchi, Ciro Cattuto, Francesco Gullo, and Alain Barrat. Relevance of temporal cores for epidemic spread in temporal networks. *Scientific Reports*, 10:12529, 2020.
- [35] Fulvio Corsi, Fabrizio Lillo, Davide Pirino, and Luca Trapin. Measuring the propagation of financial distress with granger-causality tail risk networks. *Journal of Financial Stability*, 38:18–36, 2018.
- [36] Michele Coscia and Michael Szell. Multiplex graph association rules for link prediction. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM)*, 2021.
- [37] Ana-Maria Cretu, Federico Monti, Stefano Marrone, Xiaowen Dong, Michael Bronstein, and Yves-Alexandre de Montjoye. Interaction data are identifiable even across long periods of time. *Nature Communications*, 13(1):1–11, 2022.
- [38] Laura Di Domenico, Giulia Pullano, Chiara E. Sabbatini, Pierre-Yves Boëlle, and Vittoria Colizza. Expected impact of lockdown in île-de-France and possible exit strategies. *medRxiv*, 2020.
- [39] Sergey N Dorogovtsev, José Fernando F Mendes, and Alexander N Samukhin. Structure of growing networks with preferential linking. *Physical review letters*, 85(21):4633, 2000.
- [40] Raissa M D’souza, Christian Borgs, Jennifer T Chayes, Noam Berger, and Robert D Kleinberg. Emergence of tempered preferential attachment from optimization. *Proceedings of the National Academy of Sciences*, 104(15):6112–6117, 2007.
- [41] Alexis Dudden and Andrew Marks. South Korea took rapid, intrusive measures against Covid-19 - and they worked. *The Guardian*, 20, 2020.

- [42] Daniel M Dunlavy, Tamara G Kolda, and Evrim Acar. Temporal link prediction using matrix and tensor factorizations. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(2):1–27, 2011.
- [43] Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10:255–268, 2008.
- [44] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.
- [45] European Centre for Disease Prevention and Control. Contact tracing: Public health management of persons, including healthcare workers, having had contact with covid-19 cases in the european union. <https://www.ecdc.europa.eu/sites/default/files/documents/covid-19-public-health-management-contact-novel-coronavirus-cases-EU.pdf>.
- [46] European Centre for Disease Prevention and Control. Resource estimation for contact tracing, quarantine and monitoring activities for COVID-19 cases in the eu/eea.
- [47] Neil Ferguson, Daniel Laydon, Gemma Nedjati Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, ZULMA Cucunuba Perez, Gina Cuomo-Dannenburg, et al. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. Technical report, Imperial College London, 2020.
- [48] Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, 2020.
- [49] Josh A Firth, Joel Hellewell, Petra Klepac, Stephen Kissler, Adam J Kucharski, and Lewis G Spurgin. Using a real-world network to model localized covid-19 control strategies. *Nature medicine*, pages 1–7, 2020.
- [50] Julie Fournet and Alain Barrat. Contact patterns among high school students. *PLoS ONE*, 9(9):e107878, 09 2014.
- [51] Julie Fournet and Alain Barrat. Epidemic risk from friendship network data: an equivalence with a non-uniform sampling of contact networks. *Scientific reports*, 6(1):1–11, 2016.

- [52] Christophe Fraser, Steven Riley, Roy M. Anderson, and Neil M. Ferguson. Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences*, 101(16):6146–6151, 2004.
- [53] Tapiwa Ganyani, Cécile Kremer, Dongxuan Chen, Andrea Torneri, Christel Faes, Jacco Wallinga, and Niel Hens. Estimating the generation interval for coronavirus disease (covid-19) based on symptom onset data, march 2020. *Eurosurveillance*, 25(17):2000257, 2020.
- [54] Marino Gatto, Enrico Bertuzzo, Lorenzo Mari, Stefano Miccoli, Luca Carraro, Renato Casagrandi, and Andrea Rinaldo. Spread and dynamics of the covid-19 epidemic in italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences*, 117(19):10484–10491, 2020.
- [55] Laetitia Gauvin, Mathieu Génois, Márton Karsai, Mikko Kivelä, Taro Takaguchi, Eugenio Valdano, and Christian L Vestergaard. Randomized reference models for temporal networks. *arXiv preprint arXiv:1806.04032*, 2018.
- [56] Valeria Gelardi, Didier Le Bail, Alain Barrat, and Nicolas Claidiere. From temporal network data to the dynamics of social relationships. *Proceedings of the Royal Society B*, 288(1959):20211164, 2021.
- [57] Mathieu Génois and Alain Barrat. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science*, 7(1):11, May 2018.
- [58] 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.
- [59] Hossein Gorji, Markus Arnoldini, David F Jenny, Wolf-Dietrich Hardt, and Patrick Jenny. Stecc: Smart testing with contact counting enhances covid-19 mitigation by bluetooth app based contact tracing. *medRxiv*, 2020.
- [60] Trisha Greenhalgh, Manuel B Schmid, Thomas Czypionka, Dirk Bassler, and Laurence Gruer. Face masks for the public during the covid-19 crisis. *Bmj*, 369, 2020.
- [61] Martin Grohe and Pascal Schweitzer. The graph isomorphism problem. *Commun. ACM*, 63(11):128–134, oct 2020.
- [62] Saket Gurukar, Sayan Ranu, and Balaraman Ravindran. Commit: A scalable approach to mining communication motifs from dynamic networks. In *Proceedings of*

- the 2015 ACM SIGMOD International Conference on Management of Data*, pages 475–489, 2015.
- [63] Giorgio Guzzetta, Flavia Riccardo, Valentina Marziano, Piero Poletti, Filippo Trentini, Antonino Bella, Xanthi Andrianou, Martina Del Manso, Massimo Fabiani, Stefania Bellino, et al. Impact of a nationwide lockdown on sars-cov-2 transmissibility, italy. *Emerging infectious diseases*, 27(1):267, 2021.
- [64] Jing-Dong J. Han, Nicolas Bertin, Tong Hao, Debra S. Goldberg, Gabriel F. Berriz, Lan V. Zhang, Denis Dupuy, Albertha J. M. Walhout, Michael E. Cusick, Frederick P. Roth, and Marc Vidal. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430:88–93, 2004.
- [65] Xi He, Eric H. Y. Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y. Wong, Yujuan Guan, Xinghua Tan, Xiaoneng Mo, Yanqing Chen, Baolin Liao, Weilie Chen, Fengyu Hu, Qing Zhang, Mingqiu Zhong, Yanrong Wu, Lingzhai Zhao, Fuchun Zhang, Benjamin J. Cowling, Fang Li, and Gabriel M. Leung. Author correction: Temporal dynamics in viral shedding and transmissibility of covid-19. *Nature Medicine*, 26(9):1491–1493, Sep 2020.
- [66] Xi He, Eric HY Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y Wong, Yujuan Guan, Xinghua Tan, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, pages 1–4, 2020.
- [67] Healthtech. How the NHS COVID-19 app is making the most of cutting-edge global technology, 2020.
- [68] Laurent Hébert-Dufresne, Benjamin M Althouse, Samuel V Scarpino, and Antoine Allard. Beyond R0: Heterogeneity in secondary infections and probabilistic epidemic forecasting. *medRxiv*, 2020.
- [69] Joel Hellewell, Sam Abbott, Amy Gimma, Nikos I Bosse, Christopher I Jarvis, Timothy W Russell, James D Munday, Adam J Kucharski, W John Edmunds, Fiona Sun, Stefan Flasche, Billy J Quilty, Nicholas Davies, Yang Liu, Samuel Clifford, Petra Klepac, Mark Jit, Charlie Diamond, Hamish Gibbs, Kevin [van Zandvoort], Sebastian Funk, and Rosalind M Eggo. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 8(4):e488 – e496, 2020.
- [70] Sara Heydari, Sam G.B. Roberts, Robin I.M. Dunbar, and Jari Saramäki. Multi-channel social signatures and persistent features of ego networks. *Applied Network Science*, 3(1):1–13, 2018.

- [71] Robert Hinch, Will Probert, Anel Nurtay, Michelle Kendall, Chris Wymant, Matthew Hall, Katrina Lythgoe, Ana Bulas Cruz, Lele Zhao, Andrea Stewart, Michael Ferretti, Luca Parker, Ares Meroueh, Bryn Mathias, Scott Stevenson, Daniel Montero, James Warren, Nicole K Mather, Anthony Finkelstein, Lucie Abeler-Dörner, and Christophe Bonsall, David Fraser. Effective configurations of a digital contact tracing app: A report to nhsx, 2020.
- [72] Peter Holme and Jari Saramaki. Temporal networks. *Physics Reports*, 519(3):97–125, 2012.
- [73] Petter Holme. Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88(9):234, 2015.
- [74] Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, 2012.
- [75] Yuriy Hulovatyy, Huili Chen, and Tijana Milenković. Exploring the structure and function of temporal networks with dynamic graphlets. *Bioinformatics*, 31(12):i171–i180, 2015.
- [76] ImmuniApp. The numbers of immuni. <https://www.immuni.italia.it/dashboard.html>, 2020.
- [77] Lorenzo Isella, Mariateresa Romano, Alain Barrat, Ciro Cattuto, Vittoria Colizza, Wouter Van den Broeck, Francesco Gesualdo, Elisabetta Pandolfi, Lucilla Ravà, Caterina Rizzo, and Alberto Eugenio Tozzi. Close encounters in a pediatric ward: Measuring face-to-face proximity and mixing patterns with wearable sensors. *PLoS One*, 6(2):e17144, 2011.
- [78] Ali Jazayeri and Christopher C Yang. Motif discovery algorithms in static and temporal networks: A survey. *arXiv preprint arXiv:2005.09721*, 2020.
- [79] Jayson S. Jia, Xin Lu, Yun Yuan, Ge Xu, Jianmin Jia, and Nicholas A. Christakis. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature*, 2020.
- [80] Gabriel Kaptschuk, Eszter Hargittai, and Elissa M Redmiles. How good is good enough for covid19 apps? the influence of benefits, accuracy, and privacy on willingness to adopt. *arXiv preprint arXiv:2005.04343*, 2020.
- [81] Riivo Kikas, Marlon Dumas, and Márton Karsai. Bursty egocentric network evolution in skype. *Social Network Analysis and Mining*, 3(4):1393–1401, 2013.

- [82] Stephen M. Kissler, Christine Tedijanto, Edward Goldstein, Yonatan H. Grad, and Marc Lipsitch. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science*, 2020.
- [83] Don Klinkenberg, Christophe Fraser, and Hans Heesterbeek. The effectiveness of contact tracing in emerging epidemics. *Plos One*, 1(1):1–7, 12 2006.
- [84] Michael Klompas, Charles A Morris, Julia Sinclair, Madelyn Pearson, and Erica S Shenoy. Universal masking in hospitals in the covid-19 era. *New England Journal of Medicine*, 382(21):e63, 2020.
- [85] Sadamori Kojaku, Laurent Hébert-Dufresne, and Yong-Yeol Ahn. The effectiveness of contact tracing in heterogeneous networks. *arXiv preprint arXiv:2005.02362*, 2020.
- [86] Joel R Koo, Alex R Cook, Minah Park, Yinxiaohe Sun, Haoyang Sun, Jue Tao Lim, Clarence Tam, and Borame L Dickens. Interventions to mitigate early spread of SARS-CoV-2 in singapore: a modelling study. *The Lancet Infectious Diseases*, 2020.
- [87] Gregory Kossinets, Jon Kleinberg, and Duncan Watts. The structure of information pathways in a social communication network. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 435–443, 2008.
- [88] Gregory Kossinets and Duncan Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.
- [89] Orestis Kostakis, Nikolaï Tatti, and Aristides Gionis. Discovering recurring activity in temporal networks. *Data Mining and Knowledge Discovery*, 31(6):1840–1871, 2017.
- [90] Chrysanthi Kosyfaki, Nikos Mamoulis, Evaggelia Pitoura, and Panayiotis Tsaparas. Flow motifs in interaction networks. *arXiv preprint arXiv:1810.08408*, 2018.
- [91] Lauri Kovanen, Márton Karsai, Kimmo Kaski, János Kertész, and Jari Saramäki. Temporal motifs in time-dependent networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(11):P11005, 2011.
- [92] Paul L Krapivsky, Sidney Redner, and Francois Leyvraz. Connectivity of growing random networks. *Physical review letters*, 85(21):4629, 2000.

- [93] Adam J Kucharski, Petra Klepac, Andrew JK Conlan, Stephen M Kissler, Maria L Tang, Hannah Fry, Julia R Gog, W John Edmunds, Jon C Emery, Graham Medley, et al. Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of sars-cov-2 in different settings: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(10):1151–1160, 2020.
- [94] Veronica Lachi, Giovanna Maria Dimitri, Alessandro Di Stefano, Pietro Liò, Monica Bianchini, and Chiara Mocenni. Impact of the covid 19 outbreaks on the italian twitter vaccination debat: a network based analysis, 2023.
- [95] Renaud Lambiotte, Martin Rosvall, and Ingo Scholtes. From networks to optimal higher-order models of complex systems. *Nature Physics*, 15:313–320, 2019.
- [96] Enrico Lavezzo, Elisa Franchin, Constanze Ciavarella, Gina Cuomo-Dannenburg, Luisa Barzon, Claudia Del Vecchio, Lucia Rossi, Riccardo Manganelli, Arianna Loregian, Nicolò Navarin, et al. Suppression of COVID-19 outbreak in the municipality of Vo, Italy. *medRxiv*, 2020.
- [97] Le Conseil Général de l’Economie, de l’Industrie, de l’Energie et des Technologies (CGE), l’Autorité de Régulation des Communications Electroniques et des Postes (ARCEP) et l’Agence du numérique. Baromètre du numérique 2019. <https://www.credoc.fr/publications/barometre-du-numerique-2019>, 2019.
- [98] Douglas J Leith and Stephen Farrell. Contact tracing app privacy: What data is shared by europe’s gaen contact tracing apps. *Testing Apps for COVID-19 Tracing (TACT)*, 2020.
- [99] Douglas J Leith and Stephen Farrell. Coronavirus contact tracing app privacy: What data is shared by the singapore opentrace app? In *International Conference on Security and Privacy in Communication Systems*, pages 80–96. Springer, 2020.
- [100] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es, 2007.
- [101] Lun Li, David Alderson, John C Doyle, and Walter Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2(4):431–523, 2005.
- [102] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy S.M. Leung, Eric H.Y. Lau, Jessica Y. Wong, Xuesen Xing, Nijuan Xiang, Yang Wu, Chao Li, Qi Chen, Dan Li, Tian Liu, Jing Zhao, Man Liu, Wenxiao Tu,

- Chuding Chen, Lianmei Jin, Rui Yang, Qi Wang, Suhua Zhou, Rui Wang, Hui Liu, Yinbo Luo, Yuan Liu, Ge Shao, Huan Li, Zhongfa Tao, Yang Yang, Zhiqiang Deng, Boxi Liu, Zhitao Ma, Yanping Zhang, Guoqing Shi, Tommy T.Y. Lam, Joseph T. Wu, George F. Gao, Benjamin J. Cowling, Bo Yang, Gabriel M. Leung, and Zijian Feng. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 382(13):1199–1207, 2020. PMID: 31995857.
- [103] Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science*, 368(6490):489–493, 2020.
- [104] Ruoran Li, Caitlin Rivers, Qi Tan, Megan B Murray, Eric Toner, and Marc Lipsitch. The demand for inpatient and ICU beds for COVID-19 in the US lessons from Chinese cities. *medRxiv*, 2020.
- [105] Paul Liu, Austin R Benson, and Moses Charikar. Sampling methods for counting temporal motifs. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 294–302, 2019.
- [106] Penghang Liu, Valerio Guarrasi, and Ahmet Erdem Sariyüce. Temporal network motifs: Models, limitations, evaluation. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):945–957, 2021.
- [107] Yang Liu, Li-Meng Yan, Lagen Wan, Tian-Xin Xiang, Aiping Le, Jia-Ming Liu, Malik Peiris, Leo LM Poon, and Wei Zhang. Viral dynamics in mild and severe cases of covid-19. *The Lancet Infectious Diseases*, 2020.
- [108] Antonio Longa, Steve Azzolin, Gabriele Santin, Giulia Cencetti, Pietro Liò, Bruno Lepri, and Andrea Passerini. Explaining the explainers in graph neural networks: a comparative study. *arXiv preprint arXiv:2210.15304*, 2022.
- [109] Antonio Longa, Giulia Cencetti, Sune Lehmann, Andrea Passerini, and Bruno Lepri. Neighbourhood matching creates realistic surrogate temporal networks. *arXiv preprint arXiv:2205.08820*, 2022.
- [110] Antonio Longa, Giulia Cencetti, Bruno Lepri, and Andrea Passerini. An efficient procedure for mining egocentric temporal motifs. *Data Mining and Knowledge Discovery*, 2021.
- [111] Antonio Longa, Veronica Lachi, Gabriele Santin, Monica Bianchini, Bruno Lepri, Pietro Lio, Franco Scarselli, and Andrea Passerini. Graph neural networks for tem-

- poral graphs: State of the art, open challenges, and opportunities. *arXiv preprint arXiv:2302.01018*, 2023.
- [112] Jesús A Moreno López, Beatriz Arregui-García, Piotr Bentkowski, Livio Bioglio, Francesco Pinotti, Pierre-Yves Boëlle, Alain Barrat, Vittoria Colizza, and Chiara Poletto. Anatomy of digital contact tracing: role of age, transmission setting, adoption and case detection. *medRxiv*, 2020.
- [113] Lars Lorch, William Trouleau, Stratis Tsirtsis, Aron Szanto, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. A spatiotemporal epidemic model to quantify the effects of contact tracing. *Testing, and Containment*, 2020.
- [114] Massimiliano Luca, Gian Maria Campedelli, Simone Centellegher, Michele Tizzoni, and Bruno Lepri. Crime, inequality and public health: A survey of emerging trends in urban data science. *Frontiers in Big Data*, 6:50, 2023.
- [115] Massimiliano Luca, Bruno Lepri, Enrique Frias-Martinez, and Andra Lutu. Modeling international mobility using roaming cell phone traces during covid-19 pandemic. *EPJ Data Science*, 11(1):22, 2022.
- [116] R Duncan Luce and Albert D Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.
- [117] Shujuan Ma, Jiayue Zhang, Minyan Zeng, Qingping Yun, Wei Guo, Yixiang Zheng, Shi Zhao, Maggie H Wang, and Zuyao Yang. Epidemiological parameters of covid-19: Case series study. *Journal of medical Internet research*, 22(10):e19994, 2020.
- [118] Ying Mao, Susiyan Jiang, Daniel Nametz, Yuxin Lin, Jake Hack, John Hensley, Ryan Monaghan, and Tess Gutenbrunner. Data-driven analytical models of covid-2019 for epidemic prediction, clinical diagnosis, policy effectiveness and contact tracing: A survey, 2020.
- [119] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.
- [120] 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), 2015.
- [121] 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), 2015.

- [122] Giovanni Mauro, Massimiliano Luca, Antonio Longa, Bruno Lepri, and Luca Pappalardo. Generating mobility networks with generative adversarial networks. *EPJ data science*, 11(1):58, 2022.
- [123] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.
- [124] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.
- [125] Kenji Mizumoto, Katsushi Kagaya, Alexander Zarebski, and Gerardo Chowell. Estimating the asymptomatic proportion of coronavirus disease 2019 (covid-19) cases on board the diamond princess cruise ship, yokohama, japan, 2020. *Eurosurveillance*, 25(10):2000180, 2020.
- [126] Antoine Moinet, Michele Starnini, and Romualdo Pastor-Satorras. Burstiness and aging in social temporal networks. *Physical review letters*, 114(10):108701, 2015.
- [127] Antoine Moinet, Michele Starnini, and Romualdo Pastor-Satorras. Aging and percolation dynamics in a non-poissonian temporal network model. *Physical Review E*, 94(2):022316, 2016.
- [128] Anders Mollgaard, Sune Lehmann, and Joachim Mathiesen. Correlations between human mobility and social interaction reveal general activity patterns. *PLoS One*, 12(12):e0188973, 2017.
- [129] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180, 1995.
- [130] Enys Mones, Arkadiusz Stopczynski, Alex ‘Sandy’ Pentland, Nathaniel Hupert, and Sune Lehmann. Optimizing targeted vaccination across cyber-physical networks: an empirically based mathematical simulation study. *Journal of The Royal Society Interface*, 15(138):20170783, 2018.
- [131] Jessica Morley, Josh Cows, Mariarosaria Taddeo, and Luciano Floridi. Ethical guidelines for covid-19 tracing apps, 2020.
- [132] Mark Newman. *Network: An introduction*. Oxford University Press, 2010.
- [133] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.
- [134] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2):026118, 2001.

- [135] Hiroshi Nishiura, Tetsuro Kobayashi, Takeshi Miyama, Ayako Suzuki, Sungmok Jung, Katsuma Hayashi, Ryo Kinoshita, Yichi Yang, Baoyin Yuan, Andrei R Akhmetzhanov, et al. Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). *medRxiv*, 2020.
- [136] Nuria Oliver, Bruno Lepri, Harald Sterly, Renaud Lambiotte, Sébastien Delataille, Marco De Nadai, Emmanuel Letouzé, Albert Ali Salah, Richard Benjamins, Ciro Cattuto, Vittoria Colizza, Nicolas de Cordes, Samuel P. Fraiberger, Till Koebe, Sune Lehmann, Juan Murillo, Alex Pentland, Phuong N Pham, Frédéric Pivetta, Jari Saramäki, Samuel V. Scarpino, Michele Tizzoni, Stefaan Verhulst, and Patrick Vinck. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Science Advances*, 2020.
- [137] Daniel Oran and Eric Topol. Prevalence of asymptomatic sars-cov-2 infection. *Annals of Internal Medicine*, 0(0):null, 0. PMID: 32491919.
- [138] Fragkiskos Papadopoulos, Maksim Kitsak, M Serrano, Marián Boguná, and Dmitri Krioukov. Popularity versus similarity in growing networks. *Nature*, 489(7417):537–540, 2012.
- [139] Ashwin Paranjape, Austin R Benson, and Jure Leskovec. Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 601–610, 2017.
- [140] Nicola Perra, Bruno Gonçalves, Romualdo Pastor-Satorras, and Alessandro Vespignani. Activity driven modeling of time varying networks. *Scientific reports*, 2(1):1–7, 2012.
- [141] Francesco Pinotti, Laura Di Domenico, Ernesto Ortega, Marco Mancastropa, Giulia Pullano, Eugenio Valdano, Pierre-Yves Boelle, Chiara Poletto, and Vittoria Colizza. Lessons learnt from 288 COVID-19 international cases: importations over time, effect of interventions, underdetection of imported cases. *medRxiv*, 2020.
- [142] Charley Presigny, Petter Holme, and Alain Barrat. Building surrogate temporal network data from observed backbones. *Physical Review E*, 103(5):052304, 2021.
- [143] Sumit Purohit, Lawrence B Holder, and George Chin. Temporal graph generation based on a distribution of temporal motifs. In *Proceedings of the 14th International Workshop on Mining and Learning with Graphs*, volume 7, 2018.
- [144] Benjamin Rader, Laura F White, Michael R Burns, Jack Chen, Joseph Brilliant, Jon Cohen, Jeffrey Shaman, Larry Brilliant, Moritz UG Kraemer, Jared B Hawkins,

- et al. Mask-wearing and control of sars-cov-2 transmission in the usa: a cross-sectional study. *The Lancet Digital Health*, 2021.
- [145] Anatol Rapoport, Albert M Chammah, and Carol J Orwant. *Prisoner's dilemma: A study in conflict and cooperation*, volume 165. University of Michigan press, 1965.
- [146] Ramesh Raskar, Isabel Schunemann, Rachel Barbar, Kristen Vilcans, Jim Gray, Praneeth Vepakomma, Suraj Kapa, Andrea Nuzzo, Rajiv Gupta, Alex Berke, et al. Apps gone rogue: Maintaining personal privacy in an epidemic. *arXiv preprint arXiv:2003.08567*, 2020.
- [147] Abhik Ray, Larry Holder, and Sutanay Choudhury. Frequent subgraph discovery in large attributed streaming graphs. In *Proceedings of the 3rd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, pages 166–181, 2014.
- [148] E Rea, J Lafleche, S Stalker, BK Guarda, H Shapiro, I Johnson, SJ Bondy, R Upshur, ML Russell, and M Eliasziw. Duration and distance of exposure are important predictors of transmission among community contacts of ontario sars cases. *Epidemiology & Infection*, 135(6):914–921, 2007.
- [149] Luis EC Rocha and Vincent D Blondel. Bursts of vertex activation and epidemics in evolving networks. *PLoS computational biology*, 9(3):e1002974, 2013.
- [150] Luis EC Rocha, Naoki Masuda, and Petter Holme. Sampling of temporal networks: Methods and biases. *Physical Review E*, 96(5):052302, 2017.
- [151] Ryan A. Rossi and Nesreen K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, 2015.
- [152] Polina Rozenshtein, Giulia Preti, Aristides Gionis, and Yannis Velegrakis. Mining dense subgraphs with similar edges. *arXiv preprint arXiv:2007.03950*, 2020.
- [153] Polina Rozenshtein, Nikolaj Tatti, and Aristides Gionis. Finding dynamic dense subgraphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(3):1–30, 2017.
- [154] Henrik Salje, Cécile Tran Kiem, Noémie Lefrancq, Noémie Courtejoie, Paolo Bosetti, Juliette Paireau, Alessio Andronico, Nathanaël Hozé, Jehanne Richet, Claire-Lise Dubost, Yann Le Strat, Justin Lessler, Daniel Levy-Bruhl, Arnaud Fontanet, Lulla Opatowski, Pierre-Yves Boelle, and Simon Cauchemez. Estimating the burden of SARS-CoV-2 in France. *Science*, 2020.

- [155] Hazem Peter Samoaa, Antonio Longa, Mazen Mohamad, Morteza Haghiri Chehrehgani, and Philipp Leitner. Tep-gnn: Accurate execution time prediction of functional tests using graph neural networks. In *Product-Focused Software Process Improvement: 23rd International Conference, PROFES 2022, Jyväskylä, Finland, November 21–23, 2022, Proceedings*, pages 464–479. Springer, 2022.
- [156] Piotr Sapiezynski, Arkadiusz Stopczynski, David Dreyer Lassen, and Sune Lehmann. Interaction data from the copenhagen networks study. *Scientific Data*, 6(1):1–10, 2019.
- [157] Piotr Sapiezynski, Arkadiusz Stopczynski, David Dreyer Lassen, and Sune Lehmann. Interaction data from the Copenhagen Networks Study. *Scientific Data*, 6(315), 2019.
- [158] Jari Saramäki, Elizabeth A. Leicht, Eduardo López, Sam G.B. Roberts, Felix Reed-Tsochas, and Robin I.M. Dunbar. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences*, 111(3):942–947, 2014.
- [159] Wolfgang E Schlauch and Katharina A Zweig. Influence of the null-model on motif detection. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 514–519, 2015.
- [160] Vedran Sekara and Sune Lehmann. The strength of friendship ties in proximity sensor data. *Plos One*, 9(7):e100915, 2014.
- [161] Timo Smieszek. A mechanistic model of infection: why duration and intensity of contacts should be included in models of disease spread. *Theoretical Biology and Medical Modelling*, 6(1):25, 2009.
- [162] 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):341, 2016.
- [163] Michele Starnini, Andrea Baronchelli, Alain Barrat, and Romualdo Pastor-Satorras. Random walks on temporal networks. *Physical Review E*, 85(5):056115, 2012.
- [164] Michele Starnini and Romualdo Pastor-Satorras. Topological properties of a time-integrated activity-driven network. *Physical Review E*, 87(6):062807, 2013.
- [165] Statista.de. Anzahl der Downloads der Corona-Warn-App über den Apple App Store und den Google Play Store in Deutschland von Juni bis November

2020. <https://de.statista.com/statistik/daten/studie/1125951/umfrage/downloads-der-corona-warn-app/>, 2020.
- [166] Juliette Stehlé, Alain Barrat, and Ginestra Bianconi. Dynamical and bursty interactions in social networks. *Physical review E*, 81(3):035101, 2010.
- [167] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Vittoria Colizza, Lorenzo Isella, Corinne Régis, Jean-François Pinton, Nagham Khanafer, Wouter Van den Broeck, et al. Simulation of an seir infectious disease model on the dynamic contact network of conference attendees. *BMC medicine*, 9(1):1–15, 2011.
- [168] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one*, 6(8):e23176, 2011.
- [169] Chayant Tantipathananandh, Tanya Berger-Wolf, and David Kempe. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 717–726, 2007.
- [170] Carmela Troncoso, Mathias Payer, Jean-Pierre Hubaux, Marcel Salathé, James Larus, Edouard Bugnion, Wouter Lueks, Theresa Stadler, Apostolos Pyrgelis, Daniele Antonioli, et al. Decentralized privacy-preserving proximity tracing. *arXiv preprint arXiv:2005.12273*, 2020.
- [171] Jay J Van Bavel, Katherine Baicker, Paulo S Boggio, Valerio Capraro, Aleksandra Cichocka, Mina Cikara, Molly J Crockett, Alia J Crum, Karen M Douglas, James N Druckman, et al. Using social and behavioural science to support covid-19 pandemic response. *Nature Human Behaviour*, pages 1–12, 2020.
- [172] Philippe Vanhems, Alain Barrat, Ciro Cattuto, Jean-François Pinton, Nagham Khanafer, Corinne Régis, Byeul-a Kim, Brigitte Comte, and Nicolas Voirin. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS One*, 8(9):e73970, 2013.
- [173] Philippe Vanhems, Nicolas Voirin, Sylvain Roche, Vanessa Escuret, Corinne Régis, Christine Gorain, Silène Pires-Cronenberger, Marine Giard, Bruno Lina, Fatiha Najjoulah, et al. Risk of influenza-like illness in an acute health care setting during community influenza epidemics in 2004-2005, 2005-2006, and 2006-2007: a prospective study. *Archives of internal medicine*, 171(2):151–157, 2011.

- [174] Christian L Vestergaard, Mathieu Génois, and Alain Barrat. How memory generates heterogeneous dynamics in temporal networks. *Physical Review E*, 90(4):042805, 2014.
- [175] Patrick Vinck, Phuong N Pham, Kenedy K Bindu, Juliet Bedford, and Eric J Nilles. Institutional trust and misinformation in the response to the 2018–19 ebola outbreak in north kivu, dr congo: a population-based survey. *The Lancet Infectious Diseases*, 19(5):529–536, 2019.
- [176] Jingjing Wang, Yanhao Wang, Wenjun Jiang, Yuchen Li, and Kian-Lee Tan. Efficient sampling algorithms for approximate temporal motif counting. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1505–1514, 2020.
- [177] Stanley Wasserman, Katherine Faust, et al. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.
- [178] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’ networks. *nature*, 393(6684):440–442, 1998.
- [179] Giselle Zeno, Timothy La Fond, and Jennifer Neville. Dymond: Dynamic motif-nodes network generative model. In *Proceedings of the Web Conference 2021*, pages 718–729, 2021.
- [180] Juanjuan Zhang, Maria Litvinova, Yuxia Liang, Yan Wang, Wei Wang, Shanlu Zhao, Qianhui Wu, Stefano Merler, Cécile Viboud, Alessandro Vespignani, Marco Ajelli, and Hongjie Yu. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science*, 2020.
- [181] Juanjuan Zhang, Maria Litvinova, Wei Wang, Yan Wang, Xiaowei Deng, Xinghui Chen, Mei Li, Wen Zheng, Lan Yi, Xinhua Chen, et al. Evolving epidemiology of novel coronavirus diseases 2019 and possible interruption of local transmission outside Hubei Province in China: a descriptive and modeling study. *medRxiv*, 2020.
- [182] Xiao Zhang, Cristopher Moore, and Mark EJ Newman. Random graph models for dynamic networks. *The European Physical Journal B*, 90(10):1–14, 2017.
- [183] Kun Zhao, Juliette Stehlé, Ginestra Bianconi, and Alain Barrat. Social network dynamics of face-to-face interactions. *Physical review E*, 83(5):056109, 2011.
- [184] Qiankun Zhao, Yuan Tian, Qi He, Nuria Oliver, Ruoming Jin, and Wang-Chien Lee. Communication motifs: a tool to characterize social communications. In

Proceedings of the 19th ACM international conference on Information and knowledge management, pages 1645–1648, 2010.

- [185] Dawei Zhou, Lecheng Zheng, Jiawei Han, and Jingrui He. A data-driven graph generative model for temporal interaction networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 401–411, 2020.

Chapter 7

Appendix

7.1 Digital Contact Tracing

7.1.1 Characteristic parameters of the disease

In this section we provide details on the various parameters that represent the epidemic spread in both the continuous model and the network model. Moreover, we demonstrate that the model is robust with respect to the choice of the infectiousness probability as a function of the time since infection.

Infectiousness parameters in the continuous model

The choice of the infectiousness function and the epidemic parameters that describe the COVID-19 spreading in the continuous model follow the work of Ferretti et al. (48), with some modifications that we describe here and summarize in Supplementary Table 7.2.

The infectiousness $\omega(\tau)$ is a function of the days since infection, proposed by Ferretti et al. (48). It takes into account four different contributions: asymptomatic, pre-symptomatic and symptomatic infectiousness, plus environmental transmission representing the indirect contagion occurring for instance via contaminated surfaces. The symptomatic infectiousness has been obtained by Ferretti et al. by making use of generation time data. The pre-symptomatic infectiousness is assumed to be equal to the symptomatic one, while the asymptomatic individuals are considered to have only 10% of the infection potential, according to the recent literature (107; 103). An alternative shape of the curve $\omega(\tau)$ is discussed in 7.1.1. The infectiousness is a probability distribution and as such it is nor-

malized to one. It appears in the model equation (3.1) in the main text multiplied by R_0 , that we consider equal to 3 when no measure is implemented. All the analyses are however performed using reduced values, $R_0 = 1.2, 1.5, 2.0$, which take into account the combined effect of all the alternative measures (masks, physical distancing, etc.) in a range suggested by recent literature (33; 144; 84; 60).

For the cumulative distribution $s(\tau)$ of onset times (i.e. time between infection and appearance of symptoms), we adopt the assumptions of Ferretti et al. (48) with two modifications. This function actually gives the fraction of the infected population that becomes known as infected by the health authorities, and does not distinguish between symptomatic individuals and asymptomatics identified by randomized testing. This is the same assumption as in Ferretti et al. (48), and it is motivated by the fact that the tracing and quarantining policy is activated independently of the source of knowledge of the infected status. The first modification to the onset time is that we rescale the function s so that its cumulative probability $s(\tau)$ reaches $p = 0.8$ at large times instead of 1. This models our assumption that even at infinite time only 80% of the infected population is detected, instead of 100%. This describes a situation in which 60% of infected are symptomatic, and additionally 50% of asymptomatics are identified by randomized testing, or equivalently to a situation with 80% symptomatics and no randomized testing. The second modification is that we shift the symptom onset forward in time by 2 days, modelling a delay in the functioning of the testing and reporting policy. Different assumptions on this delay are discussed in 7.1.2.

Name	Inputs	Definition	Description
$\omega(\tau)$	time τ (days)	Weibull distribution with shape = 2.826 and scale = 5.665.	Probability for an infected individual to transmit the disease at time τ .
R_0		1.2, 1.5, 2	Reproductive number.
onset_time(τ)	time τ (days)	Lognormal distribution with $\mu = 1.54$, $\sigma = 0.47$, shifted by the delay of 2 days, and scaled in $[0, 0.8]$.	Probability for an infected individual to be detected exactly at time τ .
$s(\tau)$	time τ (days)	Cumulative distribution of onset_time(τ).	Probability for an infected individual to be detected within time τ .

Table 7.2: Characteristic parameters of the disease that are used in the continuous model.

Parameter tuning to validate the infection probabilities

As mentioned in the main text in Section 3.2.2, the CNS data set provides us with the opportunity to explore the dependence of the infectiousness from duration and proximity, a question to which the literature is not yet able to express a specific answer. We rely on some simplifying assumptions by supposing that in occasion of a contact between an infected and a susceptible person the contagion probability depends only on their proximity, on the duration of the contact and on the time since the infectious individual has been infected. We moreover assume that those probabilities are independent from each other and require that, if simulated on the CNS data set without any restriction, the resulting reproductive number is equal to $R_0 = 3$, in agreement with recent literature on the COVID-19. Given a choice of the infectiousness parameters, the corresponding value of R_0 is estimated by computing an empirical value R_0^{data} . This is obtained by numerically simulating the epidemic spreading, assuming one random individual initially infected, and counting the number of secondary infections caused by this patient zero (167). The average of this value over multiple independent runs is the estimated value R_0^{data} .

The infectiousness function is thus defined as:

$$\omega_{\text{data}}(\tau, e, s_s) = r_{R_0} \cdot p_{R_0} \cdot \omega(\tau) \cdot \omega_{\text{exposure}}(e) \cdot \omega_{\text{dist}}(s_s) \quad (7.1)$$

where $\omega(\tau)$ is the probability for an infected individual to transmit the disease at time τ after its own infection, $\omega_{\text{exposure}}(e)$ is the probability to transmit the disease given the duration e of a contact, and $\omega_{\text{dist}}(s_s)$ is the probability as a function of the signal strength s_s of the contact. The constant r_{R_0} is a reduction factor that can be tuned to obtain the desired value of R_0 , and p_{R_0} is a scaling factor. Using two distinct scaling factors allows us to decouple the estimate of the parameters to obtain the target value of $R_0 = 3$, and the computation of the reduction factor needed to obtain a smaller value.

Considering everything fixed except for $\omega_{\text{exposure}}(e)$ and $\omega_{\text{dist}}(s_s)$ we can play with the free parameters of these functions so as to explore different scenarios while keeping a balance between time and space dependencies corresponding to an R_0 around 3 (with $r_{R_0} = 1$).

The shape of ω_{exposure} has been inspired by the literature (51; 167; 13):

$$\omega_{\text{exposure}}(e) = (1 - \beta_0)^{e/dt}, \quad (7.2)$$

where dt is a time step and β_0 a free parameter. The value of β_0 can be set by requiring that a specific probability σ for an infected individual to transmit the disease is reached for a given contact duration e_σ :

$$\omega_{\text{exposure}}(e_\sigma) = \sigma. \quad (7.3)$$

The parameter β_0 can thus be expressed as a function of e_σ and σ as:

$$\beta_0(e_\sigma, \sigma) = 1 - (1 - \sigma)^{dt/e_\sigma}. \quad (7.4)$$

Supplementary Table 7.3 reports some examples. For instance, to obtain a 90% probability of infection for contacts of 1 hour, the parameter β_0 needs to be set equal to 0.038.

	e_σ [hours]	σ	β_0
●	1.0	0.9	0.038
●	2.0	0.9	0.019
●	4.0	0.9	0.010

Table 7.3: Numerical values for β_0 for three different sets of physical scenarios (e_σ, σ) . The value of β_0 highlighted in bold is the one chosen for the simulations reported in all the other sections.

The term $\omega_{\text{dist}}(s_s)$ instead depends on the Bluetooth signal strength (RSSI), expressed in dBm, which is considered as a proxy for the distance between individuals. We thus define the function $\tilde{\omega}(x) = \omega_{\text{dist}}(s_s(x))$, where x indicates distances in meters. We emphasize here again that the relationship between RSSI and distance is far from trivial (160; 130), so in the main text we will rely on signal strength as a proxy for distance.

To our knowledge, the literature on COVID-19 has not yet produced some evidence regarding the probability of contagion as a function of the distance between an infected individual and a susceptible one. We make the realistic assumption that infectiousness is large when the individuals are in close proximity and that it decreases with distance. In particular we hypothesize that it follows a sigmoid function:

$$\tilde{\omega}_{\text{dist}}(x) = \frac{s}{\log(1 + e^b)} \left(1 - \frac{1}{1 + e^{b-sx}} \right), \quad (7.5)$$

where s and b are free parameters. As we have two parameters, we need to specify two physical conditions to find their values. We then require that the probability for an infected individual to transmit the disease to a contact within a distance x_i ($i = 1, 2$) should be w_i ($i = 1, 2$):

$$\begin{cases} \int_0^{x_1} \tilde{\omega}_{\text{dist}}(x) dx = w_1 \\ \int_0^{x_2} \tilde{\omega}_{\text{dist}}(x) dx = w_2 . \end{cases} \quad (7.6)$$

Computing explicitly the integrals using Eq. (7.5), we obtain

$$\begin{cases} 1 - \frac{\log(1 + e^{b-sx_1})}{\log(1 + e^b)} = w_1 \\ 1 - \frac{\log(1 + e^{b-sx_2})}{\log(1 + e^b)} = w_2 \end{cases} \quad (7.7)$$

which is a transcendental system, that can be numerically solved once we have set the two couples $(x_i, w_i)_{i=1,2}$. Some examples are reported in Supplementary Table 7.4.

	$x_1 [m]$	w_1	$x_2 [m]$	w_2	$s [m^{-1}]$	b
●	1.7	0.5	6.0	0.99	1.16	3.65
●	2.5	0.5	7.0	0.99	1.34	6.67
●	4.0	0.5	10.0	0.99	1.16	9.31

Table 7.4: Numerical solutions (s, b) for the system (7.7) for three different sets of physical requests $(x_i, w_i)_{i=1,2}$. The values of s and b highlighted in bold are the ones chosen for the simulations reported in all the other sections.

The three curves that we obtain using the values in Supplementary Table 7.3 and Supplementary Table 7.4 are shown in Supplementary Fig. 7.1.

While the reproductive number of COVID-19 is estimated to be around 3 (30), there is small evidence for the dependence on proximity and duration. Therefore, we combine the two functions $\omega_{\text{exposure}}(e)$ and $\tilde{\omega}_{\text{dist}}(x)$ and choose the parameters β_0 , b and s to obtain $R_0 = 3$ in each combination. In particular, given a possible choice for (β_0, b, s) , we run a set of 800 simulations on the CNS data set without any restrictive policy, i.e. with $\varepsilon_I = 0$ and one initial infected. We then count the number of secondary infections caused by this first individual and average this number on all the 800 simulations to obtain an estimate of R_0 . The constraint $R_0 = 3$ requires to find a balance between ω_{exposure} and $\tilde{\omega}_{\text{dist}}$ and combine the parameters accordingly. If for instance we suppose that infectiousness decreases slowly even at long distances (like in the last row of Supplementary Table 7.4) we should set β_0 such that the infectiousness of contacts has a slow increase with duration (like in the last row of Supplementary Table 7.3), in order not to have a huge R_0 , and we obtain the pink curves in Supplementary Fig. 7.1. Vice-versa, if $\tilde{\omega}_{\text{dist}}$ is adjusted such that only close contacts are contagious, we should give more importance to duration and suppose that also short durations are at risk (e.g. blue curves in Supplementary Fig. 7.1).

In the numerical simulations discussed in the main text, we use the intermediate curves in Figure 7.1 (in orange) as infectiousness functions. We report in Supplementary Fig. 7.2 some results obtained by using in the simulations the two other sets of curves. The left and central panels represent the growth or decrease of the epidemic with the different policies assuming respectively the pink curves (thus assuming that contagion can take place even at long distance but only for long contact duration) and the blue ones (assuming contagion even for short durations but only at close proximity). We observe that for what concerns the controllability of the epidemics the two choices of proximity-duration dependence of infectiousness do not bring significantly different results. Nevertheless, the right panel in Supplementary Fig. 7.2 shows effectiveness and cost of each policy for the three proposed curves of infectiousness, and we notice that circles and diamonds have a similar

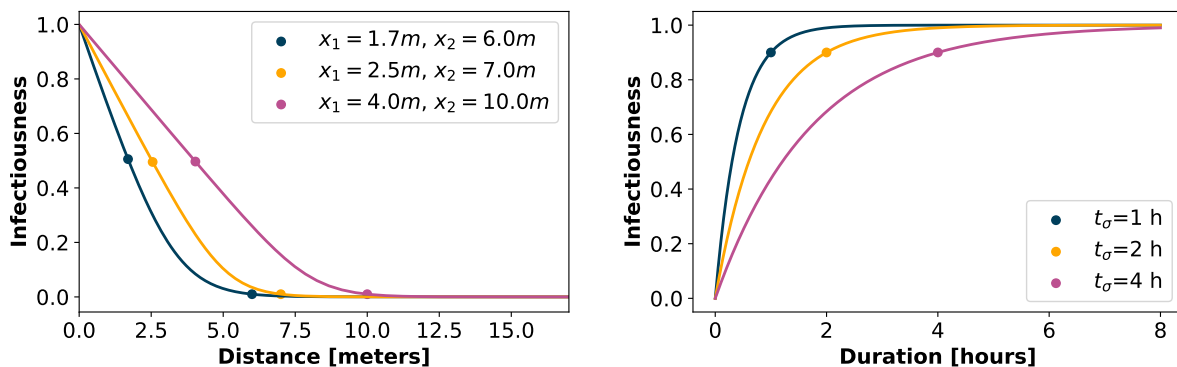


Figure 7.1: Infectiousness as a function of distance (left panel) or duration (right panel) of the contact, for three different parameters configurations. By combining the two curves corresponding to each color we obtain $R_0 = 3$ in each case. The blue configuration implies an infectiousness increasing rapidly with duration but decreasing fast with distance. On the contrary, the pink curves correspond to an infectiousness that increases slowly with contact duration but has a broader spatial range. All the simulation results in the manuscript are obtained assuming the infectiousness to be ruled by the intermediate orange configuration.

trend (respectively corresponding to orange and blue curves in Supplementary Fig. 7.1), the choice of the pink curve (triangular symbols) would lead to a more optimized balance between cost and effectiveness, with lower numbers of both false negatives and total quarantined for each policy. This strengthens the idea that a better knowledge of infectiousness as a function of duration and proximity of contacts would be fundamental to devise appropriate policies to fight the pandemic.

It is worth mentioning the two constant factors p_{R_0} and r_{R_0} that appear in Eq. (7.1). The first one is just a scaling factor, that we fix to the same constant value in all settings. The second one instead plays a pivotal role. Indeed, the procedure described above for parameters' setting is aimed to reconstruct a scenario without restrictions, where the epidemic of COVID-19 is free to spread with $R_0 = 3$. In this work, we analyze the effect of isolation and tracing in a context where other protective measures contribute to mitigate the spreading. These general precautions are described in our model as an overall reduction of R_0 , obtained by using the reduction factor $r_{R_0} \in [0, 1]$, with values reported in Supplementary Table 7.5. The chosen reduced values of R_0 take into account the combined effect of all the alternative measures in a range suggested by recent literature (33; 144; 84; 60).

Let us notice that the two functions ω_{exposure} and ω_{dist} are in principle defined as two

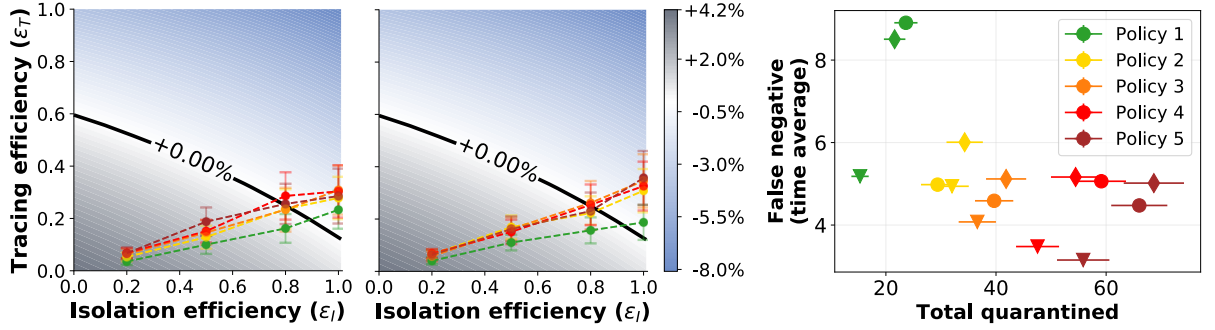


Figure 7.2: Left and central panels: Growth or decrease rate of the number of newly infected individuals for each policy, assuming respectively that the dependence of infectiousness from duration and proximity follows the pink curves and the blues curves of Supplementary Fig. 7.1. The reducing factor r_{R_0} is set to have $R_0 = 1.5$ and we assume 40% app adoption. All the points have been obtained as mean values over $n = 200$ simulations and the error bars represent the standard error. Right panel: corresponding average values of false negatives vs total quarantines for the different policies assuming for infectiousness the curves in pink (triangles), in orange (circles), and in blue (diamonds) of Supplementary Fig. 7.1.

independent functions reflecting respectively the dependency from duration and proximity. We however chose to set their free parameters simultaneously combining these two effects so as to explore how their mutual contributions change in shaping the contagions, while keeping p_{R_0} fixed.

R_0	3.0	2.0	1.5	1.2
r_{R_0}	1.0	0.53	0.39	0.26

Table 7.5: In the first row the desired values of R_0 are reported, while the second row shows the corresponding values of the reduction factor r_{R_0} needed to obtain them, with a scaling factor $p_{R_0} = 60$.

Robustness of the model with respect to the definition of the infectiousness probability

We consider here another infectiousness curve that has been derived in the recent literature by He et al. (66). We follow here the author-correction version (65), that followed a critic and correction suggestion (8) on the first version.

We show that, although this curve is different from the curve ω that we use in this paper, the predictions of the model do not change significantly, showing their robustness with respect to changes in the infectiousness curve.

In the cited works the infectiousness is defined by means of two probability density functions (PDFs): The incubation time $g(t)$ (probability of symptom onset as a function of

the time t since infection), and the infectiousness probability $f(t)$, which is a function of the time t elapsed since the symptom onset (t can take negative values because of pre-symptomatic infectiousness). In more details, the function g is in turn taken from Li et al. (102), and it is a lognormal distribution with mean 1.434065 and std 0.6612. The function f is instead estimated by He et al. (65): it is assumed to be a gamma distribution, and via a max-likelihood approach it is estimated to have shape 20.516508 and scale 1.592124, and to be shifted by an offset 12.272481. A numerical PDF of the two distributions, computed over 10^5 samples, and the analytical expression of the two PDFs are shown in Supplementary Fig. 7.3a.

From these g, f , we can reconstruct a PDF $\omega_{\text{He}}(\tau)$ to be used in our model. This can be done simply by sampling two values from g and f and adding them (the total time from infection to secondary infection is simply split into two intervals separated by the time of symptoms onset). A numerical PDF of this distribution ω_{He} , computed over the same 10^5 samples, is in Supplementary Fig. 7.3b. This function ω_{He} may also be obtained analytically by convolution as

$$\omega_{\text{He}}(\tau) = \int_{-\infty}^{\infty} f(\tau - t)g(t)dt,$$

using the analytically known f and g . The discretized convolution is also shown in Supplementary Fig. 7.3b, and it coincides indeed with the numerical values of ω_{He} .

Observe that this distribution assigns a positive probability (6.01%, see below) also to infectiousness at negative times (i.e. an individual may infect another one before being itself infected). We assume that this is due to the fact that the two distributions f and g are estimated from two different populations (65), and thus statistical errors may be present. For our aims this is not a limitation, as it just mean that the (cumulative) probability of infection at zero is strictly positive.

Supplementary Fig. 7.3b shows also the PDF ω that we used in the paper. Both distributions peak roughly at the same time (ω at 5 days, while ω_{He} at 4 days). On the other hand, ω_{He} has a wider support and a larger right tail, meaning that it models a non negligible probability of secondary infection also several days after the infection of the spreader.

To have an analytical expression of ω_{He} we try to fit shifted lognormal, gamma, and Weibull distributions to ω_{He} by least-squares minimization over the PDF obtained by convolution. The best results are obtained with a gamma distribution with density $h(\tau) = \frac{p_2^{p_1}}{\Gamma(p_1)} \tau^{p_1-1} e^{-p_2\tau}$ with parameters $p_1 = 5.73$, $p_2 = 0.55$, and shifted by 4.67, which is plotted in Supplementary Fig. 7.3c. This allows also to derive an explicit cumulative density

function CDF_{He} of ω_{He} , which gives an estimate of $\text{CDF}_{\text{He}}(0) = 0.0601$ (the fraction of negative-time infections).

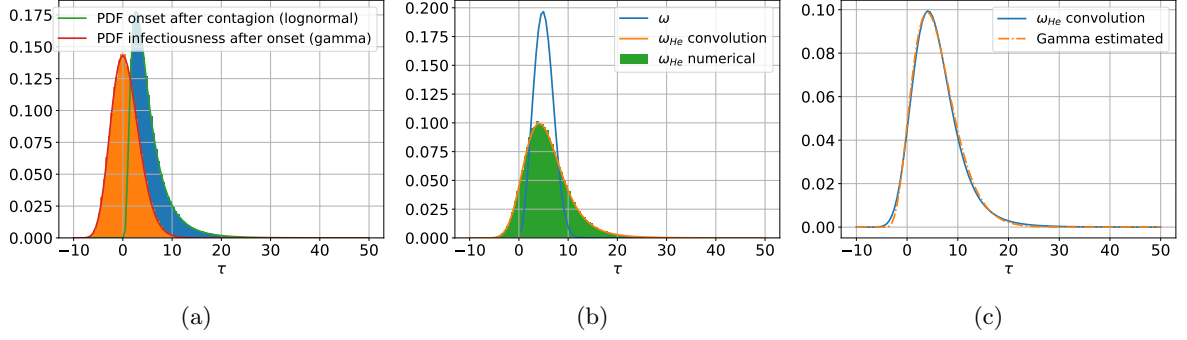


Figure 7.3: Visualization and estimation of the infectiousness probability density function (PDF) ω_{He} . PDFs f and g (Supplementary Fig. 7.3a); estimated PDF ω_{He} , and PDF ω_{He} (Supplementary Fig. 7.3b); fit of ω_{He} with a gamma distribution (Supplementary Fig. 7.3c).

We can now use this modified infectiousness ω_{He} in our model and compare the results with the ones of Fig. 3.5 of the main text. First, we estimate again the reduction parameter defining ω_{data} (see Section 3.2.2 of the main text), and we get $r_{R_0} = 0.35$.

Using this functional form of ω_{He} in the model, we obtain the results of Supplementary Fig. 7.4 (see central panel in Fig. 3.5 of the main text for the corresponding results with ω). It is clear that the difference is quite limited since only Policy 1 and Policy 2 for $\varepsilon_I = 0.8$ move from being ineffective (Fig. 3.5, main text) to being effective. We can thus conclude that no significant change in our conclusions would be introduced by adopting this alternative infectiousness function in place of the current one. In particular, the predictions using ω appear to be less optimistic in the prediction of the policies' effectiveness, since they estimate that not all policies are successful for $\varepsilon_I = 0.8$.

Contact patterns in the CNS data set

To further guarantee the reproducibility of the results of this paper, we provide additional details on the CNS data set.

As mentioned before, the CNS data set (157) contains one month of data that is used here as it is. Thus, for any detail we refer to the cited paper, and we only visualize in Supplementary Fig. 7.5a the temporal distribution of the total number of contacts contained in the data set. It is immediate to observe that the number of contacts has a periodical behavior that reflects the day/night periods and the days of the week. Moreover, a certain uniformity is present between different weeks.

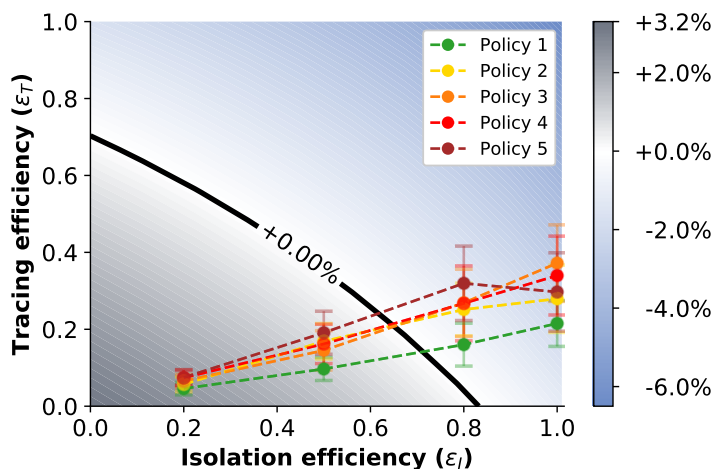


Figure 7.4: **Tracing policy efficiency for alternative infectiousness.** Growth or decrease rate of the number of newly infected individuals using the modified infectiousness curve ω_{He} . The points correspond to the parameter pairs such that ε_I is an input and ε_T an output of the simulations on real contact data, for the policies of Fig. 3.3. Here $R_0 = 1.5$ with 40% app adoption. All the points have been obtained as mean values over $n = 200$ simulations and the error bars represent the standard error.

For the simulations discussed in SI 7.1.2 we need to use a longer time period, that is extracted from data that are not publicly shared in the CNS data set (157). We extract the period from the 1st of September to the 30th of November 2013, and remove the week between 7th and 13th of October, since it corresponds to a holiday week with very few contacts. In this way, the whole timespan used for the simulations has an amount of contacts that remains on average homogeneous in time. Supplementary Fig. 7.5b shows the distribution of contacts in this case.

7.1.2 Evaluation of additional containment measures and refined policies

Longer and shorter tracing memory

We explore here how the outcomes of the different policies depend on the memory length of the contact history, which has been set to 7 days in the previous simulations (see Supplementary Notes 3.2.3 of the main text).

First, to understand whether or not an increased memory would improve the effectiveness of each policy, we repeat the experiments assuming that the contacts of each individual are recorded for 15 days in the past, and report the results in Supplementary Fig. 7.6. When comparing Supplementary Fig. 7.6a with the original setting, it is clear that the increased memory brings a negligible advantage. This is confirmed by the total number

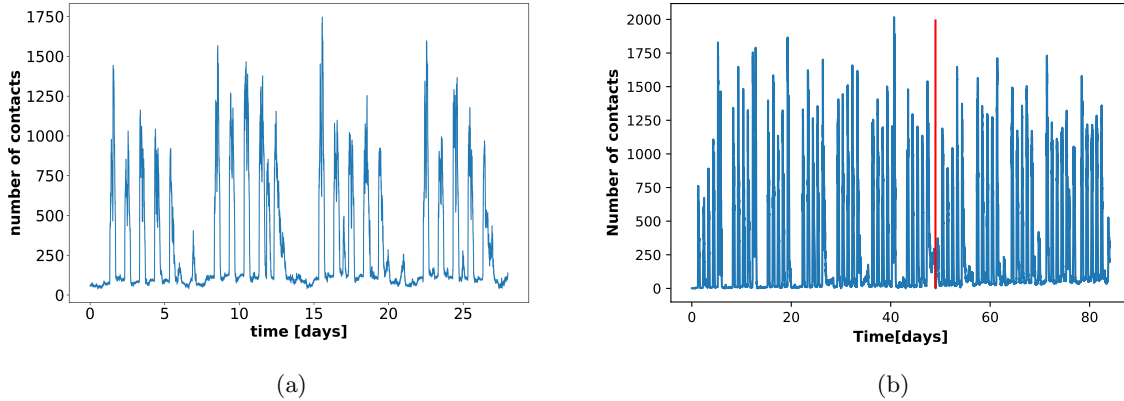


Figure 7.5: **Temporal distribution of the total number of contacts in the CNS data set.** The figures show the total number of contacts in the CNS data set (Supplementary Figure 7.5a), and in the extended version (Supplementary Figure 7.5b) as a function of time. The vertical red line represents the cut of the holiday week. The aggregation is computed with a temporal gap of 300 seconds.

of false negatives in Supplementary Fig. 7.6b if compared with Fig. 3.6 of the main text, and this is at the price of increased storage requirements, see total quarantines.

Second, it is worth investigating if a shorter tracing memory would give improvements in terms of the numbers of false positives. We thus repeat the simulations assuming that the memory is reduced to 2 days (still including the 2 days delay in the case reporting as in all other settings). Supplementary Fig. 7.7 shows that the shorter memory reduces the effectiveness of the policies of a significant amount, none of them crossing the black line for $\varepsilon_I = 0.8$. Apparently, storing only 2 days of contacts reduces too much the number of quarantined individuals (see Supplementary Fig. 7.7b), affecting the effectiveness.

Longer delay

The implemented model, for the sake of realism, includes a variable delay between the instant when a person is recognized as infected and the instant when that person is isolated. We set the delay to 2 days in all the other simulations and we test here the effect of a longer delay: 3 days, which is a good estimate for a system which is overburdened but not close to collapse. From Supplementary Fig. 7.8a we observe that even one additional day of delay has a strong impact on the behavior of the epidemic, with none of the proposed policies able to cross the threshold of controllability, even for maximal isolation efficiency. Moreover Supplementary Fig. 7.8b shows that high levels of false negatives are reached for each policy, around twice those obtained with only two days of

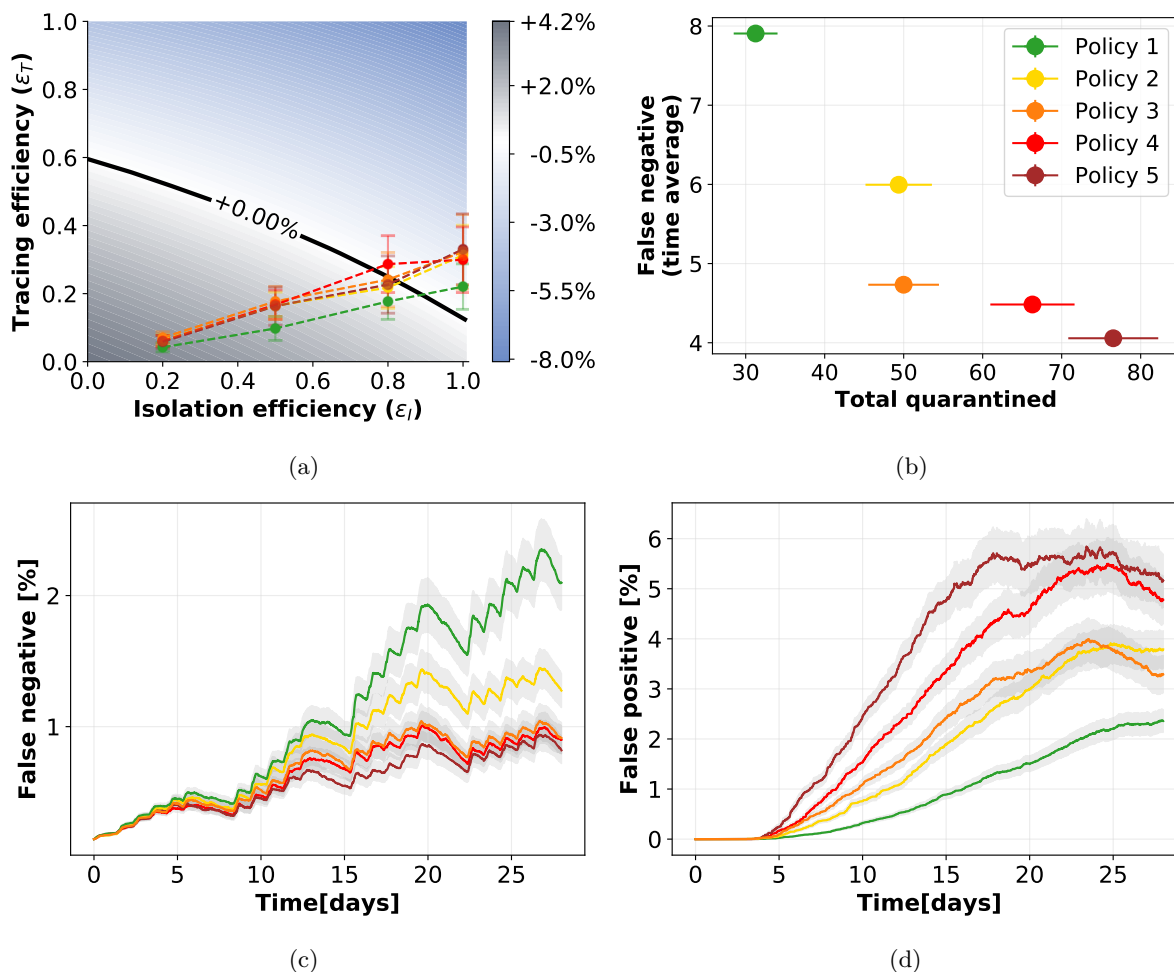


Figure 7.6: **Tracing policy efficiency with longer contact memory: 15 (instead of 7) days.** 7.6a: Growth or decrease rate of the number of newly infected individuals and efficiency of the containment policies. 7.6b: Cross plot of the cost (number of quarantines) versus the effectiveness (low number of false negatives) for each policy. 7.6c and 7.6d: Temporal evolution of respectively the percentages of false negatives, i.e. infected individuals not quarantined, and false positives, i.e. not infected individuals quarantined, over the entire population, assuming an isolation efficiency of $\epsilon_I = 0.8$, a reproductive number $R_0 = 1.5$, and 40% app adoption. The points in the first two panels and the curves in the last two have been obtained as mean values over 200 independent simulations, the corresponding error bars and the curve shadings represent the standard error.

delay (see Fig. 3.6 in the main text) even if the total number of people in quarantine is slightly higher.

This highlights how rapid interventions are fundamental in containment policies based on contact tracing.

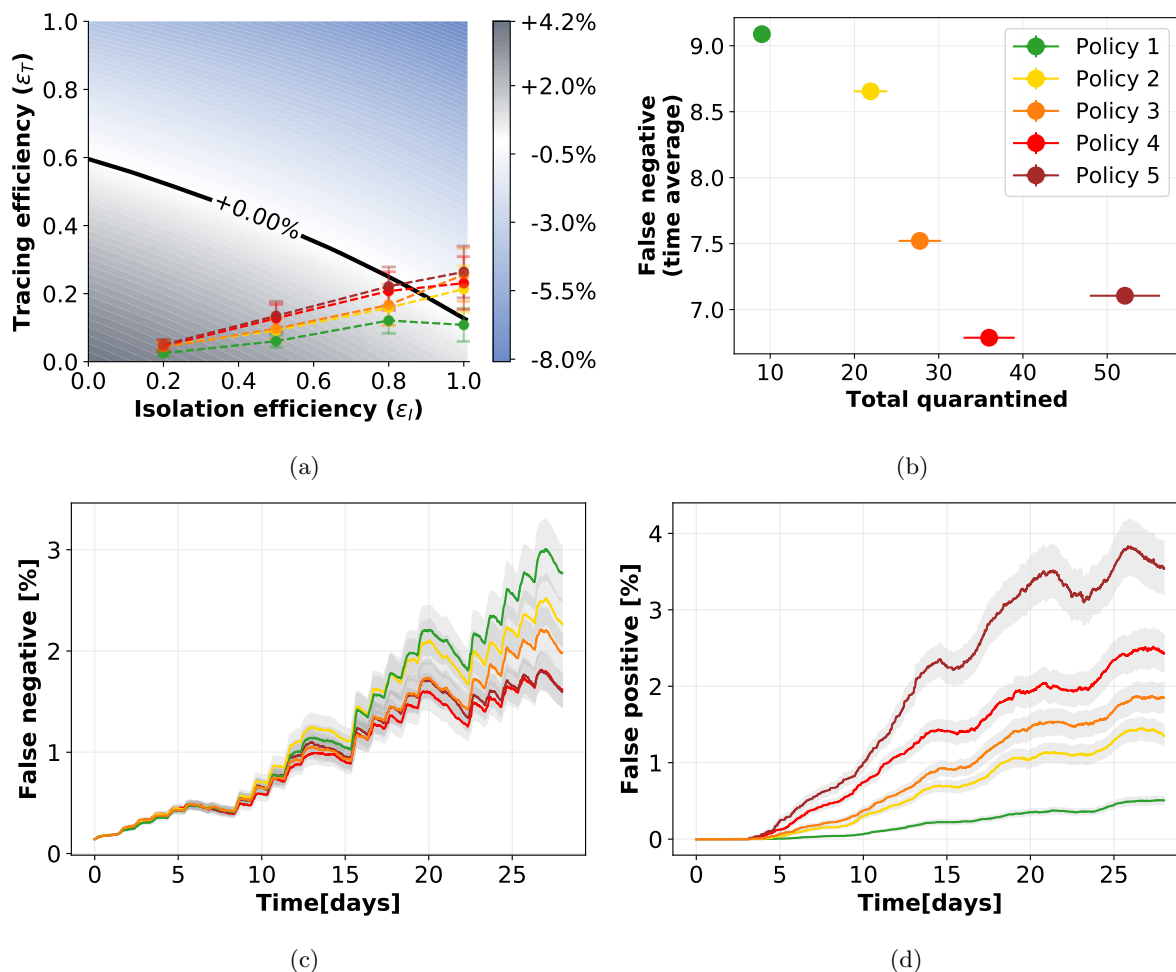


Figure 7.7: **Tracing policy efficiency with shorter contact memory: 2 (instead of 7) days.** 7.7a: Growth or decrease rate of the number of newly infected individuals and efficiency of the containment policies. 7.7b: Cross plot of the cost (number of quarantines) versus the effectiveness (low number of false negatives) for each policy. 7.7c and 7.7d: Temporal evolution of respectively the numbers of false negatives, i.e. infected individuals not quarantined, and false positives, i.e. not infected individuals quarantined, assuming an isolation efficiency of $\epsilon_I = 0.8$, a reproductive number $R_0 = 1.5$, and 40% app adoption. The points in the first two panels and the curves in the last two have been obtained as mean values over 200 independent simulations, the corresponding error bars and the curve shadings represent the standard error.

Second order tracing

We additionally explore the possibility to keep track of contacts in a recursive way. Namely, when an individual is isolated, not only its contacts are quarantined, but also its contacts' contacts. This obviously means an enhanced risk in terms of preserving the

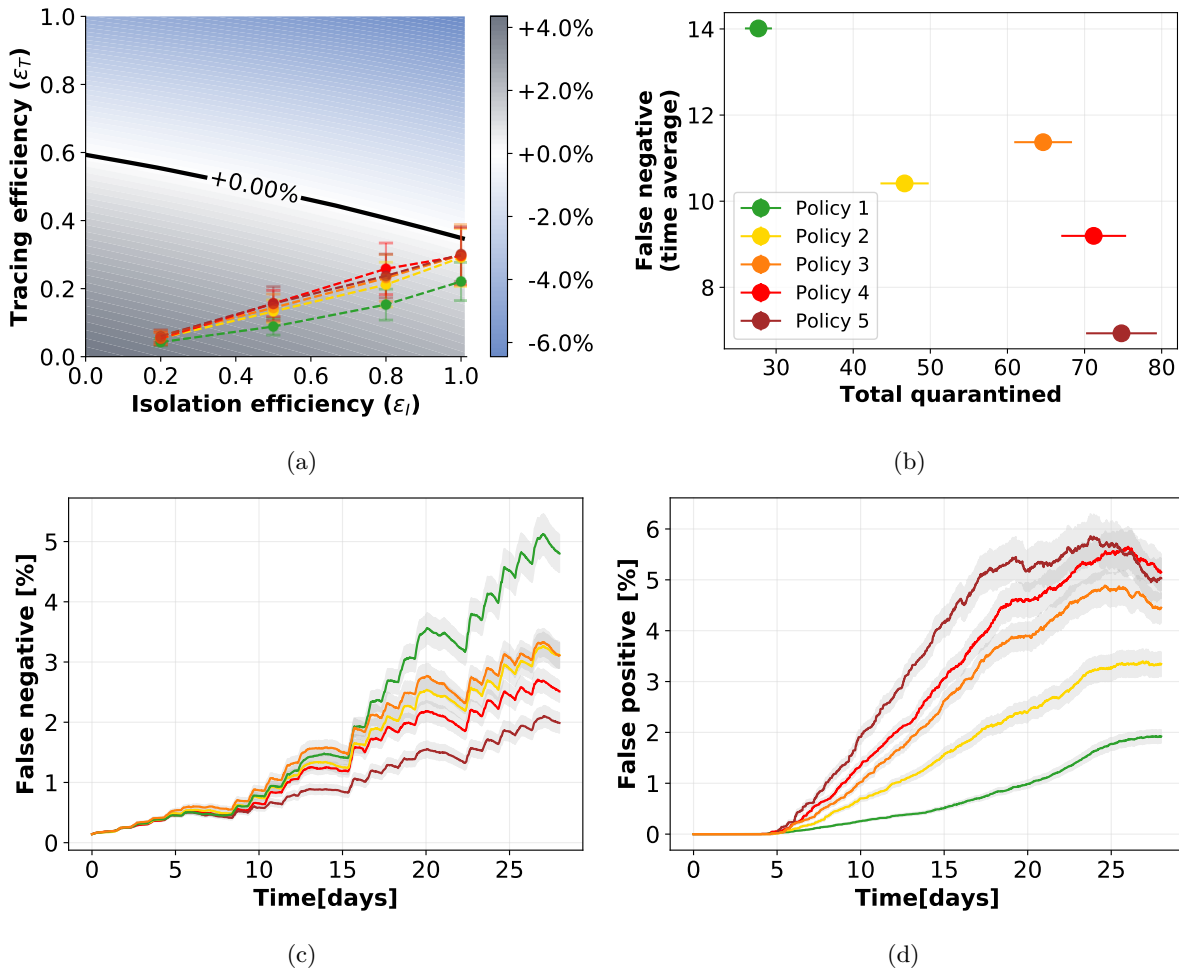


Figure 7.8: **Tracing policy efficiency with a longer reporting delay: 3 (instead of 2) days.**

7.8a: Growth or decrease rate of the number of newly infected individuals and efficiency of the containment policies. 7.8b: Cross plot of the cost (number of quarantines) versus the effectiveness (low number of false negatives) for each policy. 7.8c and 7.8d: Temporal evolution of respectively the percentages of false negatives, i.e. infected individuals not quarantined, and false positives, i.e. not infected individuals quarantined, over the entire population, assuming an isolation efficiency of $\varepsilon_I = 0.8$, a reproductive number $R_0 = 1.5$, and 40% app adoption. The points in the first two panels and the curves in the last two have been obtained as mean values over 200 independent simulations, the corresponding error bars and the curve shadings represent the standard error.

privacy of individuals, and hence the major open question regarding this kind of policies is whether or not the increased intrusiveness into an individual's social network provides a tangible improvement of the virus containment efforts.

A complete study of this scenario is beyond the scope of this paper for a specific reason:

the continuous model does not take into consideration this kind of tracing, and there is thus no way to use the information provided by the study of the data set in this framework.

Nevertheless, we find meaningful to report here the results of this additional experiment. We simulated the epidemic on the CNS data set, considering $R_0 = 1.5$, a delay of 2 days in isolating infected individuals and an app adoption of 40%. The numerical results are shown in Supplementary Fig. 7.9. We immediately notice that such intrusive tracing policy does not provide a significantly beneficial effect. Indeed, comparing Supplementary Fig. 7.9 top left and top right with respectively Fig. 3.6 top left and 3.6 top right in the main text, which are the corresponding results for first order tracing, we notice that the levels reached by both false negatives and false positives are slightly reduced with second order tracing but not of a large amount. This appears clear also observing Supplementary Fig. 7.9 (bottom left) and the table, where the values of both total false negative and total quarantines are similar to those obtained with first order tracing (see Fig. 3.6 of the main text), with a slightly higher cost (larger percentages of quarantines) and a slightly larger effectiveness (lower false negatives).

This preliminary study seems to suggest that such a high level of tracing, which implies privacy issues (possibly even leading to lower adoption and compliance levels (80)), does not seem to be worth it since it is not going to provide meaningful improvements to the tracing system. We however remark once more that the reliability of this result is limited, being linked to a specific data set and not to a general theory. For this reason we observe that the concept of second-order tracing, a topic of recent discussions, deserves further investigation and may possibly be expanded in future works.

Variations in the number of asymptomatic individuals

In order to additionally verify the robustness of our predictions with respect to the epidemiological modelling, we assume here that the number of asymptomatic individuals is 20%, and additionally that a randomized testing policy that covers 25% of the asymptomatic population is in place.

In this case, little changes in the predictions of the model (Supplementary Fig. 7.10a) with respect to the case of 40% asymptomatics that was analyzed in the main text, since all the policies are effective for $\varepsilon_I = 1$, while Policy 1 is the only one that fails to contain the epidemic for $\varepsilon_I = 0.8$. No policy is effective for lower isolation efficiency. Similarly, the quarantine dynamics (false negative and false positive, Supplementary Fig. 7.10c and 7.10d) appear to have a similar behavior as in the basic setting. Despite these seemingly

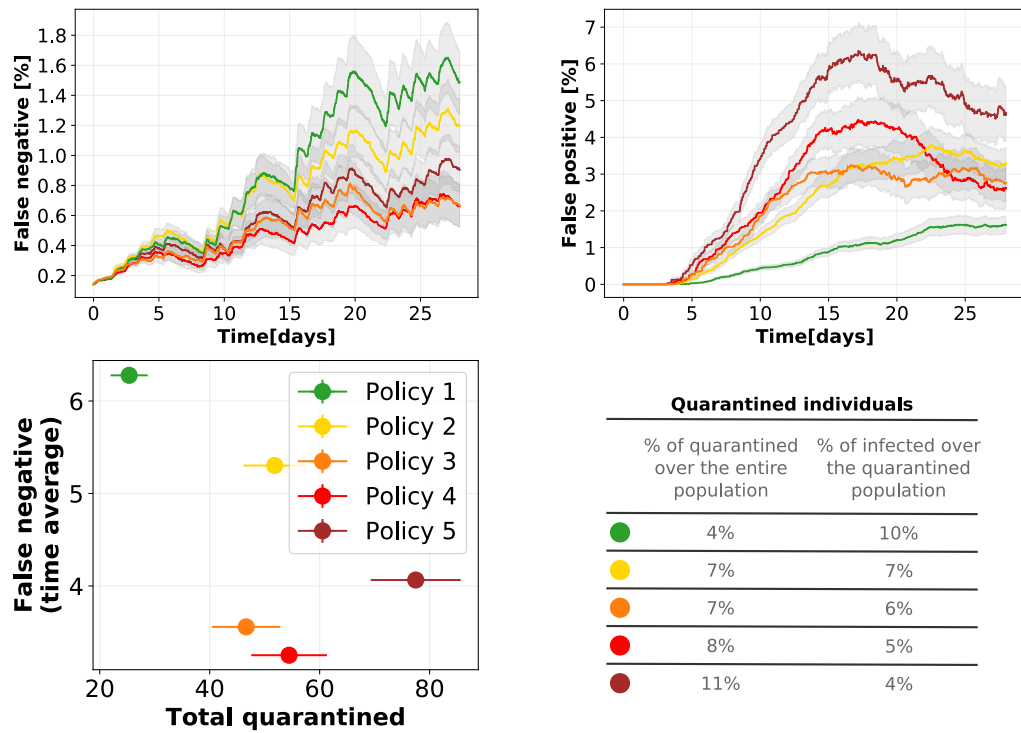


Figure 7.9: **Numerical simulations with second order tracing.** Figure at the top left and top right: Temporal evolution of percentages of false negatives, i.e. infected individuals not quarantined, and false positives, i.e. not infected individuals quarantined, assuming an isolation efficiency of $\varepsilon_I = 0.8$. Figure bottom left: plot of the effectiveness (low number of false negatives) vs. cost (total quarantines) of the policies. The parameters are set so as to have $R_0 = 1.5$ and 40% app adoption. The table reports the percentage of distinct individuals who have been quarantined over the entire population and the percentage of them who were actually infected (true positive). The curves in the first two panels and the points in the third have been obtained as mean values over 100 independent simulations, the corresponding curve shadings and error bars represent the standard error.

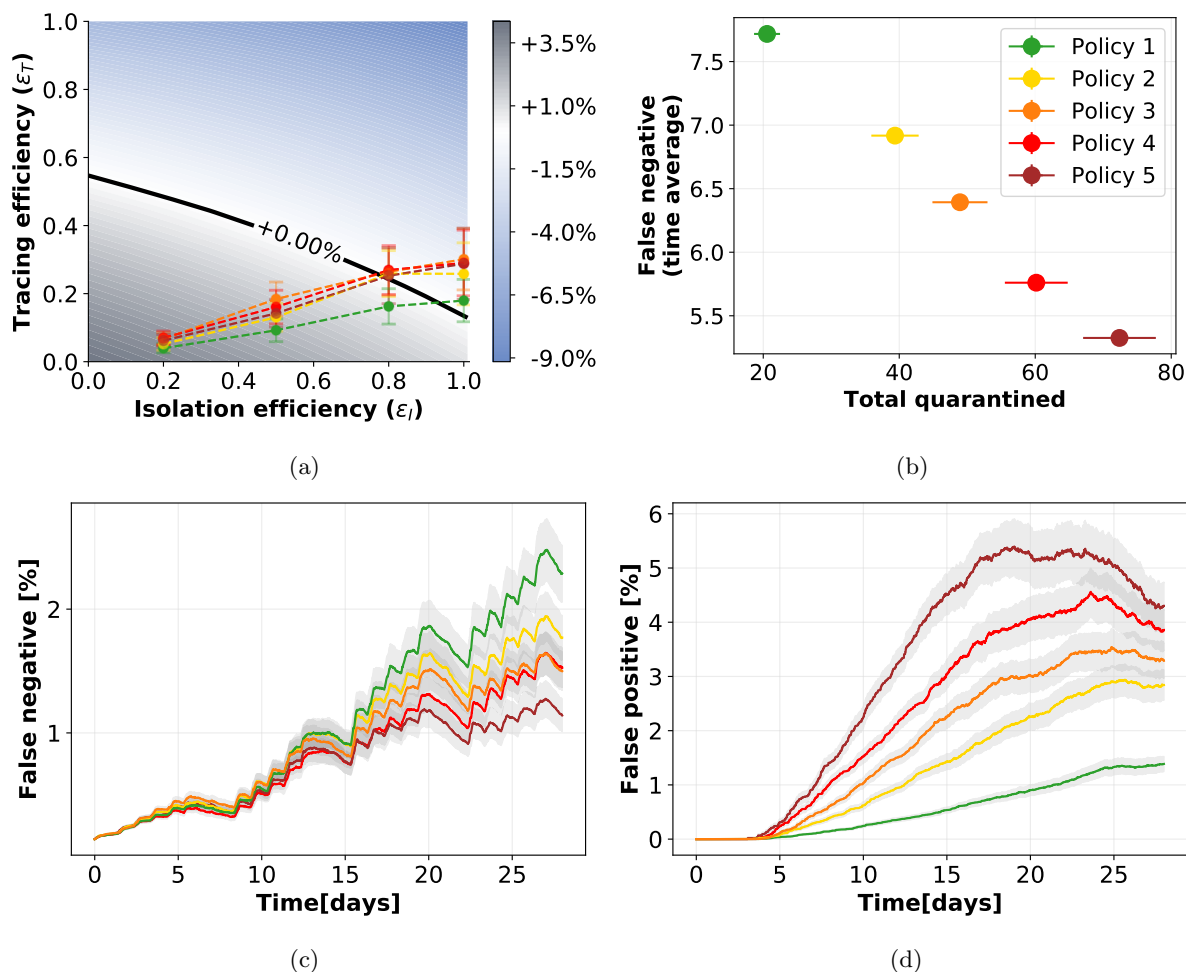


Figure 7.10: **Tracing policy efficiency with 20% asymptomatic and 25% random testing.** 7.10a: Growth or decrease rate of the number of newly infected individuals and efficiency of the containment policies, assuming that symptomatic people account for the 80% of the infected individuals, that they can be isolated and that an additional 25% of asymptomatics can be identified via randomized testing. 7.10b: Cross plot of the cost (number of quarantines) versus the effectiveness (low number of false negatives) for each policy. 7.10c and 7.10d: Temporal evolution of respectively the percentages of false positives, i.e. not infected individuals quarantined, and false negatives, i.e. infected individuals not quarantined, over the entire population, assuming an isolation efficiency of $\varepsilon_I = 0.8$, a reproductive number $R_0 = 1.5$, and 40% app adoption. The points in the first two panels and the curves in the last two have been obtained as mean values over 200 independent simulations, the corresponding error bars and the curve shadings represent the standard error.

small changes in the success of the policies and in their cost, the cross visualization of Supplementary Fig. 7.10b shows that in this scenario it is harder to find a clear tradeoff between cost and effectiveness, since the two scores change smoothly between the five

policies.

Close-range short-exposure vs long-range long-exposure interactions

We test here two additional policies obtained by mixing a low space resolution and a high time resolution, and viceversa. The policies are defined in Supplementary Fig. 7.12. Policy 6 delimits the risk to short exposure but close range interactions, while Policy 7 captures long exposure but long range interactions.

ID	Signal strength (dBm)	Duration (min)	Fraction
● Policy 6	-70	5	17.9%
● Policy 7	-91	30	2.1%

Table 7.6: Parameters defining the two additional policies, and fraction of the total number of interactions of the CNS data set that they are able to detect.

Supplementary Fig. 7.11, in analogy with Supplementary Fig. 3.4 of the main text, shows the new policies overlaid to the histograms of duration and signal strength of the CNS data set contacts.

The values of the parameters $(\varepsilon_I, \varepsilon_T)$ characterizing the numerical simulations for the new policies with $R_0 = 1.5$ are shown in Supplementary Fig. 7.12a (see Fig. 3.5 in the main text, central panel, for a comparison with the policies in Fig. 3.3, main text), and it is clear that Policy 7 is as effective as the most restrictive policies (Policy 2 to Policy 5), while Policy 6 fails to contain the virus for an isolation efficiency smaller than 1. As for the policies of Fig. 3.3, this effectiveness comes at the cost of a larger number of quarantines (Supplementary Fig. 7.12c and Supplementary Fig. 7.12d). However, Supplementary Fig. 7.12b shows that the cost of Policy 7 is in larger than the ones of Policy 2 and Policy 3, but smaller than the ones of Policy 4 and Policy 5, while achieving a similar effectiveness.

We deduce that the ability to control the contagion seems to be more sensitive to duration of contacts than to their spatial distance. Indeed, policies which capture close range but short exposure interactions happen to be less performative in quarantining people than those signaling long range interactions with long exposure. In other words, quarantining individuals who have had a short interaction with an infected one, even if at close-range, is unnecessary. On the other hand, it appears to be important to track contacts with a high spatial resolution, including the ones that happens at a rather long distance, if their

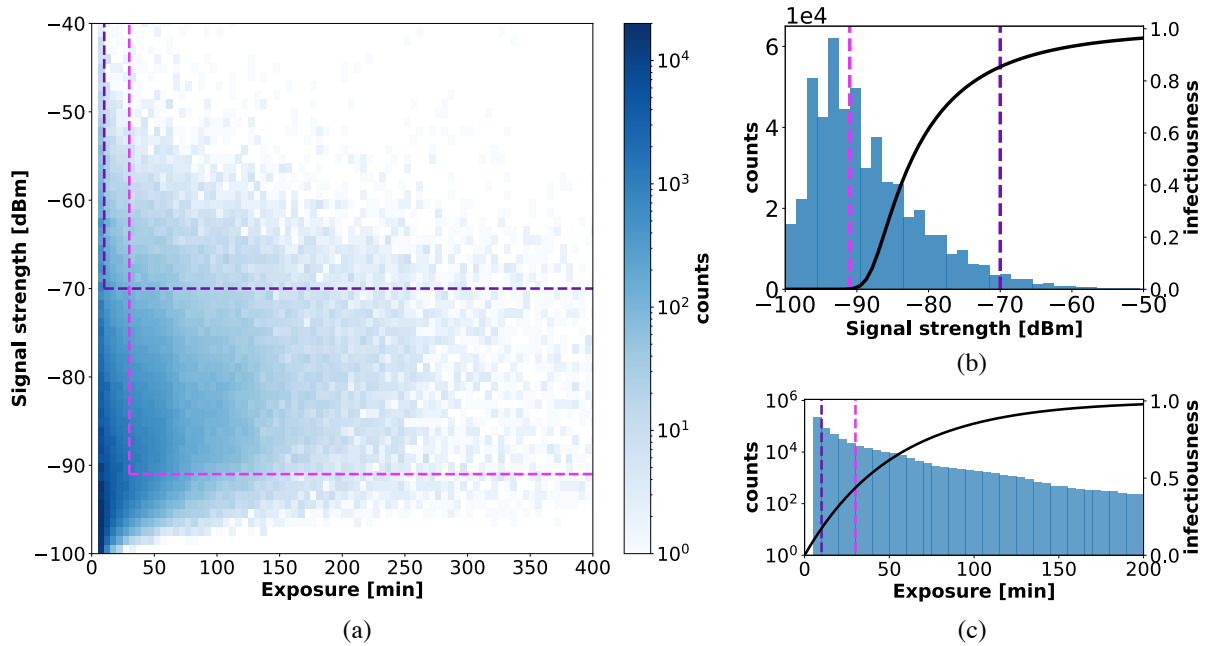


Figure 7.11: Distribution of the duration, panel (c), and signal strength (taken as a proxy for proximity), panel (b), of the contacts in the CNS data set. Panel (a) gives a scatterplot of signal strength vs duration, and displays the thresholds defining the two policies of Supplementary Table 7.6.

duration is significant.

However, we remark once more that these results are depending on the infectiousness model that we have defined here, and that they could possibly change in a different setting.

Compliance to quarantine decreases if notified multiple times

In the main text we consider compliance as encoding the compliance to all parts of the contact tracing and quarantine procedure. In other words, if some of the participants install the app but then do not quarantine if notified, then they should be counted among the non-compliant individuals since the effect would be the same than that of not adopting the app at all. The non-compliance (or impossibility) to quarantine is therefore already considered when choosing the percentage of app adoption. However, despite the fact that people who adopt the app are aware that they could be required to quarantine even if not infected, they may underestimate the possibility to be notified multiple times. A repeated quarantine could represent a relevant problem under social and economical aspects for many people, especially if unjustified. For this reason we decided to run an

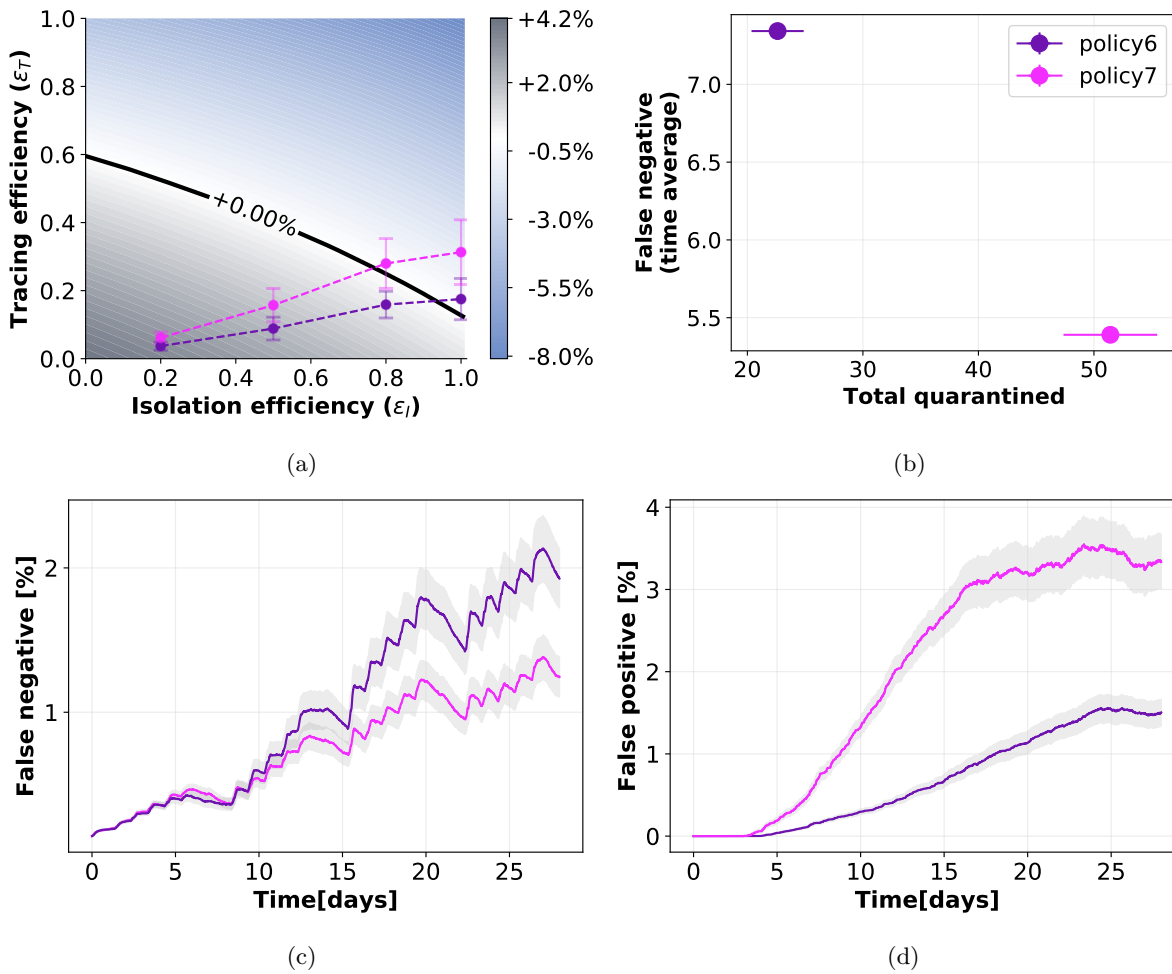


Figure 7.12: **Tracing policy efficiency with additional policies.**

7.12a: Growth or decrease rate of the number of newly infected individuals and efficiency of the containment policies. 7.12b: Cross plot of the cost (number of quarantines) versus the effectiveness (low number of false negatives) for each policy. 7.12c and 7.12d: Temporal evolution of respectively the percentages of false positives, i.e. not infected individuals quarantined, and false negatives, i.e. infected individuals not quarantined, over the entire population, assuming an isolation efficiency of $\epsilon_I = 0.8$, a reproductive number $R_0 = 1.5$, and 40% app adoption. The points in the first two panels and the curves in the last two have been obtained as mean values over 200 independent simulations, the corresponding error bars and the curve shadings represent the standard error.

additional set of simulations where adoption of the app does not necessarily coincide with compliance to quarantine, and in particular it decreases if the same person is wrongly notified multiple times.

In particular we assume that compliance to quarantine can drop due to repeated notifica-

tions because the trust in healthcare and government institutions would drop too (171; 21; 175). Therefore the progressive decrease can be roughly estimated by considering the most classical game based on trust: the prisoner’s dilemma (145; 9). We focus in particular on an experiment of repeated game (6) where people were asked to play multiple rounds, each one with a different person. The experiment showed that willingness to cooperate decreased at each round and was measured for 10 rounds in total. We consider that the same reduction in trust can be applied to the willingness to quarantine if notified. In a broad sense, these two settings are indeed similar: in the prisoner’s dilemma each person can choose to cooperate, which they know would be the best option for everybody, but they do it at their own expenses, while in alternative they can choose an egoistic strategy, putting the others at risk. In case of notification from the contact tracing app, people would undergo a sort of “quarantine dilemma”. Indeed there are two possible choices: the compliant one (for the social benefit, but possibly in detriment of their own social and economic life) and the egoistic one where a person decides not to quarantine, putting at risk all the others.

We therefore consider that the first time that people are traced and identified as possible infected they quarantine with probability 1. The second time it happens, if the person did not develop symptoms during the first quarantine, the probability drops to 0.86. The third time to 0.6, and so on, according to the values in Supplementary Table 7.7.

Previous quar.	0	1	2	3	4	5	6	7	8	9
Compliance	1	0.86	0.60	0.57	0.49	0.46	0.43	0.41	0.40	0.29

Table 7.7: The second row reports the probabilities of compliance to quarantine if notified by the app, given that the same person has already been quarantined, even if not infected, a number of times reported in the first row. The level of compliance have been chosen according to Ref. (6).

We simulated this setting on an extended version of the CNS data set, containing contacts for a period of three months instead of one, in order to be able to catch all the repeated notifications (see SI 7.1.1 for a description of the extended time period).

Notice that this modification can be inserted into the mathematical model if we consider that the ε_T , that we compute as explained in Section 3.1.3 of the main text, changes its meaning. In this case it does not represent the ability to trace people but the possibility to quarantine them, since traced individuals could refuse to quarantine. Only for this case we thus rename ε_T into ε_Q . The controllability of the epidemic is depicted in Supplementary Fig. 7.13a, while in Supplementary Fig. 7.13b we report the number of people who have been requested to quarantine as a function of the number of repetitions of these requests, for the five different policies. The time evolution of false negatives is depicted in

Supplementary Fig. 7.13c. In general, in Supplementary Fig. 7.13 we observe a similar behavior to the one obtained in the original setting (Fig. 3.5 central panel and Fig. 3.6 in the main text), with a slightly general reduction of the efficacy of containment. Indeed, only few people are asked to quarantine multiple times, as shown by Supplementary Fig. 7.13b. We can therefore assume that the original setting that we chose – and used in all other simulations – depicts a scenario which is not far from the one that we obtain with this additional characteristic making the system more realistic, thus confirming the robustness of our model.

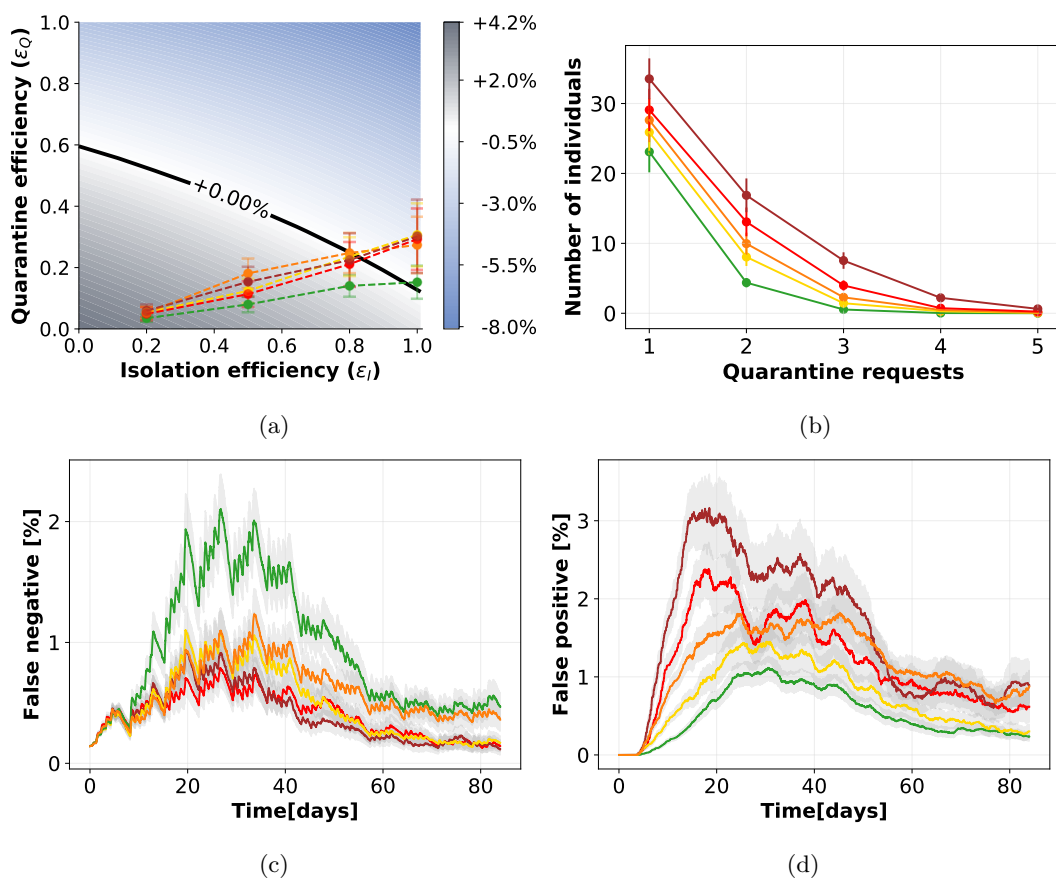


Figure 7.13: **Compliance to quarantine variable in time.** 7.13a: Growth or decrease rate of the number of newly infected individuals and efficiency of the containment policies. 7.13b: Number of people who have been requested to quarantine as a function of the number of repetitions of these requests, for the five different policies. 7.13c and 7.13d: Temporal evolution of the percentages of respectively false negatives, i.e. infected individuals not quarantined, and false positives, i.e. not infected individuals quarantined, over the entire population, assuming an isolation efficiency of $\epsilon_I = 0.8$, a reproductive number $R_0 = 1.5$, and 40% app adoption. The points in the first two panels and the curves in the last two have been obtained as mean values over $n = 100$ independent simulations, the corresponding error bars and the curve shadings represent the standard error.

The possibility to run the code on the extended data set provides in addition the possibility to observe the phenomenon of growth and decrease of the active infected, which after one month and a half dampen down, almost extinguishing the epidemic. The false negative peak is followed by the false positive and unjustified quarantines are reduced to almost zero in a couple of months (see Supplementary Fig. 7.13d).

7.2 Generating Temporal Networks

7.2.1 Execution time comparison

The egocentric perspective, that ignores interactions among neighbors of each ego node, implies a huge simplification with respect to mining standard motifs. Traditional techniques for motifs mining indeed rely on an isomorphism test for assessing sub-network equivalence, which is a major bottleneck for the entire procedure. For this reason, standard motifs mining techniques usually limit the search to small motifs containing a handful of nodes. The strength of *ETN-gen* lies in the possibility of encoding neighborhoods into a unique bit vector, boiling down sub-network equivalence to bit vector matching. This hence results in a very computationally efficient model, and the time required for network generation is drastically lower than that of the competitors. This is evident from table 7.8, where we report the time (in seconds) required to generate networks for the three face-to-face datasets with our algorithm and the competitors. *ETN-gen* is more than 15 times faster than the fastest competitor on each network, and there is a difference in time of three orders of magnitude with the slowest one.

	Hospital	Workplace	High School
<i>ETN-gen</i>	17s	52s	22s
<i>Dymond</i>	$3.6 \times 10^4 s$	$1.4 \times 10^3 s$	$3.2 \times 10^5 s$
<i>STM</i>	$1.4 \times 10^3 s$	$9.6 \times 10^2 s$	$1.6 \times 10^3 s$
<i>TagGen</i>	$2.7 \times 10^4 s$	$8.7 \times 10^3 s$	$2.4 \times 10^4 s$

Table 7.8: **Execution time.** Time in seconds required to train and generate networks with each method on three different networks.

7.2.2 Scalability

To show the scalability of our approach we extend the analysis to other seven networks, briefly described bellow.

- **High school 2 (50).** The dataset has been collected in 2012 in Lycée Thiers, Marseilles, France, over seven days (Monday to Tuesday of the following week). It contains interactions among students in five different high school classes. Number of edges: 2220, number of nodes: 180. As stated by the research group responsible for the data collection, a signed informed consent was obtained for each study

participant (all involved students were at least 18). Moreover, the study was approved by the “Commission Nationale de l’Informatique et des Libertés” (CNIL, <http://www.cnil.fr>), the French national body responsible for ethics and privacy, and by the high school authorities. More details can be found in the paper describing the data collection (50).

- **High school 3 (120).** The dataset has been collected in 2013 in Lycée Thiers, Marseilles, France, over five days in December. It contains interactions among students in nine different high school classes. Number of edges: 5818, number of nodes: 327. As stated by the research group responsible for the data collection, a signed informed consent was obtained for each study participant (all involved students were at least 18). Moreover, the study was approved by the “Commission Nationale de l’Informatique et des Libertés” (CNIL, <http://www.cnil.fr>), the French national body responsible for ethics and privacy, and by the high school authorities. More details can be found in the paper describing the data collection (120).
- **Primary school (168).** The dataset has been collected in a primary school in France, over two days in October 2009. It contains interactions among 232 children and 10 teachers. Number of edges: 8317, number of nodes: 242. As stated by the research group responsible of the data collection, the “Commission Nationale de l’Informatique et des Libertés” (CNIL, <http://www.cnil.fr>) and the “Comité de Protection des personnes” (<http://www.cppsudest2.com/>) were notified of the study. The study was also approved by the relevant academic authorities of the primary school in which the study took place. Finally, parents, teachers, and the director of the school expressed a verbal informed consent. More details can be found in the paper describing the data collection (168).
- **SMS 1 (156).** The dataset represents SMSs among university freshmen students in the Copenhagen University. Number of edges: 697, number of nodes: 568. The dataset was collected within the Copenhagen Network Study and the data collection was approved by the Danish Data Supervision Authority. Each study participant was asked to sign an informed consent.
- **SMS 2 (2).** The dataset represents SMSs among members of a young-family residential living community adjacent to a major research university in North America. Number of edges: 153, number of nodes: 85. The dataset was collected within the Friends and Family Study and the data collection was approved by the Institutional Review Board (IRB). The participation was optional and each study participant was asked to explicitly adhere.

- **Calls 1 (156).** The dataset represents phone calls among university freshmen students in the Copenhagen University. Number of edges: 605, number of nodes: 525. The dataset was collected within the Copenhagen Network Study and the data collection was approved by the Danish Data Supervision Authority. Each study participant was asked to sign an informed consent.
- **Calls 2 (2).** The dataset represents phone calls among members of a young-family residential living community adjacent to a major research university in North America. Number of edges: 432, number of nodes: 129. The dataset was collected within the Friends and Family Study and the data collection was approved by the Institutional Review Board (IRB). The participation was optional and each study participant was asked to explicitly adhere.

Each face-to-face interaction network has been aggregated with a temporal resolution of five minutes, while SMS and phone calls networks have been aggregated within ten minutes. We opt for this different aggregations due to the natural sparsity of SMS and phone calls networks.

In Figure 7.14 we show the original number of interactions (in black) and those generated by our method (in orange) for each network. The figure clearly shows the ability of our method in mimicking day/night and week/weekend periodicity. Moreover, our algorithm perfectly operates with different network sizes in both number of individuals and temporal length. Finally, our method is able to capture multiple picks within the same day, that could be associated to the period before and after lunch (i.e. high schools).

7.2.3 Varying K

In this section, we evaluate the performance of our method when k varies. In particular, we generate the hospital network using several k ($k \in \{2, 3, 4, 5\}$). As one may expect as far as k increases, the execution time increases (see table 7.9).

k	2	3	4	5
Time (seconds)	17	25	40	55

Table 7.9: Execution time (in seconds) when varying K

The first panel of figure 7.15 shows the number of interaction in the hospital network. From the figure, can be seen that for each k the periodicity is not affected. However, as

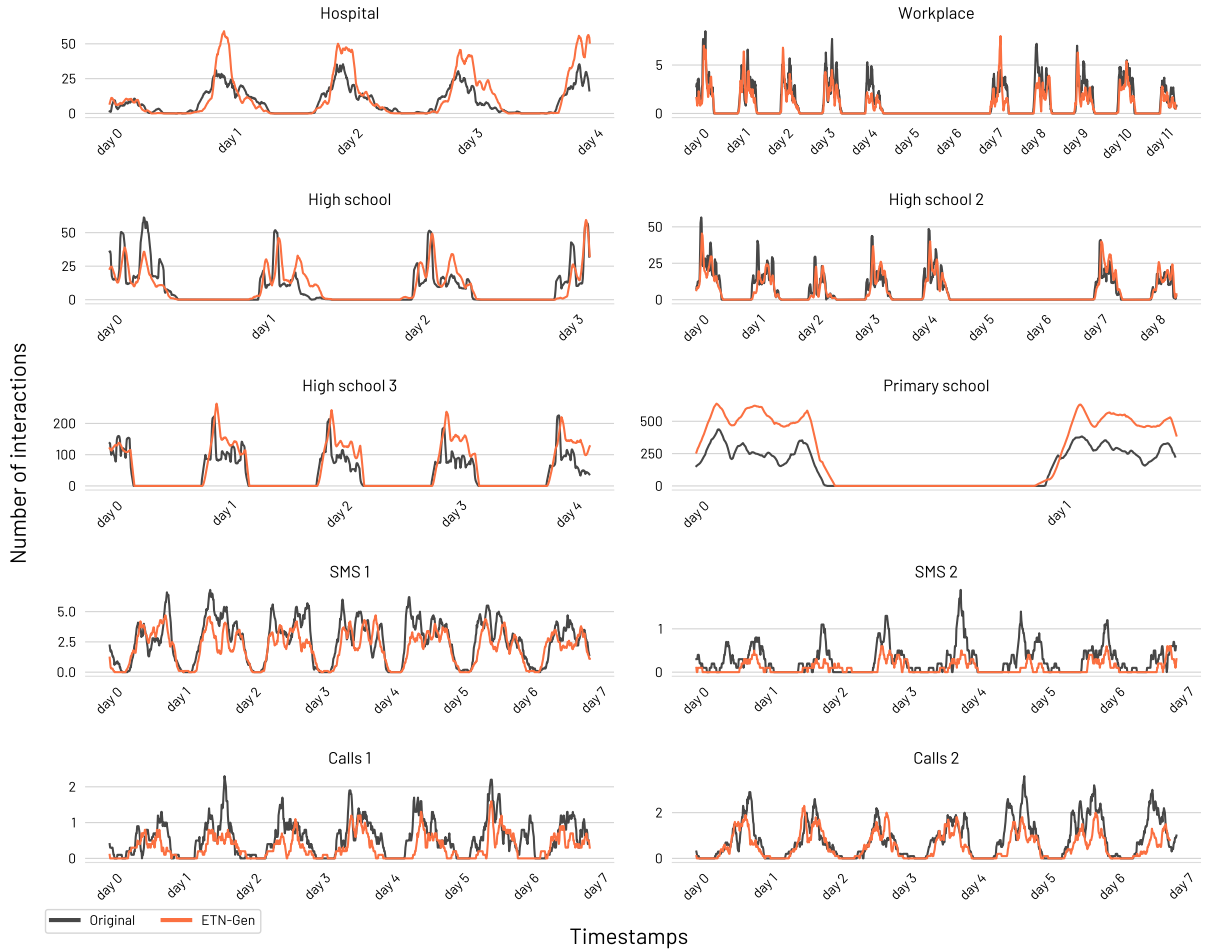


Figure 7.14: **Number of interactions in the generated network for different datasets.** Each panel shows the number of interactions of the original (black curve) and *ETN-gen* (orange curve) graphs. We use a temporal gap of 5 minutes for face-to-face interactions and 10 minutes for calls and SMS (intrinsically sparser networks).

much as k increases, the average number of interactions decreases. The second panel of figure 7.15 shows the topological similarity of several metrics with the original network. First of all, it is easy to see that for the first five metrics (number of interactions, density, interacting individuals, new conversation and s-metric) we can see a similar trend. In particular, the best results are obtained with $k \in \{2, 3\}$. Even in the edge strength and assortativity, the best score is obtained with $k = 2$. In the remaining metrics, we have different behaviours. Overall, with $k \in \{2, 3\}$ we have the best topological similarity with the original network. Finally, The last panel of figure 7.15 shows the dynamic similarity. The best coverage similarity is obtained with $k = 3$, while, the best MFPT similarity is obtained with $k = 5$. Overall, the best R_0 similarity is achieved with $k = 2$.

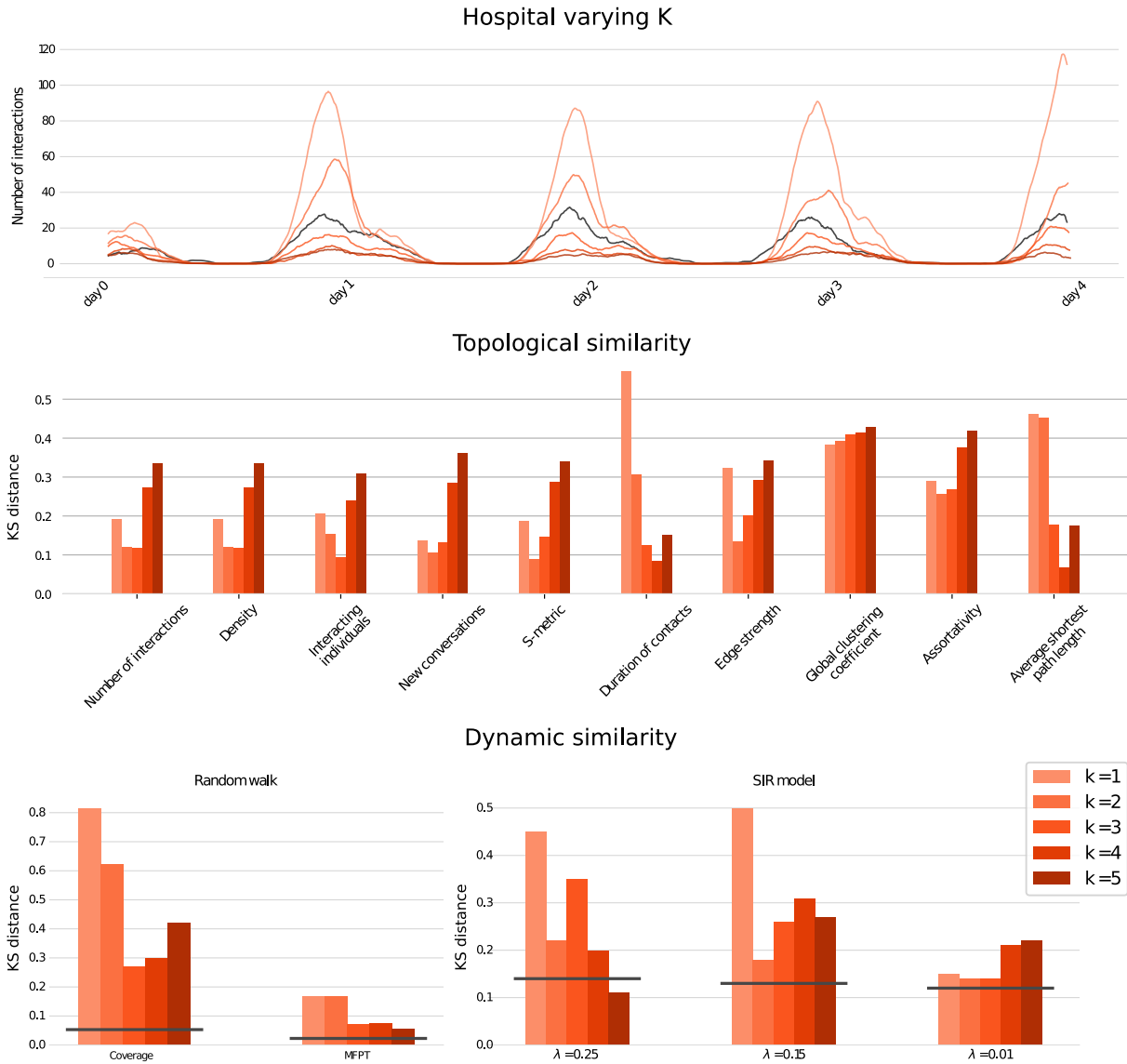


Figure 7.15: Number of interactions, topological and dynamic similarity on hospital network when K varies.

In conclusion, since we obtained similar results in others networks, in the manuscript we opt to use a fixed k equal to 2

7.2.4 Multiple versus single probabilistic model

In this section, we show that using an unique probabilistic model does not capture the daily/night periodicity. However, we are able to capture the average number of interac-

tions.

In the first panel of figure 7.16, it is shown the number of interactions of the original network (in black) the one generated by our method using multiple local probabilistic models (in orange) the number of interactions of the generated network with an unique local probabilistic model (in red) and those generated by Dymond and STM. It is worth to mention that our model with an unique probabilistic model match the average number of interactions, while competitors does not. In particular, the original average number of interactions is 8.27, while, our method has an average number of interaction equal to 8.27 and 8.19 for multiple and unique probabilistic model, respectively. On the other hand, the average number of interactions of Dymond and STM are 1.65 and 3.89.

In conclusion, it is true that using multiple probabilistic models stores more information of the input network. However, even using an unique probabilistic model, our method performs better than competitors.

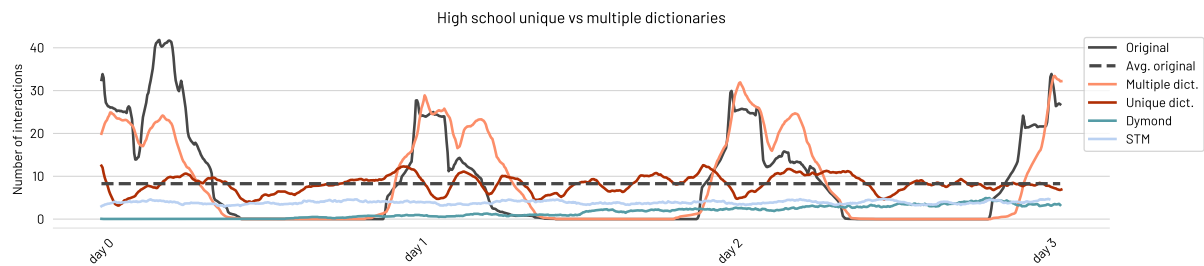


Figure 7.16: Number of interactions on high school network with multiple, an unique local probabilistic model and competitors.

