Characterization and Modeling of Time-Varying Networks

Ceria, A.

DOI
10.4233/uuid:a482f922-4c23-4b0e-bdbf-594b0c2b6142

Publication date
2023

Document Version
Final published version

Citation (APA)
Ceria, A. (2023). *Characterization and Modeling of Time-Varying Networks*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:a482f922-4c23-4b0e-bdbf-594b0c2b6142

Characterization and Modeling of Time-Varying Networks

Alberto Ceria

CHARACTERIZATION AND MODELING OF TIME-VARYING NETWORKS

CHARACTERIZATION AND MODELING OF TIME-VARYING NETWORKS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
donderdag 14 december 2023 om 10.00 uur

door

Alberto CERIA

Master of Science in Physics of Complex Systems,
University of Turin, Turin, Italy,
geboren te Turijn, Italië.

Dit proefschrift is goedgekeurd door de

promotor: Prof. dr. A. Hanjalic
copromotor: Dr. ir. H. Wang

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. A. Hanjalic,	Technische Universiteit Delft
Dr. ir. H. Wang,	Technische Universiteit Delft

Onafhankelijke leden:

Prof. dr. G. Petri	Notheastern University London
Prof. dr. R. Kooij	Technische Universiteit Delft
Prof. dr. O. Cats	Technische Universiteit Delft
Dr. L. Douw	Amsterdam Universitair Medische Centra
Dr. R. Quax	Universiteit van Amsterdam

Keywords: Temporal Networks, Higher-Order Networks, Network Models, Network Characterization Methods

Printed by: Ridderprint B.V.

ISBN 978-94-6469-703-2

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

And so it turned out that only a life similar to the life of those around us, merging with it without a ripple, is genuine life, and that an unshared happiness is not happiness. . .

Boris Pasternak , Dr. Zhivago

CONTENTS

SUMMARY

WE experience the effects of being connected parts of a whole every day. We receive news, messages or information by friends via virtual contacts on online social platforms, we can get infected by some friend while we meet them in person, we can reach far destinations by flights or experiencing the traffic congestion while driving by car in our city. In particular, epidemic and information spread through social interactions among individuals. These interactions are usually time-varying, and are represented accordingly by temporal networks. The understanding of the generating mechanism of temporal networks is thus crucial for predicting and controlling of spreading of epidemics and fake news. A first important step in addressing this challenge is the development of characterization methods able to compare different temporal networks in order to recognize common patterns or differences. The detected properties can indeed inspire the development of more realistic temporal network models able to reproduce these properties. This thesis will focus on proposing characterization methods for temporal networks. Besides this, we propose a methodology to identify the underlying spreading process of activity of nodes (or links), when the experimental record of activity is given and the process unfolds on a static network.

In Chapter 2, we analyze temporal networks to understand if contacts that occur close in time also tend to be close in topology. We explore the relationship between topological distance and temporal delay of contacts, and the temporal correlation within local neighborhood of a link. Our findings show that contacts close in time are generally close in topology, with virtual contacts exhibiting a stronger correlation. This is supported by higher local temporal correlation of contacts belonging to the neighborhood of different links. We observe this local temporal correlation in the form of long trains of consecutive contacts occurring at the neighborhood of a link. These trains are composed by the activity of many neighboring links. This is particularly evident in virtual contacts and in those social settings, such as primary school or museum, where individuals are less constrained in space. Such results may suggest that virtual communications, due to their low cost and easy access, may facilitate social contagion, i.e., the influence of a node activity on the activity of its neighbors. However, a limitation of our methodologies is that they assume interactions occur only between pairs of nodes.

In Chapter 3, we develop methods to characterize temporal higher order networks, where interactions involve groups of nodes larger than pairs. We apply these methods to collaboration and face-to-face (physical) interaction networks. Our findings show that nodes involved in many events (groups) composed of the same number of nodes, i.e., with the same order, tend to be involved in many events (groups) of different orders as well. However, significant differences emerge in the relation between topological and temporal properties of events in these two classes of networks. In physical contacts, events with different numbers of nodes occurring closely in time also tend to be close in topology. This is supported by the observation that locally close events are temporally correlated. In contrast, collaboration networks exhibit weak or absent correlation relation between topological distance and

temporal delay, and no local temporal correlation is observed. These differences likely arise from the fact that in physical contacts, differently from collaboration networks, interactions are partly driven by proximity, so that a set of individuals are close to each other, they are more likely to interact close in time among (subsets of) them.

A temporal network can be represented as a static network together with an unknown dynamical process that determines the activity of its connections. The observation that in face-to-face and virtual communication networks, contacts (and events) close in time tend to occur close in topology, seems to suggest that the occurrence of a contact/event may trigger the activity of neighboring contact events. This leads to the idea of modelling a temporal network as a spreading process of the activity, which should at least approximate the fact that the occurrence of a contact/event could trigger the activity of other contacts/events in its neighborhood. However, the assumption that the activity in temporal networks can spread among nodes/links is still not strongly supported, we apply our methodologies to a simpler case, where the spreading process is well supported by domain knowledge, i.e. the congestion contagion of airports, mediated by the air transportation network.

In Chapter 4 we proposed a methodology to identify the underlying spreading process among nodes, given the static topology on which the process occur and the experimental record of node activations. In particular, we tested the possibility of using a heterogeneous Susceptible- Infected-Susceptible (SIS) spreading process to model the congestion contagion of airports in the U.S. air transportation network to reproduce airport vulnerability, defined as their probability of being congested. We derive congestion probabilities from the U.S. Airport Network data and construct three types of airline networks to capture different flight characteristics. In our model, the infection rate of each link is proportional to its weight in the airline network, and the recovery rate of each airport depends on its node strength. Our heterogeneous model effectively reproduces nodal vulnerability and ranks airports better than a homogeneous model. We find that airports with intermediate strengths are the most vulnerable; the fact that the heterogeneous model can better capture this feature could partially explain its better performances.

The final chapter reflects on the insights of this thesis and suggests possible future directions related to our research.

1

Introduction

He who is unable to live in society, or who has no need because he is sufficient for himself, must be either a beast or a god.

Politics, Aristotle

1.1. Background

We live in a world that is increasingly interconnected. A flight can connect inhabitants of two cities that are thousands-kilometer far, the power grid allow to effectively transport electricity around different cities and online social media platforms allow communications among people living in different parts of the world. Besides this positive aspects, in our everyday life we also experience the negative effects of this increasing interconnectedness. The spread of diseases is inevitably facilitated by airline connection among cities, blackout can propagate across different cities and the spread of misinformation is facilitated by online communication. In each of these examples, we can recognize fundamental constituents such as airports (or cities), power stations and individuals that are connected in pair. Another crucial aspect of such examples is that the disease/blackout/fake news can only propagate from a constituent to another one which is connected to it. Note that these connections are not fully determined by the relative geographical location of the constituents. It seems natural then to study these systems as networks composed by nodes (representing the fundamental constituents) and links (representing the connections between the constituents).

Fueled by the increasing amount of available data, in the last 30 years, the field of network science [1–4] has shown itself as an effective way of representing and studying many different social, biological and communication systems. Initially, network science studied systems assuming that the connections among their nodes cannot change over time, i.e. static networks. The most important finding related to static network studies was the discovery that the number of links attached to a random node of a real network, i.e., its degree, is usually approximately distributed as a power-law. This fat-tailed distribution of the degree in the number of connections of a node seems ubiquitous in nature and it is shared by many real-world networks [1–4] and seems to explain many properties of real world systems, such as resilience to random failures and vulnerability to target attack [5–7] or epidemic spreading [8]. Collections of time-resolved interaction data are becoming increasingly available. For example, records of virtual messages exchange , face-to-face interactions among individuals collected in different social settings, or scientific papers (together with their corresponding author names and time of publication) are freely available on the web. In these cases, representing links that are active at a given timestamp as static links is an oversimplification. The dynamics of activation and deactivation of links affect spreading process, and other dynamical processes unfolding on the network [9–20]. Were link activations uncorrelated and uniformly spread in time, they could be included in the static description by assigning weights to the edges of the static network, so that the weights would represent the frequencies of events between nodes [21] and regulate the rate of interactions. However, human communications between pairs of individuals (or of a single individual towards their friends) usually evolves in bursts of consecutive communications followed by long period of inactivity [22–28].

As a result, in the last two decades a growing interest has been thus devoted to so called temporal networks [29–31], i.e., networks where links can be activated and deactivated over time. A link active at a given timestamp is also called contact. Large efforts were devoted to modeling temporal networks, e.g., via activity-driven models and extensions [32–36]. Such models are effective to capture certain network properties qualitatively, such as the distribution of nodes' degrees or the number of activations at each link in the time aggregated topology. Despite these benefits, however they are still far from being realistic models of temporal networks. The underlying generating mechanism of temporal networks is still unknown. This usually results in limitations in forecasting or controlling the performance of a spreading process unfolding upon the network, such as the prevalence of an epidemic spreading.

In this thesis we contribute to the grand goal of identifying the underlying mechanism of temporal networks in two main directions. First, we propose characterization methods to compare temporal networks with different number of nodes (or links) and observation time window, in order to find shared properties that can inspire the development of realistic temporal network models. Then, we propose a methodology to identify the spreading process of node/link activity, given an underlying network topology and the experimental record of node/link activity.

1.2. THESIS SCOPE AND CONTRIBUTIONS

Nowadays, we have a large availability of temporal network data. It is natural to ask what are the similarities or differences among these networks. To detect their common patterns or differences, early work have proposed characterization methods mainly focused on either temporal [22, 24, 25, 27, 28] or topological [3, 4, 21, 37–39] dimension separately. Recent studies have started to characterize both the topological and temporal properties together. For example, events of link addition and removal in online social network and communication platforms occur not randomly, but in bursts [40, 41]. Moreover, the ordered sequence of link activations among a small number of nodes conforming in a specified temporal interval have been grouped in different classes called temporal motifs, used to classify and characterize different temporal networks [42, 43]. Finally, Karsai et al. grouped the bursty train of consecutive activations of directed edges from a node to its neighbors counting the total number of activations in each train [44]. It has been shown that in the analyzed datasets (SMS exchange and phone calls) bursty trains of the total activity of links attached to a node are usually formed by the contacts of this node with a single other node. However, systematic methods to characterize simultaneously the temporal and topological relations of contacts for analyzing and better understanding real-world networks are still missing.

The first main research question that we address in this thesis is the following: *Can we propose systematic methods that can characterize simultaneously the temporal and topological relations of active connections?*

In Chapter 2, we addressed this question by proposing a method to characterize jointly topological and temporal properties of contacts, to investigate if contacts occurring within short temporal intervals tend to occur also close in their topological locations. The key positive aspects of this method are that (1) it characterizes the networks by considering both temporal and topological properties of contacts, (2) it can find similarities and differences between temporal networks with different size or observation times. We firstly examine the correlation of the time series of global activity, i.e. the time series recording the total number of contacts per each timestamp. Then we investigated the relation between the distance in topological locations of contacts and their temporal delay. Finally, we study the temporal correlation of contacts within each link egonetwork link. This is the local neighborhood centered around each link. By applying our method to several empirical networks of physical and virtual contacts, we discover that, in general, contacts close in time are close in topology. This phenomenon is particularly evident in virtual contacts, supporting the idea that virtual communications, by virtue of their low cost and easy access, permit social contagion, i.e. interactions of one individual can trigger the activity of its neighbors. A fundamental limitation of this method is that it can be applied only to traditional temporal networks, i.e., networks where nodes can only connect in pairs. However pairwise connections can only partially capture interactions among constituents of a system [45, 46].For example, a neuron may exchange signals from and to many different neighbouring neurons [47], individuals interacts in gatherings composed by many people [48], and scientific publications can be the result of the joint work of many co-authors [49]. Such interactions are named higher-order, to emphasize that they involve more than just a couple of nodes.

In Chapter 3, we generalize the characterization method of Chapter 2 to higher-order temporal networks, i.e., networks in which nodes can interact in groups composed of more than a couple. Group interactions among d nodes are then called events or hyperlink activations of size d. We focus on the interplay between topological and temporal properties of these higher order events. We applied our method to collaboration and face-to-face (physical) interaction networks. We found that in both classes of networks, nodes involved in many different groups (events) of a given order are likely involved in several different groups (events) of another order. However, substantial differences regarding the relation between topological and temporal properties of events were also found. In physical contacts, indeed, events involving different number of nodes (or equivalently with different order) occurring within short temporal delay tend to be also close in their topological location. This is also supported by the observation that, in these networks, local events that are close in topology are correlated in time. Differently in collaboration networks, the relation between topological distance and temporal delay is almost absent, and no local temporal correlation is observed. The detected differences between these two types of networks are likely due to the proximity nature of physical contact networks, which is substantially different from collaboration networks.

A fundamental grand challenge in temporal network studies is identifying the underlying generating mechanism of temporal networks. A temporal network can be considered as a static network with an unknown dynamic process unfolding on it that determines the active connections. The problem of identifying the generating mechanism of a temporal network is thus the problem of identifying the unknown dynamic process generating the activity of its connections. A key result of Chapters 2 and 3, is that, when we apply our methods to real-world physical and virtual contact networks, we observe that links (or hyperlinks) activating in short time delay tend to occur also close in topology. This seems to support the idea that the unknown dynamic process of links' activity can be (approximated as) a spreading process, where the activation of a link can trigger the activation of its neighboring links. However, in order to model the temporal network mechanism as a spreading process, we need two main ingredients: first, domain knowledge to support the assumption that the activation of a link will trigger the activation of its neighboring links; second, a methodology to identify the correct spreading process. The problem of identifying the underlying spreading process generating the activity of link of a network is equivalent to the problem of identifying the process that generates the activity of nodes in the line graph [1] of the network.

The second main question addressed in this question is thus the following: *How can we identify the underlying spreading process of node/link activity, given an underlying static*

[1] The line graph of a static network G is indeed the static network $L(G)$ in which each link in G is a node in $L(G)$ and two nodes in $L(G)$ are connected if the corresponding links in G share an ending node.

network and the the empirical records of node (or link) activity?

In Chapter 4, we address this question in the case of the process of airport congestion contagion in air transportation network, where each node is an airport and the activity of the node is the congestion of the corresponding airport. Note that this setting represents a simplified version of modeling temporal networks. In the case of airline transportation network, indeed, domain knowledge supports the assumption that the congestion state of an airport can trigger the congestion of other airports it is connected to, e.g., via delayed flights [50]. In particular, we examine if a heterogeneous Susceptible-Infected-Susceptible (SIS) spreading process on an airline network can model airport congestion contagion, and can successfully reproduce the airport vulnerability, i.e. the airport probability of being congested. We determine the vulnerability of each airport from U.S. Airport Network data. To capture the diversity of flight features, such as frequency and duration, we construct three types of weighted static airline networks. In our model, an infected node can infect a susceptible neighbor with a rate proportional to the weight of the link connecting the two nodes in the airline network. The recovery rate of each airport, which modulates the chances of the congested (infected) airport to recover to the susceptible (non-congested) state, is also heterogeneous. For each node, the recovery rate is dependent on the node strength in the network, which represents the total weight of the links connected to the node. This heterogeneity of recovery rates reflects the fact that larger airports with better infrastructure can recover faster from congestion [51]. The nodal infection probability in the meta-stable state is used as a prediction for airport vulnerability. Our model successfully reproduces the distribution of nodal vulnerability and the rank of nodes in vulnerability significantly outperforming the homogeneous SIS model with a uniform recovery rate. Interestingly, we find that the highest vulnerability is observed at airports with moderate strength in the airline network. This pattern is captured by our heterogeneous model but not by the homogeneous model, where airports with higher strength have a higher infection probability. This discrepancy partially explains the superior performance of the heterogeneous model.

1.3. PUBLICATIONS RELATED TO THIS THESIS

The following articles are produced by the author of this thesis while pursuing the doctoral degree at Delft University of Technology.

1. **Ceria, A.**, Havlin, S., Hanjalic, A., & Wang, H. (2022). *Topological–temporal properties of evolving networks.* Journal of Complex Networks, 10(5), cnac041 [**Chapter 2**].

2. **Ceria, A.**, & Wang, H. (2023). *Temporal-topological properties of higher-order evolving networks.* Scientific Reports, 13(1), 5885. [**Chapter 3**].

3. **Ceria, A.**, Köstler, K., Gobardhan, R., & Wang, H. (2021). *Modeling airport congestion contagion by heterogeneous SIS epidemic spreading on airline networks.* Plos One, 16(1), e0245043 [**Chapter 4**].

1.4. HOW TO READ THIS THESIS

The three main chapters of this thesis, i.e., Chapters 2, 3 and 4 adopt original publications. We provide the reference of each corresponding publication in the footnote of the heading page of each chapter. Each chapter correspond to an independent piece of work and can be read without reading the previous chapters. The notations may differ across the different chapters.

BIBLIOGRAPHY

[1] Réka Albert and Albert-László Barabási. “Statistical mechanics of complex networks”. In: *Reviews of modern physics* 74.1 (2002), p. 47.

[2] Mark EJ Newman. “The structure of scientific collaboration networks”. In: *Proceedings of the national academy of sciences* 98.2 (2001), pp. 404–409.

[3] Mark EJ Newman. “The structure and function of complex networks”. In: *SIAM review* 45.2 (2003), pp. 167–256.

[4] Stefano Boccaletti et al. “Complex networks: Structure and dynamics”. In: *Physics Reports* 424.4-5 (2006), pp. 175–308.

[5] Réka Albert, Hawoong Jeong, and Albert-László Barabási. “Error and attack tolerance of complex networks”. In: *nature* 406.6794 (2000), pp. 378–382.

[6] Reuven Cohen et al. “Resilience of the internet to random breakdowns”. In: *Physical review letters* 85.21 (2000), p. 4626.

[7] Duncan S Callaway et al. “Network robustness and fragility: Percolation on random graphs”. In: *Physical review letters* 85.25 (2000), p. 5468.

[8] Marián Boguá, Romualdo Pastor-Satorras, and Alessandro Vespignani. “Epidemic spreading in complex networks with degree correlations”. In: *Statistical mechanics of complex networks* (2003), pp. 127–147.

[9] Xiu-Xiu Zhan, Alan Hanjalic, and Huijuan Wang. “Information diffusion backbones in temporal networks”. In: *Scientific Reports* 9.1 (2019), pp. 1–12.

[10] Roni Parshani et al. “Dynamic networks and directed percolation”. In: *EPL (Europhysics Letters)* 90.3 (2010), p. 38004.

[11] Dávid X Horváth and János Kertész. “Spreading dynamics on networks: the role of burstiness, topology and non-stationarity”. In: *New Journal of Physics* 16.7 (2014), p. 073037.

[12] Jean-Charles Delvenne, Renaud Lambiotte, and Luis EC Rocha. “Diffusion on networked systems is a question of time or structure”. In: *Nature Communications* 6.1 (2015), pp. 1–10.

[13] Xiu-Xiu Zhan, Alan Hanjalic, and Huijuan Wang. “Suppressing Information Diffusion via Link Blocking in Temporal Networks”. In: *International Conference on Complex Networks and Their Applications*. Springer. 2019, pp. 448–458.

[14] René Pfitzner et al. “Betweenness preference: Quantifying correlations in the topological dynamics of temporal networks”. In: *Physical Review Letters* 110.19 (2013), p. 198701.

[15] Giovanna Miritello, Esteban Moro, and Rubén Lara. “Dynamical strength of social ties in information spreading”. In: *Physical Review E* 83.4 (2011), p. 045102.

[16] Mikko Kivelä et al. “Multiscale analysis of spreading in a large communication network”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.03 (2012), P03005.

[17] Márton Karsai et al. “Small but slow world: How network topology and burstiness slow down spreading”. In: *Physical Review E* 83.2 (2011), p. 025102.

[18] Oliver E Williams, Fabrizio Lillo, and Vito Latora. “How auto-and cross-correlations in link dynamics influence diffusion in non-Markovian temporal networks”. In: *arXiv preprint arXiv:1909.08134* (2019).

[19] Ville-Pekka Backlund, Jari Saramäki, and Raj Kumar Pan. “Effects of temporal correlations on cascades: Threshold models on temporal networks”. In: *Physical Review E* 89.6 (2014), p. 062815.

[20] Raj Kumar Pan and Jari Saramäki. “Path lengths, correlations, and centrality in temporal networks”. In: *Physical Review E* 84.1 (2011), p. 016105.

[21] Alain Barrat et al. “The architecture of complex weighted networks”. In: *Proceedings of the National Academy of Sciences* 101.11 (2004), pp. 3747–3752.

[22] K-I Goh and A-L Barabási. “Burstiness and memory in complex systems”. In: *EPL (Europhysics Letters)* 81.4 (2008), p. 48002.

[23] Albert-Laszlo Barabasi. “The origin of bursts and heavy tails in human dynamics”. In: *Nature* 435.7039 (2005), pp. 207–211.

[24] Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi. “Entropy of dialogues creates coherent structures in e-mail traffic”. In: *Proceedings of the National Academy of Sciences* 101.40 (2004), pp. 14333–14337.

[25] Joao Gama Oliveira and Albert-László Barabási. “Darwin and Einstein correspondence patterns”. In: *Nature* 437.7063 (2005), pp. 1251–1251.

[26] Alexei Vazquez et al. “Impact of non-Poissonian activity patterns on spreading processes”. In: *Physical Review Letters* 98.15 (2007), p. 158702.

[27] Julián Candia et al. “Uncovering individual and collective human dynamics from mobile phone records”. In: *Journal of Physics A: Mathematical and Theoretical* 41.22 (2008), p. 224015.

[28] Anders Johansen. “Probing human response times”. In: *Physica A: Statistical Mechanics and its Applications* 338.1-2 (2004), pp. 286–291.

[29] Petter Holme and Jari Saramäki. “Temporal networks”. In: *Physics Reports* 519.3 (2012), pp. 97–125.

[30] Petter Holme. “Modern temporal network theory: a colloquium”. In: *The European Physical Journal B* 88.9 (2015), p. 234.

[31] Naoki Masuda and Renaud Lambiotte. *A guide to temporal networks*. World Scientific, 2016.

[32] Nicola Perra et al. “Activity driven modeling of time varying networks”. In: *Scientific reports* 2.1 (2012), pp. 1–7.

[33] Márton Karsai, Nicola Perra, and Alessandro Vespignani. “Time varying networks and the weakness of strong ties”. In: *Scientific reports* 4.1 (2014), p. 4001.

[34] Antoine Moinet, Michele Starnini, and Romualdo Pastor-Satorras. “Burstiness and aging in social temporal networks”. In: *Physical review letters* 114.10 (2015), p. 108701.

[35] Laura Alessandretti et al. “Random walks on activity-driven networks with attractiveness”. In: *Physical Review E* 95.5 (2017), p. 052318.

[36] Giovanni Petri and Alain Barrat. “Simplicial activity driven model”. In: *Physical review letters* 121.22 (2018), p. 228301.

[37] Albert-László Barabási and Eric Bonabeau. “Scale-free networks”. In: *Scientific American* 288.5 (2003), pp. 60–69.

[38] Albert-László Barabási. “Scale-free networks: a decade and beyond”. In: *Science* 325.5939 (2009), pp. 412–413.

[39] J-P Onnela et al. “Structure and tie strengths in mobile communication networks”. In: *Proceedings of the National Academy of Sciences* 104.18 (2007), pp. 7332–7336.

[40] Hilla Brot et al. “Evolution through bursts: Network structure develops through localized bursts in time and space”. In: *Network Science* 4.3 (2016), pp. 293–313.

[41] Riivo Kikas, Marlon Dumas, and Márton Karsai. “Bursty egocentric network evolution in skype”. In: *Social Network Analysis and Mining* 3.4 (2013), pp. 1393–1401.

[42] 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*. 2017, pp. 601–610.

[43] Lauri Kovanen et al. “Temporal motifs in time-dependent networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2011.11 (2011), P11005.

[44] Márton Karsai, Kimmo Kaski, and János Kertész. “Correlated dynamics in egocentric communication networks”. In: *Plos One* 7.7 (2012), e40612.

[45] Federico Battiston et al. “Networks beyond pairwise interactions: structure and dynamics”. In: *Physics Reports* 874 (2020), pp. 1–92.

[46] Federico Battiston et al. “The physics of higher-order interactions in complex systems”. In: *Nature Physics* 17.10 (2021), pp. 1093–1098.

[47] Giovanni Petri et al. “Homological scaffolds of brain functional networks”. In: *Journal of The Royal Society Interface* 11.101 (2014), p. 20140873.

[48] Vedran Sekara, Arkadiusz Stopczynski, and Sune Lehmann. “Fundamental structures of dynamic social networks”. In: *Proceedings of the national academy of sciences* 113.36 (2016), pp. 9977–9982.

[49] Alice Patania, Giovanni Petri, and Francesco Vaccarino. “The shape of collaborations”. In: *EPJ Data Science* 6 (2017), pp. 1–16.

[50] Álvaro Rodrıguez-Sanz et al. “Assessment of airport arrival congestion and delay: Prediction and reliability”. In: *Transportation Research Part C: Emerging Technologies* 98 (2019), pp. 255–283.

[51] Richard De Neufville and Amedeo Odoni. *Airport Systems. Planning, Design and Management*. 2003.

2

Topological-Temporal Properties of Evolving Networks

[1] This chapter has been published as Ceria, A., Havlin, S., Hanjalic, A., & Wang, H. (2022). Topological–temporal properties of evolving networks. Journal of Complex Networks, 10(5), cnac041

Many real-world complex systems including human interactions can be represented by temporal (or evolving) networks, where links activate or deactivate over time. Characterizing temporal networks is crucial to compare different real-world networks and to detect their common patterns or differences. A systematic method that can characterize simultaneously the temporal and topological relations of the time specific interactions (also called contacts or events) of a temporal network, is still missing. In this chapter, we propose a method to characterize to what extent contacts that happen close in time occur also close in topology. Specifically, we study the interrelation between temporal and topological properties of the contacts from three perspectives: (1) the correlation (among the elements) of the activity time series which records the total number of contacts in a network that happen at each time step; (2) the interplay between the topological distance and time difference of two arbitrary contacts; (3) the temporal correlation of contacts within the local neighborhood centered at each link (so called ego network) to explore whether such contacts that happen close in topology are also close in time. By applying our method to 13 real-world temporal networks, we found that temporal-topological correlation of contacts is more evident in virtual contact networks than in physical contact networks. This could be due to the lower cost and easier access of online communications than physical interactions, allowing and possibly facilitating social contagion, i.e., interactions of one individual may influence the activity of its neighbors. We also identify different patterns between virtual and physical networks and among physical contact networks at, e.g., school and workplace, in the formation of correlation in local neighborhoods. Patterns and differences detected via our method may further inspire the development of more realistic temporal network models, that could reproduce jointly temporal and topological properties of contacts.

2.1. Introduction

Complex systems can be represented as networks, where nodes and links represent the components of a system and their interactions respectively. In a temporal or evolving network [1, 2], the network topology changes over time, or equivalently, pairs of nodes interact at specific time stamps. Such time-stamped interactions between nodes are called contacts or events. Early work on evolving networks and their characterization methods have mostly focused on either temporal [3–7] or topological [8–13] dimension separately but rarely on combining both [14–19]. Regarding the topological aspect, the aggregated networks, where two nodes are connected if they have at least one contact or interaction, have been characterized using classical static network analysis methods. Scaling properties such as a scale-free degree distribution have been observed in many real networks [8–11]. From the perspective of time dimension, it has been found that individuals tend to execute actions like contacts in bursts within a short time duration and such high activity periods are separated by relatively long inactive ones. The approximate scale-free distribution of the inter-event times of contacts of a node or of a system, the so-called burstiness, seems to be common in real-world temporal networks [3–7, 20, 21]. The temporal correlation of the events of a network has been measured by e.g. auto-correlation [22] and the distribution of the number of contacts in a bursty period, the so-called event train. [23].

Recent studies have started to characterize both the topological and temporal properties together. It has been observed that events of addition and removal of links by users do not occur sporadically at random nodes but rather occur in brief bursts in time and locally in topology, on both an online blogging platform and Skype [14, 15]. Temporal motifs are sets of contacts among a small number of nodes conforming to a specific pattern in topology and time ordering as well as a specific duration of time. The occurrence of diverse temporal motifs has been used to characterize and to classify evolving networks [16, 17]. Karsai et al. [18] characterized the sequence of contacts between each node and its neighbours using the distribution of the number of contacts in a bursty period, which is also called the event train size. However, it has been shown that bursty trains are usually formed by contacts between pair of nodes instead of in the aforementioned neighborhood of a node.

However, systematic methods to characterize simultaneously the temporal and topological properties of contacts/events to better understand real-world networks' differences and similarities are still missing. In this work, we aim to develop methods to characterize to what extent contacts that happen close in time (topology) are also close in topology (time). Specifically, we characterize the relationship between temporal and topological properties of the contacts in real evolving networks from the following three perspectives: (a) The auto-correlation of the activity time series which records the total number of contacts in a network that happen at each time step; (b) The interplay between the topological distance and temporal delay of two contacts; (c) The temporal correlation of contacts within local neighborhoods beyond a node pair. These perspectives characterize simultaneously both the temporal and topological interrelations of contacts from a global level to a more granular level. In order to be able to characterize and compare real-world networks, normalization and three control network randomizations have been designed in our characterization methods. We apply our method to 13 real-world physical and virtual contact networks. We find that the temporal and topological correlation tends to be more evident in virtual contact networks compared to physical contact networks. This is likely because the online communications, which are of lower cost and easier to perform than physical contacts, allows and possibly facilitates social contagion, i.e. the interaction of one individual to influence the activity of its neighbors. At the local neighborhood centered at each link, we observe long trains of events, i.e., consecutive activations of links in the neighborhood. In physical contact networks, the number of distinct links whose activations contribute to a train seems to reflect the spatial constrains of interactions. For example, the number of distinct links activated in a train is larger (smaller) in a primary school (workplace) where contacts are less (more) constrained in space.

The detected patterns and differences could further guide the development of evolving network models, pushing the boundary of temporal network models towards reproducing jointly realistic temporal and topological properties. Moreover, temporal network properties influence the dynamic process which unfolds on the network [19, 24–34]. The temporal and topological correlation in an evolving network discovered using our methods could possibly better explain the dynamic process than topological property or temporal property alone.

2.2. DEFINITIONS

2.2.1. REPRESENTATION OF A TEMPORAL NETWORK

A network whose topology vary over time is called a temporal or evolving network. It can be represented by $\mathscr{G} = (\mathscr{N}, \mathscr{L})$, where $\mathscr{N}$ is the set of nodes (with size $|\mathscr{N}| = N$), $\mathscr{L} = \{\ell(i,j,t), t \in [0,T), i,j \in \mathscr{N}\}$ is the set of contacts, and each element $\ell(i,j,t)$ indicates that a contact or an interaction between node i and j occurs at time t. A temporal network can also be represented by a 3 dimensional adjacency matrix $\mathscr{A}_{N\times N\times T}$ whose elements $\mathscr{A}(i,j,t)=1$ or $\mathscr{A}(i,j,t)=0$ represent, respectively, the presence or the absence of a contact between node i and j at time t.

We consider undirected temporal networks, where $\ell(i,j,t)=\ell(j,i,t)$ and $\mathscr{A}(i,j,t)=\mathscr{A}(j,i,t)$. By aggregating the contacts between each node pair over the whole observation time $[0,T-1]$ one obtains the time aggregated network $G_W=(\mathscr{N},\mathscr{L}_W)$. The aggregated network is static: two nodes i and j are connected, i.e., $e(i,j)\in\mathscr{L}_W$, if there is at least one contact between i and j over the observation time $[0,T-1]$. The adjacency matrix of the unweighted aggregated network is denoted by $A_{N\times N}$ whose element $A(i,j)=1$ or $A(i,j)=0$ depending whether i and j are connected or not. Each link $e(i,j)$ in $\mathscr{L}_W$ can be further associated with a weight $W(i,j)$, which represents the total number of contacts between i and j over the time window $[0,T-1]$. The corresponding weighted adjacency matrix $W_{N\times N}$ has elements $W(i,j)=\sum_{t=0}^{t=T-1}\mathscr{A}(i,j,t)$.

2.2.2. Temporal distance and topological distance between two contacts

The contacts between two arbitrary nodes i and j can be regarded as the activation of the link $e(i,j)\in\mathscr{L}_W$ at the corresponding time stamps. The activity between i and j can be represented by a time series $X_{ij}=\{x_{ij}(t)=\mathscr{A}(i,j,t), t\in[0,T-1]\}$. The link $e(i,j)$ is active at time t if there is a contact between i and j at time t, i.e. $x_{ij}(t)=\mathscr{A}(i,j,t)=1$. The total number of contacts in a network at each time stamp $Y=\{y(t)=\sum_{i,j\in\mathscr{N},i<j}x_{ij}(t), t\in[0,T-1]\}$ reflects the global activity of the temporal network over time. The temporal distance between two contacts $\ell(i,j,t)$ and $\ell(k,l,s)$ is $\mathscr{T}(\ell(i,j,t),\ell(k,l,s))=|t-s|$.

The topological distance, also called hopcount, between two nodes on a static network is the number of links contained in the shortest path between these two nodes. We define the topological distance $\eta(\ell(i,j,t),\ell(k,l,s))$ between two contacts $\ell(i,j,t)$ and $\ell(k,l,s)$ as the distance $\eta(e(i,j),e(k,l))$ between the corresponding two links $e(i,j)$ and $e(k,l)$ on the unweighted aggregated network, G_W. It can be derived as follows. The distance between the same link is zero, e.g. $\eta(e(i,j),e(i,j))=0$. The distance between two different links follows

$$\eta(e(i,j),e(k,l))=\min_{u\in\{i,j\},\ v\in\{k,l\}}(h(u,v)+1) \tag{2.1}$$

where $h(u,v)$ is the distance or hopcount between node u and v on the unweighted aggregated network G_W. The distance between two links is thus one plus the minimal distance between two end nodes of the two links. For example $\eta(e(i,j),e(i,k))=1$. Moreover, the line graph, e.g, G_W^L of a network G_W can be constructed by considering each link in G_W as a node, and two nodes are connected in G_W^L if the two corresponding links in G_W share a same end node. The distance (2.1) between two links in G_W equals the hopcount between their corresponding nodes in the line graph G_W^L.

2.2.3. Network randomization -control methods

In Section 2.4, we will explore diverse temporal-topological properties to understand the temporal and topological interrelations between contacts. However, real-world evolving networks may differ in, e.g., the number of nodes and the number contacts. In order to detect the non-trivial temporal-topological features and their interrelations in real-world networks, we compare each real-world network with its three controlled randomized networks which systematically preserve or remove specific topological and temporal correlation of contacts.

For a given temporal network $\mathscr{G}$, we introduce three randomized temporal networks $\mathscr{G}^1$, $\mathscr{G}^2$ and $\mathscr{G}^3$ respectively. Consider the set of contacts $\{\ell(i,j,t)\}$ in a temporal network $\mathscr{G}$, where each contact is described by its topological location, i.e., between pair of nodes (i,j) and its time stamp, t. Randomized network $\mathscr{G}^1$ is obtained by reshuffling the time stamps among the contacts, without changing the topological locations of the contacts. This randomization does not change the number of contacts between each node pair, only the timing is randomly changed, thus preserving the probability distribution of the topological distance of two randomly selected contacts. A temporal network can be also considered as an unweighted aggregated network and each link $e(i,j) \in \mathscr{L}_W$ is associated with its activity time series $\{\mathscr{A}(i,j,t), t \in [0, T-1]\}$. Randomized network $\mathscr{G}^2$ is obtained by iterating the step where two links are randomly selected from the aggregated network and their time series are swapped. This randomization does not change the distribution of the inter-event time of the activity of a random link, shown in Figures S3.4 (virtual contacts) and S3.5 (physical contacts). The third randomized network $\mathscr{G}^3$ is obtained by swapping the activity time series of two randomly selected links but with the same total number of contacts. This randomization preserves the number of contacts per node pair, the distribution of the inter-event time of contacts between a node pair and the distribution of the topological distance of two randomly selected contacts. The three randomized networks lead to the same unweighted aggregated network as the original network $\mathscr{G}$.

2.3. Datasets

All datasets of temporal networks are obtained from open access websites [123]. For each dataset, we consider nodes that belong to the largest connected component of the static aggregated network. The corresponding temporal network captures only the contacts between those nodes. Furthermore, we remove the long periods without any contact in the network, corresponding to e.g. night or weekend: we recognized these periods as outliers in the inter-event time [4] distribution of the global activity series Y that are far from the bulk. (see Figure 2.1). Finally, multiple contacts between the same pair of nodes at the same time step are accounted as a single contact. Details of the datasets are given in Table 3.1. In the original DNC Mail dataset [5], more than 96% of the total contacts forming the largest connected component occur in the last 33 days out of the 982 days. Hence, we include only the contacts of the last 33 days in our DNC Mail data.

[1] http://www.sociopatterns.org/

[2] http://konect.uni-koblenz.de/

[3] https://snap.stanford.edu/data/index.html

[4] The inter-event time t_{ie} is the time interval between the occurrence of two consecutive events. A global activity time series Y with total number of events $k = \sum_{t=0}^{T} y(t)$ has $k-1$ inter-event times. If two events are contemporary, their corresponding inter-event time is 0.

[5] http://konect.uni-koblenz.de/

Network	N	$\|\mathscr{L}_W\|$	$\|\mathscr{S}\|$	$\|\mathscr{L}\|$	T	dt	*contact type*
DNC Mail Part 2 (DNC_ 2 *) [35]	1598	4085	17300	30091	2861358	1	*virtual*
Manufacturing Email (ME*) [36]	167	3250	57791	82281	23430482	1	*virtual*
College Messages (CM*) [37]	1892	13833	58905	59789	16362751	1	*virtual*
Email EU (EEU*) [16, 38]	986	16025	206311	324933	44719809	1	*virtual*
Infectious (Infectious) [39]	410	2765	1392	17298	1421	20	*physical*
Primary School (PS) [40]	242	8317	3099	125771	3098	20	*physical*
High School 2012 (HS2012) [41]	180	2220	11267	45047	14114	20	*physical*
High School 2013 (HS2013) [42]	327	5818	7371	188504	7370	20	*physical*
Hypertext 2009 (HT2009) [39]	113	2196	5243	20818	7226	20	*physical*
SFHH Conference (SFHH) [43, 44]	403	9565	3508	70261	3799	20	*physical*
Workplace 2013 (WP) [45]	92	755	7095	9827	17844	20	*physical*
Workplace 2015 (WP2) [46]	217	4274	18479	78246	20946	20	*physical*
Hospital (Hospital) [47]	75	1139	9452	32424	16026	20	*physical*

Table 2.1: Basic features of the empirical networks after data processing. The number of nodes ($N = |\mathscr{N}|$), the number of links in $\mathscr{L}_W$ ($|\mathscr{L}_W|$), the number of snapshots ($|\mathscr{S}|$), the total number of contacts ($|\mathscr{L}|$), the length of the observation time window in time steps (T), the time resolution or duration of each time step (dt) in seconds and contact type are shown.

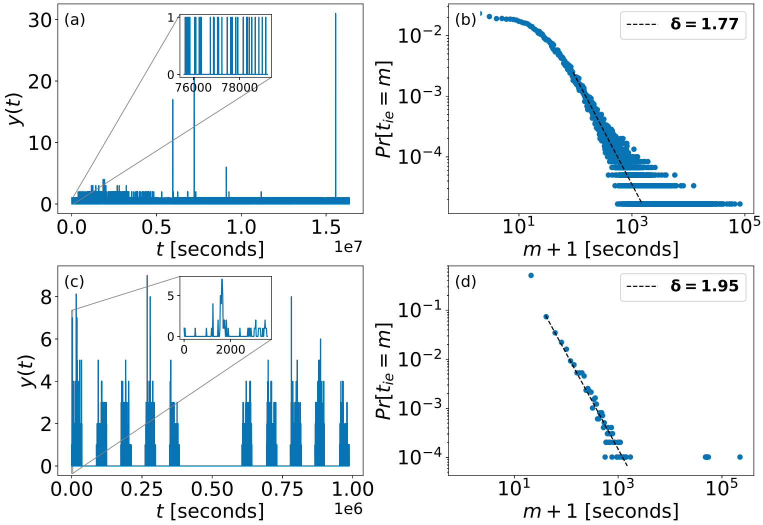

Figure 2.1: Global activity (left) and its inter-event time distribution (right) in (a-b) virtual contact network CM and (c-d) physical contact network WP. The dashed line indicates the slope δ of the power-law fit and the scaling region, obtained via Clauset's method [48]. If the goodness of the power-law fit is significantly better than the exponential fit, the value of δ is reported in bold characters [6]. Time is expressed in seconds. Values of global activity are the total number of contacts occurred in each step of dt seconds. Insets in left figures show global activity for one hour. In WP, long time periods of null global activity correspond to night and weekend periods. These periods correspond to isolated outliers in the global inter-event time distribution with $m > 10^4 s$ and are removed in the data processing.

2.4. CHARACTERIZING TOPOLOGICAL-TEMPORAL PROPERTIES OF EVOLVING NETWORKS

In this section, we propose a systematic method to characterize topological-temporal properties of the contacts in an evolving network. In Subsection 2.4.1 we focus on the characterization of temporal properties, while in Section 2.4.2 and 2.4.3 we characterize the joint topological and temporal features of contacts.

2.4.1. TEMPORAL ANALYSIS OF GLOBAL ACTIVITY

The time series of global activity $Y = \{y(t), t \in [0, T-1]\}$ records the total number of contacts at each time step $t \in [0, T-1]$. In this section, we analyze the correlation among the elements of the global activity time series.

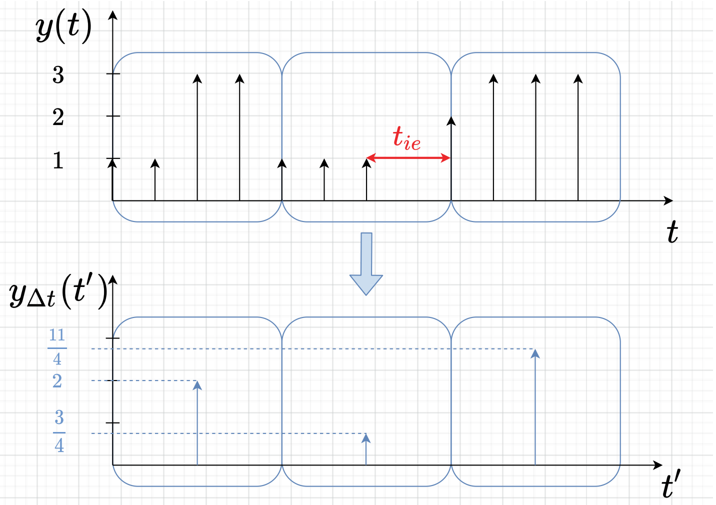

Figure 2.2: Construction of the aggregated activity series $y_{\Delta t}(t')$ from the global activity time series $y(t)$, where $\Delta t = 4$ time steps. In the top sub-figure, we present the event sequence of $y(t)$, where each vertical line indicates the timing of one (or more -depending on thickness) event(s), while t_{ie} is the inter-event time and two events happening at the same time have an inter-event time zero.

We aggregate the global activity at each time bin of duration Δt time steps as follows. The time steps $t \in [0, T-1]$ can be divided into a set of non-overlapping consecutive time bins of duration Δt. The aggregated activity $y_{\Delta t}(t')$ at a time bin $[t'\Delta t, t'\Delta t + \Delta t)$ is the average activity of $y(t)$ within the time bin $[t'\Delta t, t'\Delta t + \Delta t)\}$, as shown in Figure 2.2. Given bin duration Δt, the aggregated time series of activity is $Y_{\Delta t} = \{y_{\Delta t}(t'), 0 \le t' \le \left\lfloor \frac{T-1}{\Delta t} \right\rfloor - 1\}$.

To evaluate the correlation among the elements of the activity time series Y, we investigate $\frac{Var[Y_{\Delta t}]}{Var[Y]}$, the ratio of the variance of the aggregated $Y_{\Delta t}$ to that of the original time series Y, as a function of Δt (see Figure 2.3).

[6] This evaluation is performed via the likelihood ratio test on power-law and exponential fits. If the test indicate a better performance of the power-law fit with p-value $p < 0.05$, then the exponent of power-law fit δ is reported in bold characters.

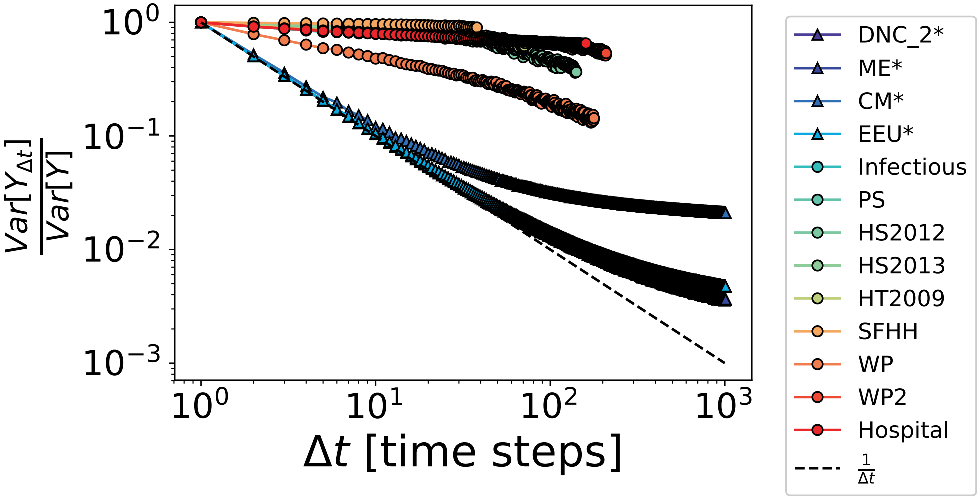

Figure 2.3: The normalized variance $\frac{Var[Y_{\Delta t}]}{Var[Y]}$ as a function of the aggregation resolution Δt. Circles correspond to physical contact temporal networks, triangles correspond to virtual contact networks (online messages and mail), while the black dashed line ($\frac{\text{Var}[Y_{\Delta t}]}{\text{Var}[Y]} = \frac{1}{\Delta t}$) represents the uncorrelated curve. The resolution Δt is in units of time steps.

Firstly, we derive $\frac{\text{Var}[Y_{\Delta t}]}{\text{Var}[Y]}$ analytically for the general case and then prove that $\frac{\text{Var}[Y_{\Delta t}]}{\text{Var}[Y]} = \frac{1}{\Delta t}$, if each element of Y is independent. The global activity Y can be considered as a realization of a set of T random variables $\{\hat{Y}(t)\}$, that are identically distributed as a variable $\hat{Y}$. Hence, $Var[\hat{Y}(t)] = Var[\hat{Y}]$, where $t \in [0, T-1]$. The variance of the aggregated activity $\hat{Y}_{\Delta t}(t')$ at a random time biin t' with $0 \le t' \le \left\lfloor \frac{T-1}{\Delta t} \right\rfloor - 1$ follows

$$\begin{aligned}
\text{Var}\Big[\hat{Y}_{\Delta t}(t')\Big] &= \text{Var}\Big[\frac{1}{\Delta t}\sum_{t=t'\Delta t}^{t'\Delta t+\Delta t-1} \hat{Y}(t)\Big] \\
&= \frac{1}{(\Delta t)^2}\sum_{t=t'\Delta t}^{t'\Delta t+\Delta t-1}\Big(\text{Var}[\hat{Y}(t)] + 2\sum_{t'\Delta t\le s<k\le t'\Delta t+\Delta t-1} \text{Cov}[\hat{Y}(s),\hat{Y}(k)]\Big) \\
&= \frac{\text{Var}[\hat{Y}]}{\Delta t} + \frac{2}{(\Delta t)^2}\sum_{t=t'\Delta t}^{t'\Delta t+\Delta t-1}\sum_{t'\Delta t\le s<k\le t'\Delta t+\Delta t-1}\text{Cov}[\hat{Y}(s),\hat{Y}(k)]
\end{aligned} \tag{2.2}$$

When the activity $\hat{Y}(t)$ at each time t is independently distributed, i.e. the set $\{\hat{Y}(t)\}$ are independent, the second term is zero and we have $\frac{\text{Var}[\hat{Y}_{\Delta t}(t')]}{\text{Var}[\hat{Y}]} = \frac{1}{\Delta t}$.

This explains why $\frac{\text{Var}[Y_{\Delta t}]}{\text{Var}[Y]} = \frac{1}{\Delta t}$ in Figure 2.3 when we randomly re-shuffle the global activity $Y = \{y(t), t \in [0, T-1]\}$ in each of the thirteen temporal networks. Figure 2.3 shows that $\frac{\text{Var}[Y_{\Delta t}]}{\text{Var}[Y]} > \frac{1}{\Delta t}$ in all real-world temporal networks, suggesting the correlation among the number of contacts per time step at different time steps. Moreover, it is seen that the physical contact networks are further away from $\frac{\text{Var}[Y_{\Delta t}]}{\text{Var}[Y]} = \frac{1}{\Delta t}$ compared to the virtual contact networks, reflecting higher correlation in physical contacts than in virtual activities. The higher correlation in global activity in physical contacts networks than in virtual networks seems to be supported by the relatively higher probability for the global inter-event time to be relatively small in physical contact networks (see Figure S3.3) than that in virtual contact ones (see Figure S3.2).

2.4.2. Topological and temporal distances between two contacts

Next we wish to explore the relation between the topological distance and temporal distance of two contacts. Firstly, we explore whether contacts that are close in time are also close in topology. Contacts of temporal networks are measured at discrete time steps. The duration of each time step is either 1 or 20 seconds in the datasets listed in Table 3.1. To compare physical and virtual contact networks, we present the time distance between any two contacts in units of seconds.

We analyze the average topological distance $E[\eta(\ell,\ell')|\mathscr{T}(\ell,\ell') < \Delta t]$ of two contacts given that their temporal distance is less than Δt. If topological and temporal distances of two contacts are independent, $E[\eta(\ell,\ell')|\mathscr{T}(\ell,\ell') < \Delta t] = E[\eta(\ell,\ell')]$ does not depend on the temporal distance Δt. Figures 2.4 and 2.5 show that $E[\eta(\ell,\ell')|\mathscr{T}(\ell,\ell') < \Delta t]$ increases with Δt in real-world temporal networks. That is, contacts that are close in time are typically also close in topology. Such an increasing trend or correlation between temporal and topological distances in each real-world temporal network is evidently higher than that in the corresponding three randomized networks. Network $\mathscr{G}^3$ (swapping the activity time series of the two randomly selected links but with the same total number of contacts) preserves more properties of the original temporal network compared to $\mathscr{G}^1$ (swapping timestamps among contacts) and $\mathscr{G}^2$ (swapping the activity time series of two randomly selected links). Consistently, the increasing trend between temporal and topological distances is significantly reduced in $\mathscr{G}^3$, reduced further in $\mathscr{G}^2$ and disappears in $\mathscr{G}^1$. The slight initial decrease and afterwards increase of $E[\eta(\ell,\ell')|\mathscr{T}(\ell,\ell') < \Delta t]$ with Δt in $\mathscr{G}^2$ can be explained by the changes of the probability that a couple of contacts with temporal distance smaller than Δt are activations of the same link with Δt (see the detailed discussion in Section 2.6.2 of Appendix). The increase of $E[\eta(\ell,\ell')|\mathscr{T}(\ell,\ell') < \Delta t]$ with Δt in $\mathscr{G}$, in comparison with that in $\mathscr{G}^2$, is more significant in virtual contact networks and physical contact networks Infectious. In these networks, contacts that occur close in time tend to be close in topology. The stronger correlation in virtual networks and the physical network Infectious has also been observed when the other methods are used to characterize the temporal-topological correlation of contacts (see Subsections 2.4.3 and 2.4.3). The high correlation in network Infectious is related to the specific properties of this network. Network Infectious records the contacts among visitors of a museum and only people that visit the museum at a similar time could have contacts [39].

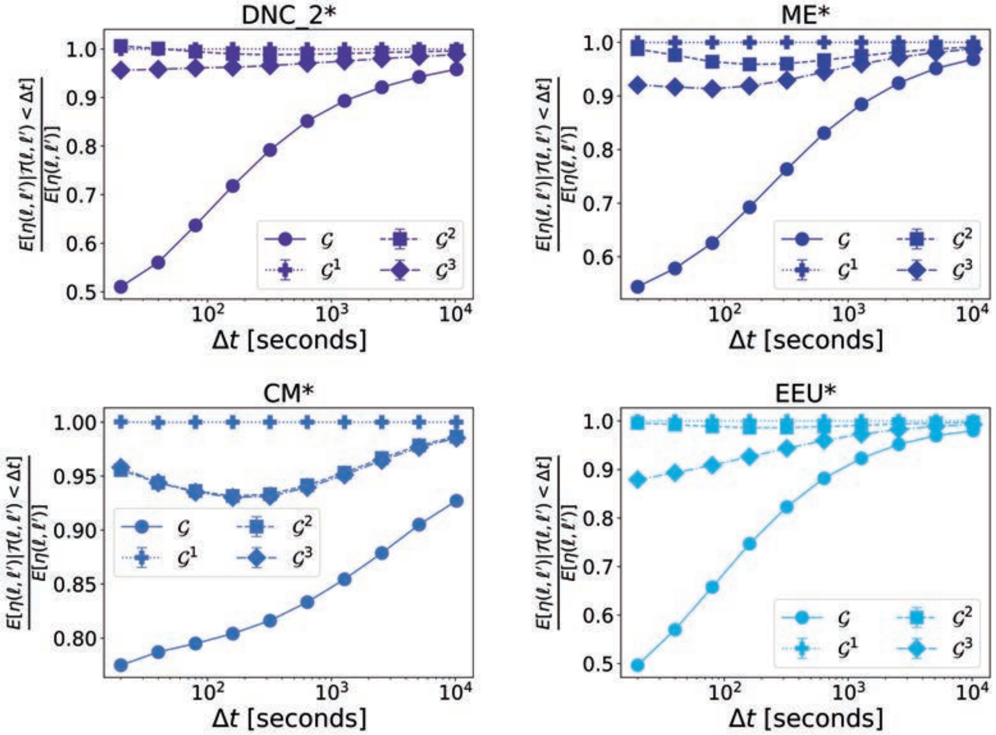

Figure 2.4: $\frac{E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell')<\Delta t]}{E[\eta(\ell,\ell')]}$ as a function of Δt for real world network $\mathcal{G}$ (points, solid line) and the three randomized reference models $\mathcal{G}^1$ (pluses, dotted line), $\mathcal{G}^2$ (squares, dash line) and $\mathcal{G}^3$ (diamonds, dash-dotted line) in each virtual contact dataset. When $\frac{E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell')<\Delta t]}{E[\eta(\ell,\ell')]} = 1$, topological and temporal distances are independent. Moreover, $\lim_{\Delta t\to\infty} E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell') < \Delta t] = E[\eta(\ell,\ell')]$. The results for each of the three randomized networks are obtained from 10 independent realizations of the randomized network. Note that the horizontal axis is presented in logarithmic scale.

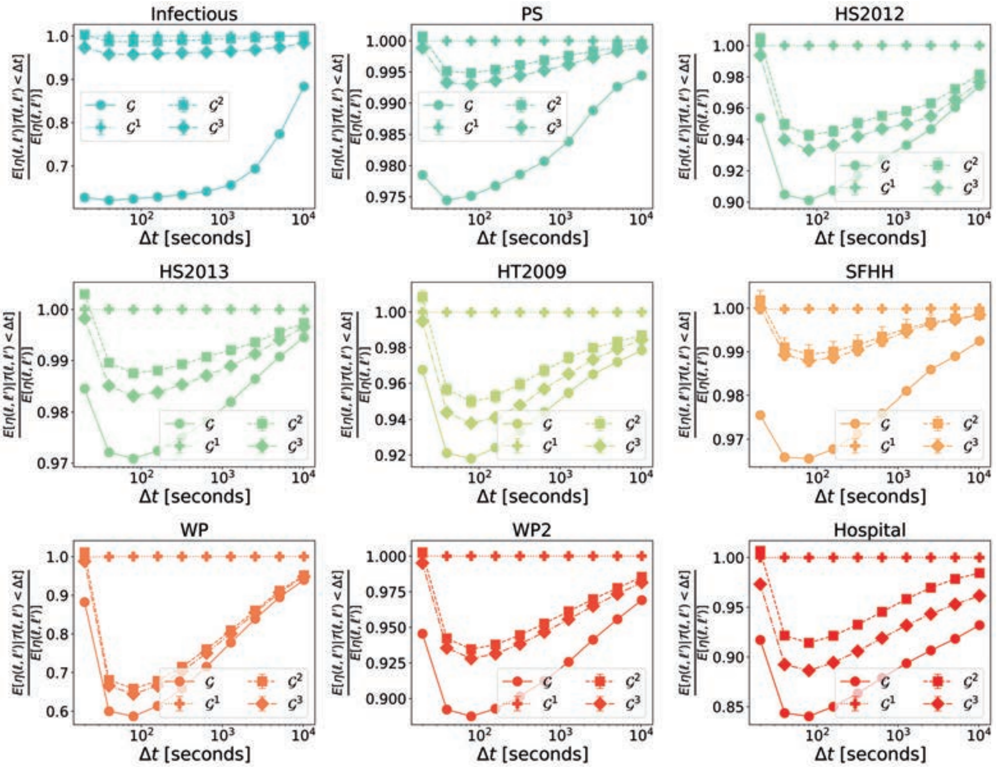

Figure 2.5: $\frac{E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell')<\Delta t]}{E[\eta(\ell,\ell')]}$ as a function of Δt for real world network $\mathcal{G}$ (points, solid line) and the three randomized reference models $\mathcal{G}^1$ (pluses, dotted line), $\mathcal{G}^2$ (squares, dash line) and $\mathcal{G}^3$ (diamonds, dash-dotted line) in each physical contact dataset. When $\frac{E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell')<\Delta t]}{E[\eta(\ell,\ell')]} = 1$, topological and temporal distances are independent. Moreover, $\lim_{\Delta t\to\infty} E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell') < \Delta t] = E[\eta(\ell,\ell')]$. For each of the three randomized models, the lines and corresponding error bars correspond to the average and standard deviation of the results obtained from 10 independent realizations. Note that the horizontal axis is presented in logarithmic scale.

We could also identify the temporal and topological correlation of contacts via $E[\mathcal{T}(\ell,\ell')|\eta(\ell,\ell') = j]$, the average temporal distance of two contacts given that their topological distance is j. However, this measure could be limited in distinguishing the difference among networks due to the small diameter, i.e., the maximal hopcount of real-world networks.

2.4.3. Local events

In this section, we explore the temporal correlation of contacts that happen within (at any link of) a local neighborhood in the aggregated network. The local neighborhood refers to the ego network $ego(e(i,j))$ centered at a link $e(i,j)$ which consists of the link itself and all its neighboring links that share a common node with the link $e(i,j)$. The objective is to understand whether and how local events are correlated in time, in forming trains (bursts) of events, where events within a burst have short inter-event times and trains are separated by a long inactive period.

Temporal correlation of contacts at an ego network

We will analyze the activity (sequence) of an $ego(e(i,j))$ which records the number of contacts that happen within the ego network at each time step, and equals the sum of the activity time series of every link in $ego(e(i,j))$ (see Figure 2.6).

To evaluate correlation of local events in forming trains of events, we study the train size distribution [18] of the activity sequence of an ego network. A train of events is a sequence of consecutive contacts/events whose inter-event times are shorter than or equal to Δt and separated from the other contacts by an inter-event time larger than Δt. Given a Δt, trains can be identified for each ego network activity sequence (see Figure 2.6). Given a Δt and a temporal network, the train size distribution $Pr[\mathscr{E}_{\Delta t} = s]$, i.e., the probability that the size $\mathscr{E}_{\Delta t}$ of a random train is s can be derived from the trains of all the ego networks centered at every link. The train size distribution is compared between each real-world network and its three randomized networks.

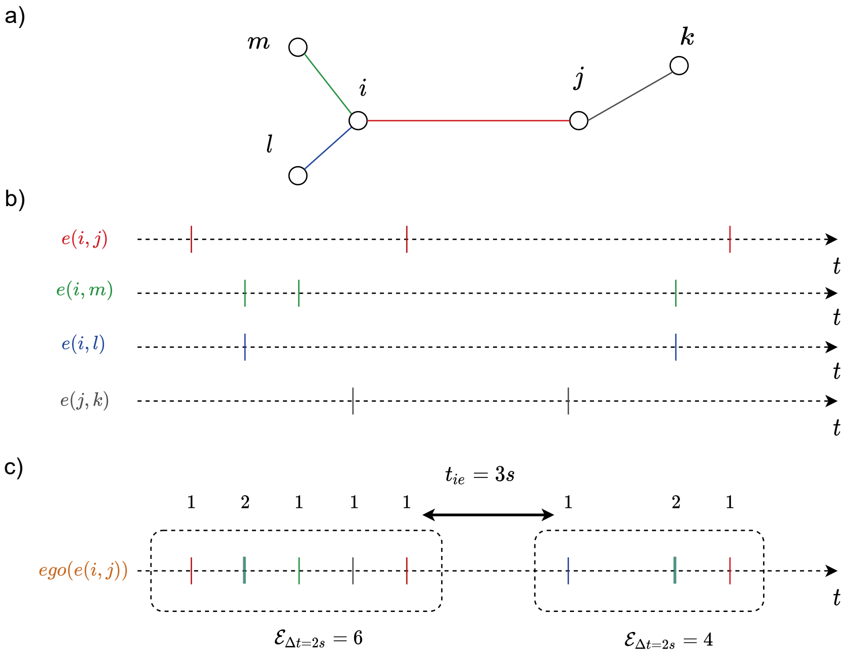

Figure 2.6: Schematic representation of a) the ego network of the link $e(i,j)$, i.e. $ego(e(i,j))$, b) the time series associated to each links in $ego(e(i,j))$, c) the activity time series of $ego(e(i,j))$, which is the sum of the time series of links belonging to the ego network, and its event trains when $\Delta t = 2s$.

We find that the train size distribution when $\Delta t = 60s$ in a real-world network has an evidently higher tail than that of the corresponding randomized networks in the four real-world virtual contact network (Figure 2.7) and the physical contact network Infectious (Figure 2.8). Later, we will explain why $\Delta t = 60s$ is representative. In these real-world networks, local events have a higher chance to form long trains, than in their corresponding randomized networks. Randomized network $\mathscr{G}_2$ is obtained by shuffling the activity sequences among the links, thus preserving the set of link activity sequences but removing their correlation with the network topology. The difference between the train size distribution in the ego networks of $\mathscr{G}_2$ and an exponential distribution (the train size distribution when the inter-event times in the activity sequence are independent[7]) reflects solely the temporal correlation of events in a link activity sequence in real-world networks. The different train size distributions in real-world networks $\mathscr{G}$ and their corresponding randomized networks $\mathscr{G}_2$ in Figures 2.7 and 2.8 indicate that temporal correlation of activities at each link is insufficient to explain the temporal correlation of contacts at ego networks. Instead, the correlation between the activity sequences and topology also contributes. Such temporal correlation of local activities suggests that neighboring nodes tend to have contacts or activities within a short time. The evidently stronger correlation observed in virtual networks and the physical network Infectious is in line with the finding in Section 2.4.2.

The choice of Δt is non-trivial. Karsai et al. [18] have observed a power-law train size distribution $Pr[\mathscr{E}_{\Delta t} = s] \sim s^{-\beta}$ of the activity of a link with $\beta = 0.39$ (0.42) in voice calls (SMS) temporal contact network and found that the power-law exponent remains approximately the same when Δt varies within a broad range. Our comparison of train size distribution with different Δt for virtual (Figure S3.6) and physical contact datasets (Figure S3.7) shows that when Δt is small ($\Delta t \leq 120s$), the distribution is fat-tailed. The exponent β of the power-law fit seems more stable across different values of Δt in virtual contacts (FigureS3.6) datasets than in physical contact ones (Figure S3.7). The changes in the shape of the train size distribution of physical contact datasets are likely due to finite size effects which emerge because of limited duration of empirical temporal networks' observation window. When Δt is sufficiently large, for example, any ego network has a single train, whose size is the total number of contacts that occur within the ego network. Figures 2.4 and 2.5 show that the positive correlation between topological and time distances (in linear scale) of two contacts is more evident when the time distance is small. Moreover, the observation time windows of temporal networks, especially physical contact networks, are short in duration. All these perspectives motivate us to consider a small Δt, e.g. $\Delta t = 60s$. Moreover, our observations are similar when $\Delta t = 120s$ and when $\Delta t = 60s$ for all the analysis. Hence, we focus our discussion on $\Delta t = 60s$ and all the results when $\Delta t = 120s$ are given in the Appendix.

[7] The train size distribution follows an exponential function $Pr[\mathscr{E}_{\Delta t} = s] = Pr[t_{ie} \leq \Delta t]^{(s-1)}(1 - Pr[t_{ie} \leq \Delta t])$ when the inter-event times in the activity sequence are independent. Such exponential function is the product of the probability of observing s-1 inter-event times shorter than or equal to Δt, and a single inter-event time longer than Δt.

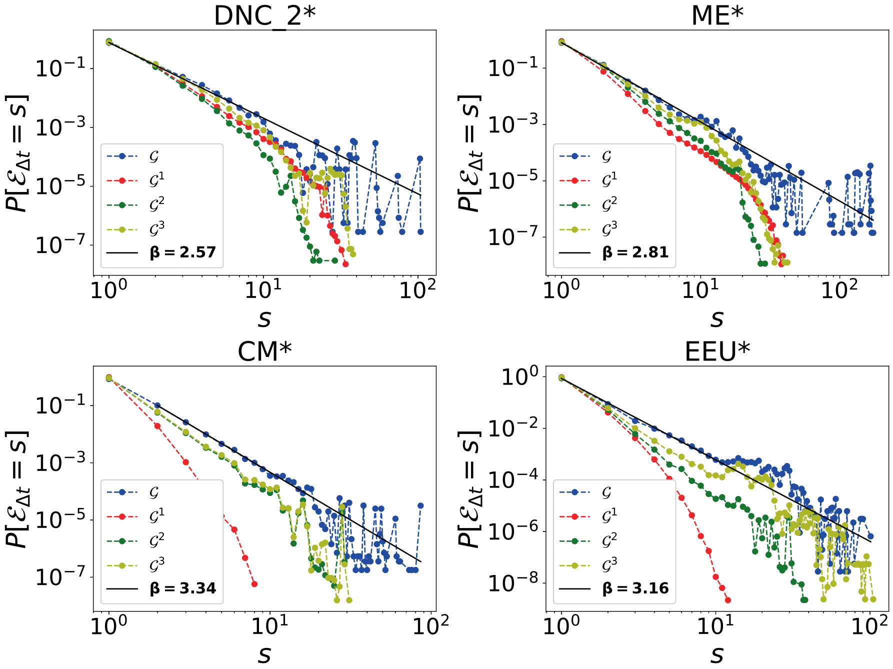

Figure 2.7: Train size distribution ($\Delta t = 60s$) of ego network activity for $\mathscr{G}$ (blue), $\mathscr{G}_1$ (red), $\mathscr{G}_2$ (green), $\mathscr{G}_3$ (yellow) of virtual contact datasets. The black solid line represents the fit $P[\mathscr{E}_{\Delta t} = s] \sim s^{-\beta}$ to the distribution of the train size of $\mathscr{G}$ with $\Delta t = 60s$. The power law fit and its fitting region were computed with Clauset's method [48]. If the goodness of the power-law fit is significantly better than the exponential fit (likelihood ratio test with p-value $p < 0.05$), the value of β is reported in bold characters.

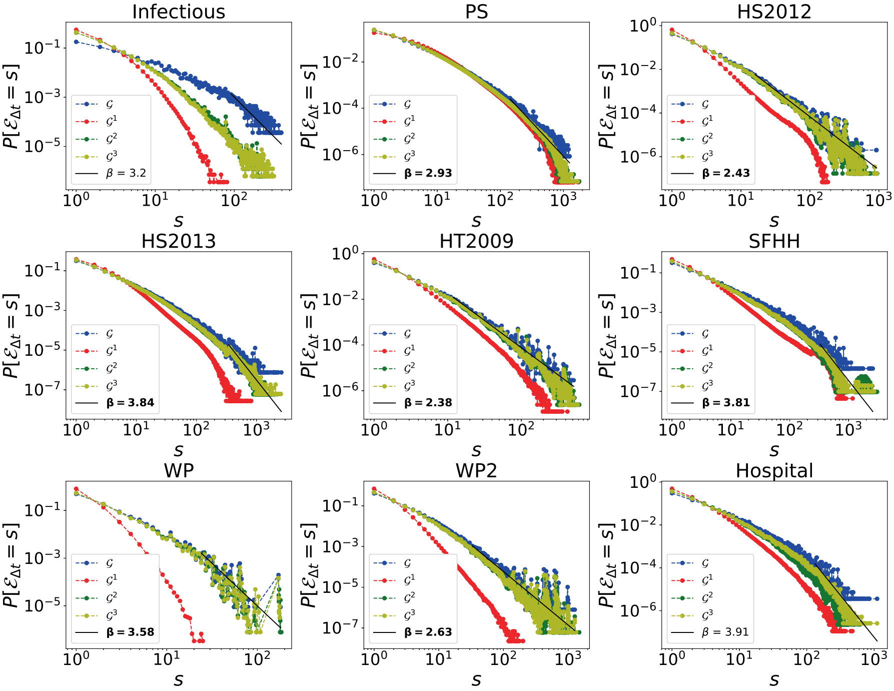

Figure 2.8: Train size distribution ($\Delta t = 60s$) of ego network activity for $\mathscr{G}$ (blue), $\mathscr{G}_1$ (red), $\mathscr{G}_2$ (green), $\mathscr{G}_3$ (yellow) of physical contact datasets. The black solid line represents the fit $P[\mathscr{E}_{\Delta t} = s] \sim s^{-\beta}$ to the distribution of the train size of $\mathscr{G}$ with $\Delta t = 60s$. The power law fit and its fitting region were computed with Clauset's method [48]. If the goodness of the power-law fit is significantly better than the exponential fit (likelihood ratio test with p-value $p < 0.05$), the value of β is reported in bold characters.

EGO NETWORK ACTIVITY VERSUS LINK ACTIVITY

We investigate further whether the temporal correlation of contacts that occur within an ego network in forming long event trains could be explained or introduced by the temporal correlation of contacts at each single link.

Firstly, we explore whether each activity train of an ego network contains the activities (contacts) of a single link or multiple links in the ego network. We examine the number $\mathscr{M}$ of distinct active links that a train of an ego network involves. Specifically, each identified train of an ego network is composed of a set of contacts, occurring at a subset of links within the ego network, the so-called active links. For each real-world network and given $\Delta t = 60s$, trains are identified for every ego network centered at each link, and the number $\mathscr{M}$ of distinct active links of each train is counted. Figure 2.9 illustrates the average number of active links $\frac{E[\mathscr{M}|\mathscr{E}_{\Delta t}=s]}{s}$ for trains with size $\mathscr{E}_{\Delta t} = s$, normalized by the train size s, for virtual and physical contact networks, respectively. In all networks the fraction of active links $\frac{E[\mathscr{M}|\mathscr{E}_{\Delta t}=s]}{s}$ is above $\frac{E[\mathscr{M}|\mathscr{E}_{\Delta t}=s]}{s} = 1/s$ suggesting that a train usually involves far more than 1 active link. Interestingly, we observe in all 9 physical contact networks a seemingly power-law decay $\frac{E[\mathscr{M}|\mathscr{E}_{\Delta t}=s]}{s} \sim s^{-\alpha}$ (right plot of Figure 2.9). In contrast, $\alpha \approx 0$, or equivalently $E[\mathscr{M}|\mathscr{E}_{\Delta t} = s] \sim s$ in virtual contact networks, especially mail dataset, i.e. EEU, ME and DNC2 (left plot of Figure 2.9). This suggests that, in virtual contact networks, each train is mostly composed of the activities of many links in an ego network.

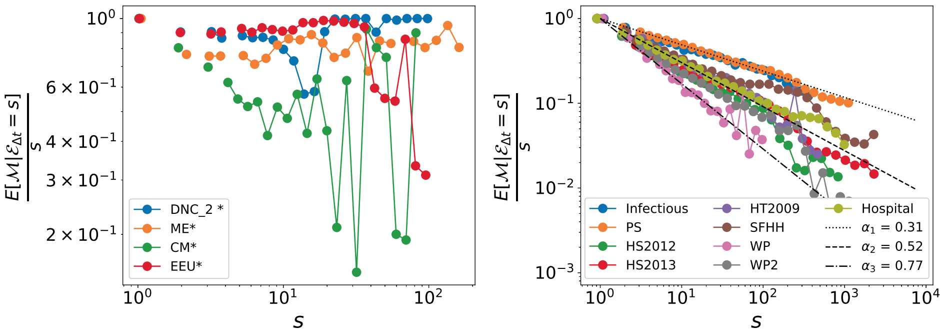

Figure 2.9: The average number of active links $\frac{E[\mathcal{M}|\mathcal{E}_{\Delta t}=s]}{s}$ for trains with size $\mathcal{E}_{\Delta t} = s$, normalized by the train size s of the ego networks for virtual (left) and physical (right) contact datasets, when $\Delta t = 60s$. The three reference lines in right plot indicate $\frac{E[\mathcal{M}|\mathcal{E}_{\Delta t}=s]}{s} = s^{-\alpha}$ with slope $\alpha_1 = 0.31$ (dotted), $\alpha_2 = 0.52$ (dashed) and $\alpha_3 = 0.77$ (dash-dot). Note that the horizontal and vertical axes are presented in logarithmic scales. In total 30 logarithmic bins are split within the interval $[1, s_{max}]$, where s_{max} is the largest train size observed in the considered real temporal network.

We compare further the physical contact networks. Their power-law exponents are within $0.31 \le \alpha \le 0.77$. The slope of the power-law decay seems to be influenced by the type of human interaction and spacial constraints of the contact environment. Networks that lead to the slowest decay, i.e. $\alpha \approx 0.31$ are Infectious and PS datasets, which are contact networks in a museum and primary school respectively. The two contact networks of employees at a work place, WP and WP2 have the largest slope $\alpha \approx 0.77$. The other networks, i.e., contacts of high school students, conference participants have a power-law exponent in between $0.31 \le \alpha \le 0.77$. A similar trend has been observed when $\Delta t = 120s$ (see Figure S3.12 in Appendix). These observations could be explained by the spatial constraints of contacts and the nature that younger students tend to interact with many others in an active period. The bursty events of a train tend to engage the largest number of links in an ego network in network Infectious and PS than the other physical contact networks. This could be due to the freedom for individuals to move in the museum and in the primary school (relative to the small museum/class room) and the tendency that primary school students interact with many others in an active period. The other way round, employees at a work place are confined in space (their offices) and tend to interact with limited number of colleagues during a train of activities. In this sense, virtual contacts are the least confined to space, leading thus to a larger number of active links than physical contacts.

Whether each activity train of an ego network contains the activities of a single link or of multiple links could also be reflected via the train size distribution in an ego network versus the train size distribution in a link. In Figures 2.10 and 2.11, we compare the train size distribution (with $\Delta t = 60s$) of the activity sequence of single links, of the most active single links (top 10% of links with the largest number of contacts) and of ego networks. The trains of ego networks tend to be longer than those of single links and the most active single links, in all networks except for WP. Therefore, the trains of the ego network are usually the results of the activity of more than one link. The same observations are obtained when $\Delta t = 120s$ (see Figures S3.10 and S3.11 in Appendix). The similar train size distribution in ego networks and in links in the WP dataset is consistent with the largest power-exponent observed in Figure 2.9. In WP, a train of an ego network is composed of the activity of relatively few links.

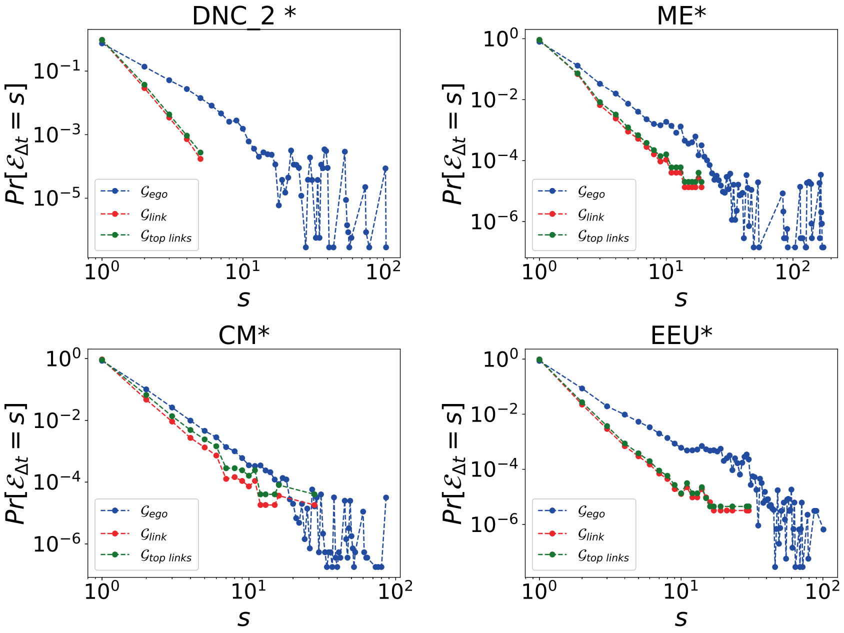

Figure 2.10: Train size distribution ($\Delta t = 60s$) of ego network activity (blue), single link activity (red), most active link activity (green) of virtual contact datasets. Note that the horizontal and vertical axes are presented in logarithmic scales.

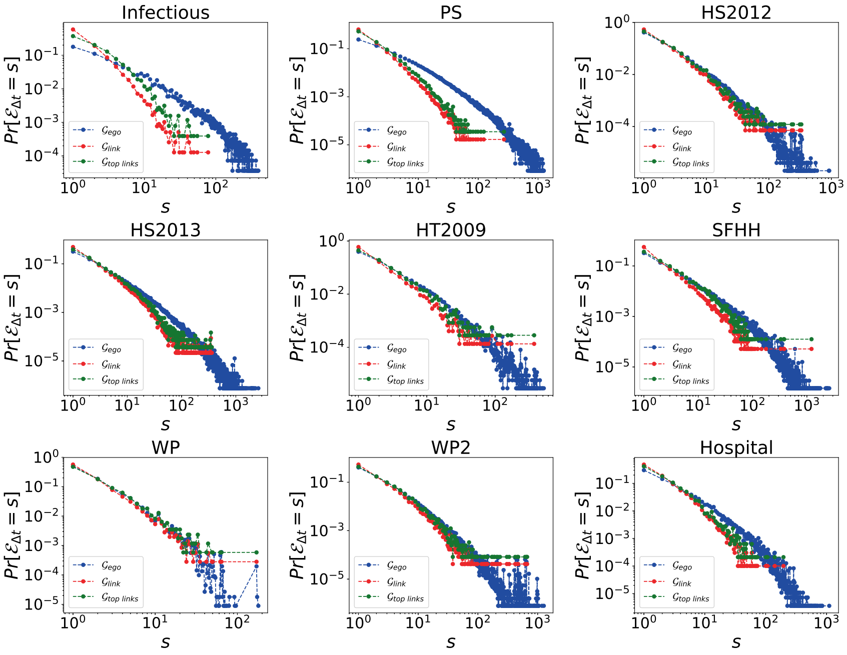

Figure 2.11: Train size distribution ($\Delta t = 60s$) of ego network activity (blue), single link activity (red), most active link activity (green) of physical contact datasets. Note that the horizontal and vertical axes are presented in logarithmic scales.

Ego network activity versus node activity

We further address the question whether a train at an ego network $ego(e(i,j))$ involves the activity of both end nodes i and j, or only one of them. Event trains at ego networks engaging activities of both end nodes may suggest a possible social contagion in activity between nodes.

For each train of an ego network $ego(e(i,j))$, we consider the events that associate with only one end node but not both. Among these events, we count the fraction of events ϕ_i and ϕ_j that associate with end node i and j respectively and $\phi_i + \phi_j = 1$. The maximum of the two fractions $B = max(\phi_i, \phi_i)$ quantifies how unbalanced the activities of the two end nodes i and j contribute to a train and is called the activity balance of a train.

Table 2.2 shows the average activity balance $E[B]$ and the probability $Pr[B \leq 0.95]$ of the activity balance for all contact networks, accounting all trains (whose sizes are larger than 1) of all ego networks. We find that the average activity $E[B] < 1$, suggesting that an activity train in an ego network $ego(e(i,j))$ engages in the activity of both end nodes i and j. This is in line with the previous finding that the activity correlation in an ego network cannot be explained by the activity of a single link. Moreover, the activity is found to be larger, thus more unbalanced, in virtual contacts than in physical contact networks. This is likely because, in a virtual contact network like email contact network, an individual tends to contact many others at a similar time. A train of events at an ego network $ego(e(i,j))$ in a virtual network contains mainly the activities of a single end node i or j.

Virtual Contacts			Physical Contacts		
Dataset	$P[B \leq 0.95]$	$E[B]$	Dataset	$P[B \leq 0.95]$	$E[B]$
DNC 2*	0.05	0.98	Infectious	0.38	0.87
ME*	0.12	0.95	PS	0.44	0.86
CM*	0.03	0.99	HS2012	0.16	0.95
EEU*	0.12	0.94	HS2013	0.25	0.93
			HT2009	0.21	0.94
			SFHH	0.19	0.95
			WP	0.06	0.98
			WP2	0.1	0.97
			Hospital	0.12	0.97

Table 2.2: Probability $P[B \leq 0.95]$ and average $E[B]$ of the activity balance B in virtual contact (left) and physical contact (right) networks.

2.5. CONCLUSIONS

In this chapter, we developed systematically methods to characterize jointly the topological and temporal properties of contacts in a time-evolving network, ranging from global network level to local neighborhoods. Via applying these methods to real-world networks, we identified substantial differences between virtual and physical contact networks.

We find that contacts that occur close in time tend to be close in topology and this trend is more evident in virtual contact networks compared to physical contact networks. This is in line with the observation that the contacts within an ego networks tend to have a higher chance to form long trains and thus happen closely in time in real-world networks. Such activity correlation is more evident in virtual contact networks. Moreover, an event train of an ego network $ego(e(i,j))$ is mostly composed of the activities of multiple component links. Interestingly, more links tend to be engaged in e.g., virtual networks and physical contact network primary school where the contacts are less constrained in space, in contrast to e.g., the contact network at workplace. These may suggest that contacts with a low cost may better facilitate social contagion, i.e. influence between neighboring nodes in the activity. Finally, an event train of an ego network $ego(e(i,j))$ usually contains the activity of both ends, node i and j. Two connected nodes, thus, tend to have contacts with their neighbors close in time. The two end nodes' contributions are more unbalanced in virtual contacts than in physical contacts, likely driven by the nature that in a virtual (e.g. email) contact network, an individual tends to contact many others close in time.

Our methods are confined to undirected networks. A full-fledged directed temporal network characterization method is deemed as promising to develop. The application of these methods may enhance our understanding of diverse time-evolving systems and allow exploration of the influence of detected properties/patterns on a dynamic process upon the network. Finally, the detected patterns may further inspire the development of more realistic temporal network models that reproduce key realistic temporal and topological properties of contacts.

2.6. APPENDIX

2.6.1. DATASETS DESCRIPTION (* INDICATES VIRTUAL CONTACTS)

- **Manifacturing Email (ME) ***: Emails exchanged between 167 employees of a mid-size company in Poland, observation time: 270 days, time resolution 1 s

- **European Union Mail(EEU) ***: Emails exchanged between 986 accounts of a large European research institution during a period from October 2003 to May 2005 (18 months), time resolution 1 s

- **Democratic National Committee Mail ***: Emails of 1900 members (1598 after preprocessing) of the Democratic National Committee, in our case only final 33 days were considered, because they are more than 95% of the entire corpus of email, time resolution 1 s

- **College Messages (CM)** *: messages from an online community of 1899 (1892 after preprocessing) students at the University of California, Irvine. Time span of approximately 6 months, time resolution 1 s

- **Hypertext 2009 (HT09)** face-to-face interactions (Rfid sensors, range of 1.5-2 m, time resolution of 20s) of the 113 participants to Hypertext conference, during 3 days.

- **Infectious (Science Gallery, Dublin)** face-to-face interactions (Rfid sensors, range of 1.5-2 m, time resolution of 20s) of 14000 visitors (410 after preprocessing) at the Science Gallery of Dublin, during 3 months of observation (after preprocessing, i.e. selecting the largest connected component, 1 day). Community structure linked to time of visit (only visitors present at the same time can interact)

- **Workplace (WP)** face-to-face interactions (Rfid sensors, range of 1.5-2 m, time resolution of 20s) of 92 employees in one of the two office buildings of the InVS, located in Saint Maurice near Paris, France, during two weeks. Each participant belongs to a department (5 in total), so the network has community structure.

- **Workplace (WP2)** Second deployment of WP, same details as WP, but larger number of participants (217) and more departments included (12). Each participant belongs to a department, so the network has community structure.

- **SFHH Conference (SFHH)** face-to-face interactions (Rfid sensors, range of 1.5-2 m, time resolution of 20s) of 403 participants to the 2009 SFHH conference in Nice, France (June 4-5, 2009).

- **Primary School (PS)** face-to-face interactions (Rfid sensors, range of 1.5-2 m, time resolution of 20s) of 242 individuals (232 children and 10 teachers) in a primary school in Lyon, France during two days in October 2009. Each kid or teacher belongs to a class, so the network has community structure.

- **High school 2012 (HS2012)** face-to-face interactions (Rfid sensors, range of 1.5-2 m, time resolution of 20s) of 180 students of five classes of a high school in Marseilles, France, during 7 days (from a Monday to the Tuesday of the following week) in Nov. 2012. Each student belongs to a class, so the network has community structure.

- **High school 2013 (HS2013)** face-to-face interactions (Rfid sensors, range of 1.5-2 m, time resolution of 20s) of 327 students of nine classes of a high school in Marseilles, France, during 5 days in Dec. 2013. Each student belongs to a class, so the network has community structure.

- **Hospital** face-to-face interactions (Rfid sensors, range of 1.5-2 m, time resolution of 20s) between patients, patients and health-care workers (HCWs) and among HCWs in a hospital ward in Lyon, France, from Monday, December 6, 2010 at 1:00 pm to Friday, December 10, 2010 at 2:00 pm. The study included 46 HCWs and 29 patients.

2.6.2. $E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell') < \Delta t]$ in $\mathcal{G}^2$

In this subsection, we will explain the initial decreasing trend of $E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell') < \Delta t]$ with Δt observed in $\mathcal{G}^2$ in every considered dataset. In general,

$$E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell') < \Delta t] = E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell') < \Delta t, \eta(\ell,\ell') > 0]Pr[\eta(\ell,\ell') > 0|\mathcal{T}(\ell,\ell') < \Delta t] \tag{2.3}$$

where $E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell') < \Delta t, \eta(\ell,\ell') > 0]]$ is the average topological distance of couple of contacts ℓ,ℓ' that are not activations of the same link ($\eta(\ell,\ell') > 0$), given that their temporal distance $\mathcal{T}(\ell,\ell') < \Delta t$. In $\mathcal{G}^2$, $E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell') < \Delta t, \eta(\ell,\ell') > 0] \approx E[\eta(\ell,\ell')]$, i.e., the average topological distance of a couple of random contacts ℓ,ℓ' which are not activations of the same link and have temporal distance $\mathcal{T}(\ell,\ell') < \Delta t$ does not depend on Δt, as shown in Figure S3.1. By substituting this approximation in Equation 2.3, we obtain $E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell') < \Delta t] \approx E[\eta(\ell,\ell')]Pr[\eta(\ell,\ell') > 0|\mathcal{T}(\ell,\ell') < \Delta t]$. As shown in Figure S3.1, $Pr[\eta(\ell,\ell') > 0|\mathcal{T}(\ell,\ell') < \Delta t]$ and $E[\eta(\ell,\ell')|\mathcal{T}(\ell,\ell') < \Delta t]$ as a function of Δt follow the same trend and obtain the minimum at the same value of Δt. This is likely due to the relatively bursty activation patterns of single links, i.e., the high chance of observing small inter-event times in the time series of activations of single links in the considered temporal networks (see Figures S3.4 and S3.5))

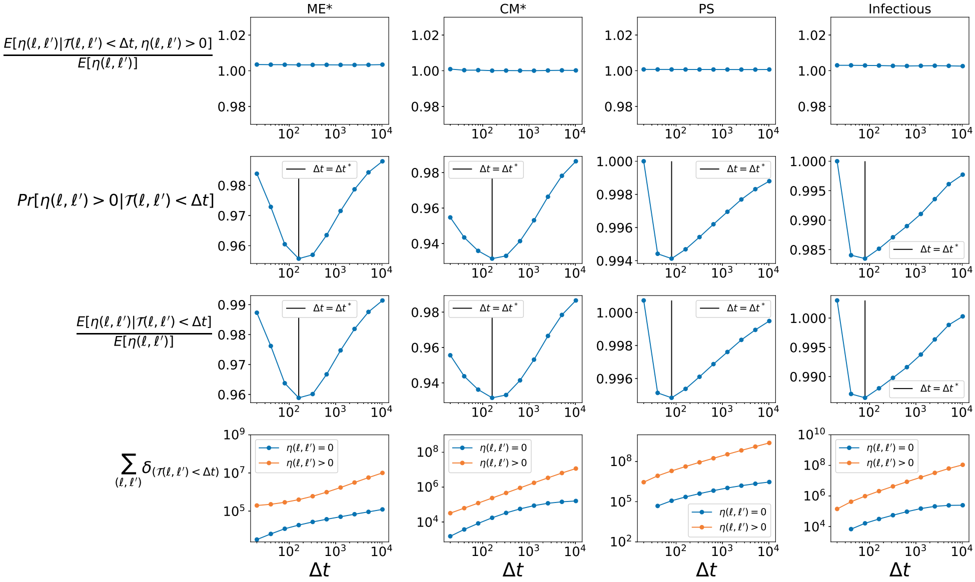

Fig. S3.1: The average topological distance $E[\eta(\ell,\ell')|\mathscr{T}(\ell,\ell') < \Delta t, \eta(\ell,\ell') > 0]$ of two random contacs ℓ, ℓ' which are not activations of the same link and have temporal distance $\mathscr{T}(\ell,\ell') < \Delta t$ (first row), the probability $Pr[\eta(\ell,\ell') > 0|\mathscr{T}(\ell,\ell') < \Delta t]$ of observing two random contacts ℓ, ℓ' which are not activations of the same link and have temporal distance $\mathscr{T}(\ell,\ell') < \Delta t$ (second row), the average topological distance $E[\eta(\ell,\ell')|\mathscr{T}(\ell,\ell') < \Delta t]$ of a couple of random contacts ℓ, ℓ' with temporal distance $\mathscr{T}(\ell,\ell') < \Delta t$, and the number of couples of contacts (ℓ, ℓ') with temporal distance $\mathscr{T}(\ell,\ell') < \Delta t$ (fourth row) and topological distance $\eta(\ell,\ell') = 0$ (blue) or $\eta(\ell,\ell') > 0$ (yellow) as a function of Δt in randomized reference model $\mathscr{G}^2$ for two examples of virtual (ME, CM) and physical (Infectious, PS) contact datasets. First and third row vertical axes are presented normalized by the average topological distance of contacts $E[\eta(\ell,\ell')]$. In second and third row the value $\Delta t = \Delta t^*$ where $Pr[\eta(\ell,\ell') > 0|\mathscr{T}(\ell,\ell') < \Delta t]$ and $E[\eta(\ell,\ell')|\mathscr{T}(\ell,\ell') < \Delta t]$ reach their minimum is highlighted. The results are the average of 10 independent realizations of randomized network $\mathscr{G}^2$. Horizontal axes are presented in logarithmic scale.

2.6.3. Global probability distribution of inter-event times

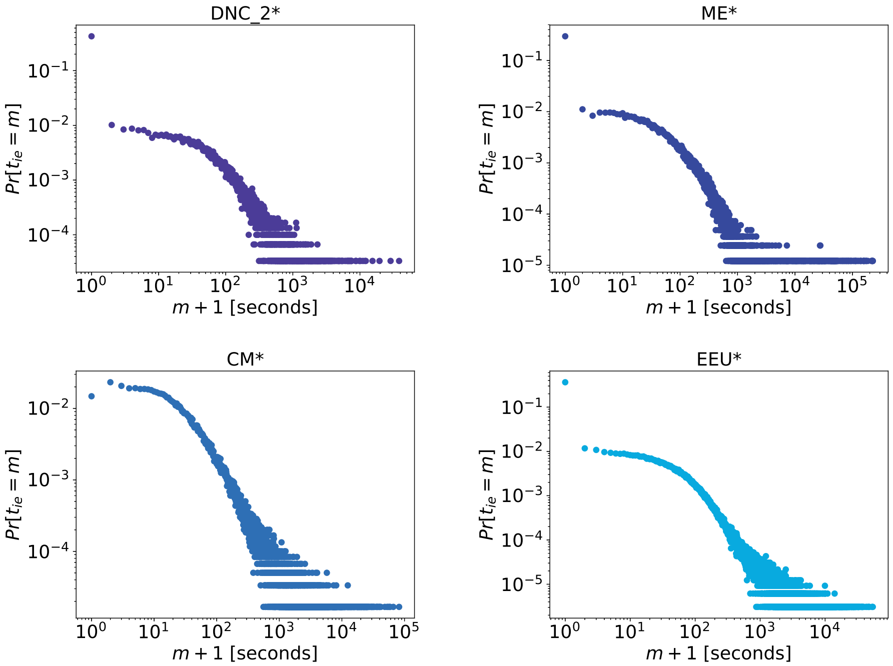

Fig. S3.2: Probability distribution $Pr[t_{ie} = m]$ of the inter-event time of the global activity of virtual contact temporal networks. Inter-event times are reported in seconds.

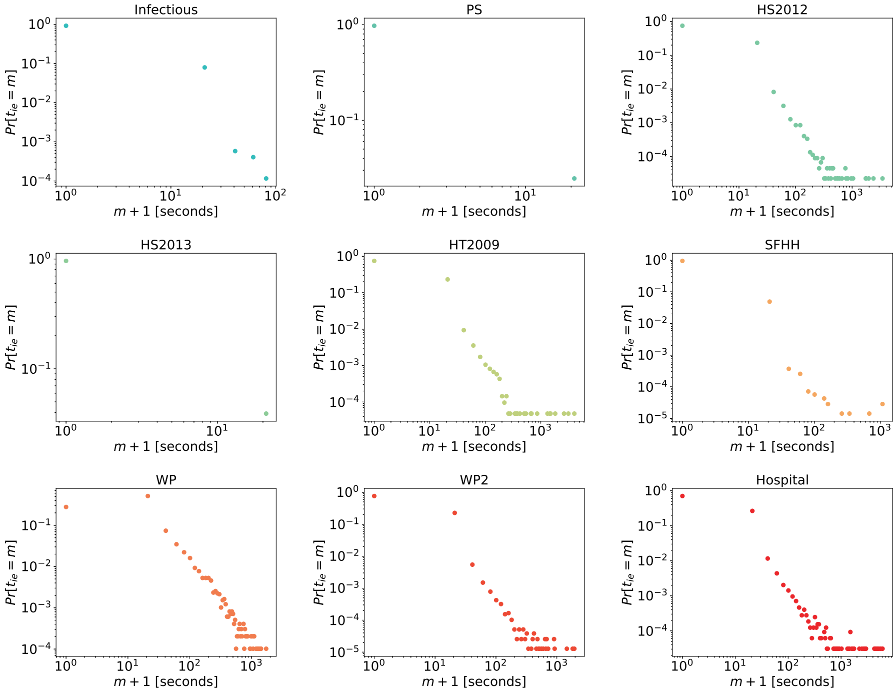

Fig. S3.3: Probability distribution $Pr[t_{ie} = m]$ of the inter-event time of the global activity of physical contact temporal networks. Inter-event times are reported in seconds.

2.6.4. Inter event time distribution of links

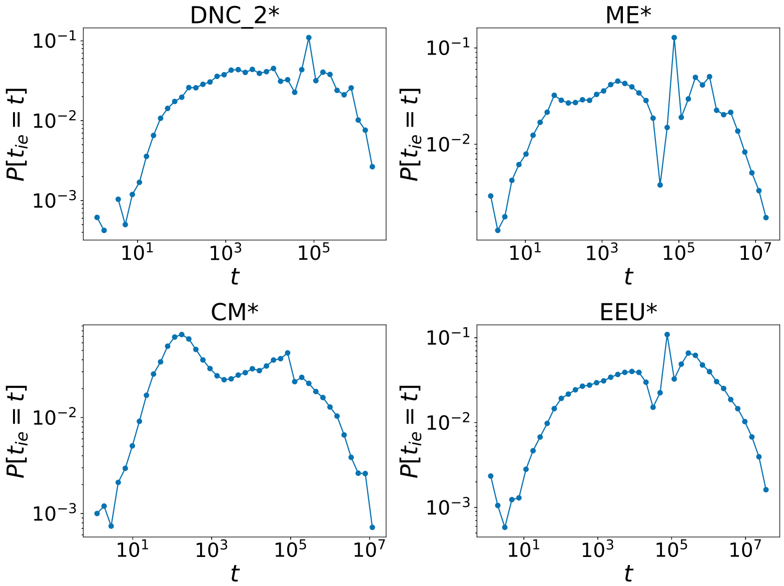

Fig. S3.4: Inter-event time distribution of single link activity of virtual contact datasets. Note that the horizontal and vertical axes are presented in logarithmic scales. Inter-event times are measured in seconds. In total 40 logarithmic bins are split within the interval $[t_{min}, t_{max}]$ where t_{min} and t_{max} are, respectively, the minumum and maximum inter-event time observed in the considered dataset.

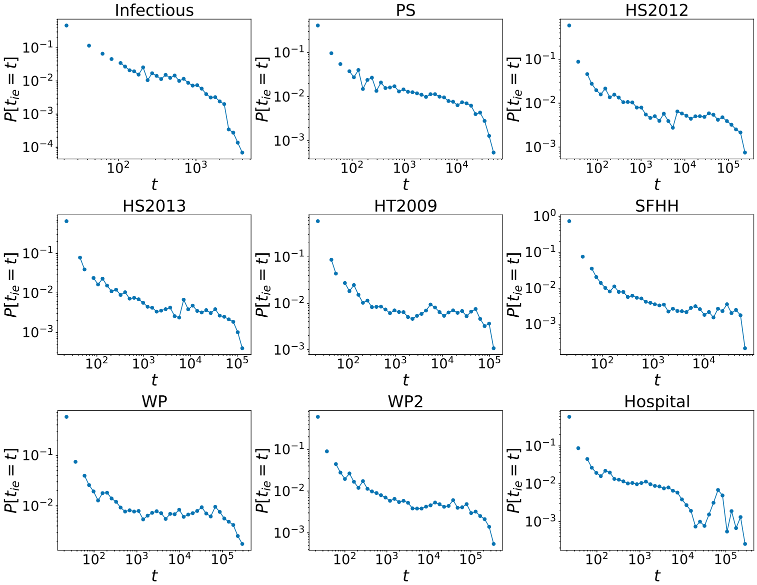

Fig. S3.5: Inter-event time distribution of single link activity of physical contact datasets. Note that the horizontal and vertical axes are presented in logarithmic scales. Inter-event times are measured in seconds. In total 40 logarithmic bins are split within the interval $[t_{min}, t_{max}]$ where t_{min} and t_{max} are, respectively, the minumum and maximum inter-event time observed in the considered dataset.

2.6.5. Temporal correlation of local events, additional figures

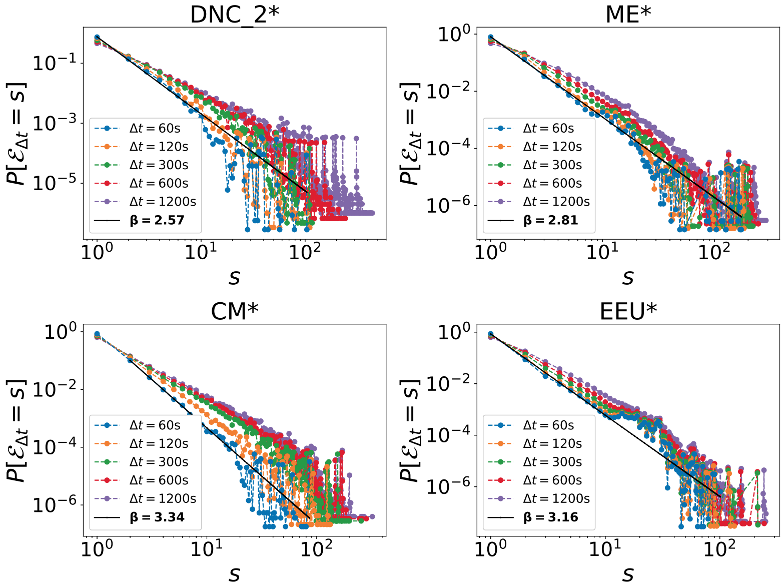

Fig. S3.6: Train size distributions of ego network activity of $\mathscr{G}$ of virtual contact datasets with $\Delta t = 60$ (blue), 120 (red), 300 (green), 600 (yellow), 1200 (purple) seconds. The black solid line represents the fit $P[\mathscr{E}_{\Delta t} = s] \sim s^{-\beta}$ to the distribution of the train size of $\mathscr{G}$ with $\Delta t = 60s$. The power law fit and its fitting region were computed with Clauset's method [48]. If the goodness of the power-law fit is significantly better than the exponential fit (likelihood ratio test with p-value $p < 0.05$), the value of β is reported in bold characters.

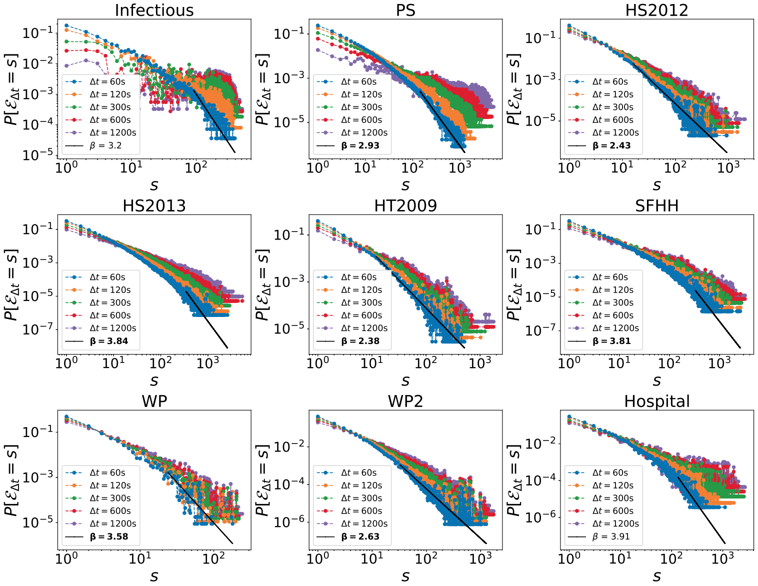

Fig. S3.7: Train size distributions of ego network activity of $\mathscr{G}$ of physical contact datasets with $\Delta t = 60$ (blue), 120 (red), 300 (green), 600 (yellow), 1200 (purple) seconds. The black solid line represents the fit $P[\mathscr{E}_{\Delta t} = s] \sim s^{-\beta}$ to the distribution of the train size of $\mathscr{G}$ with $\Delta t = 60s$. The power law fit and its fitting region were computed with Clauset's method [48]. If the goodness of the power-law fit is significantly better than the exponential fit (likelihood ratio test with p-value $p < 0.05$), the value of β is reported in bold characters.

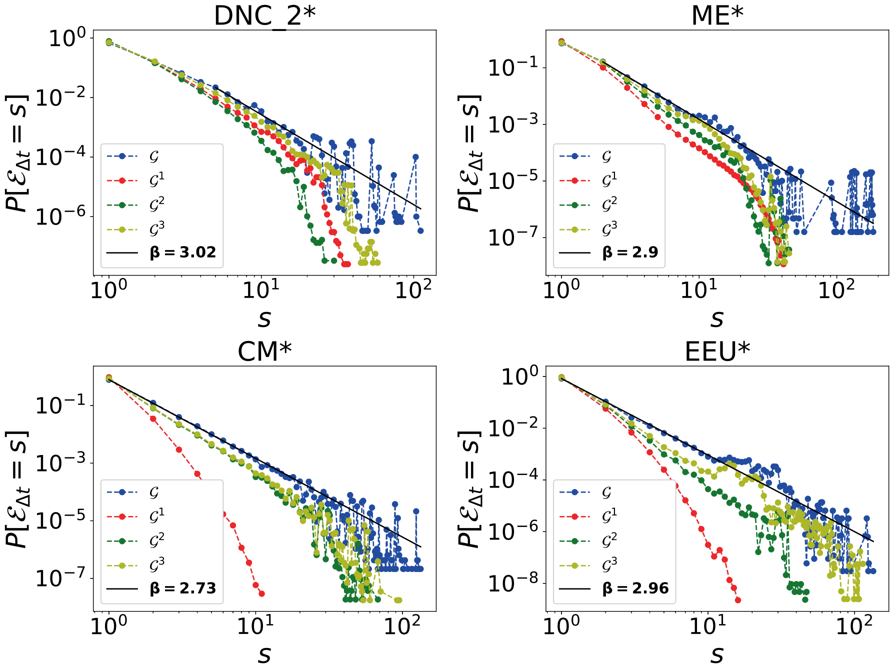

Fig. S3.8: Train size distribution ($\Delta t = 120s$) of ego network activity for $\mathcal{G}$ (blue), $\mathcal{G}_1$ (red), $\mathcal{G}_2$ (green), $\mathcal{G}_3$ (yellow) of virtual contact datasets. The black solid line represents the fit $P[\mathcal{E}_{\Delta t} = s] \sim s^{-\beta}$ to the distribution of the train size of $\mathcal{G}$ with $\Delta t = 120s$. The power law fit and its fitting region were computed with Clauset's method [48]. If the goodness of the power-law fit is significantly better than the exponential fit (likelihood ratio test with p-value $p < 0.05$), the value of β is reported in bold characters.

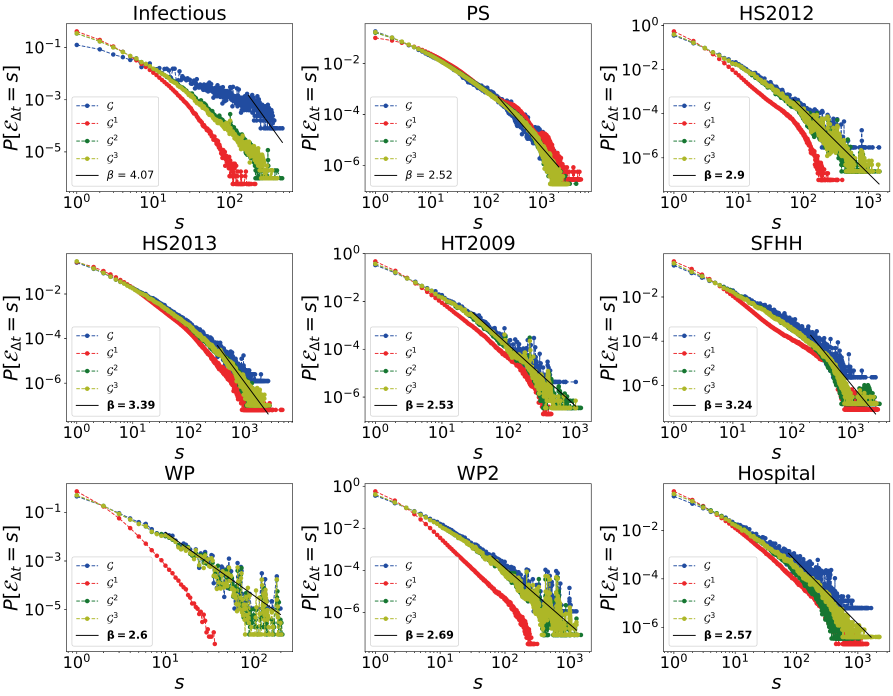

Fig. S3.9: Train size distribution (Δt = 120s) of ego network activity for $\mathcal{G}$ (blue), $\mathcal{G}_1$ (red), $\mathcal{G}_2$ (green), $\mathcal{G}_3$ (yellow) of physical contact datasets. The black solid line represents the fit $P[\mathcal{E}_{\Delta t} = s] \sim s^{-\beta}$ to the distribution of the train size of $\mathcal{G}$ with Δt = 120. The power law fit and its fitting region were computed with Clauset's method [48]. If the goodness of the power-law fit is significantly better than the exponential fit (likelihood ratio test with p-value $p < 0.05$), the value of β is reported in bold characters.

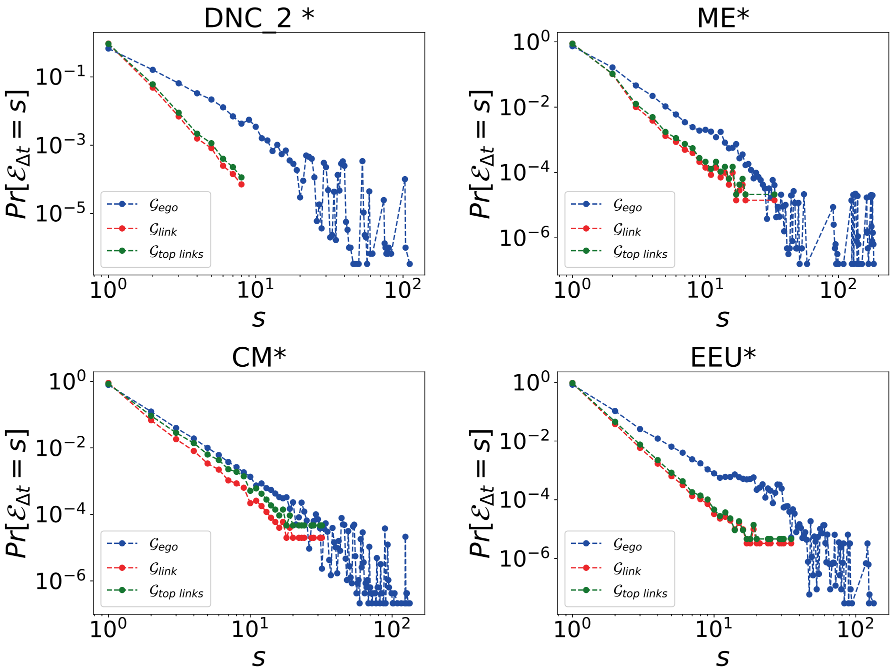

Fig. S3.10: Train size distribution ($\Delta t = 120s$) of ego network activity (blue), single link activity (red), most active link activity (green) of virtual contact datasets. Note that the horizontal and vertical axes are presented in logarithmic scales.

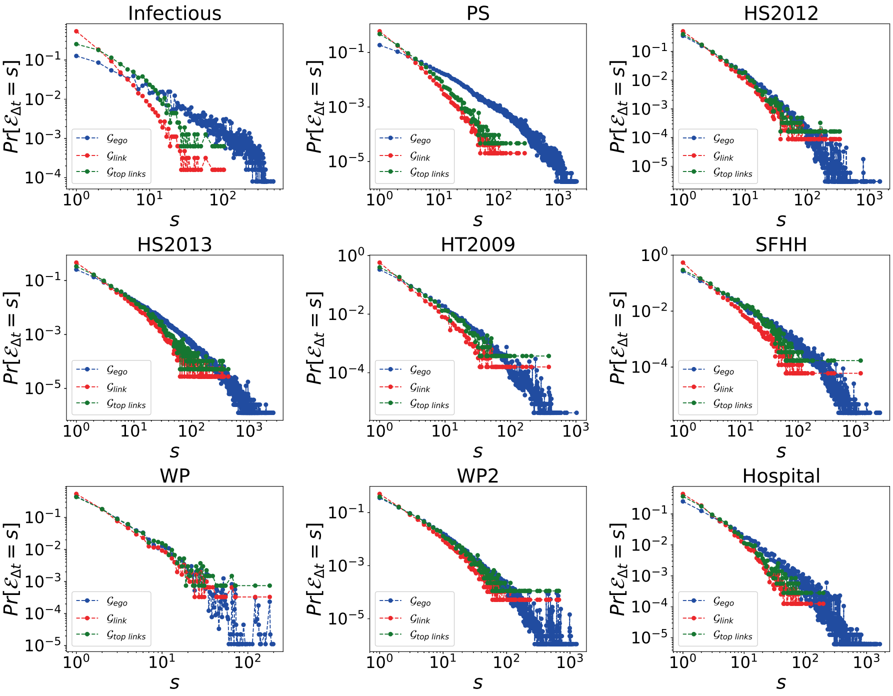

Fig. S3.11: Train size distribution ($\Delta t = 120s$) of ego network activity (blue), single link activity (red), most active link activity (green) of physical contact datasets. Note that the horizontal and vertical axes are presented in logarithmic scales.

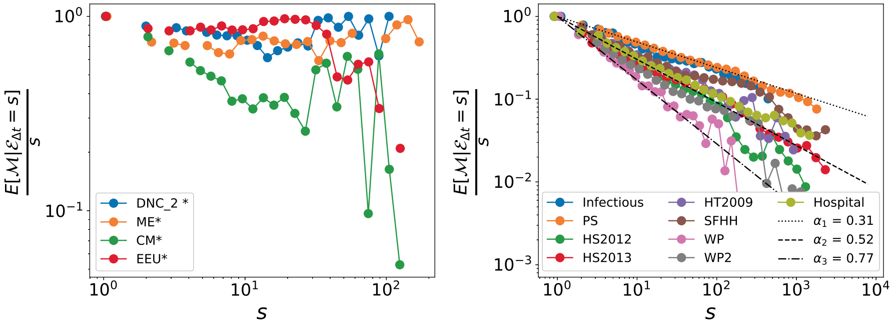

Fig. S3.12: The average number of active links $\frac{E[\mathcal{M}|\mathcal{E}_{\Delta t}=s]}{s}$ for trains with size $\mathcal{E}_{\Delta t} = s$ ($\Delta t = 120s$), normalized by the train size s of the ego networks for virtual (left) and physical (right) contact datasets. The three reference lines in right plot indicate $\frac{E[\mathcal{M}|\mathcal{E}_{\Delta t}=s]}{s} = s^{-\alpha}$ with slope $\alpha_1 = 0.31$ (dotted), $\alpha_2 = 0.52$ (dashed) and $\alpha_3 = 0.77$ (dash-dot). Note that the horizontal and vertical axes are presented in logarithmic scales. In total 30 logarithmic bins are split within the interval $[1, s_{max}]$, where s_{max} is the largest train size observed in the considered real temporal network.

BIBLIOGRAPHY

[1] Petter Holme and Jari Saramäki. "Temporal networks". In: *Physics Reports* 519.3 (2012), pp. 97–125.

[2] Petter Holme. "Modern temporal network theory: a colloquium". In: *The European Physical Journal B* 88.9 (2015), p. 234.

[3] K-I Goh and A-L Barabási. "Burstiness and memory in complex systems". In: *EPL (Europhysics Letters)* 81.4 (2008), p. 48002.

[4] Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi. "Entropy of dialogues creates coherent structures in e-mail traffic". In: *Proceedings of the National Academy of Sciences* 101.40 (2004), pp. 14333–14337.

[5] Joao Gama Oliveira and Albert-Lászlό Barabási. "Darwin and Einstein correspondence patterns". In: *Nature* 437.7063 (2005), pp. 1251–1251.

[6] Julián Candia et al. "Uncovering individual and collective human dynamics from mobile phone records". In: *Journal of Physics A: Mathematical and Theoretical* 41.22 (2008), p. 224015.

[7] Anders Johansen. "Probing human response times". In: *Physica A: Statistical Mechanics and its Applications* 338.1-2 (2004), pp. 286–291.

[8] Albert-László Barabási and Eric Bonabeau. "Scale-free networks". In: *Scientific American* 288.5 (2003), pp. 60–69.

[9] Albert-László Barabási. "Scale-free networks: a decade and beyond". In: *Science* 325.5939 (2009), pp. 412–413.

[10] Mark EJ Newman. "The structure and function of complex networks". In: *SIAM review* 45.2 (2003), pp. 167–256.

[11] Stefano Boccaletti et al. "Complex networks: Structure and dynamics". In: *Physics Reports* 424.4-5 (2006), pp. 175–308.

[12] Alain Barrat et al. "The architecture of complex weighted networks". In: *Proceedings of the National Academy of Sciences* 101.11 (2004), pp. 3747–3752.

[13] J-P Onnela et al. "Structure and tie strengths in mobile communication networks". In: *Proceedings of the National Academy of Sciences* 104.18 (2007), pp. 7332–7336.

[14] Hilla Brot et al. "Evolution through bursts: Network structure develops through localized bursts in time and space". In: *Network Science* 4.3 (2016), pp. 293–313.

[15] Riivo Kikas, Marlon Dumas, and Márton Karsai. "Bursty egocentric network evolution in skype". In: *Social Network Analysis and Mining* 3.4 (2013), pp. 1393–1401.

[16] 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*. 2017, pp. 601–610.

[17] Lauri Kovanen et al. “Temporal motifs in time-dependent networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2011.11 (2011), P11005.

[18] Márton Karsai, Kimmo Kaski, and János Kertész. “Correlated dynamics in egocentric communication networks”. In: *Plos One* 7.7 (2012), e40612.

[19] Raj Kumar Pan and Jari Saramäki. “Path lengths, correlations, and centrality in temporal networks”. In: *Physical Review E* 84.1 (2011), p. 016105.

[20] Albert-Laszlo Barabasi. “The origin of bursts and heavy tails in human dynamics”. In: *Nature* 435.7039 (2005), pp. 207–211.

[21] Alexei Vazquez et al. “Impact of non-Poissonian activity patterns on spreading processes”. In: *Physical Review Letters* 98.15 (2007), p. 158702.

[22] Diego Rybski et al. “Scaling laws of human interaction activity”. In: *Proceedings of the National Academy of Sciences* 106.31 (2009), pp. 12640–12645.

[23] Márton Karsai et al. “Universal features of correlated bursty behaviour”. In: *Scientific Reports* 2 (2012), p. 397.

[24] Xiu-Xiu Zhan, Alan Hanjalic, and Huijuan Wang. “Information diffusion backbones in temporal networks”. In: *Scientific Reports* 9.1 (2019), pp. 1–12.

[25] Roni Parshani et al. “Dynamic networks and directed percolation”. In: *EPL (Europhysics Letters)* 90.3 (2010), p. 38004.

[26] Dávid X Horváth and János Kertész. “Spreading dynamics on networks: the role of burstiness, topology and non-stationarity”. In: *New Journal of Physics* 16.7 (2014), p. 073037.

[27] Jean-Charles Delvenne, Renaud Lambiotte, and Luis EC Rocha. “Diffusion on networked systems is a question of time or structure”. In: *Nature Communications* 6.1 (2015), pp. 1–10.

[28] Xiu-Xiu Zhan, Alan Hanjalic, and Huijuan Wang. “Suppressing Information Diffusion via Link Blocking in Temporal Networks”. In: *International Conference on Complex Networks and Their Applications*. Springer. 2019, pp. 448–458.

[29] René Pfitzner et al. “Betweenness preference: Quantifying correlations in the topological dynamics of temporal networks”. In: *Physical Review Letters* 110.19 (2013), p. 198701.

[30] Giovanna Miritello, Esteban Moro, and Rubén Lara. “Dynamical strength of social ties in information spreading”. In: *Physical Review E* 83.4 (2011), p. 045102.

[31] Mikko Kivelä et al. “Multiscale analysis of spreading in a large communication network”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.03 (2012), P03005.

[32] Ingo Scholtes et al. “Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks”. In: *Nature Communications* 5.1 (2014), pp. 1–9.

[33] Oliver E Williams, Fabrizio Lillo, and Vito Latora. “How auto-and cross-correlations in link dynamics influence diffusion in non-Markovian temporal networks”. In: *arXiv preprint arXiv:1909.08134* (2019).

[34] Ville-Pekka Backlund, Jari Saramäki, and Raj Kumar Pan. “Effects of temporal correlations on cascades: Threshold models on temporal networks”. In: *Physical Review E* 89.6 (2014), p. 062815.

[35] Jérôme Kunegis. “Konect: the koblenz network collection”. In: *Proceedings of the 22nd international conference on World Wide Web*. 2013, pp. 1343–1350.

[36] Radosław Michalski, Sebastian Palus, and Przemysław Kazienko. “Matching Organizational Structure and Social Network Extracted from Email Communication”. In: *Lecture Notes in Business Information Processing*. Vol. 87. Springer Berlin Heidelberg, 2011, pp. 197–206.

[37] Pietro Panzarasa, Tore Opsahl, and Kathleen M Carley. “Patterns and dynamics of users’ behavior and interaction: Network analysis of an online community”. In: *Journal of the American Society for Information Science and Technology* 60.5 (2009), pp. 911–932.

[38] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. “Graph evolution: Densification and shrinking diameters”. In: *ACM transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007), 2–es.

[39] Lorenzo Isella et al. “What’s in a crowd? Analysis of face-to-face behavioral networks”. In: *Journal of theoretical biology* 271.1 (2011), pp. 166–180.

[40] Juliette Stehlé et al. “High-resolution measurements of face-to-face contact patterns in a primary school”. In: *PloS one* 6.8 (2011), e23176.

[41] Julie Fournet and Alain Barrat. “Contact patterns among high school students”. In: *PloS one* 9.9 (2014), e107878.

[42] 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”. In: *PloS one* 10.9 (2015), e0136497.

[43] Ciro Cattuto et al. “Dynamics of person-to-person interactions from distributed RFID sensor networks”. In: *PloS one* 5.7 (2010), e11596.

[44] Juliette Stehlé et al. “Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees”. In: *BMC medicine* 9.1 (2011), pp. 1–15.

[45] Mathieu Génois et al. “Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers”. In: *Network Science* 3.3 (2015), pp. 326–347.

[46] Mathieu Génois and Alain Barrat. “Can co-location be used as a proxy for face-to-face contacts?” In: *EPJ Data Science* 7.1 (2018), pp. 1–18.

[47] Philippe Vanhems et al. “Estimating potential infection transmission routes in hospital wards using wearable proximity sensors”. In: *PloS one* 8.9 (2013), e73970.

[48] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. “Power-law distributions in empirical data”. In: *SIAM review* 51.4 (2009), pp. 661–703.

3

Temporal-Topological Properties of Evolving Higher-Order Networks

This chapter has been published as Ceria, A., & Wang, H. (2023). Temporal-topological properties of higher-order evolving networks. Scientific Reports, 13(1), 5885.

Human social interactions are typically recorded as time-specific dyadic interactions, and represented as evolving (temporal) networks, where links are activated/deactivated over time. However, individuals can interact in groups of more than two people. Such group interactions can be represented as higher-order events of an evolving network. Here, we propose methods to characterize the temporal-topological properties of higher-order events to compare networks and identify their (dis)similarities. We analyzed 8 real-world physical contact networks, finding the following: a) Events of different orders close in time tend to be also close in topology; b) Nodes participating in many different groups (events) of a given order tend to involve in many different groups (events) of another order; Thus, individuals tend to be consistently active or inactive in events across orders; c) Local events that are close in topology are correlated in time, supporting observation a). Differently, in 5 collaboration networks, observation a) is almost absent; Consistently, no evident temporal correlation of local events has been observed in collaboration networks. Such differences between the two classes of networks may be explained by the fact that physical contacts are proximity based, in contrast to collaboration networks. Our methods may facilitate the investigation of how properties of higher-order events affect dynamic processes unfolding on them and possibly inspire the development of more refined models of higher-order time-varying networks.

3.1. Introduction

Interactions among individuals are usually experimentally measured as time-resolved records of face-to-face contacts between couples of people in controlled social setting such as workplaces, hospitals, schools and conferences. These time specific records are thus collected in the form of dyadic interactions, and have been effectively studied in the framework of evolving (temporal) networks, where each link between two nodes is activated only when the node pair interacts [1–3]. The temporal patterns of link activations (or contacts) in real-world networks are far from being fully random nor deterministic [4]. Contacts between a pair of nodes usually occur in bursts of many contacts close in time followed by a long period of inactivity [5] and the time between two consecutive interactions is usually fat-tailed distributed [6–8]. Such temporal properties of contacts influence the dynamic processes unfolding on the network [9–17]. Despite these tremendous advances in the last decade, studies on temporal networks have traditionally focused on pairwise interactions only. However pairwise interactions can only partially capture interactions among constituents of a system [18, 19]. For example, a neuron may receive the output from or send a signal to many different neighbouring neurons [20], individuals may gather in groups [21], and scientific collaborations are not limited to couples of authors [22]. Such interactions are named higher-order, to emphasize that they involve more than just a couple of nodes. Benson et al.[23] showed that a generalization of triadic closure seems to lead the first activation of a given hyperlink. On the other hand, Cencetti et al. [24] focused on temporal inhomogeneities of activations of the same hyperlink. The focus so far is on the prediction of hyperlink activations [23] or on pure temporal properties of higher-order events [24]. However, the interplay between temporal and topological properties of higher-order events, e.g. if higher-order events close in time tend to occur also close in topology, remains far from well understood. Hence, this work aims to systematically characterize the relation between temporal and topological properties of higher-order events to compare higher-order temporal networks. Inspired by our recent work that characterizes temporal and topological properties of dyadic interactions in temporal networks [25], we redesign the characterization method for higher-order events. In particular, we are going to explore such properties from three perspectives: 1) The interrelation between the distance in topology and the temporal delay of events, 2) Their correlation or overlap in topological location 3) The temporal correlation of local events that overlap in component nodes. In order to compare real-world networks with different sizes, we design null models where temporal and topological properties of events of an arbitrary order are systematically destroyed or preserved. We applied our methods to 8 real-world physical contact networks and 5 collaboration networks. We show that, in physical contacts, events of different orders with short temporal delay tend to be close in topology too. We then investigate the correlation of events in topology and discover that events of different orders are likely to overlap in component nodes. In particular, nodes who participate in many different groups (events) of a given order are likely to be involved in many different groups (events) of another order. Individuals do not reduce their number of interactions of one order due to frequent interactions of another order. Finally, we show that those local events that overlap in component nodes are correlated in time, which supports the finding that events close in time are also close in topology. In collaboration networks, we observe that events also overlap in component nodes. However, the correlation between topological distance and temporal delay of events are usually either weak or absent. Coherently, in collaboration networks, the temporal correlation of local events that overlap in component nodes is almost absent. Such differences between physical contacts and collaboration networks may be due to the fact that physical interactions are partly driven by proximity, so that a set of individuals close to each other tend to interact close in time among (subsets of) them.

Our methods can be applied to compare real-world higher-order networks and to investigate how the properties of their events affects the dynamic processes unfolding on them. More realistic models of higher-order evolving networks can be further developed to reproduce specific properties of the higher-order interactions observed in this chapter.

3.2. DEFINITIONS

3.2.1. HIGHER-ORDER EVOLVING NETWORKS

Time-varying social interactions or contacts have been mostly measured pairwise and studied with the formalism of (pairwise) temporal networks. A temporal network observed at discrete time within $[0,T)$ can be described by $\mathscr{G} = (\mathscr{N}, \mathscr{C})$, where $\mathscr{N}$ is the set of nodes or individuals, $\mathscr{C}$ is the set of pairwise interactions. If node u and v have a contact at time step $0 \le t \le T-1$, $(\ell, t) \in \mathscr{C}$, where $\ell = \ell(u,v)$ is the link connecting the pair of nodes between which the contact occurs. The contact $(\ell(u,v), t)$ can be regarded as the activation of the link $\ell(u,v)$ at time t. This traditional temporal network representation records social contacts as a set of pair-wise interactions. However, individuals may gather in larger groups, so that more than two people interact with each other at the same time. For example, an interaction $(h(i,j,k), t)$ among three nodes at time t is usually measured and recorded as three pair-wise interactions $(\ell(i,j),t)$, $(\ell(j,k),t)$ and $(\ell(i,k),t)$. Social interactions can be more precisely represented as a higher-order evolving network $\mathscr{H} = (\mathscr{N}, \mathscr{E})$ (or temporal hypergraph, following the definition of Cencetti et al. [24]), where $\mathscr{E}$ is the set of events of arbitrary orders. Such group interaction or higher-order event $(h(u_1, \dots u_d), t)$ can be regarded as the activation of the corresponding hyperlink $h(u_1, \dots u_d)$ at t. The size or order of the interaction is d, where d is the size of the group. The pairwise time aggregated network of a traditional pairwise temporal network is $G = (\mathcal{N}, \Lambda)$, where any couple of nodes (i, j) is connected by a link $\ell(i, j) \in \Lambda$ if $\ell(i, j)$ has been active at least once during the entire observation time $[0, T)$. Consistently, the higher-order time aggregated network is $H = (\mathcal{N}, \mathcal{L})$, where any set $\{u_1, \dots u_d\}$ of d nodes are connected by a hyperlink $h(u_1, \dots u_d) \in \mathcal{L}$ with size d if $h(u_1, \dots u_d)$ has been activated at least once. The activity of each hyperlink h can be represented by a time series $X_h = \{x_h(t), 0 \le t < T\}$ where $x_h(t) = 1$ only if the hyperlink h is active at time t, i.e., $e = (h, t) \in \mathcal{E}$.

3.2.2. Temporal and topological distance of events

The temporal distance or delay between two events $e_1 = (h_1, t)$ and $e_2 = (h_2, s)$ is $\mathcal{T}(e_1, e_2) = |t - s|$.

The topological distance, also called hop-count, between two nodes on a pair-wise static network is the number of links contained in the shortest path between these two nodes. We define the topological distance $\eta(e_1, e_2)$ between two events $e_1 = (h_1, t)$ and $e_2 = (h_2, s)$ as the topological distance between the corresponding two hyperlinks h_1 and h_2, which is further defined as follows. The distance between the same hyperlink is zero, e.g., $\eta((h_1, t), (h_1, s)) = 0$. The distance between two different hyperlinks $h(u_1, \dots, u_d)$ and $h(v_1, \dots, v_{d'})$ with size d and d', respectively, follows

$$\eta((h(u_1, \dots, u_d), t), (h(v_1, \dots, v_{d'}), s)) = min_{u \in \{u_1, \dots, u_d\}, v \in \{v_1, \dots, v_{d'}\}} (\delta(u, v) + 1) \tag{3.1}$$

where $\delta(u, v)$ is the distance or hop-count between node u and v on the unweighted pairwise time aggregated network G. The distance between two events is thus one plus the minimal distance between two component nodes from the two events respectively. For example, the distance between events $e_1 = (h(i, j, k), t)$ and $e_2 = (h(i, m, n), s)$ is $\eta(e_1, e_2) = 1$.

3.2.3. Network randomization - control methods

To detect non-trivial temporal and topological patterns of events, we compare properties obtained from real-world higher-order temporal networks with those of designed null models. We generalize the randomized reference models of pairwise evolving networks which gradually preserve and destroy temporal and topological properties of pairwise interactions [25–27] for higher-order temporal networks. Given a higher-order evolving network $\mathcal{H}$ and any given order d of events, we introduce 3 randomized null models $\mathcal{H}_d^1$, $\mathcal{H}_d^2$ and $\mathcal{H}_d^3$ which systematically randomize order d events only, without changing events of any other order $d' \neq d$. We denote as $\mathcal{E}_d$ the set of events with the same size d. Randomized network $\mathcal{H}_d^1$ is obtained by randomly re-shuffling the time stamps of the events in $\mathcal{E}_d$, without changing the topological locations of these events. This randomization does not change the total number of activations of each hyperlink, nor the probability distribution of the topological distance of two randomly selected events. Null model $\mathcal{H}_d^1$ randomizes the time stamps of order d events. As a consequence, the distribution of the inter-event time of order d events, i.e., the time between two consecutive activations of a random order d hyperlink, in $\mathcal{H}_d^1$ tends to be less heterogeneous than that in $\mathcal{H}$. As mentioned in Subsection 3.2.2, the activations of a given hyperlink h can be represented by a time series X_h. The randomized network $\mathcal{H}_d^2$ is obtained by iteratively swapping the time series of two randomly selected order d hyperlinks . In $\mathcal{H}_d^2$, the inter-event time distribution of order d events is preserved as in the original network $\mathcal{H}$, while the time series of activations of a given order d hyperlink are independent from its topological location. The third randomized network $\mathcal{H}_d^3$

is obtained by swapping the activity time series of two randomly selected order d hyperlinks with the same total number of activations. This randomization does not change the number of activations of any hyperlink, the distribution of the topological distance of two random events, nor the inter-event (order d events) time distribution. The pairs of order d hyperlinks with the same number of events can be few in number in real-world temporal networks, such that the difference between a real-world network and its randomized network $\mathscr{H}_d^3$ is small. This is especially the case when the order d is large, thus the number of hyperlinks is small. These three randomized models preserve the unweighted higher-order time aggregated network H and the probability distribution of the temporal distance of two random events of size d.

3.3. DATASETS

We will apply our method to 13 real-world datasets of human physical interactions and scientific collaborations. The first 8 datasets are collections of face-to-face interactions at a distance smaller than 2 m in several social contexts such as conferences (HT2009, SFHH)[28, 29], hospital[30], primary school (PS) [31, 32], high schools (HS2012 ,HS2013)[33, 34], workplace (WP2)[29] and museum (Infectious)[28]. Face-to-face interactions are recorded as a set of pair-wise interactions. Based on them, we deduce group interactions, by promoting each set of $\binom{d}{2}$ dyadic interactions occurring at the same time and forming a fully connected clique of d nodes to an event of size d. Since a clique of order d contains all its sub-cliques of order $d' < d$, only the maximal clique is promoted to a higher-order event, whereas sub-cliques are ignored. For example, 3 pairwise contacts $(\ell(i,j),t)$, $(\ell(j,k),t)$ and $(\ell(i,k),t)$ occurring at the same time t are regarded as a single event of order 3 i.e., $(h(i,j,k),t)$ without any order 2 event. This method has been already used by Cencetti et al [24]. to deduce higher-order interactions from datasets of human face-to-face interactions. We further preprocess these datasets by removing nodes which are not connected to the largest connected component in the pairwise time-aggregated network. We also remove long periods of inactivity, when no event occurs in the network. Such periods usually correspond, e.g., to night and weekends, and are recognized as outliers in the inter-event time distribution of the time series which records the total number of events per timestamp. Such data pre-processing method has also been used in our recent work [25]. The other 5 higher-order collaborations networks are obtained based on scientific papers recorded in the arxiv in various fields: lattice high energy physics (hep-lat), theoretical nuclear physics (nucl-th), quantitative biology (q-bio), quantitative finance (q-fin) and quantum physics (quant-ph). In a collaboration network, each node represents an author, and an event of order d occurs at time t if a paper co-authored by d authors is published at t. Assigning papers to the correct authors is not easy. The same author can be named differently, e.g., using the full or initial of the first name and typographic errors may be present. Thus, we applied standard text preprocessing methods to authors' name, and we identify each author by the initials of their first names, together with their surname according to the method of Newman et al.[35]. The total number of events of each order in each real-world temporal network is shown in Figures S3.1 and S3.2 in Appendix. In each dataset, the number of events with order $2 \leq d \leq 4$ is not negligible; however events with an order larger than 4 are rare (if not absent) in most of the physical contact datasets. Details of the datasets after preprocessing are given in Table 3.1.

Network	$\lvert\mathcal{N}\rvert$	$\lvert\mathcal{L}\rvert$	$\lvert\mathcal{E}\rvert$	T	dt	*contact type*
Primary School (PS)	242	12704	106877	3099	20 s	*physical*
High School 2013 (HS2013)	327	7818	172031	7371	20 s	*physical*
Hypertext 2009 (HT2009)	113	2434	19037	7227	20 s	*physical*
Infectious (Infectious)	410	3350	14275	1422	20 s	*physical*
Workplace 2015 (WP2)	217	4909	73820	20947	20 s	*physical*
SFHH Conference (SFHH)	403	10541	54306	3800	20 s	*physical*
Hospital (Hospital)	75	1825	27835	16027	20 s	*physical*
High School 2012 (HS2012)	180	2645	42105	14115	20 s	*physical*
High energy physics, lattice (hep-lat)	10598	11588	18267	10809	1 d	*collaboration*
Nuclear physics, theory (nucl-th)	25246	27094	39511	10620	1 d	*collaboration*
Quantitative biology (q-bio)	45645	22978	25973	10704	1 d	*collaboration*
Quantitative finance (q-fin)	7509	6192	7577	9027	1 d	*collaboration*
Quantum physics (quant-ph)	56036	70119	88769	10600	1 d	*collaboration*

Table 3.1: Basic features of the empirical higher-order time-evolving networks after data processing. The number of nodes ($|\mathcal{N}|$), the number of hyperlinks ($|\mathcal{L}|$), the total number of events ($|\mathcal{E}|$), the length of the observation time window in time steps (T), the time resolution or duration of each time step (dt) in seconds or days and the contact type are shown.

3.4. Characterizing temporal-topological properties of networks

In this section we introduce a systematic characterization method of higher-order temporal networks. We characterize the temporal and topological properties of events from three different perspectives. In Subsection 3.4.1, we analyze the interrelation between the temporal and topological distance of two arbitrary events of different orders. In Subsection 3.4.2, we study the topological correlation of events, i.e., how events of different orders overlap in component nodes. Finally, Subsection 3.4.3 introduces a method to characterize the temporal correlation of events occurring close in topology.

3.4.1. Correlation of temporal and topological distance of events

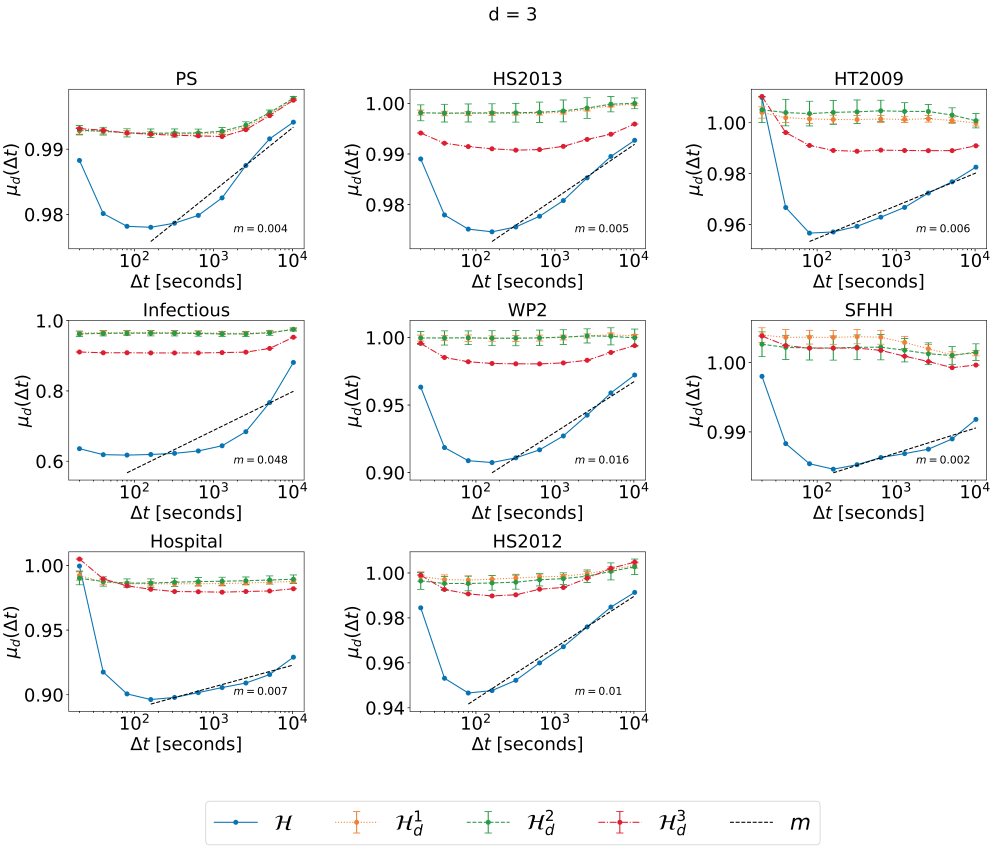

Fig. 3.1 : The normalized average topological distance $\mu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\backslash\mathcal{E}_d]}{E[\eta(e,e')|\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\backslash\mathcal{E}_d]}$, between an order $d = 3$ event and an event of a different order, in each physical contact network and its corresponding three randomized null models $\mathcal{H}_d^1$ (yellow), $\mathcal{H}_d^2$ (green) and $\mathcal{H}_d^3$ (red), which preserve or destroy specific properties of order $d = 3$ events. $\lim_{\Delta t\to\infty} E[\eta(e,e')|\mathcal{T}(e,e') < \Delta t,\ e \in \mathcal{E}_d,\ e' \in \mathcal{E} \setminus \mathcal{E}_d] = E[\eta(e,e')|\ e \in \mathcal{E}_d,\ e' \in \mathcal{E} \setminus \mathcal{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure corresponds to the linear fit (with slope m) of $\mu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathcal{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

In this subsection we investigate how temporal and topological distance of events are related to each other. Specifically, we aim to understand to what extent events close in time are also close in topology. In our previous work [25], we considered all interactions in a temporal network as pairwise interactions alone and found in real-world physical and virtual contact networks that pairwise interactions that are close in time tend to be close in topology (in the pairwise time aggregated network). Here, we generalize the method of characterizing the relation between topological and temporal distance of two dyadic interactions to that of two higher-order events with different orders. In this analysis, normalizations in topological distance and randomizations in networks have been applied so that we can compare real-world temporal networks with different properties in e.g., the number of nodes and contacts. We take order $d = 3$ as an example to illustrate our method and observations. In Figures 3.1 and 3.2 we investigate the average topological distance $E[\eta[(e,e')|\mathcal{T}(e,e') < \Delta t, e \in \mathcal{E}_d,\ e' \in \mathcal{E} \setminus \mathcal{E}_d]$ between two events (e,e') with different orders $d \neq d'$, given that their temporal distance is smaller than Δt in physical contact and collaboration networks, respectively. In physical contact networks (Figure 3.1), we observe in general an increasing trend of the normalized average topological distance $\mu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\setminus\mathcal{E}_d]}{E[\eta(e,e')|\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\setminus\mathcal{E}_d]}$ between between events of different orders with their conditional temporal distance Δt, except that the topological distance decreases with Δt when Δt is small, approximately when $\Delta t \leq 100s$. Usually, events of different orders that occur relatively close in time tend to be also close in topology. The decrease of the average distance $\mu_d(\Delta t)$ with Δt when Δt is small is introduced by the way how higher-order physical contact networks are constructed. In these networks higher-order events are inferred from their contact records, so that if a higher-order event that involves a set of d nodes occur at a given timestamp, no event of an order d' smaller than d involving only a subset of these d nodes can occur at the same timestamp. This explains why as Δt decreases further when it is small, the topological distance $\mu_d(\Delta t)$ does not decrease anymore. This is not the case in collaboration networks, where when a group of scientists collaborate in a paper, a subgroup could co-author another paper at the same time. Accordingly, we do not observe the decreasing trend of the $\mu_d(\Delta t)$ with Δt when Δt is small in collaboration networks. Besides this initial decreasing trend, we observe an increasing trend of $\mu_d(\Delta t)$ between events with their conditional temporal distance in every physical contact networks, but this is generally much less evident in collaboration networks. The slope of the increase of $\mu_d(\Delta t)$ with the conditional temporal distance Δt indicates the relative strength of temporal-topological correlation of events. In Figures 3.1 and 3.2 we show the slope of the linear fit of $\mu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ for the part of the curve that has an increasing trend. In physical contacts, the highest slopes are observed in Infectious and Workplace (WP2) networks. Moreover, in each dataset we observe an increasing trend with slope larger than 0. In contrast, this slope is small around zero in the corresponding randomized network $\mathcal{H}_d^1$, $\mathcal{H}_d^2$ and $\mathcal{H}_d^3$. This means the set of activity time series of each order 3 hyperlink of a higher-order network $\mathcal{H}$, which is preserved in the corresponding randomized network $\mathcal{H}_d^2$ and $\mathcal{H}_d^3$ does not contribute to the correlation between topological and temporal distance of events of different orders.

Differently, in collaboration networks, the increasing trend is usually either very weak (nucl-th, quant-ph) or absent (q-bio and q-fin), with the only exception of hep-lat dataset. The temporal-topological correlation of events tends to disappear in collaboration networks.

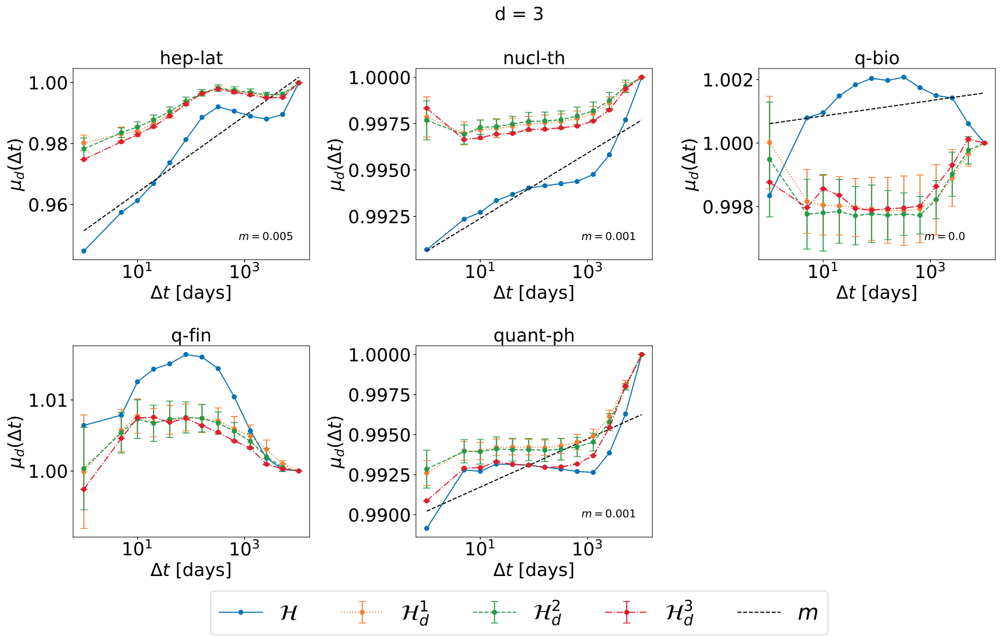

Fig. 3.2 : The normalized average topological distance $\mu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\backslash\mathcal{E}_d]}{E[\eta(e,e')|\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\backslash\mathcal{E}_d]}$, between an order $d = 3$ event and an event of a different order, in each collaboration network and its corresponding three randomized null models $\mathcal{H}_d^1$ (yellow), $\mathcal{H}_d^2$ (green) and $\mathcal{H}_d^3$ (red), which preserve or destroy specific properties of order $d = 3$ events. $\lim_{\Delta t\to\infty} E[\eta(e,e')|\mathcal{T}(e,e') < \Delta t,\ e \in \mathcal{E}_d,\ e' \in \mathcal{E} \setminus \mathcal{E}_d] = E[\eta(e,e')|\ e \in \mathcal{E}_d,\ e' \in \mathcal{E} \setminus \mathcal{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure corresponds to the linear fit (with slope m) of $\mu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathcal{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

Conclusions drawn from the discussion of Figures 3.1 and 3.2 hold for the other orders $d = 2$ (see Figures S3.5 and S3.6 in Appendix) and $d = 4$ (see Figures S3.7 and S3.8 in Appendix). The only exceptions are observed in datasets HT2009 and WP2 when $d = 4$: in this case indeed the trend of $\mu_d(\Delta t)$ in three randomized reference models seems to partially re produce the increasing trend observed in $\mathcal{H}$. This is likely due to the low number of hyperlinks of order 4 in these two networks.

We focus on the analysis of events of different orders. We have also analyzed events of the same order and obtain similar observations. As an example, Figures 3.3 and 3.4 , show the normalized average topological distance $\nu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e,\ e'\in\mathcal{E}_d]}{E[\eta(e,e')|\ e,\ e'\in\mathcal{E}_d]}$ of events of the same order $d = 3$ with a temporal delay smaller than Δt. The temporal-topological correlation is observed in physical contact networks but not collaboration networks. In contrast to events of different orders, in physical contacts, events of the same order demonstrate similar temporal-topological correlation in randomized networks $\mathcal{H}_d^2$ and $\mathcal{H}_d^3$ as in the corresponding real-world network $\mathcal{H}$, reflected the similar slope of the increase of the topological distance with Δt in these three networks. Randomized network $\mathcal{H}_d^2$ and $\mathcal{H}^3$ preserve the same set of activity time series of each single order d hyper link. The burstiness property, i.e. the frequent activation of the same hyperlink within a short time followed by a long resting period of an activity time series contributes to the temporal-topological correlation observed in real-world physical networks. These conclusions hold also for the analysis for orders $d = 2$ (Figures S3.9 and S3.10 in Appendix) and 4 (Figures S3.11 and S3.12 in Appendix). The only exception is that no evident increase of $\nu_d(\Delta t)$ with Δt is observed when $d = 4$ in Workplace and Hypertext 09, likely due to the low number of order $d = 4$ events observed in these two networks. In this work, we focus on the analysis of events of different orders, whose temporal-topological correlation cannot be explained by the burstiness of the activations of each hyperlink.

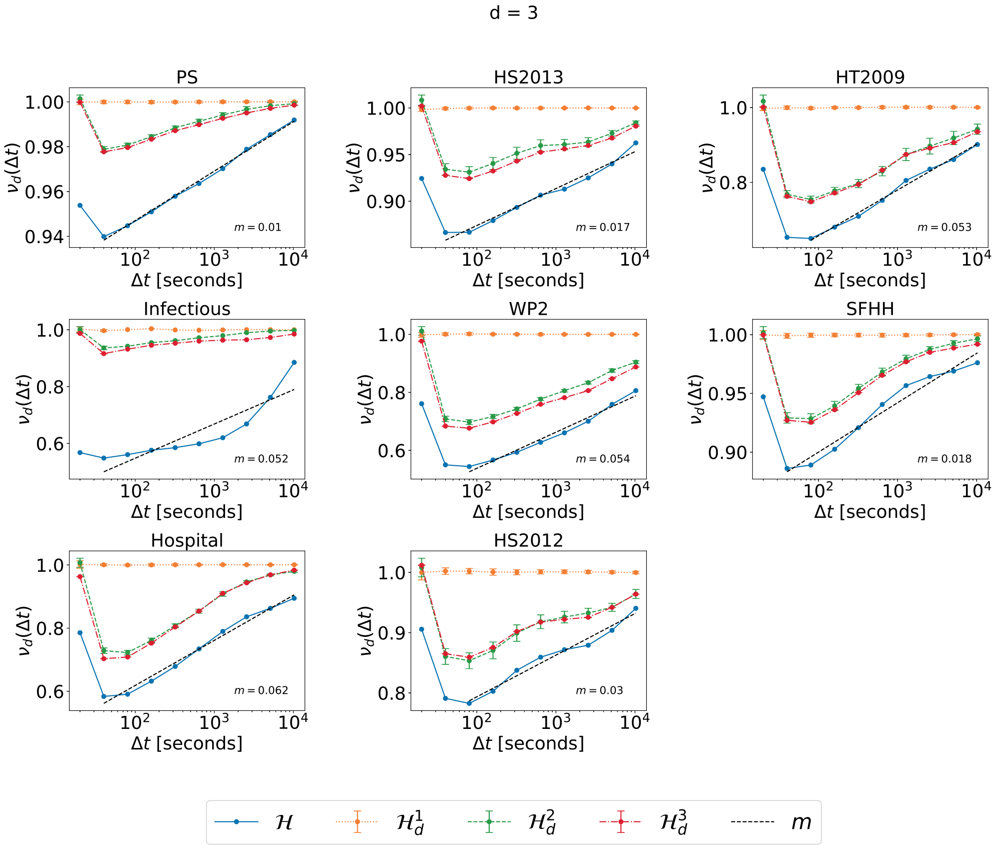

Fig. 3.3 : The normalized average topological distance $\nu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e,\ e'\in\mathcal{E}_d]}{E[\eta(e,e')|\ e,\ e'\in\mathcal{E}_d]}$, between two order $d = 3$ events, in each physical contact network and its corresponding three randomized null models $\mathcal{H}_d^1$ (yellow), $\mathcal{H}_d^2$ (green) and $\mathcal{H}_d^3$ (red), which preserve or destroy specific properties of order $d = 3$ events. $\lim_{\Delta t\to\infty} E[\eta(e,e')|\mathcal{T}(e,e') < \Delta t,\ e,\ e' \in \mathcal{E}_d] = E[\eta(e,e')|\ e,\ e' \in \mathcal{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure corresponds to the linear fit (with slope m) of $\nu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathcal{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

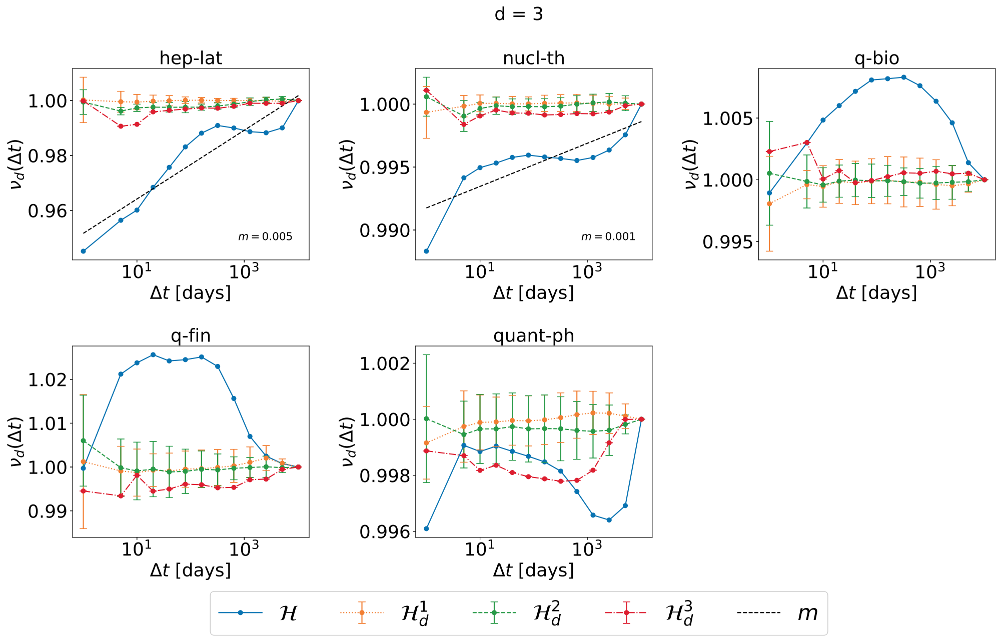

Fig. 3.4 : The normalized average topological distance $\nu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e,\ e'\in\mathcal{E}_d]}{E[\eta(e,e')|\ e,\ e'\in\mathcal{E}_d]}$, between two order $d = 3$ events, in each collaboration network and its corresponding three randomized null models $\mathcal{H}_d^1$ (yellow), $\mathcal{H}_d^2$ (green) and $\mathcal{H}_d^3$ (red), which preserve or destroy specific properties of order $d = 3$ events. $\lim_{\Delta t\to\infty} E[\eta(e,e')|\mathcal{T}(e,e') < \Delta t,\ e,\ e' \in \mathcal{E}_d] = E[\eta(e,e')|\ e,\ e' \in \mathcal{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure corresponds to the linear fit (with slope m) of $\nu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathcal{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

3.4.2. TOPOLOGICAL CORRELATION OF EVENTS WITH DIFFERENT ORDERS

To better understand the observed correlation between temporal and topological distance of events, we explore further whether higher-order events overlap in component nodes (correlation in topology) in this subsection and whether events that overlap in topology are correlated in time in Subsection 3.4.3. Higher-order events that overlap in component nodes and occur close in time may partially explain the observed temporal and topological correlation between events. Would a node that belongs to many hyperlinks of order d, also be connected to many hyperlinks of order $d' \neq d$? To investigate this question, we examine the number of hyperlinks of each order that a node belongs to in the unweighted higher-order time aggregated network. The total number of order d hyperlinks that the node v is connected to, denoted as $k_d(v)$, is also called the d-degree of node v. In Figure 3.5 (3.6), we compare the d-degree and the d'-degree of a node when (d', d) is equal to (3,2), (4,2) and (4,3) respectively in each physical contact (collaboration) network. All three randomized networks $\mathcal{H}_d^1$, $\mathcal{H}_d^2$ and $\mathcal{H}_d^3$ have the same higher-order time-aggregated unweighted network as the corresponding real-world network $\mathcal{H}$. Hence, the $d-$degree and $d'-$degree of each node remain the same in the randomized networks as in the real-world network.We focus on the case when (d', d) is equal to (3,2), as an example. We observe that the d'-degree of a node is an increasing function of the d-degree of the node in every considered collaboration and physical contact networks. Hence, a node that participates in many groups of order 3, tends to involve in many groups of order 2. When (d', d) equals to (4,2) and (4,3), such trend is less evident in physical networks (especially in WP2, HS2012, Infectious and HT2009) and remains evident in collaboration networks. This is likely because the number of order 4 hyperlinks is generally low (see Figure S3.3 in Appendix) in physical contact networks, but not in collaboration networks (see Figure S3.4 in Appendix).

Furthermore, we investigate whether a node that involves in many order d events tends to join many order d' interactions. The number of order d events that a node v is involved in, denoted by $s_d(v)$, is also called the d-strength of node v. The $d-$strength of a node is actually the sum of the weights of order d hyperlinks that a node belong to in the weighted higher-order time aggregated network. The weight of each hyperlink represents the number of events/activations of the hyperlink. Similar to our analysis of the d-degree and d'-degree of node, we find the d-strength and d'-strength of a node are also positively correlated when (d', d) equal to (3,2) in each temporal network, as shown in Figures 3.7 and 3.8 . This trend is less evident only in physical contacts that have few order 4 events, when (d', d) is equal to (4,3) and (4,2). This suggests that an individual's large number of interactions of one order would not reduce his or her number of events of another order. Individuals tend to be consistently active or inactive in events across orders.

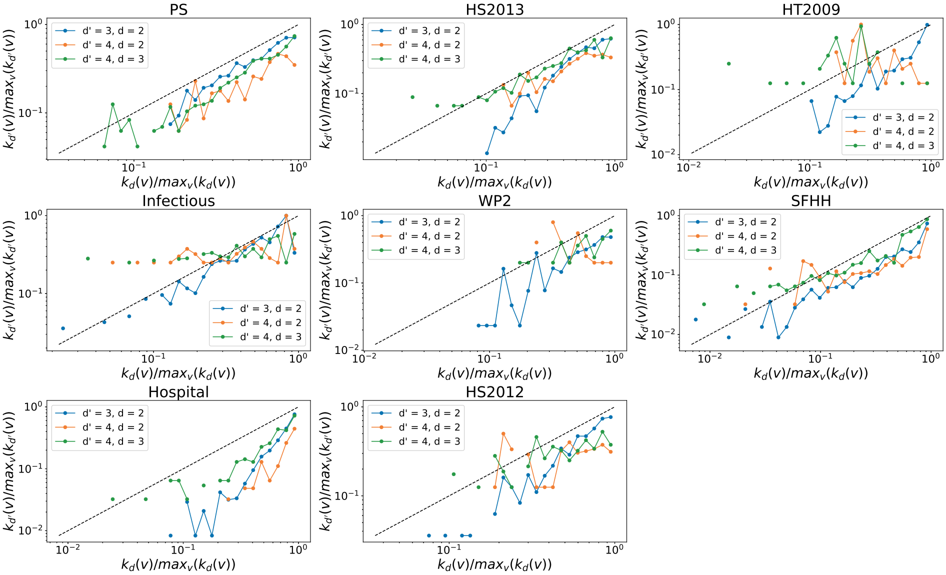

Fig. 3.5 : The d'-degree $k_{d'}(v)$ versus the the d-degree $k_d(v)$ of a node v when (d', d) is equal to (3,2) (blue line), (4,2) (yellow line) and (4,3) (green line) respectively in each physical contact network. Each axis (e.g., $k_d(v)$) has been normalized by its maximum (e.g., $max_v(k_d(v))$). Only nodes whose d-degree and d'-degree are both non-zero are considered. The dashed line represent the reference case $\frac{k_{d'}(v)}{max_v(k_{d'}(v))} = \frac{k_d(v)}{max_v(k_d(v))}$. Note that both axes are presented in logarithmic scales. In total 30 logarithmic bins are split for horizontal axis.

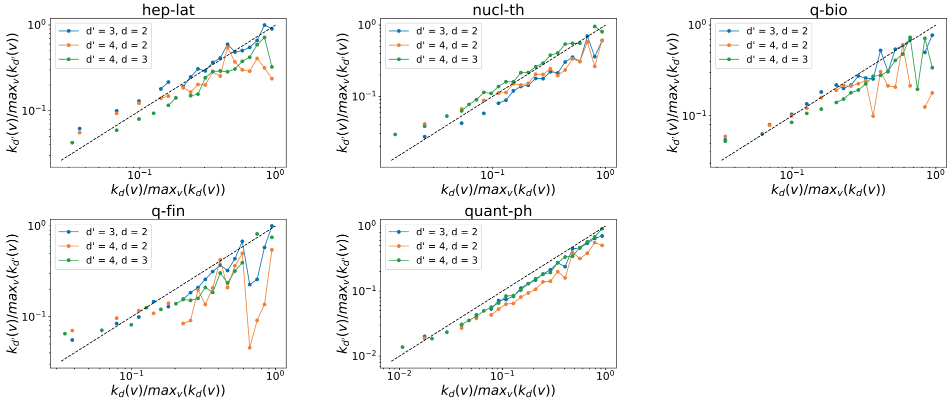

Fig. 3.6 : The d'-degree $k_{d'}(v)$ versus the the d-degree $k_d(v)$ of a node v when (d', d) is equal to (3,2) (blue line), (4,2) (yellow line) and (4,3) (green line) respectively in each collaboration network. Each axis (e.g., $k_d(v)$) has been normalized by its maximum (e.g., $max_v(k_d(v))$). Only nodes whose d-degree and d'-degree are both non-zero are considered. The dashed line represent the reference case $\frac{k_{d'}(v)}{max_v(k_{d'}(v))} = \frac{k_d(v)}{max_v(k_d(v))}$. Note that both axes are presented in logarithmic scales. In total 30 logarithmic bins are split for horizontal axis.

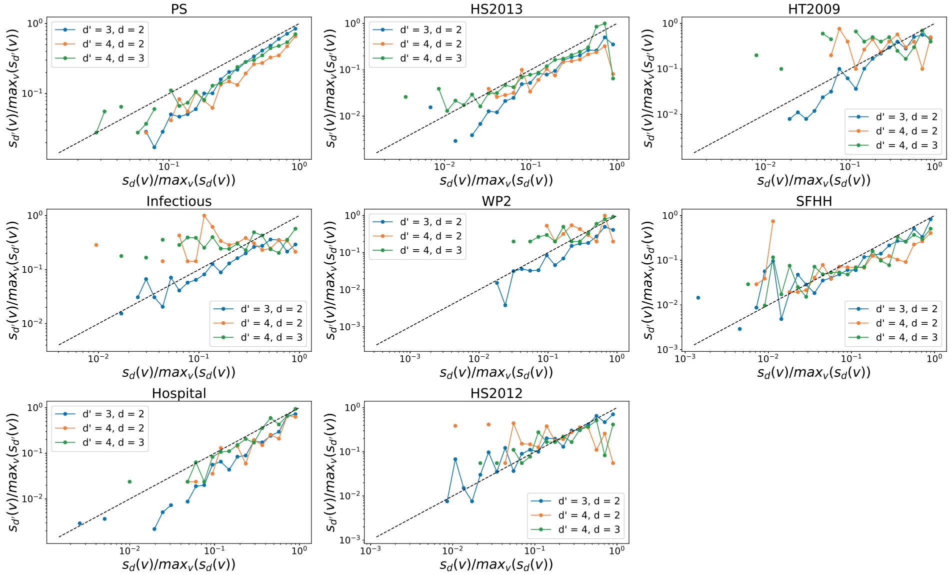

Fig. 3.7 : The d'-strength $s_{d'}(v)$ versus the the d-strength $s_d(v)$ of a node v when (d', d) is equal to (3,2) (blue line), (4,2) (yellow line) and (4,3) (green line) respectively in each physical contact network. Each axis (e.g., $s_d(v)$) has been normalized by its maximum (e.g., $max_v(s_d(v))$). Only nodes whose d-strength and d'-strength are both non-zero are considered. The dashed line represent the reference case $\frac{s_{d'}(v)}{max_v(s_{d'}(v))} = \frac{s_d(v)}{max_v(s_d(v))}$, where d'-strength is a linear function of the d-strength of nodes. Note that both axes are presented in logarithmic scales. In total 30 logarithmic bins are split for horizontal axis.

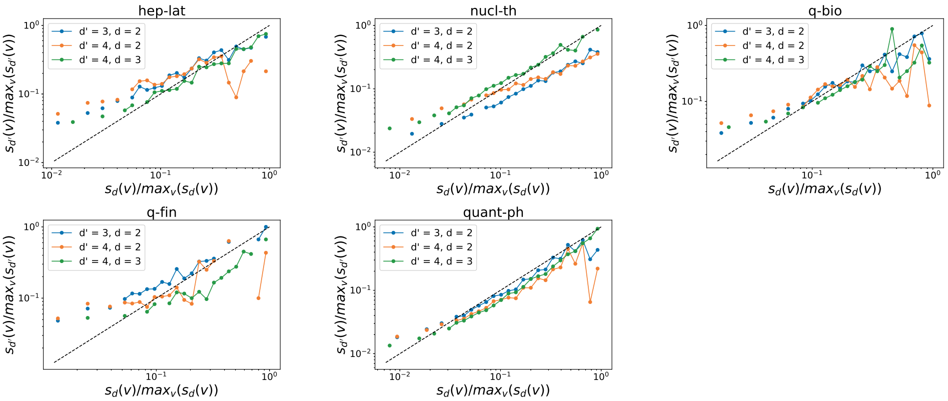

Fig. 3.8 : The d'-strength $s_{d'}(v)$ versus the the d-strength $s_d(v)$ of a node v when (d', d) is equal to (3,2) (blue line), (4,2) (yellow line) and (4,3) (green line) respectively in each collaboration network. Each axis (e.g., $s_d(v)$) has been normalized by its maximum (e.g., $max_v(s_d(v))$). Only nodes whose d-strength and d'-strength are both non-zero are considered. The dashed line represent the reference case $\frac{s_{d'}(v)}{max_v(s_{d'}(v))} = \frac{s_d(v)}{max_v(s_d(v))}$, where d'-strength is a linear function of the d-strength of nodes. Note that both axes are presented in logarithmic scales. In total 30 logarithmic bins are split for horizontal axis.

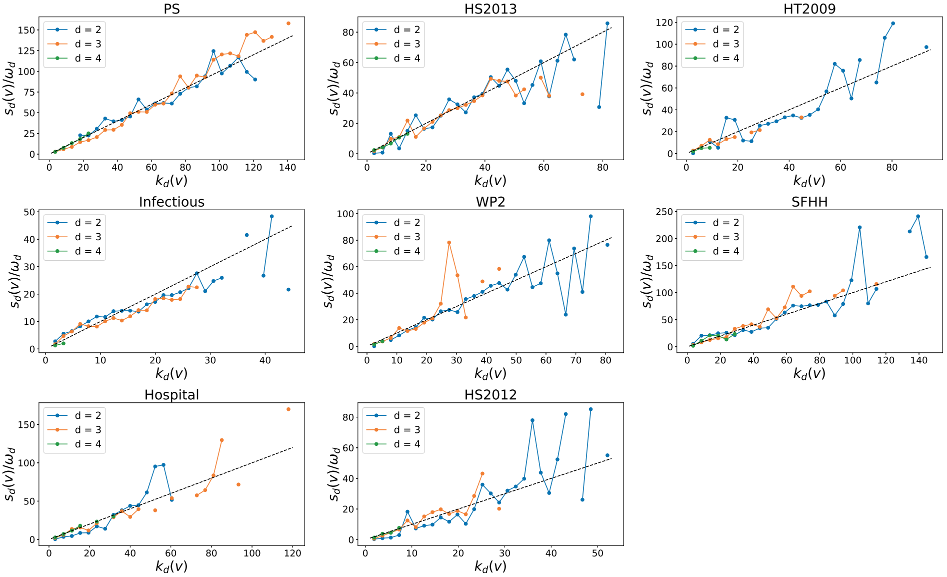

Fig. 3.9 : The d-strength $s_d(v)$ versus the the d-degree $k_d(v)$ of a node v when d is equal to 2 (blue line), 3 (yellow line) and 4 (green line) respectively in each physical contact network. The dashed line represent the reference case $s_d(v) = \omega_d * k_d(v)$, where ω_d is the average number of activations of a hyperlink of order d. In total 30 linear bins are split for horizontal axis.

To explain the positive correlation observed both in the degree of a node between two different orders and in the strength of a node between two different orders, we investigated the correlation between the d-strength and d-degree of a node, in every dataset as shown in Figures 3.9 and 3.10 . We find that the d-strength of a node is approximately a linear function of the d-degree of the node at each order. In particular, we found that, given a node v, $s_d(v) \approx \omega_d \ * k_d(v)$, where ω_d is the average number of activations of a hyperlink of order d.

The degree and strength of each node for any order remain the same in a real-world network and its three randomized networks except that the strength of nodes in $\mathscr{H}_d^2$ differs from that in the other networks. In $\mathscr{H}_d^2$, $s_d(v) = \omega_d \ * k_d(v)$ is expected for each order d and confirmed in Figures S3.13 and S3.14 (in Appendix), since the time series of order d hyperlinks are swapped in $\mathscr{H}_d^2$. This linear function $s_d(v) = \omega_d \ * k_d(v)$ observed in each real-world network approximately, means that the average number of times a node interacts with an order d group (the ratio of the d-strength to the d-degree of the node) is a constant, independent of the number of distinct order d groups the node interacts with. Thus, engaging in more groups of a given order d will not affect an individual's average number of interactions per group. The positive correlation in the degree of a node between two different orders, together with the linear relation found between the d-strength and d-degree of a node, explains the positive correlation found in the strength of a node between two different orders.

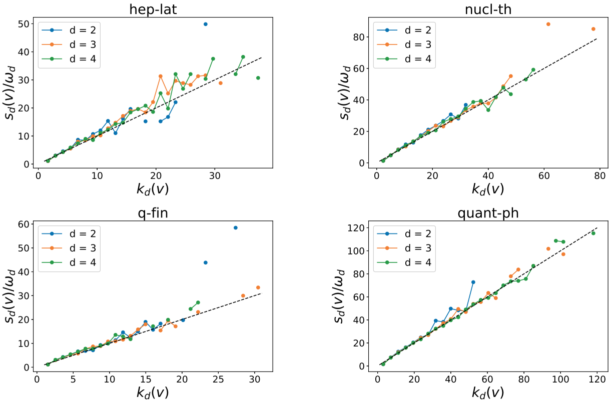

Fig. 3.10 : The d-strength $s_d(v)$ versus the the d-degree $k_d(v)$ of a node v when d is equal to 2 (blue line), 3 (yellow line) and 4 (green line) respectively in each collaboration network. The dashed line represent the reference case $s_d(v) = \omega_d * k_d(v)$, where ω_d is the average number of activations of a hyperlink of order d. In total 30 linear bins are split for horizontal axis.

3.4.3. TEMPORAL CORRELATION OF EVENTS AT A LOCAL EGO NETWORK

Since higher-order events overlap in topology, e.g., the component nodes of a higher-order event tend to participate in events of a lower order, we explore further the temporal correlation of events that occur locally in topology. The topological neighborhood of a hyperlink h_d of order d, so called the ego network $ego(h_d)$ centered at h_d, is defined as the union of the hyperlink h_d and all hyperlinks with an order lower than d that share at least one node with h_d in the higher-order aggregated network. We construct the time series of the aggregated activity of an ego network $ego(h_d)$, as the sum of the time series of hyperlinks belonging to $ego(h_d)$, as shown in Figure 3.11 . We then evaluate the temporal correlation of the time series of an ego network $ego(h_d)$, to understand whether the activation of the center hyperlink h_d tend to cluster in time with the activation of the other low order hyperlinks

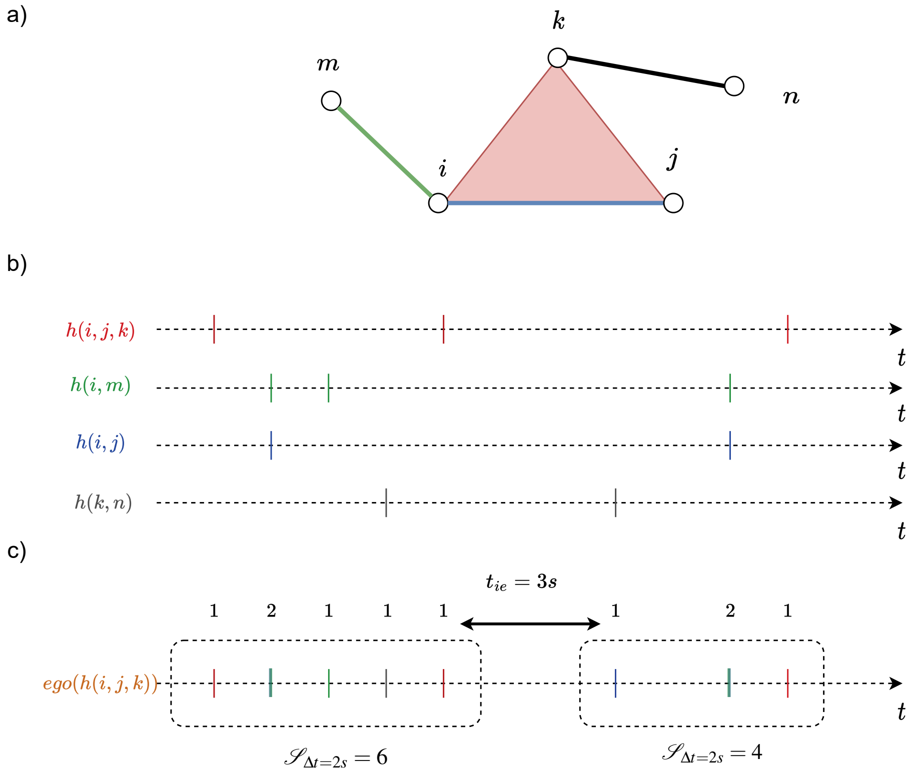

Fig. 3.11 : Schematic representation of a) the ego network of the hyperlink $h(i, j, k)$, i.e. $ego(h(i, j, k))$, b) the time series associated to links belonging to $ego(h(i, j, k))$, c) the time series of the activity of $ego(h(i, j, k))$, which is the sum of the time series of hyperlinks belonging to the ego network, and its event trains identified when $\Delta t = 2s$.

in the ego network $ego(h_d)$.

Our analysis method is based on the concept of event trains, proposed by Karsai et al. [5]. A train of events is a sequence of consecutive events whose inter-event times are shorter than or equal to a reference temporal interval Δt and separated from the other contacts by an inter-event times larger than Δt. Given a Δt and an activity time series of an ego network $ego(h_d)$, trains can be identified, as exemplified in Figure 3.11 . Given Δt and an order d, we identify all the trains for each activity series of the ego network centered at each order d hyperlink. The size of a train is the number of events the train contains. Then, we examine the size distribution $Pr[\mathscr{S}_d^* = s]$ of the identified trains in which a center hyperlink has been activated at least once. The timescales of physical contacts and collaboration networks are different. The two classes are measured per step of seconds and day respectively. To illustrate our method and findings we consider $\Delta t = 60s$ $(60d)$ in physical contact (collaboration) networks to identify the trains in each ego network. The choice $\Delta t = 60s$ is also motivated by the observation in Figure 3.1 that we start to observe the positive temporal and topological correlation of higher-order events since Δt is about $100s$ in physical contact networks. Moreover, we observe the same when $\Delta t = 120s$ $(120d)$ in physical contact (collaboration) networks in the coming analysis.

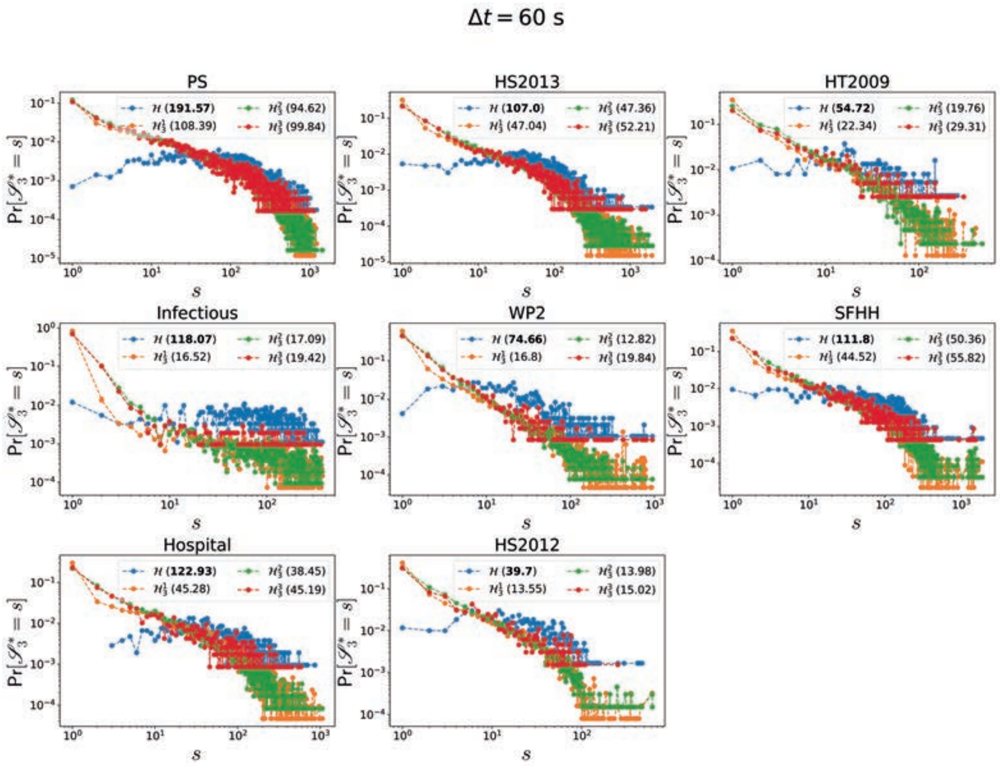

Fig. 3.12 : Probability distribution $Pr[\mathscr{S}_3^* = s]$ of the size $\mathscr{S}_3^*$ of trains (obtained from the activity series of ego networks centered at each order 3 hyperlink), where a center link is activated at least once, in each physical contact network $\mathscr{H}$ (blue) and its three randomized reference models $\mathscr{H}_3^1$ (yellow), $\mathscr{H}_3^2$ (green) and $\mathscr{H}_3^3$ (red). To identify the trains, we consider $\Delta t = 60s$. For each network, the average size of the trains is reported. The maximum average size among network $\mathscr{H}$, $\mathscr{H}_3^1$, $\mathscr{H}_3^2$ and $\mathscr{H}_3^3$ is in bold. The horizontal and vertical axes are presented in logarithmic scale.

Figure 3.12 and 3.13 show the train size distribution $Pr[\mathscr{S}_3^* = s]$ of the ego networks centered at each order 3 hyperlink in each physical and collaboration network $\mathscr{H}$ and its three null models $\mathscr{H}_3^1$, $\mathscr{H}_3^2$, $\mathscr{H}_3^3$. Only order 3 events have been randomized in the three randomized reference models $\mathscr{H}_3^1$, $\mathscr{H}_3^2$, and $\mathscr{H}_3^3$ while the set of events of any other order $d' \neq 3$ remain unchanged in each real-world network and its corresponding randomized network $\mathscr{H}_3^1$, $\mathscr{H}_3^2$, $\mathscr{H}_3^3$. In physical contact networks, the train size is evidently larger on average than that in their corresponding randomized networks. This indicates that an order 3 event tend to occur close in time with many local order 2 events, forming large trains. The trains in collaboration networks are, however, not evidently longer than those in randomized reference models on average. We found similar when considering $\Delta t = 120s$ for physical contacts and $\Delta t = 120d$ for collaboration networks (see Figures S3.15 and S3.16 in Appendix).

The temporal correlation analysis of local events helps to explain the interrelation of

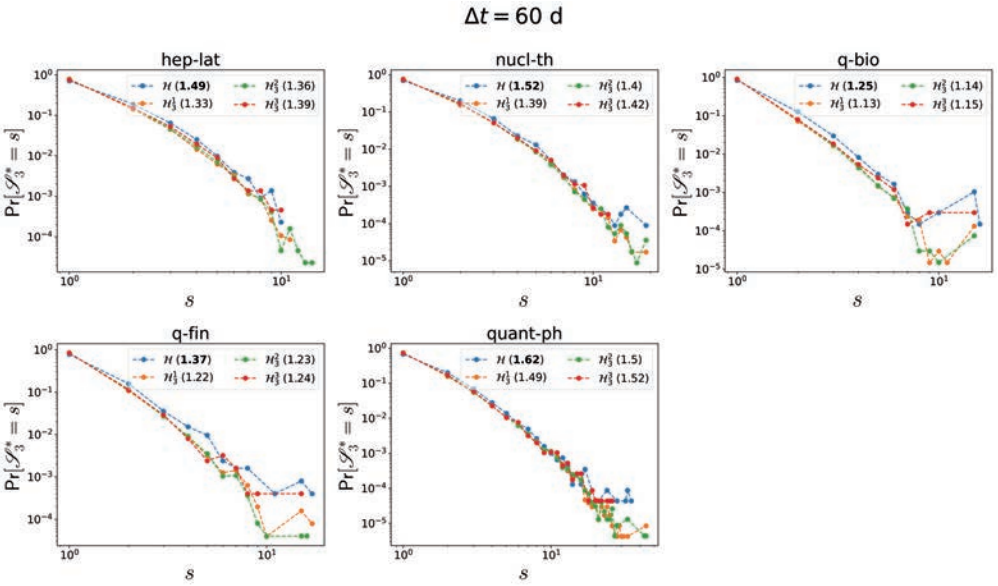

Fig. 3.13 : Probability distribution $Pr[\mathscr{S}_3^* = s]$ of the size $\mathscr{S}_3^*$ of trains (obtained from the activity series of ego networks centered at each order 3 hyperlink), where a center link is activated at least once, in each collaboration network $\mathscr{H}$ (blue) and its three randomized reference models $\mathscr{H}_3^1$ (yellow), $\mathscr{H}_3^2$ (green) and $\mathscr{H}_3^3$ (red). To identify the trains, we consider $\Delta t = 60s$. For each network, the average size of the trains is reported. The maximum average size among network $\mathscr{H}$, $\mathscr{H}_3^1$, $\mathscr{H}_3^2$ and $\mathscr{H}_3^3$ is in bold. The horizontal and vertical axes are presented in logarithmic scale.

topological and temporal distance of higher-order events discovered in Subsection 3.4.1. In physical contact (collaboration) networks, we observe evident (no evident) correlation between topological and temporal distance of events with different orders. Consistently, whereas events overlap in component nodes in both types of networks, local events, thus events close in topology, are strongly (weakly or not) correlated in time, in forming long trains, in physical contact (collaboration) networks. In networks where the interrelation between topological and temporal distance of events is more evident (e.g., Infectious and WP2), the correlation of local events in time also tends to be stronger (average train size observed in real-work network is evidently larger than that of randomized reference models). We observe similar results also for the distribution $Pr[\mathscr{S}_4^* = s]$ of the size $\mathscr{S}_4^*$ of trains obtained from the activity series of ego networks centered at each order 4 hyperlink, as shown in Figures S3.17, S3.18, S3.19 and S3.20 in Appendix.

The detected differences between physical contact and collaboration networks may be explained by the fact that physical interactions are driven by physical proximity. For example, individuals that have a group interaction are close in physical distance, which may facility the interaction of a subgroup, resulting in events close in time and topology.

Finally, we discuss briefly whether our finding of the temporal-topological correlation in higher-order temporal networks is still valid taking into account that the higher-order temporal networks we constructed is likely imprecise. The physical contact networks measured are possibly incomplete, influencing the resultant higher-order temporal networks. If the

$\binom{d}{2}$ pair-wise contacts of an order d event are not observed completely but with one contact missing, the observed higher-order network would be composed of two order $d-1$ events. Hence, we will add such potential missing contacts back to our pair-wise physical contact networks, re-construct the corresponding higher-order networks and explore whether similar temporal-topological correlation could be still be observed. We examine each pair-wise physical contact network at each time step, identify all subgraphs that are composed of a clique of size $d > 3$ with one missing link, add such missing links to original pair-wise physical contact networks and construct the corresponding higher-order networks $\mathscr{H}_{miss}$ as described in Section 3.3. Figure S3.21 (in Appendix) shows the slight change in the number of events of each order in $\mathscr{H}_{miss}$ symbol where the missing links have been added. The general observation of the temporal-topological correlation and Infectious and WP2 being among the networks with the strongest correlation holds also for $\mathscr{H}_{miss}$, as shown in Figures S3.22 and S3.23 (Appendix) for order $d = 3$ and $d = 4$, respectively.

3.5. CONCLUSION

In this chapter, we have proposed a method to systematically characterize temporal and topological properties of events of arbitrary orders. We applied our methods to 8 physical contact and 5 collaboration higher-order evolving networks and observe their difference. In physical contacts, events relatively close in time tend to occur also close in topology. Moreover, events usually overlap in component nodes and these local events overlapping in component nodes are also usually correlated in time. Such temporal correlation of local events supports again the correlation between temporal and topological distances of events observed in our first analysis. Differently, in collaboration networks, the temporal and topological correlation of events is either weak or absent. Despite events also overlap in component nodes, their temporal correlation almost disappears in collaboration networks. The detected dissimilarities between physical contacts and collaboration networks could be related to a fundamental difference between the two kind of networks. In physical contacts individuals participate in events driven by physical proximity. The physical proximity of individuals that participate in a higher-order event may facilitate interaction of them or a subgroup in the near future. The time of scientific collaborations are likely driven more by their content and creation process.

Via our analysis of the topological overlap of events with different orders in component nodes, we also observe similarities between the two kinds of networks. Nodes that participate in many events (groups) of a given order tend to interact in many events (groups) of a different order. Hence, nodes are consistent in interactions with respect to frequency and diversity across different orders.

Our method explores the temporal and topological relation of the basic building block of events, the activations of fully connected cliques. A promising direction could be generalizing this method to the activations of relevant motifs, and to investigate the interplay between topological location and temporal delay of such structures. Beyond, our method can be applied to compare different classes of networks (e.g. biological, brain or collaboration networks) and to explore how detected properties/patterns of a network can influence the dynamic processes unfolding on the network. Finally, the topological and temporal properties of events detected in this chapter could foster higher-order evolving network models that better reproduce patterns observed so far.

3.6. Appendix

3.6.1. General Statistics

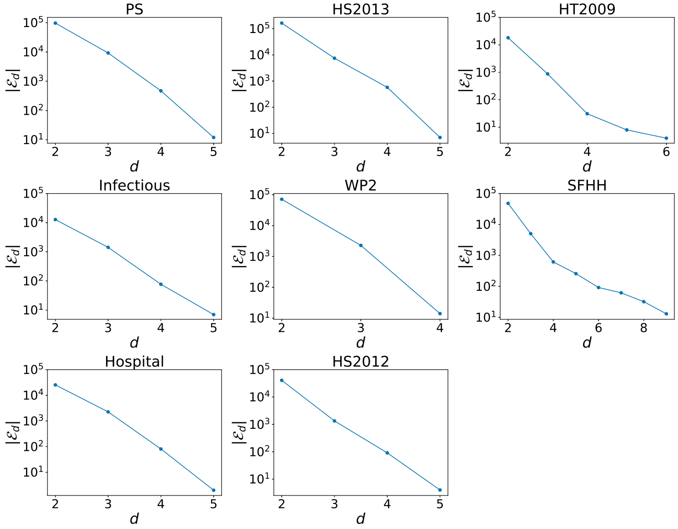

Fig. S3.1: Total number of events ($|\mathcal{E}_d|$) for each order d in physical contact networks. Vertical axis is presented in logarithmic scale.

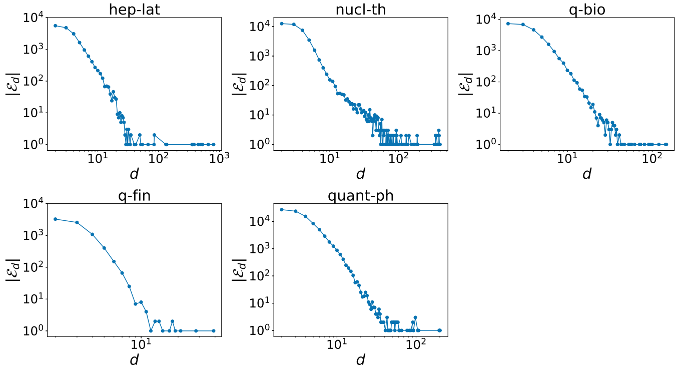

Fig. S3.2: Total number of events ($|\mathcal{E}_d|$) for each order d in collaboration networks. Vertical and horizontal axes are presented in logarithmic scale.

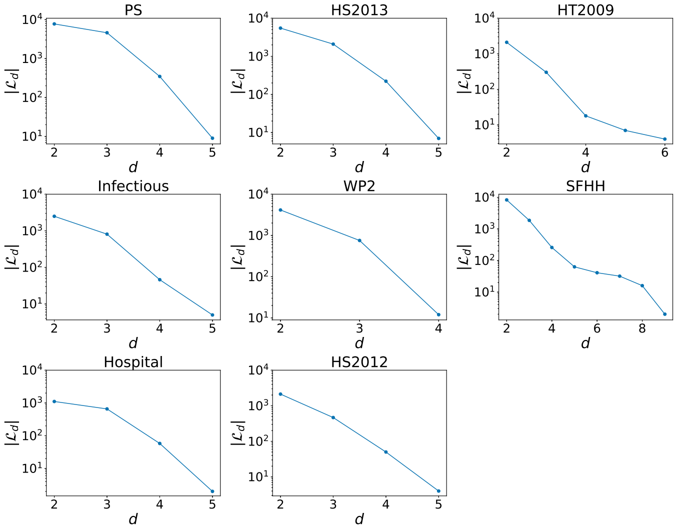

Fig. S3.3: Total number of hyperlinks ($|\mathscr{L}_d|$) in the time aggregated higher-order network for each order d for physical contact datasets. Vertical axis is presented in logarithmic scale.

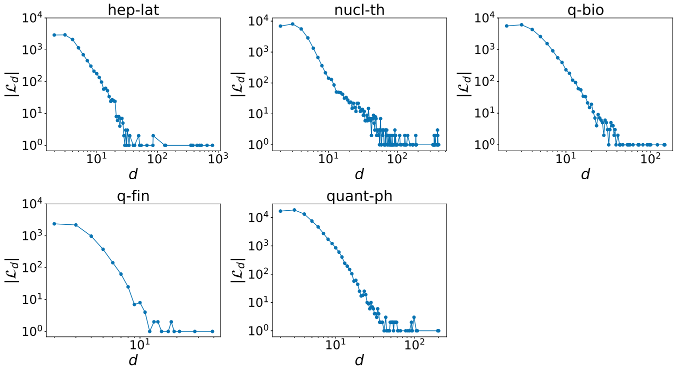

Fig. S3.4: Total number of hyperlinks ($|\mathscr{L}_d|$) for each order d in collaboration networks. Vertical and horizantal axes are presented in logarithmic scale.

3.6.2. TEMPORAL-TOPOLOGICAL CORRELATION OF EVENTS

Correlation of temporal and topological distance of events

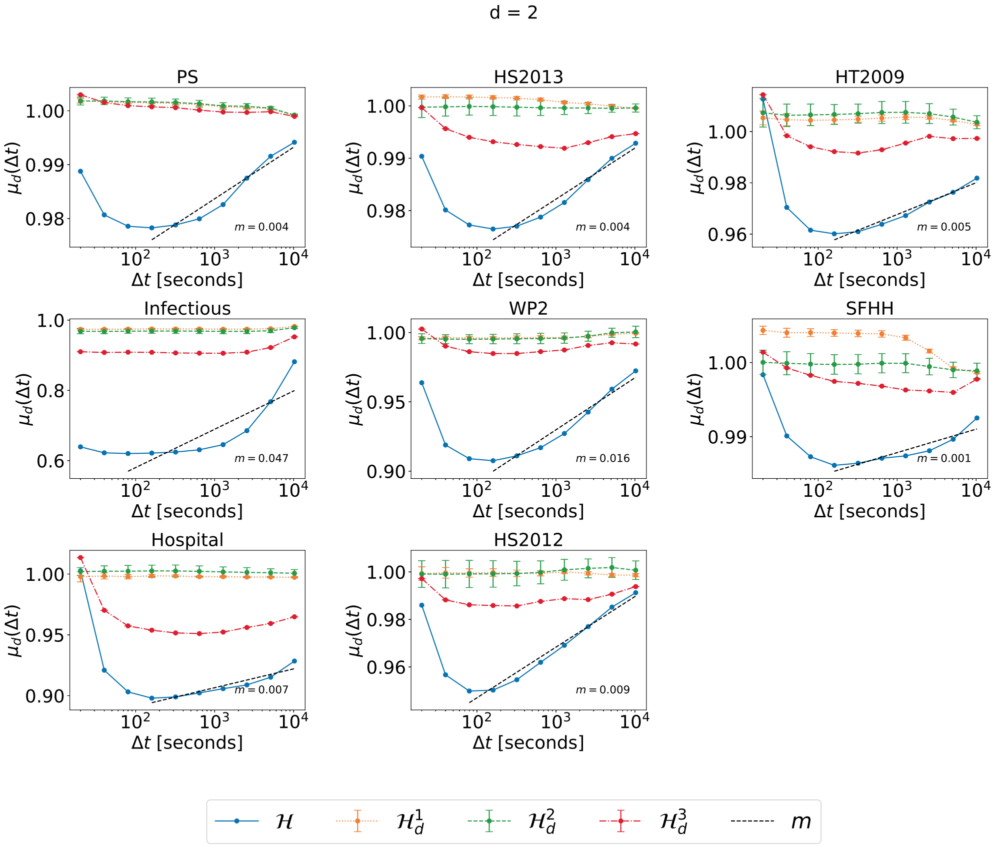

Fig. S3.5: The normalized average topological distance $\mu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\, e\in\mathcal{E}_d,\, e'\in\mathcal{E}\setminus\mathcal{E}_d]}{E[\eta(e,e')|\, e\in\mathcal{E}_d,\, e'\in\mathcal{E}\setminus\mathcal{E}_d]}$, between an order $d = 2$ event and an event of a different order, in each physical contact network and its corresponding three randomized null models $\mathcal{H}_d^1$ (yellow), $\mathcal{H}_d^2$ (green) and $\mathcal{H}_d^3$ (red), which preserve or destroy specific properties of order $d = 3$ events. $\lim_{\Delta t\to\infty} E[\eta(e,e')|\mathcal{T}(e,e') < \Delta t,\ e \in \mathcal{E}_d,\ e' \in \mathcal{E} \setminus \mathcal{E}_d] = E[\eta(e,e')|\ e \in \mathcal{E}_d,\ e' \in \mathcal{E} \setminus \mathcal{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure corresponds to the linear fit (with slope m) of $\mu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathcal{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

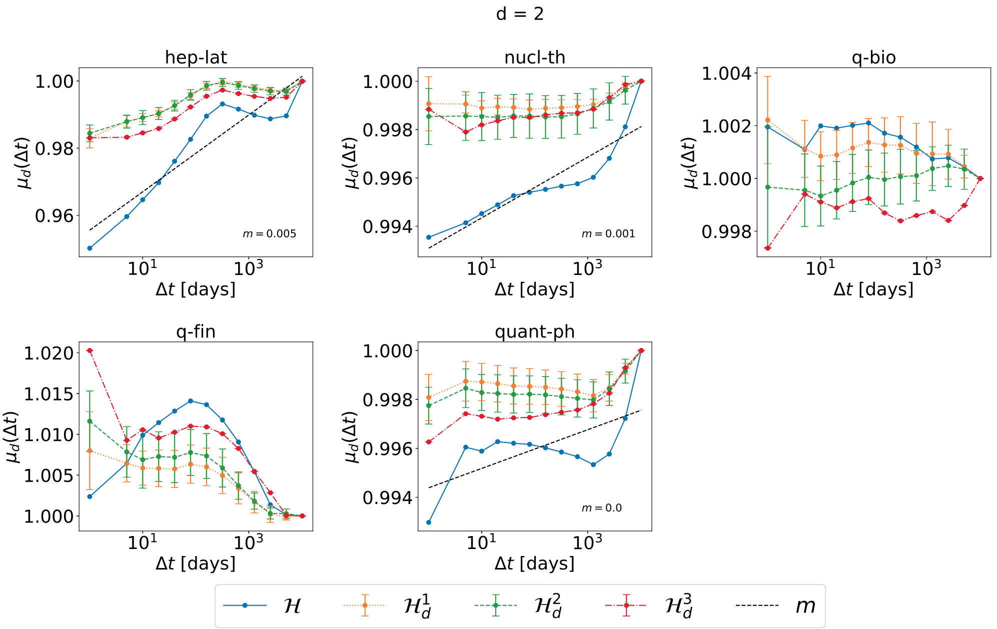

Fig. S3.6: The normalized average topological distance $\mu_d(\Delta t) = \frac{E[\eta(e,e')|\mathscr{T}(e,e')<\Delta t,\ e\in\mathscr{E}_d,\ e'\in\mathscr{E}\setminus\mathscr{E}_d]}{E[\eta(e,e')|\ e\in\mathscr{E}_d,\ e'\in\mathscr{E}\setminus\mathscr{E}_d]}$, between an order $d = 2$ event and an event of a different order, in each collaboration network and its corresponding three randomized null models $\mathscr{H}_d^1$ (yellow), $\mathscr{H}_d^2$ (green) and $\mathscr{H}_d^3$ (red), which preserve or destroy specific properties of order $d = 2$ events. $\lim_{\Delta t\to\infty} E[\eta(e,e')|\mathscr{T}(e,e') < \Delta t,\ e \in \mathscr{E}_d,\ e' \in \mathscr{E} \setminus \mathscr{E}_d] = E[\eta(e,e')|\ e \in \mathscr{E}_d,\ e' \in \mathscr{E} \setminus \mathscr{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure corresponds to the linear fit (with slope m) of $\mu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathscr{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

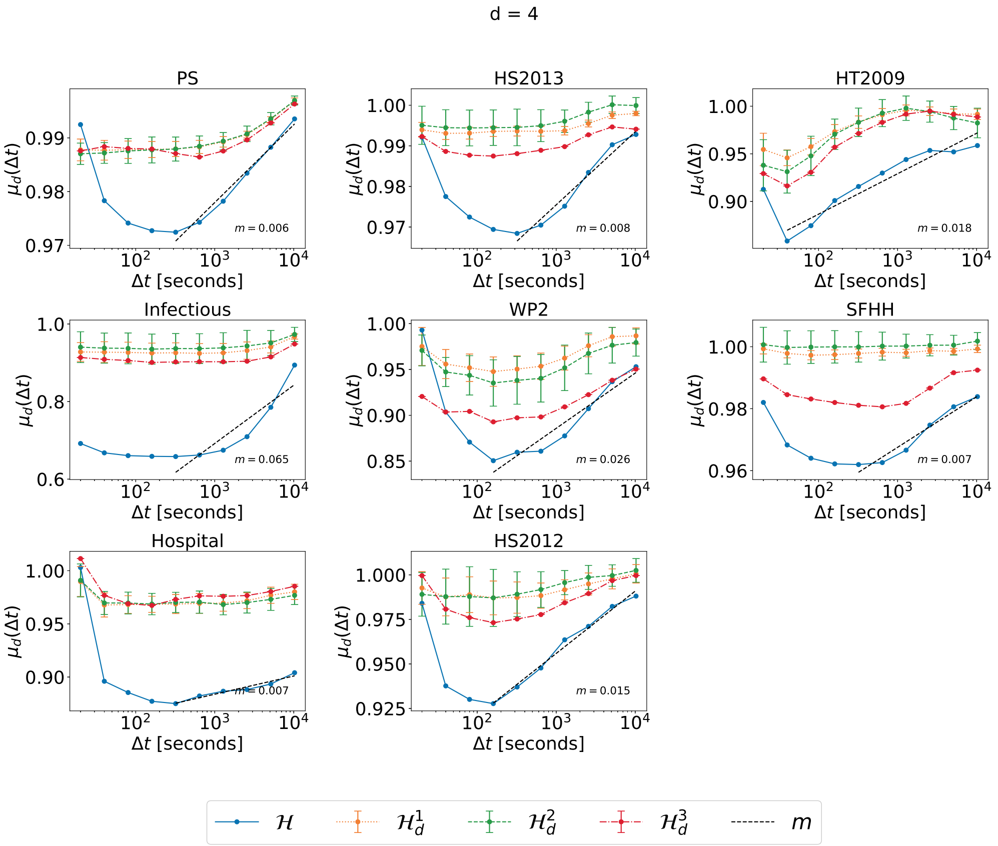

Fig. S3.7: The normalized average topological distance $\mu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\setminus\mathcal{E}_d]}{E[\eta(e,e')|\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\setminus\mathcal{E}_d]}$, between an order $d = 4$ event and an event of a different order, in each physical contact network and its corresponding three randomized null models $\mathcal{H}^1_d$ (yellow), $\mathcal{H}^2_d$ (green) and $\mathcal{H}^3_d$ (red), which preserve or destroy specific properties of order $d = 4$ events. $\lim_{\Delta t\to\infty} E[\eta(e,e')|\mathcal{T}(e,e') < \Delta t,\ e \in \mathcal{E}_d,\ e' \in \mathcal{E} \setminus \mathcal{E}_d] = E[\eta(e,e')|\ e \in \mathcal{E}_d,\ e' \in \mathcal{E} \setminus \mathcal{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure corresponds to the linear fit (with slope m) of $\mu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathcal{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

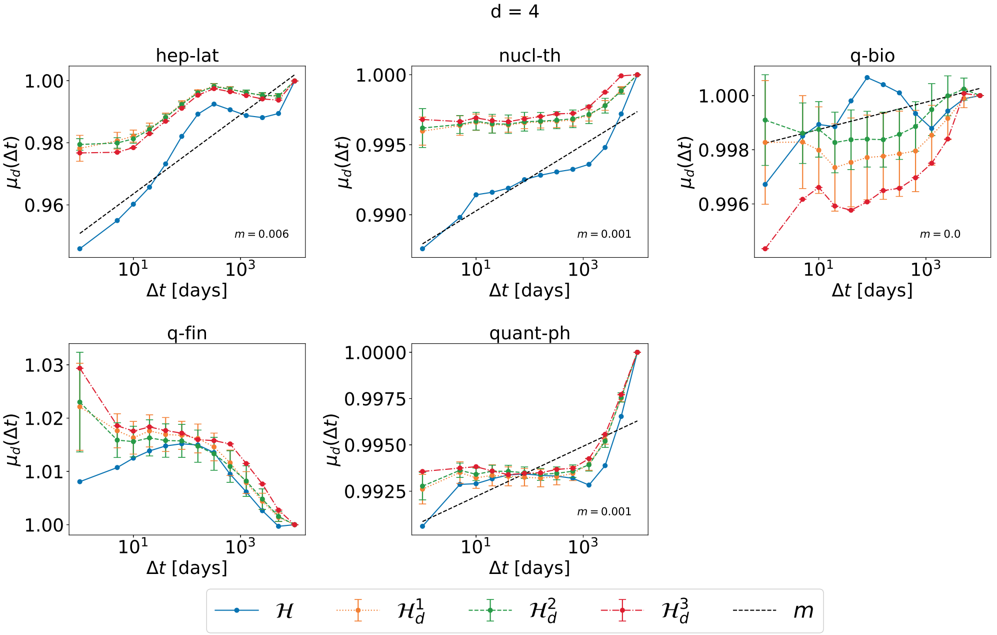

Fig. S3.8: The normalized average topological distance $\mu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\setminus\mathcal{E}_d]}{E[\eta(e,e')|\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\setminus\mathcal{E}_d]}$, between an order $d = 4$ event and an event of a different order, in each collaboration network and its corresponding three randomized null models $\mathcal{H}_d^1$ (yellow), $\mathcal{H}_d^2$ (green) and $\mathcal{H}_d^3$ (red), which preserve or destroy specific properties of order $d = 3$ events. $\lim_{\Delta t\to\infty} E[\eta(e,e')|\mathcal{T}(e,e') < \Delta t,\ e \in \mathcal{E}_d,\ e' \in \mathcal{E}\setminus\mathcal{E}_d] = E[\eta(e,e')|\ e \in \mathcal{E}_d,\ e' \in \mathcal{E}\setminus\mathcal{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure corresponds to the linear fit (with slope m) of $\mu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathcal{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

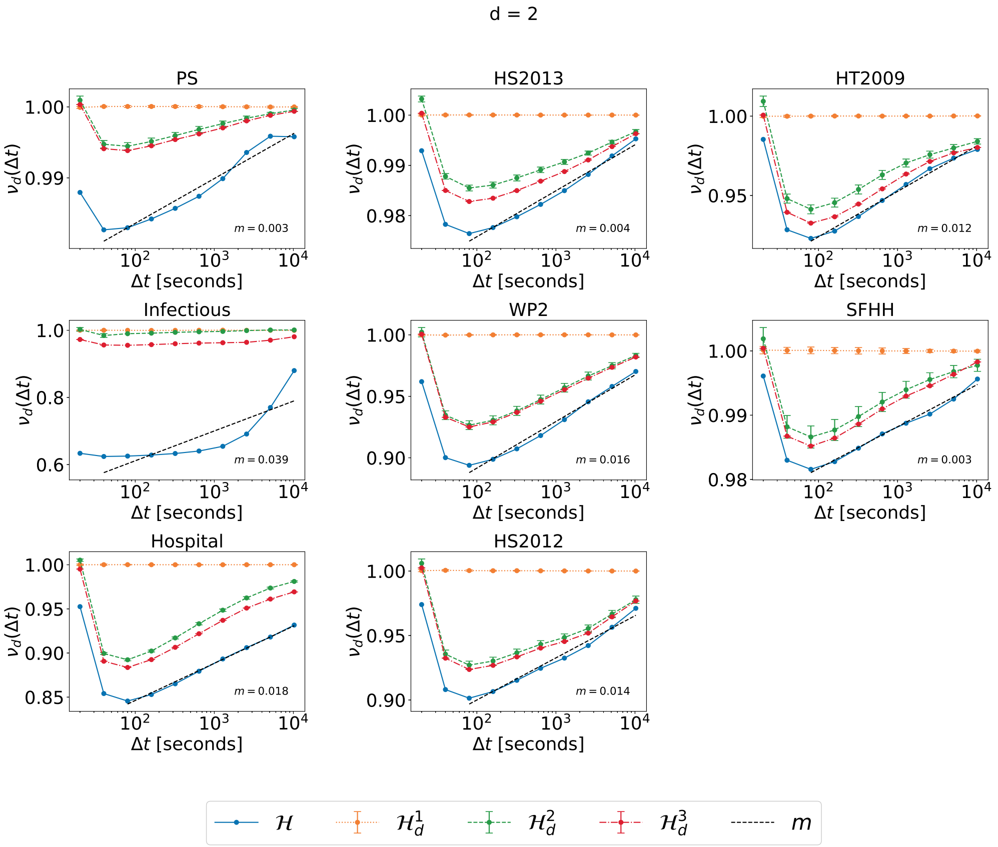

Fig. S3.9: The normalized average topological distance $\nu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e,\ e'\in\mathcal{E}_d]}{E[\eta(e,e')|\ e,\ e'\in\mathcal{E}_d]}$, between two order $d = 2$ events, in each physical contact network and its corresponding three randomized null models $\mathcal{H}^1_d$ (yellow), $\mathcal{H}^2_d$ (green) and $\mathcal{H}^3_d$ (red), which preserve or destroy specific properties of order $d = 2$ events. $\lim_{\Delta t \to \infty} E[\eta(e,e')|\mathcal{T}(e,e') < \Delta t,\ e,\ e' \in \mathcal{E}_d] = E[\eta(e,e')|\ e,\ e' \in \mathcal{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure corresponds to the linear fit (with slope m) of $\nu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathcal{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

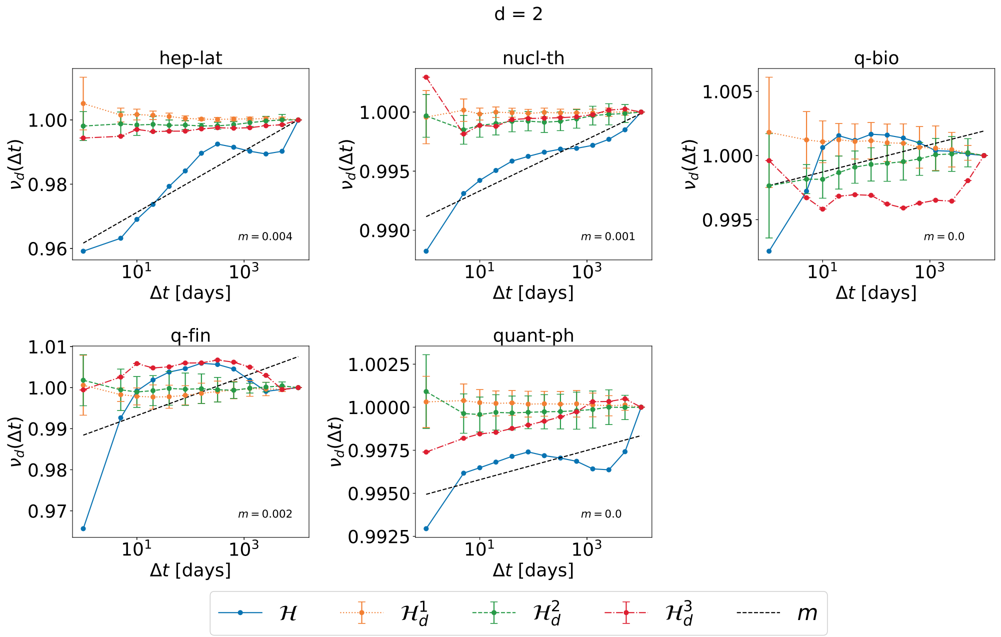

Fig. S3.10: The normalized average topological distance $\nu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e,\ e'\in\mathcal{E}_d]}{E[\eta(e,e')|\ e,\ e'\in\mathcal{E}_d]}$, between two order $d = 2$ events, in each collaboration network and its corresponding three randomized null models $\mathcal{H}_d^1$ (yellow), $\mathcal{H}_d^2$ (green) and $\mathcal{H}_d^3$ (red), which preserve or destroy specific properties of order $d = 2$ events. $\lim_{\Delta t\to\infty} E[\eta(e,e')|\mathcal{T}(e,e') < \Delta t,\ e,\ e' \in \mathcal{E}_d] = E[\eta(e,e')|\ e,\ e' \in \mathcal{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure corresponds to the linear fit (with slope m) of $\nu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathcal{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

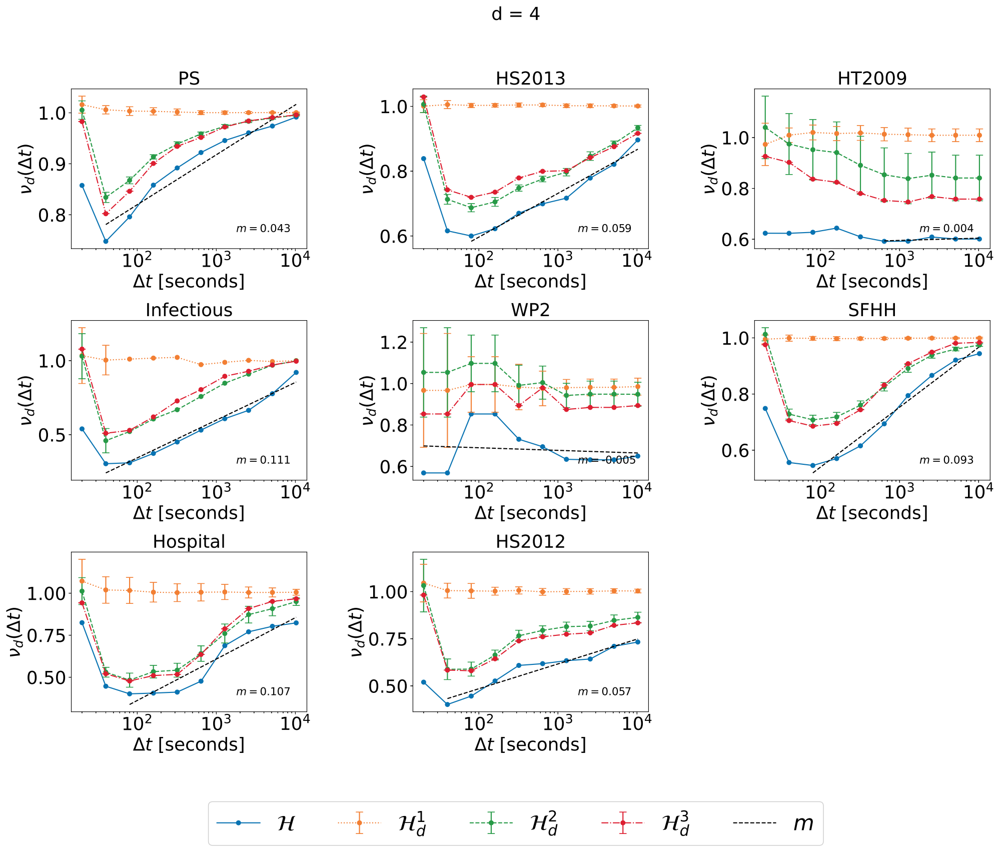

Fig. S3.11: The normalized average topological distance $\nu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e,\ e' \in \mathcal{E}_d]}{E[\eta(e,e')|\ e,\ e' \in \mathcal{E}_d]}$, between two order $d = 4$ events, in each physical contact network and its corresponding three randomized null models $\mathscr{H}_d^1$ (yellow), $\mathscr{H}_d^2$ (green) and $\mathscr{H}_d^3$ (red), which preserve or destroy specific properties of order $d = 4$ events. $\lim_{\Delta t \to \infty} E[\eta(e,e')|\mathcal{T}(e,e') < \Delta t,\ e,\ e' \in \mathcal{E}_d] = E[\eta(e,e')|\ e,\ e' \in \mathcal{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure corresponds to the linear fit (with slope m) of $\nu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathscr{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

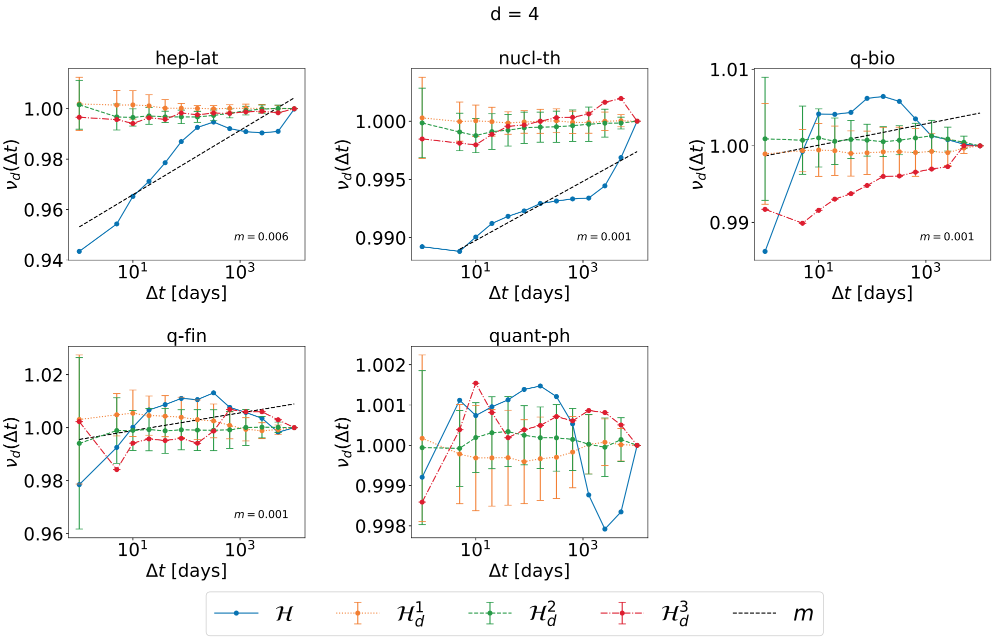

Fig. S3.12: The normalized average topological distance $\nu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e,\ e'\in\mathcal{E}_d]}{E[\eta(e,e')|\ e,\ e'\in\mathcal{E}_d]}$, between two order $d = 4$ events, in each collaboration network and its corresponding three randomized null models $\mathcal{H}_d^1$ (yellow), $\mathcal{H}_d^2$ (green) and $\mathcal{H}_d^3$ (red), which preserve or destroy specific properties of order $d = 4$ events. $\lim_{\Delta t\to\infty} E[\eta(e,e')|\mathcal{T}(e,e') < \Delta t,\ e,\ e' \in \mathcal{E}_d] = E[\eta(e,e')|\ e,\ e' \in \mathcal{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure corresponds to the linear fit (with slope m) of $\nu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathcal{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

Topological correlation of events

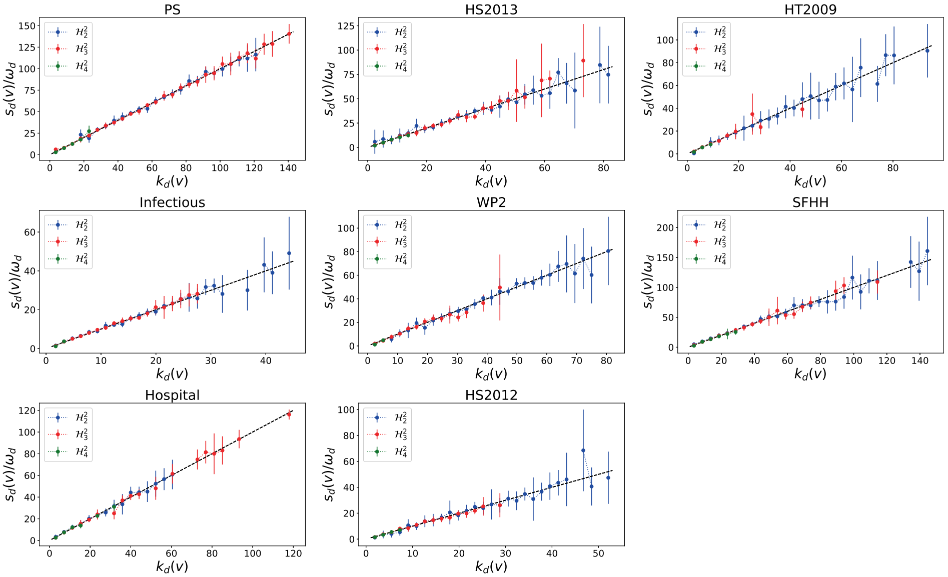

Fig. S3.13: The d-strength $s_d(v)$ versus the the d-degree $k_d(v)$ of a node v in the randomized reference model $\mathscr{H}_d^2$ obtained from each real-world physical contact network, when d is equal to 2 (blue dashed line), 3 (red dashed line) and 4 (green dashed line). The black dashed line represents the reference case $s_d(v) = \omega_d * k_d(v)$, where ω_d is the average number of activations of a hyperlink of order d. The error bar correspond to the standard deviation, centered in the mean value of 10 independent realizations of randomized reference model $\mathscr{H}_d^2$. In total 30 linear bins are split for horizontal axis.

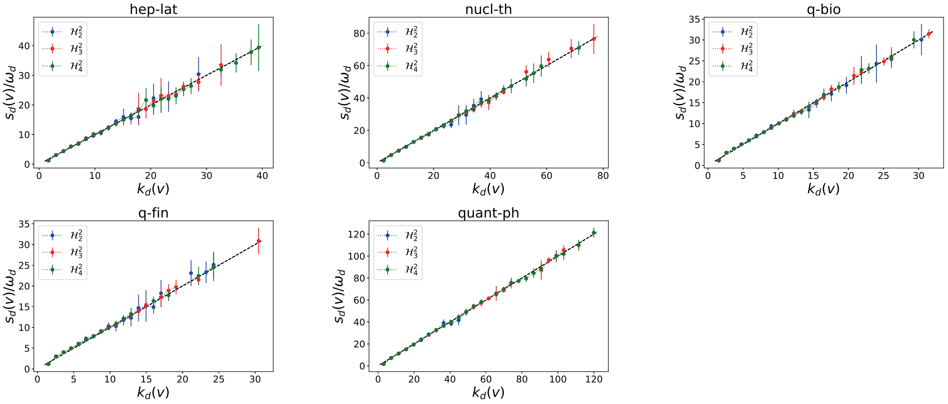

Fig. S3.14: The d-strength $s_d(v)$ versus the the d-degree $k_d(v)$ of a node v of the randomized reference model $\mathscr{H}_d^2$ obtained from each real-world collaboration network, when d is equal to 2 (blue dashed line), 3 (red dashed line) and 4 (green dashed line). The black dashed line represents the reference case $s_d(v) = \omega_d * k_d(v)$, where ω_d is the average number of activations of a hyperlink of order d. The errorbar correspond to the standard deviation, centered in the mean value of 10 independent realizations of randomized reference model $\mathscr{H}_d^2$. In total 30 linear bins are split for horizontal axis.

TEMPORAL CORRELATION OF EVENTS AT A LOCAL EGO NETWORK

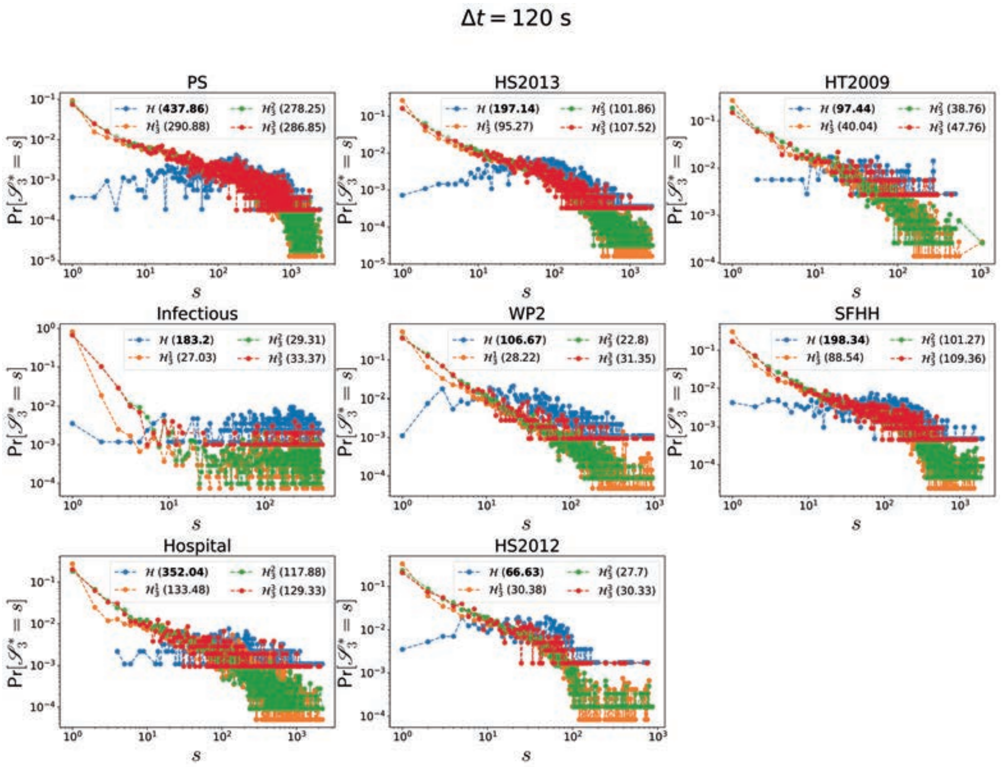

Fig. S3.15: Probability distribution $Pr[\mathscr{S}_3^* = s]$ of the size $\mathscr{S}_3^*$ of trains (obtained from the activity series of ego networks centered at each order 3 hyperlink), where a center link is activated at least once, in each physical contact network $\mathscr{H}$ (blue) and its three randomized reference models $\mathscr{H}_3^1$ (yellow), $\mathscr{H}_3^2$ (green) and $\mathscr{H}_3^3$ (red). To identify the trains, we consider $\Delta t = 120s$. For each network, the average size of the trains is reported. The maximum average size among network $\mathscr{H}$, $\mathscr{H}_3^1$, $\mathscr{H}_3^2$ and $\mathscr{H}_3^3$ is in bold. The horizontal and vertical axes are presented in logarithmic scale.

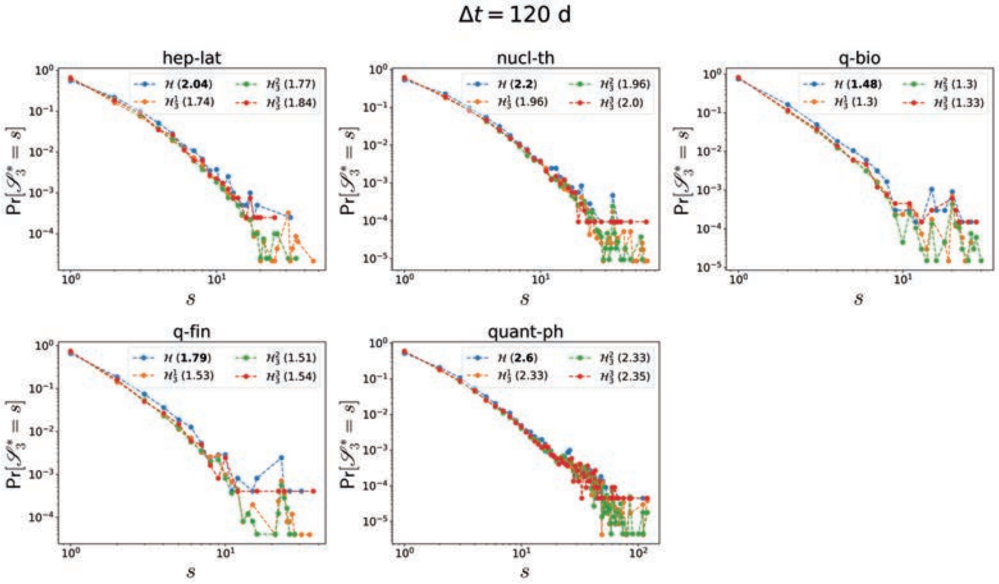

Fig. S3.16: Probability distribution $Pr[\mathcal{S}_3^* = s]$ of the size $\mathcal{S}_3^*$ of trains (obtained from the activity series of ego networks centered at each order 3 hyperlink), where a center link is activated at least once, in each collaboration network $\mathcal{H}$ (blue) and its three randomized reference models $\mathcal{H}_3^1$ (yellow), $\mathcal{H}_3^2$ (green) and $\mathcal{H}_3^3$ (red). To identify the trains, we consider $\Delta t = 120d$. For each network, the average size of the trains is reported. The maximum average size among network $\mathcal{H}$, $\mathcal{H}_3^1$, $\mathcal{H}_3^2$ and $\mathcal{H}_3^3$ is in bold. The horizontal and vertical axes are presented in logarithmic scale.

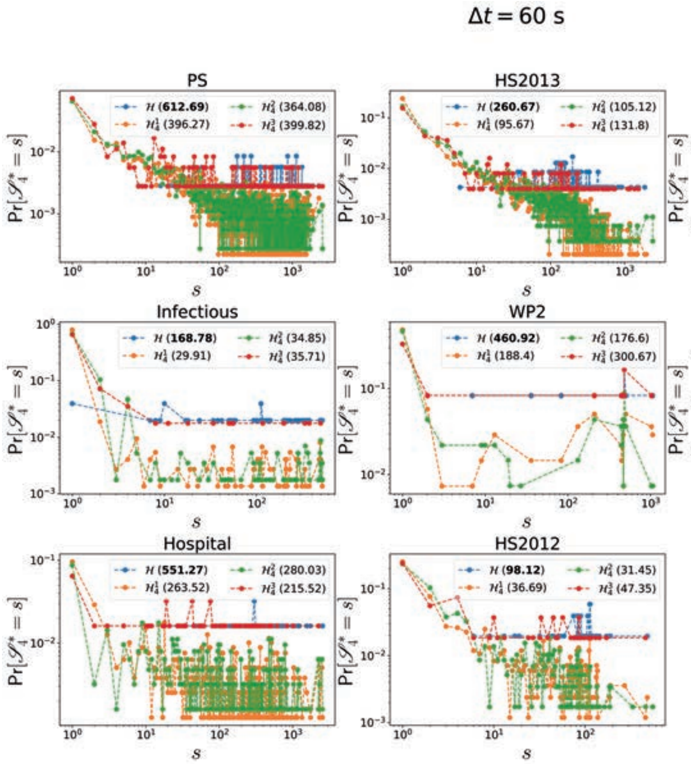

Fig. S3.17: Probability distribution $Pr[\mathscr{S}_4^* = s]$ of the size $\mathscr{S}_4^*$ of trains (obtained from the activity series of ego networks centered at each order 4 hyperlink), where a center link is activated at least once, in each physical contact network $\mathscr{H}$ (blue) and its three randomized reference models $\mathscr{H}_4^1$ (yellow), $\mathscr{H}_4^2$ (green) and $\mathscr{H}_4^3$ (red). To identify the trains, we consider $\Delta t = 60s$. For each network, the average size of the trains is reported. The maximum average size among network $\mathscr{H}$, $\mathscr{H}_4^1$, $\mathscr{H}_4^2$ and $\mathscr{H}_4^3$ is in bold. The horizontal and vertical axes are presented in logarithmic scale.

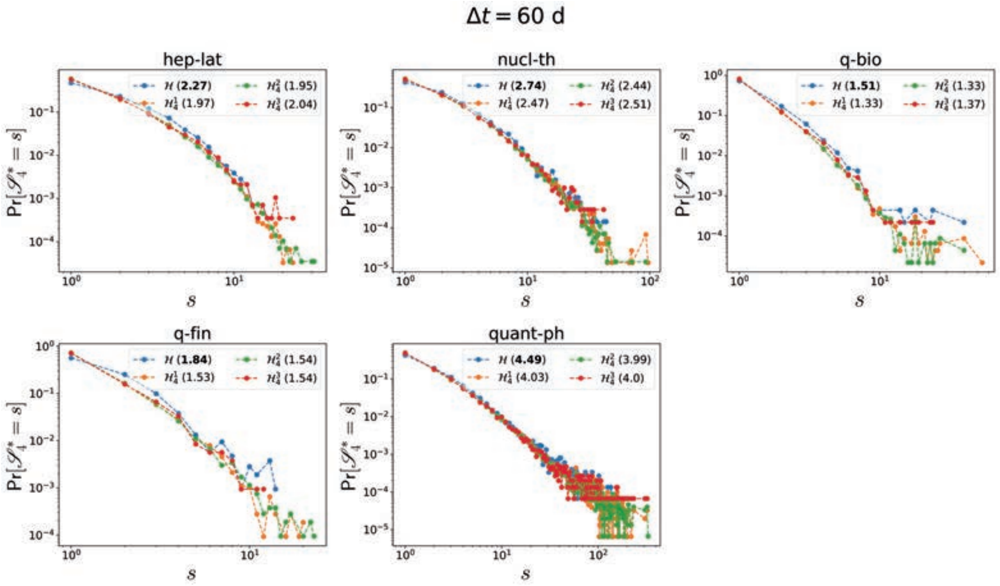

Fig. S3.18: Probability distribution $Pr[\mathcal{S}_4^* = s]$ of the size $\mathcal{S}_4^*$ of trains (obtained from the activity series of ego networks centered at each order 4 hyperlink), where a center link is activated at least once, in each collaboration network $\mathcal{H}$ (blue) and its three randomized reference models $\mathcal{H}_4^1$ (yellow), $\mathcal{H}_4^2$ (green) and $\mathcal{H}_4^3$ (red). To identify the trains, we consider $\Delta t = 60d$. For each network, the average size of the trains is reported. The maximum average size among network $\mathcal{H}$, $\mathcal{H}_4^1$, $\mathcal{H}_4^2$ and $\mathcal{H}_4^3$ is in bold. The horizontal and vertical axes are presented in logarithmic scale.

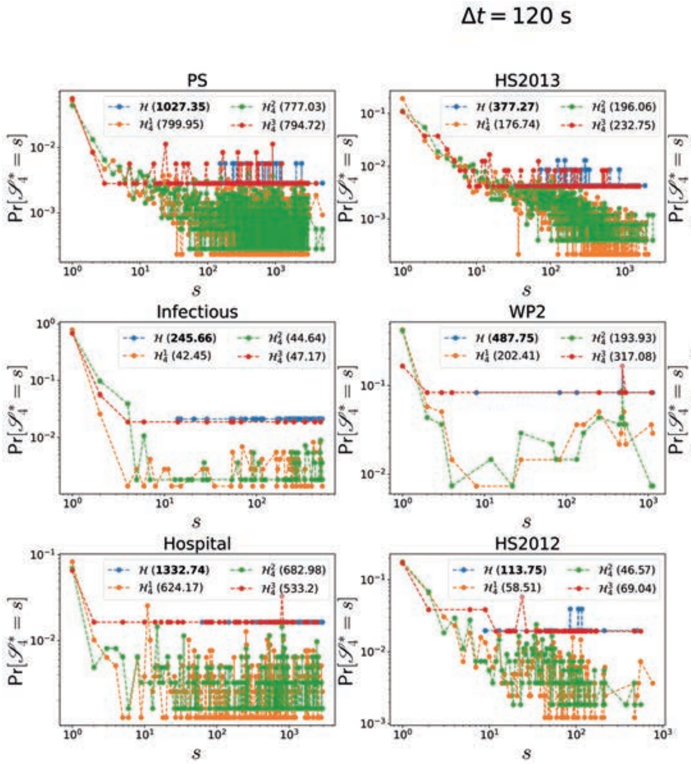

Fig. S3.19: Probability distribution $Pr[\mathscr{S}_4^* = s]$ of the size $\mathscr{S}_4^*$ of trains (obtained from the activity series of ego networks centered at each order 4 hyperlink), where a center link is activated at least once, in each physical contact network $\mathscr{H}$ (blue) and its three randomized reference models $\mathscr{H}_4^1$ (yellow), $\mathscr{H}_4^2$ (green) and $\mathscr{H}_4^3$ (red). To identify the trains, we consider $\Delta t = 120s$. For each network, the average size of the trains is reported. The maximum average size among network $\mathscr{H}$, $\mathscr{H}_4^1$, $\mathscr{H}_4^2$ and $\mathscr{H}_4^3$ is in bold. The horizontal and vertical axes are presented in logarithmic scale.

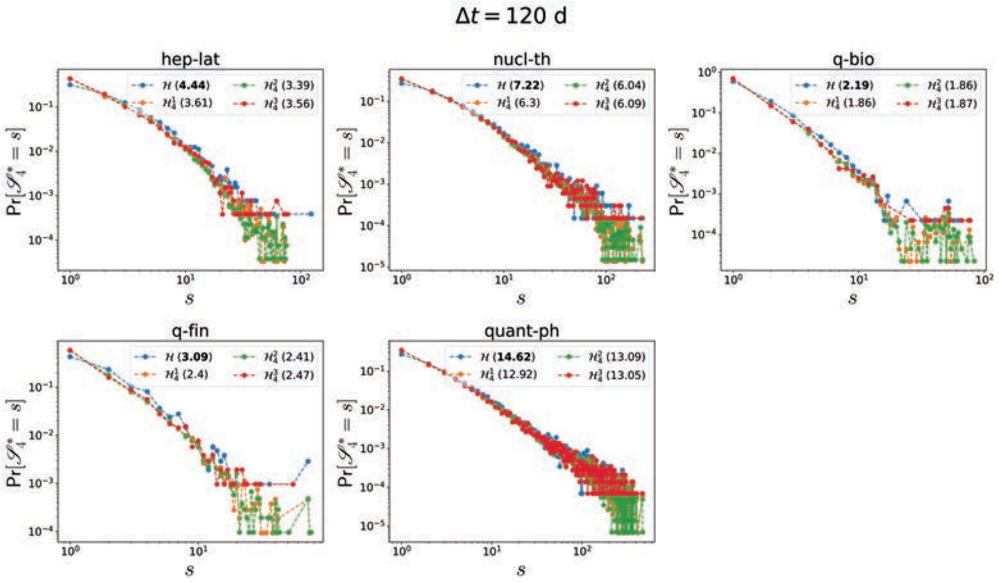

Fig. S3.20: Probability distribution $Pr[\mathscr{S}_4^* = s]$ of the size $\mathscr{S}_4^*$ of trains (obtained from the activity series of ego networks centered at each order 4 hyperlink), where a center link is activated at least once, in each collaboration network $\mathscr{H}$ (blue) and its three randomized reference models $\mathscr{H}_4^1$ (yellow), $\mathscr{H}_4^2$ (green) and $\mathscr{H}_4^3$ (red). To identify the trains, we consider $\Delta t = 120d$. For each network, the average size of the trains is reported. The maximum average size among network $\mathscr{H}$, $\mathscr{H}_4^1$, $\mathscr{H}_4^2$ and $\mathscr{H}_4^3$ is in bold. The horizontal and vertical axes are presented in logarithmic scale.

3.6.3. INCOMPLETE HIGHER-ORDER EVENTS

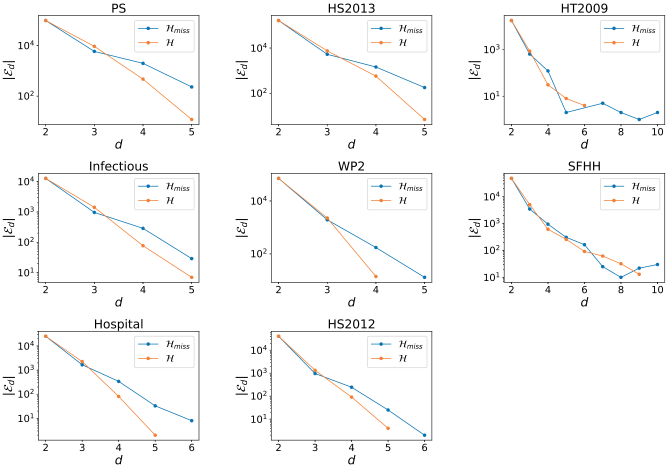

Fig. S3.21: Total number of events ($|\mathscr{E}_d|$) in original network $\mathscr{H}$ and $\mathscr{H}_{miss}$for each order d in physical contact networks. Vertical axis is presented in logarithmic scale.

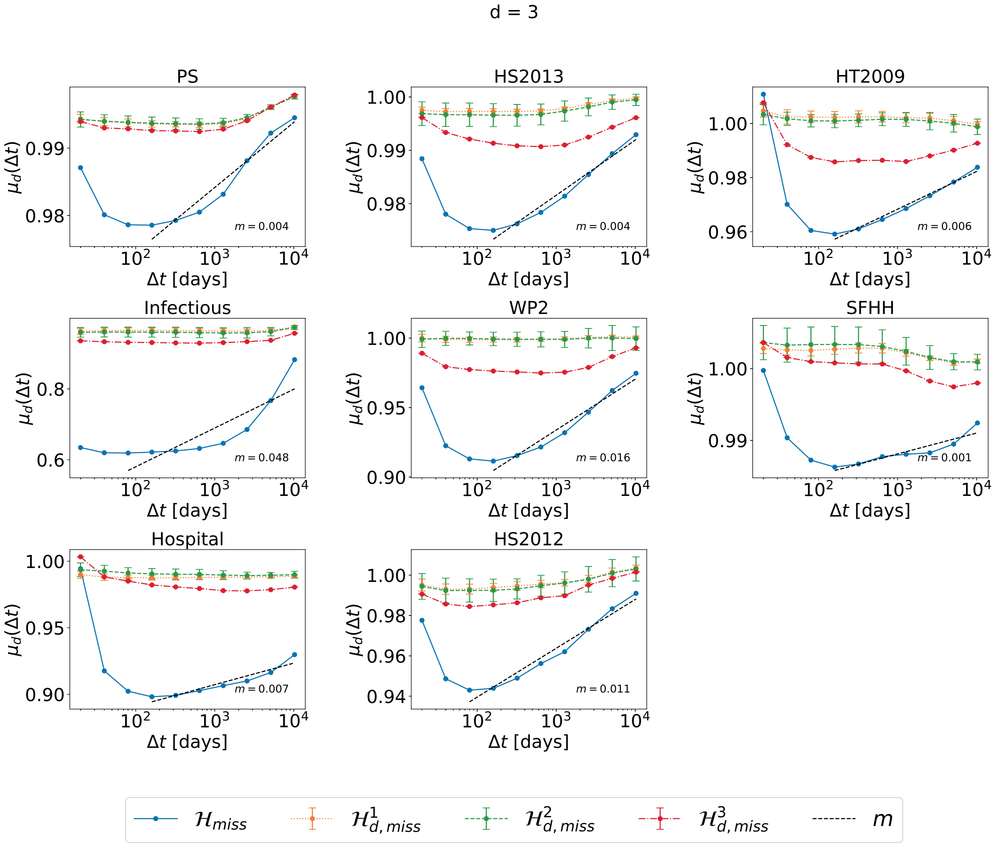

Fig. S3.22: The normalized average topological distance $\mu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\setminus\mathcal{E}_d]}{E[\eta(e,e')|\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\setminus\mathcal{E}_d]}$, between an order $d = 3$ event and an event of a different order, in each physical contact network$\mathscr{H}_{miss}$ and its corresponding three randomized null models $\mathscr{H}^1_{d,miss}$ (yellow), $\mathscr{H}^2_{d,miss}$ (green) and $\mathscr{H}^3_{d,miss}$ (red), which preserve or destroy specific properties of order $d = 3$ events. $\lim_{\Delta t\to\infty} E[\eta(e,e')|\mathcal{T}(e,e') < \Delta t,\ e \in \mathcal{E}_d,\ e' \in \mathcal{E} \setminus \mathcal{E}_d] = E[\eta(e,e')|\ e \in \mathcal{E}_d,\ e' \in \mathcal{E} \setminus \mathcal{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure corresponds to the linear fit (with slope m) of $\mu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathscr{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

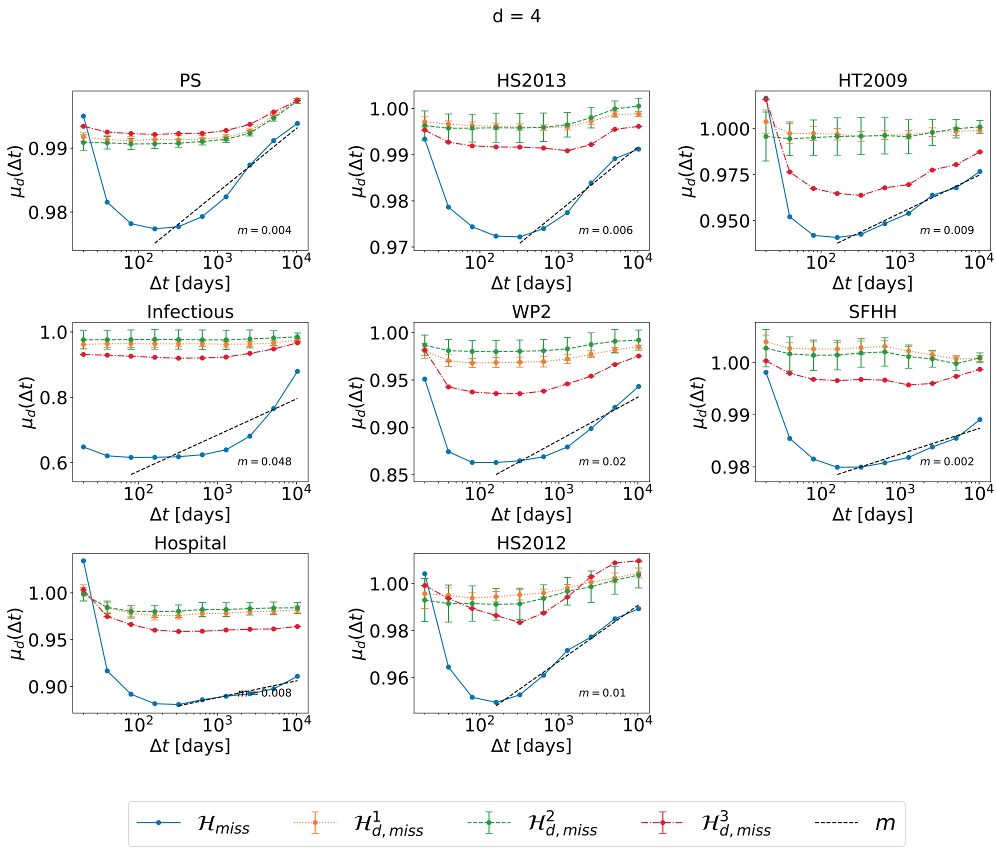

Fig. S3.23: The normalized average topological distance $\mu_d(\Delta t) = \frac{E[\eta(e,e')|\mathcal{T}(e,e')<\Delta t,\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\setminus\mathcal{E}_d]}{E[\eta(e,e')|\ e\in\mathcal{E}_d,\ e'\in\mathcal{E}\setminus\mathcal{E}_d]}$, between an order $d = 4$ event and an event of a different order, in each physical contact network$\mathcal{H}_{miss}$ and its corresponding three randomized null models $\mathcal{H}^1_{d,miss}$ (yellow), $\mathcal{H}^2_{d,miss}$ (green) and $\mathcal{H}^3_{d,miss}$ (red), which preserve or destroy specific properties of order $d = 3$ events. $\lim_{\Delta t\to\infty} E[\eta(e,e')|\mathcal{T}(e,e') < \Delta t,\ e \in \mathcal{E}_d,\ e' \in \mathcal{E}\setminus\mathcal{E}_d] = E[\eta(e,e')|\ e \in \mathcal{E}_d,\ e' \in \mathcal{E}\setminus\mathcal{E}_d]$ for any d. The horizontal axes are presented in logarithmic scale. The dashed line in each figure correspond to the linear fit (with slope m) of $\mu_d(\Delta t)$ as a function of $log_{10}(\Delta t)$ in $\mathcal{H}$, for the part that the curve has an increasing trend. For each dataset, the results of the three corresponding randomized models are obtained from 10 independent realizations.

BIBLIOGRAPHY

[1] Petter Holme and Jari Saramäki. "Temporal networks". In: *Physics Reports* 519.3 (2012), pp. 97–125.

[2] Petter Holme. "Modern temporal network theory: a colloquium". In: *The European Physical Journal B* 88.9 (2015), p. 234.

[3] Naoki Masuda and Renaud Lambiotte. *A guide to temporal networks*. World Scientific, 2016.

[4] Márton Karsai, Hang-Hyun Jo, Kimmo Kaski, et al. *Bursty human dynamics*. Springer, 2018.

[5] Márton Karsai et al. "Universal features of correlated bursty behaviour". In: *Scientific Reports* 2 (2012), p. 397.

[6] K-I Goh and A-L Barabási. "Burstiness and memory in complex systems". In: *EPL (Europhysics Letters)* 81.4 (2008), p. 48002.

[7] Joao Gama Oliveira and Albert-László Barabási. "Darwin and Einstein correspondence patterns". In: *Nature* 437.7063 (2005), pp. 1251–1251.

[8] Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi. "Entropy of dialogues creates coherent structures in e-mail traffic". In: *Proceedings of the National Academy of Sciences* 101.40 (2004), pp. 14333–14337.

[9] Xiu-Xiu Zhan, Alan Hanjalic, and Huijuan Wang. "Information diffusion backbones in temporal networks". In: *Scientific Reports* 9.1 (2019), pp. 1–12.

[10] Xiu-Xiu Zhan, Alan Hanjalic, and Huijuan Wang. "Suppressing Information Diffusion via Link Blocking in Temporal Networks". In: *International Conference on Complex Networks and Their Applications*. Springer. 2019, pp. 448–458.

[11] Giovanna Miritello, Esteban Moro, and Rubén Lara. "Dynamical strength of social ties in information spreading". In: *Physical Review E* 83.4 (2011), p. 045102.

[12] Dávid X Horváth and János Kertész. "Spreading dynamics on networks: the role of burstiness, topology and non-stationarity". In: *New Journal of Physics* 16.7 (2014), p. 073037.

[13] Ville-Pekka Backlund, Jari Saramäki, and Raj Kumar Pan. "Effects of temporal correlations on cascades: Threshold models on temporal networks". In: *Physical Review E* 89.6 (2014), p. 062815.

[14] Oliver E Williams, Fabrizio Lillo, and Vito Latora. "How auto-and cross-correlations in link dynamics influence diffusion in non-Markovian temporal networks". In: *arXiv preprint arXiv:1909.08134* (2019).

[15] Márton Karsai et al. "Small but slow world: How network topology and burstiness slow down spreading". In: *Physical Review E* 83.2 (2011), p. 025102.

[16] Jean-Charles Delvenne, Renaud Lambiotte, and Luis EC Rocha. “Diffusion on networked systems is a question of time or structure”. In: *Nature Communications* 6.1 (2015), pp. 1–10.

[17] Samuel Unicomb et al. “Dynamics of cascades on burstiness-controlled temporal networks”. In: *Nature communications* 12.1 (2021), pp. 1–10.

[18] Federico Battiston et al. “Networks beyond pairwise interactions: structure and dynamics”. In: *Physics Reports* 874 (2020), pp. 1–92.

[19] Federico Battiston et al. “The physics of higher-order interactions in complex systems”. In: *Nature Physics* 17.10 (2021), pp. 1093–1098.

[20] Giovanni Petri et al. “Homological scaffolds of brain functional networks”. In: *Journal of The Royal Society Interface* 11.101 (2014), p. 20140873.

[21] Vedran Sekara, Arkadiusz Stopczynski, and Sune Lehmann. “Fundamental structures of dynamic social networks”. In: *Proceedings of the national academy of sciences* 113.36 (2016), pp. 9977–9982.

[22] Alice Patania, Giovanni Petri, and Francesco Vaccarino. “The shape of collaborations”. In: *EPJ Data Science* 6 (2017), pp. 1–16.

[23] Austin R Benson et al. “Simplicial closure and higher-order link prediction”. In: *Proceedings of the National Academy of Sciences* 115.48 (2018), E11221–E11230.

[24] Giulia Cencetti et al. “Temporal properties of higher-order interactions in social networks”. In: *Scientific reports* 11.1 (2021), pp. 1–10.

[25] Alberto Ceria et al. “Topological–temporal properties of evolving networks”. In: *Journal of Complex Networks* 10.5 (2022), cnac041.

[26] Laetitia Gauvin et al. “Randomized reference models for temporal networks”. In: *arXiv preprint arXiv:1806.04032* (2018).

[27] Kazuki Nakajima, Kazuyuki Shudo, and Naoki Masuda. “Randomizing hypergraphs preserving degree correlation and local clustering”. In: *arXiv preprint arXiv:2106.12162* (2021).

[28] Lorenzo Isella et al. “What’s in a crowd? Analysis of face-to-face behavioral networks”. In: *Journal of theoretical biology* 271.1 (2011), pp. 166–180.

[29] Mathieu Génois and Alain Barrat. “Can co-location be used as a proxy for face-to-face contacts?” In: *EPJ Data Science* 7.1 (2018), pp. 1–18.

[30] Philippe Vanhems et al. “Estimating potential infection transmission routes in hospital wards using wearable proximity sensors”. In: *PloS one* 8.9 (2013), e73970.

[31] Valerio Gemmetto, Alain Barrat, and Ciro Cattuto. “Mitigation of infectious disease at school: targeted class closure vs school closure”. In: *BMC infectious diseases* 14.1 (2014), pp. 1–10.

[32] Juliette Stehlé et al. “High-resolution measurements of face-to-face contact patterns in a primary school”. In: *PloS one* 6.8 (2011), e23176.

[33] Julie Fournet and Alain Barrat. “Contact patterns among high school students”. In: *PloS one* 9.9 (2014), e107878.

[34] 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”. In: *PloS one* 10.9 (2015), e0136497.

[35] Mark EJ Newman. “The structure of scientific collaboration networks”. In: *Proceedings of the national academy of sciences* 98.2 (2001), pp. 404–409.

4

Modeling Airport Congestion Contagion by Heterogeneous SIS Epidemic Spreading on Airline Networks

This chapter has been published as: Ceria, A., Köstler, K., Gobardhan, R., & Wang, H. (2021). Modeling airport congestion contagion by heterogeneous SIS epidemic spreading on airline networks. Plos One, 16(1), e0245043

In this chapter, we explore the possibility of using a heterogeneous Susceptible- Infected-Susceptible SIS spreading process on an airline network to model airport congestion contagion with the objective to reproduce airport vulnerability. We derive the vulnerability of each airport from the US Airport Network data as the congestion probability of each airport. In order to capture diverse flight features between airports, e.g. frequency and duration, we construct three types of airline networks. The infection rate of each link in the SIS spreading process is proportional to its corresponding weight in the underlying airline network constructed. The recovery rate of each node is also heterogeneous, dependent on its node strength in the underlying airline network, which is the total weight of the links incident to the node. Such heterogeneous recovery rate is motivated by the fact that large airports may recover fast from congestion due to their well-equipped infrastructures. The nodal infection probability in the meta-stable state is used as a prediction of the vulnerability of the corresponding airport. We illustrate that our model could reproduce the distribution of nodal vulnerability and rank the airports in vulnerability evidently better than the SIS model whose recovery rate is homogeneous. The vulnerability is the largest at airports whose strength in the airline network is neither too large nor too small. This phenomenon can be captured by our heterogeneous model, but not the homogeneous model where a node with a larger strength has a higher infection probability. This explains partially the out-performance of the heterogeneous model. This proposed congestion contagion model may shed lights on the development of strategies to identify vulnerable airports and to mitigate global congestion by e.g. congestion reduction at selected airports.

4.1. Introduction

Networks, ranging from social, transportation to physical contact networks, support the diffusion of information, transportation of goods and spreading of epidemics. Therefore, networks and processes that unfold on them have been investigated in a wide range of fields such as mathematics, engineering and social sciences [1–5]. The Susceptible-Infected-Susceptible (SIS) epidemic spreading process is one of the most studied dynamic processes on networks [6–14]. The classic homogeneous SIS spreading process has been defined as follows. At any time t, a node is either susceptible S or infected I. A susceptible node can be infected by each of its infected neighbors with an infection rate β. Each infected node recovers to be susceptible again with a recovery rate δ. Both the infection and recovery processes are independent Poisson processes. For a given network upon which the SIS process is deployed, a critical epidemic threshold τ_c exists. When the effective spreading rate $\tau = (\beta/\delta) > \tau_c$, a non-zero fraction of infected nodes persists in the meta-stable state. When $\tau < \tau_c$, the epidemic dies out. The vulnerability of a network to an epidemic is estimated by the prevalence, defined as the average fraction of infected nodes in the meta-stable state. The infection probability $v_{i\infty}$ of a node i indicates its vulnerability to the epidemic. Recent studies have focused on the influence of the underlying network topology and heterogeneous infection/recovery rates on the epidemic threshold, the prevalence [15, 16] and nodal infection probabilities [17]. Epidemic spreading processes have been developed to model e.g. the propagation of epidemic, information, failures and computer worms.

A fundamental question is to what extent an abstract process like the epidemic spreading process could model a generic complex system, i.e. reproduce the key properties of the system. This question is motivated at least from the following perspective. The operating mechanisms of many complex systems like social systems and the brain are far from well understood. A model that could well reproduce the key properties of a complex system may unravel the possible operating mechanism. The operating mechanisms of many complex systems are possibly known, however, too complex to derive optimization/control solutions. In this case, an abstract model that well captures the key features of the system may possibly facilitate the development of optimization solutions.

For airline transportation networks, initial effort has been devoted to the analysis of their topologies, demonstrating properties such as the small-world and scale-free degree distribution [18, 19]. Topological properties of subsets of a network based on geography and airlines/alliances have also been explored [20, 21]. Recent investigations have focused on network resilience and vulnerability regarding random failures [22, 23]. The performance or state of an airport (e.g. congested or not and the average delay per hour) is not independent of the states of other airports. The delay propagation between airports has been studied via e.g. the correlation or causality measures between the time series (average delay per hour) of airports [24–28]. One of the main reasons why delay propagates is that each aircraft has a flight sequence where it travels between possibly multiple airports a day. The congestion at an airport can be introduced by local factors such as the slow boarding of passengers, the mechanical issues of an aircraft at the airport. Beyond, delayed flights that depart from a congested airport could cause an overcharge at the arrival airports. The Air-Traffic Flow Management systems use strategies such as ground holding (intentionally delaying an aircraft's takeoff) and re-routing to reduce overload [29].The weather condition could lead to the congestion of several nearby airports, which may further cascade to more airports due the rescheduling or re-routing of aircraft. These perspectives imply the possible contagion of congestion between airports. Airline congestion has been studied via network dynamics like queuing models [30]. Epidemic spreading process has been recently used to model the spreading of traffic jams in urban networks, assuming both homogeneous infection and recovery rate and homogeneous mixing approximation in network topology [31]. The possibility of modeling congestion contagion on an airline network using epidemic spreading process has been barely explored, not to mention how to develop a full-fledged heterogeneous spreading model.

In this chapter, we explore the possibility and limits of modeling airport congestion contagion by a heterogeneous SIS spreading process on an airline network in reproducing or predicting airport vulnerability. We consider the US Airport Network data [32]. The airport vulnerability is defined as the ratio of the duration of traffic congestion over the total operation time and derived from data. We construct three types of airport networks to capture diverse features such as the frequency and duration of flights. In the heterogeneous SIS model that we proposed, the infection rate of a link is proportional to the weight of the link, as defined in each of the three airline networks. Moreover, the recovery rate of a node is also heterogeneous, dependent on the strength of the node in the underlying network. We use the nodal infection probability in the meta-stable state as an estimation of the corresponding airport's vulnerability, which will be further compared with the airport vulnerability derived from the US Airport dataset to evaluate our model. Specifically, our model is evaluated according to its capability to reproduce the distribution of the vulnerability of a node and the ranking of nodes in vulnerability. The modeling of airport congestion contagion by the SIS process, where the infection rate of a link is proportional to the weight of the link and the recovery rate is homogeneous, has been explored in [33]. That SIS process is a special case of our heterogeneous model and is called the homogeneous SIS model in this chapter to emphasize its homogeneous recovery rate. We illustrate that the heterogeneous SIS model evidently outperforms the homogeneous model according both aforementioned evaluation perspectives. Our further exploration of the infection probability in relation to the node strength of an airport explains the better performance of the heterogeneous model in reproducing the ranking of nodes in vulnerabilities.

We propose and illustrate the basic method to model a complex system by an epidemic spreading process, via the airline system. The relatively good performance of the model does not imply that the derived model is the precise mechanism of congestion contagion. Further verification of the contagion mechanism is needed, e.g. regarding whether nodes with a large strength recover faster. The derived model may inspire the development of strategies to identify vulnerable airports and to mitigate global congestion by e.g. reducing congestion at selected airports.

The content of this chapter is arranged as follows. Firstly we define, derive and characterize the airport vulnerability derived from data. Furthermore we introduce the heterogeneous SIS spreading model and network construction. Afterwards, the methods to evaluate the model are presented. In results, we compare the performance of our model with the homogeneous model. The final section summarizes our key findings and discusses possible future work.

4.2. Materials and methods

4.2.1. Traffic vulnerability of an airport

Firstly, we describe the US Airport Network data. Airport vulnerability and its distribution are further defined and derived respectively. Airport vulnerability obtained from data will be adopted as a benchmark to evaluate the performance of our model.

Data

We obtain the U.S. airport dataset from the Bureau of Transportation Statistics (BTS). This data set includes detailed information about the U.S. flight schedules since 1987 [32]. The computer reservation system (CRS) further distinguishes flight schedules as the planned schedule under optimal operation conditions, and the actual schedule. In order to demonstrate our modeling approach, we use the data spanning the high season period from July 1st 2018 to July 14th 2018, since flight schedule and rotations periodically repeat. In total $N = 349$ airports and $E = 645299$ flights have been considered. This data set contains as well extra information for each flight e.g. Tail-number, Origin and Destination, Date, the actual and scheduled Departure/Arrival Times.

Definition and statistical properties

The vulnerability of an airport is defined as its duration of traffic congestion over its total operation time, which is its probability of being congested. Per hour, an airport's declared capacity corresponds approximately to the number of movements (the total number of departure and arrival flights) planned for that hour, such that a reasonable level of service (LOS) can be ensured. Delay is the principal indicator of LOS. Usually the declared capacity of an airport is up to $85-95\%$ of its maximum throughput capacity, which is the maximal number of movements per hour that the airport's runway system allows according to air traffic management rules and assuming continuous aircraft demands. An airport is considered congested if its actual number of movements per hour during operation is greater than its declared capacity (the planned number of movements) divided by a parameter α, where $0.85 \le \alpha \le 1$. We consider $\alpha = 0.9$ as an example to illustrate our methods. The state of each airport i at each hour t is derived from U.S. airport dataset as follows: the airport is congested ($X_i(t) = 1$) if the actual number of movements is larger than the number of movement planned at time t divided by 0.9. If this condition is not satisfied, the airport is not congested ($X_i(t) = 0$). Airport i's vulnerability $\phi_i = \frac{1}{m}\sum_{t=1}^{m} X_i(t)$ is the fraction of time that airport i is congested. We considered all hours in the previously specified two week's interval (excluding hours between 0 and 6 of each day due to their low number of movements). The hours considered are indexed as $[1, 2, ..., m]$, where $m = 18 \cdot 14 = 252$.

In this chapter, we confine ourselves to this limited definition of airport vulnerability to start and to illustrate our method. The definition could be further generalized to capture the level of congestion per hour. The declared capacity can also also be better estimated based on airport characteristics (e.g. active runways, taxiways, etc.) and weather conditions, beyond flight schedule.

Figure 4.1 shows the distribution of airport vulnerability, whose average is 0.15 and variance is 0.01. The vulnerabilities of all the airports are within the range $[0, 0.4]$.

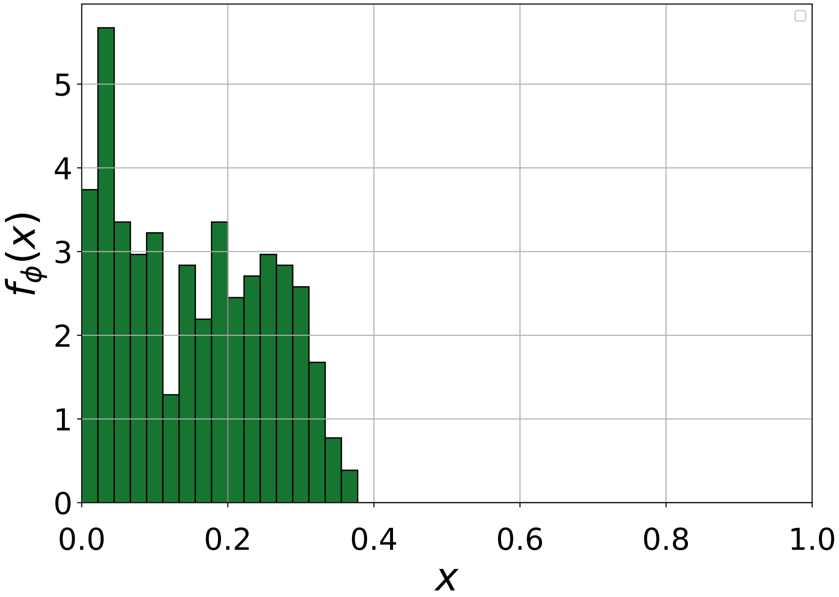

Fig. 4.1: **Probability density function $f_\phi(x)$ of the vulnerability ϕ of an airport**. The average vulnerability is $E[\phi] = 0.15$ and the variance is $Var[\phi] = 0.01$. In total 45 bins are split within the interval $[0, 1]$ with the same bin size. The probability density $f_\phi(x)$ at a given bin x equals the percentage of the airports whose vulnerability falls within the bin normalized by the bin size $1/45$.

4.2.2. HETEROGENEOUS SIS SPREADING MODEL ON AIRLINE NETWORKS

We model the contagion of airport congestion as a heterogeneous SIS spreading process on an airline network. Firstly we introduce how to construct the three types of airline networks. Secondly, we propose the heterogeneous SIS spreading model. The last subsection illustrates the individual-based mean-field approximation to compute nodal infection probabilities in the meta-stable state, given the underlying network and the model parameters.

NETWORK CONSTRUCTION AND PROPERTIES

We derive three types of undirected networks from the U.S. Airport Network data over the two weeks' period in order to capture various flight properties. This is motivated by the fact that the SIS spreading process unfolds differently on different underlying networks. Network G_1 is unweighted: two nodes (airports) are connected if at least one direct flight exists in between. Each existing link has a weight $w_{ij} = 1$. Network G_2 and G_3 are both weighted and have the same network topology as network G_1. It is assumed that the infection rate along a link is proportional to the link's weight. In G_2, the link which connects node i and j has weight $w^*_{ij} = F_{ij} + F_{ji}$, which is the sum of the total number F_{ij} of flights from i to j and the number F_{ji} of flights from j to i in the two weeks' period. We motivate this weight definition by the assumption that frequent flights between two airports correspond to a high chance that congestion spreads from one airport to the other. Furthermore, congestion propagation may be affected also by the duration of flights between airports. An airplane that has departed with a delay in time, in fact, can adapt its speed to respect its scheduled arrival time at the destination airport. In order to capture these effects we introduce Network G_3. This network is defined by assigning to each link (i, j) the weight $w^*_{ij} = \frac{1}{E[T_{ij}]}$, which is the inverse of the average flight time between airport i and j. We adopt the convention that the flight time between airports not connected by any direct flights is infinite: this ensures that the weight of non-existing links is always null. A smaller average flight time may result in a higher chance that flights delayed at the departure airport would affect the arrival time at the destination airport. This situation may be less likely in the case of a larger average flight time, when there is more room for airplanes to re-optimize the flight velocity.

Finally, the weights in Networks G_2 and G_3 are respectively normalized as

$$w_{ij} = \left(\frac{w^*_{ij}}{\max\limits_{k,l} w^*_{k,l}} \right).$$

The normalization by the maximum link weight $\max\limits_{k,l} w^*_{k,l}$ in each network leads to the normalized link weights within the range $(0, 1]$. Since there is no self-loop, $w_{ii} = 0 \; \forall i$.

Heterogeneous infection rate and recovery rate (link weight) have been shown to influence the nodal infection probabilities [11, 17]. Since the infection rate of a link and the recovery rate will later be defined as a function of the link weight and node strength of a node respectively, we examine the distribution of the link weight and node strength (the total weight of the links incident to a node) in 4.2. Network G_2 and G_3 manifest different link weight and node strength distributions, which motivate again the consideration of the three types of networks that capture different features of the airline system.

We explore relation between the strength of a node and other centrality metrics that describe varies topological properties of a node via the linear correlation coefficient. The following centrality metrics have been considered:

- *Clustering Coefficient.* In an unweighted network, the clustering coefficient is the probability that two random neighbors of a node are connected. In a weighted network, a generalized definition for clustering coefficient has been introduced by [34]. The intensity of a triangle among node i, j and k is defined as $\sqrt[3]{w_{ij} w_{jk} w_{ki}}$. The clustering

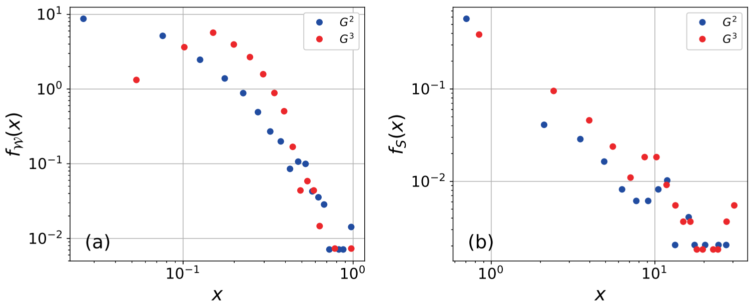

Fig. 4.2: **The probability density functions $f_{\mathscr{W}}(x)$ of the weight $\mathscr{W}$ of a link (a) and $f_S(x)$ of the strength S of a node (b) in network G_2 (blue points) and G_3 (red points).** Horizontal and vertical axes are presented in logarithmic scale. The horizontal axis is split into 20 bins, each with the same bin size in the linear scale. The probability density $f_{\mathscr{W}}(x)$ ($f_S(x)$) at a given bin x is equal to the fraction of the links (nodes) whose weight (strength) falls within the bin normalized by the bin size. Both link weight and node strength have, respectively, a higher average in G_3 and higher coefficient of variation (the ratio of standard deviation over the average) in G_2.

coefficient of a node i is then defined as the sum of the intensities of the triangles that i resides in, normalized by the maximum possible number of triangles that i could reside in, i.e. $\frac{1}{2}d_i(d_i - 1)$, where d_i is the degree of node i.

- *Betweenness Centality.* The betweenness centrality of a node is the fraction of the shortest paths between all possible node pairs that pass through the node. To compute the shortest path between a node pair, we define the distance of each link in the underlying network as the reciprocal of its link weight [35].

- *Closeness Centrality.* The closeness centrality is the average hopcount of a node to any other node. The hopcount between two nodes is the number of links of the shortest path, which is computed as described in betweenness.

- *Principal Eigenvector Component* The principal eigenvector component of a node is its corresponding component in the principal eigenvector of the weighted adjacency matrix. The principal eigenvector is the one corresponding to the largest eigenvalue.

The linear correlation coefficient between node strength and each of centrality metric in the three networks constructed are shown in 4.1

Network	Clustering	Betweenness	Closeness	Eigenvector
G_1	-0.09	0.80	0.81	0.95
G_2	0.00	0.81	0.54	0.98
G_3	-0.17	0.81	0.44	0.92

Table 4.1: The linear correlation coefficient of node strength with clustering coefficient, betweenness, closeness and eigenvector centrality respectively in network G_1, G_2 and G_3.

Node strength is strongly correlated with all the centrality metrics that describe a given importance of node in the whole network except for the clustering coefficient, a nodal prop-

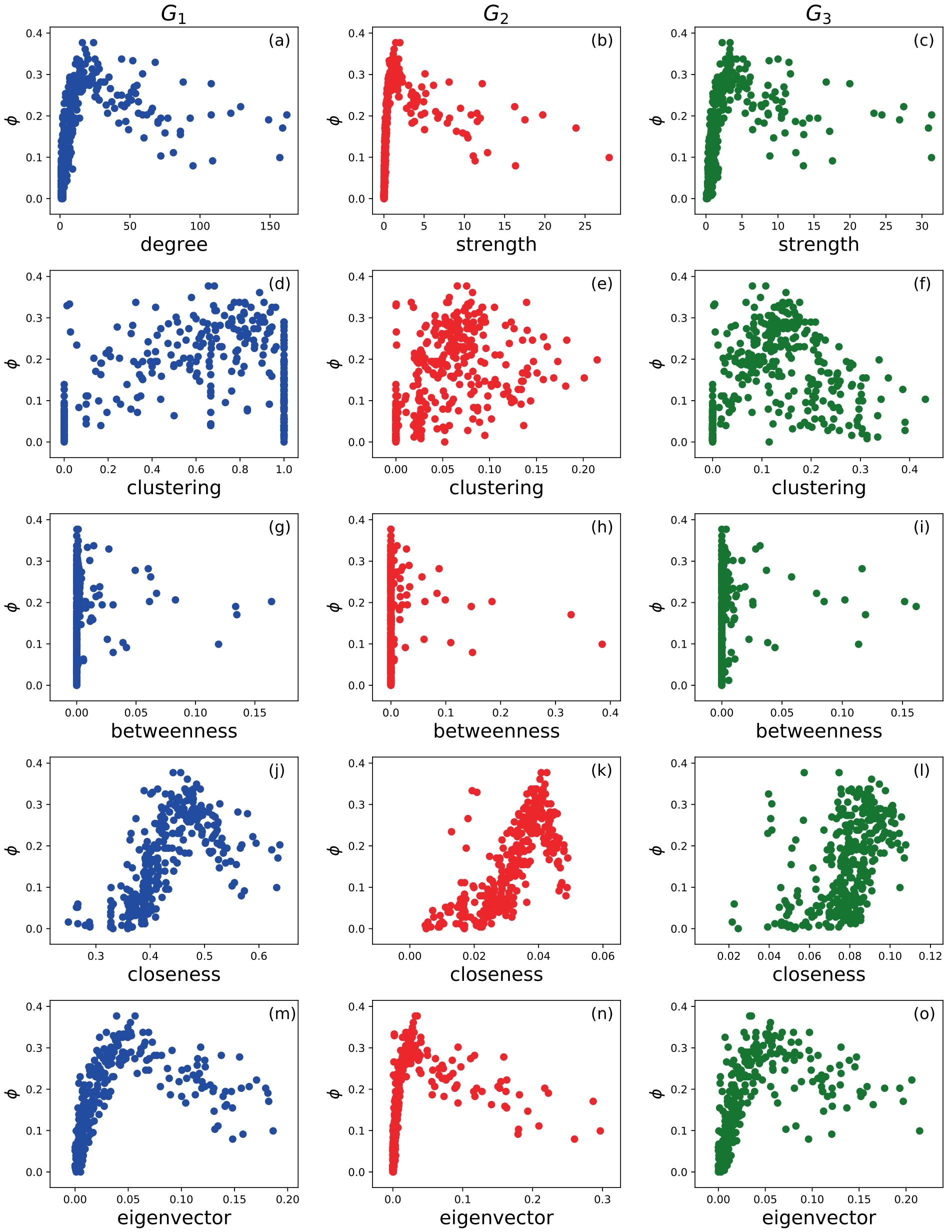

Fig. 4.3: **Airport vulnerability versus nodal centrality measures**. The scatter plot of airport vulnerability ϕ versus node strength (a,b,c), clustering coefficient (d,e,f), betweenness (g,h,i), closeness (j,k,l) and eigenvector (m,n,o) centrality in network G_1 (first column, blue color), G_2 (second column, red color) and G_3 respectively.

erty derived from local network connections. Hence, node strength that will be used to define the nodal recovery rate in the epidemic spreading model, captures as well nodal properties like betweenness, closeness and principal eigenvector component.

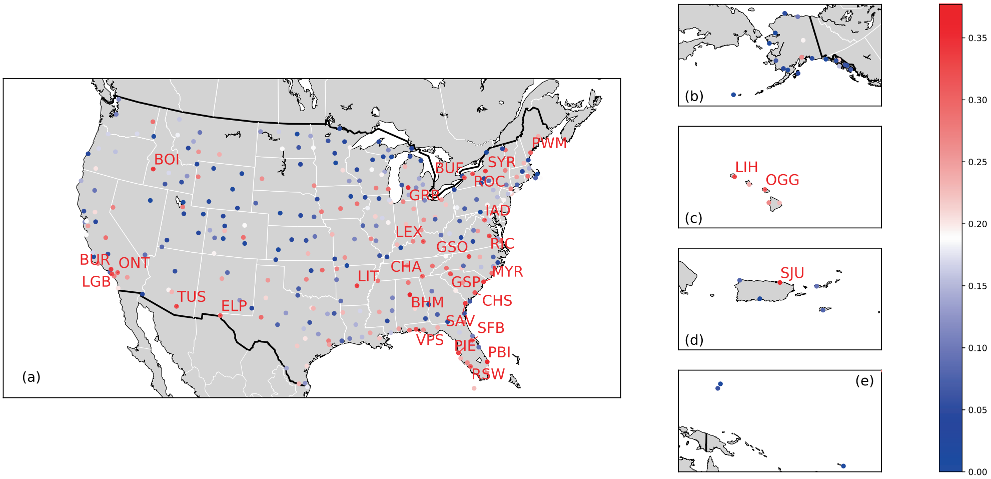

Fig. 4.4: **Geographic location and vulnerability of U.S. airports.** The geographic location and vulnerability of an airport in U.S. mainland (a), Alaska (b), Hawaii Islands (c), Puerto Rico (d), American Samoa and Guam (e). The nodes/airports are color-coded according to their airport vulnerability ϕ. We show the names of the top 30 most vulnerable airports.

Furthermore, we study the relation between the vulnerability ϕ of an airport and a given centrality metric of the corresponding node in each of the three underlying networks. This helps us to evaluate the possibility of using a nodal centrality measure to estimate nodal vulnerability. In the scatter plot in 4.3, we do not observe any monotonic trend between the vulnerability ϕ of an airport and the centrality metric of the corresponding node. This implies that centrality metrics can not be used as a good estimation of airport vulnerability. Our previous work [33] illustrated as well the worse performance of vulnerability prediction via centrality metrics than that via the homogeneous SIS model. Hence, we will compare performance of the heterogeneous SIS model with that of the homogeneous SIS model but not of the centrality metrics.

The networks we constructed have not taken the geographical locations of the airports explicitly into account. One may wonder whether the vulnerability of an airport may strongly correlate with its location, thus can be possibly estimated by its location. 4.4 shows that vulnerable airports are scattered in location and no evident relation between vulnerability and location.

THE HETEROGENEOUS SIS MODEL

We model the airport congestion dynamics as a heterogeneous SIS spreading process, where both the infection rate per link and the recovery rate per node are heterogeneous. The infection rate of a link with weight w_{ij} is $\beta_{ij} = \beta w_{ij}$. In network G_1, which is unweighted, the infection rate is homogeneous. The heterogeneous recovery rate is motivated by the fact that airports with a larger declared capacity may recovery faster i.e. are more capable to deal with operational delay and congestion due to their better infrastructure. The declared capacity of an airport is affected by the number and geometric layout of the runways, type and location of taxiway exits from the runway and the ATM system. The primary factor in determining the capacity is the number of simultaneous active runways. The selection of runways to be operated depends on demand, weather conditions (visibility, wind speed/direction) and noise restrictions. During periods of high congestion, a large airport can decide to keep more runways active to match the demand, however, a small airport does not have that option. Furthermore, a large airport with several runways will have even more runway configurations, which is a combination of simultaneous active runways, weather conditions and assignment of aircraft types and movements (arrival/departures). This makes larger airports more suitable to handle congestion [29]. Similarly, recent studies showed that large airports are less likely to propagate delay [27, 28]. In the three networks we constructed, the node strength tends to be a good proxy of the declared capacity and it is strongly correlated with several other nodal centrality metrics. Hence, we define the recovery rate δ_i of a node as a function of its node strength:

$$\delta_i = \delta\left(c + \left(\frac{s_i}{s_{max}}\right)^{\theta}\right) \tag{4.1}$$

where node i's strength is $s_i = \sum_j w_{ij}$ and $s_{max} = \max_{1 \le i \le N} \{s_i\}$. In the unweighted network topology G_1, the strength s_i of a node i corresponds to its degree. The parameter c is a constant. The scaling factor $\theta \ge 0$ regulates to what extent the recovery rate of a node depends on the normalized node strength $\frac{s_i}{s_{max}}$. A large c results in a more homogeneous recovery rate, whereas a large θ leads to a high heterogeneity in recovery rate. When $\theta > 0$ a node with a higher strength has a larger recovery rate. The heterogeneous SIS model coincides with the homogeneous one when $\theta = 0$. The definition of the heterogeneous recovery rate 4.1 is generic in the sense that it is a polynomial function of the node strength where the extent of homogeneity or heterogeneity can be tuned via parameter c and θ. The parameter set (δ, c, θ) will be calibrated or identified as the set that best reproduced the properties of the vulnerability of airports, as described in subsection Experiment description. The normalization by s_{max} in 4.1 has no influence on the performance of the model but may ease the choice of the search space of c when we calibrate the parameters.

Individual-Based mean-field approximation of the heterogeneous SIS model

We derive nodal infection probabilities via mean-field approximation instead of simulating the SIS stochastic process for computational efficiency. The N-Intertwined Mean-Field Approximation (NIMFA) is one of the most precise individual-based mean-field approximations [9]. Different from homogeneous or degree-based mean-field approximations where only the degree of a node is taken into account, NIMFA preserves the whole network topology in its governing equations, coupling the infection probability of neighboring nodes. It further assumes that the states of neighboring nodes are uncorrelated. Under NIMFA, the governing equation for a node i in our heterogeneous SIS spreading model is

$$\frac{\mathrm{d}v_i(t)}{\mathrm{d}t} = -\delta_i v_i(t) + (1 - v_i(t)) \sum_{j=1}^{N} \beta_{ij} v_j(t) \tag{4.2}$$

where $v_i(t)$ is the infection probability of node i at time t, and $\beta_{ij} = \beta w_{ij}$ is the infection rate associated to the link (*i,j*). In the meta-stable state, $\frac{\mathrm{d}V(t)}{\mathrm{d}t} = 0$, where $V(t) = [v_1(t)\ v_2(t) \cdots v_N(t)]^T$, $\lim_{t\to\infty} v_i(t) = v_{i\infty}$ and $\lim_{t\to\infty} V(t) = V_\infty$. The infection probability of each node V_∞ in the meta-stable state can be derived. The trivial all-zero solution corresponds to the absorbing state where all nodes are susceptible. The non-zero solution of V_∞, if exists, indicates the existence of a meta-stable state with a non-zero fraction of infected nodes. Or else, the meta-stable state is 0 or not-existent. Given θ, c and the underlying network, the infection probability of each node remains the same if $\frac{\beta}{\delta}$ does not change. Without loosing the generality, we consider $\beta = 1$.

In a heterogeneous SIS model, the condition for the epidemic to spread out on a given network G is $\mathrm{Re}(\lambda_1(\bar{A}) > 0$ where $\mathrm{Re}(\lambda_1(\bar{A}))$ is the real part of the largest eigenvalue of the matrix $\bar{A}$, with its elements $\bar{a}_{ij} = \beta_{ij}$ if $i \neq j$ and $\bar{a}_{ii} = -\delta_i$ [36]. In particular, in our model $\beta_{ij} = w_{ij}$, hence $\bar{A} = W - diag(\delta_i)$. δ_i is defined according to 4.1. Furthermore, the three network topologies G_1, G_2 and G_3 are undirected: thus $\bar{A}$ is real and symmetric and $\lambda_1(\bar{A})$ is real. The condition $\mathrm{Re}(\lambda_1(\bar{A})) > 0$ becomes

$$\lambda_1\Big(\frac{1}{\delta}W - diag\Big(\frac{s_i}{s_{max}}\Big)^{\theta}\Big) > c \tag{4.3}$$

4.2.3. EVALUATION METHODS

We evaluate our model via its capacity to capture: (a) the probability distribution of airport vulnerability and (b) the rank of airports in vulnerability.

SIMILARITY OF VULNERABILITY AND INFECTION PROBABILITY DISTRIBUTION

We firstly quantify the similarity of the probability distribution of nodal infection probability obtained from the heterogeneous SIS model with that of airport vulnerability via the Jensen Shannon divergence JSD. Given two discrete probability distributions $P = (p_1, p_2, \ldots, p_K)$ and $Q = (q_1, q_2, \ldots, q_K)$ where $K \geq 2$, the Jensen-Shannon divergence(JSD) [37] measures the similarity of P and Q. We define the mixture of P and Q as $M = (m_1, m_2, \ldots, m_K)$ where $m_i = \frac{p_i + q_i}{2}$, $i \in \{1, 2, \ldots, K\}$. The Shannon's entropy of of a distribution e.g. P is denoted as $H(P) = -\sum_{j=1}^{K} p_j \log_2 p_j$. Jensen Shannon divergence measures the difference between the Shannon entropy of the mixture $M = \frac{1}{2}(P+Q)$ and the average Shannon entropy of P and Q, i.e.

$$JSD(P,Q) = H(M) - \frac{1}{2}(H(P) + H(Q)) \tag{4.4}$$

The Jensen-Shannon divergence is symmetric $0 \leq JSD(P,Q) \leq 1$. A smaller JSD(P,Q) indicates a high similarity between the two distribution P and Q.

AIRPORT RANKING IN VULNERABILITY

From the application perspective, the identification of the most vulnerable airports is crucial. We can evaluate the quality of using nodal infection probability to rank airports in vulnerability as follows. A node with a high infection probability is supposed to correspond to an airport with a high vulnerability. We rank the nodes (airports) according to their infection probability and vulnerability respectively. These two rankings are recorded by two vectors $R^v = [R^v_{(1)}, R^v_{(2)}, \ldots, R^v_{(N)}]$ and $R^\phi = [R^\phi_{(1)}, R^\phi_{(2)}, \ldots, R^\phi_{(N)}]$ where $R^v_{(i)}$ is the index of the $i-th$ highest node in infection probability and $R^\phi_{(i)}$ is the index of the $i-th$ most vulnerable airport. The performance of using nodal infection probability to identify the top f fraction most vulnerable airports can be quantified by the top f recognition rate

$$r_{\phi v}(f) = \frac{|R^\phi_f \cap R^v_f|}{|R^\phi_f|} \tag{4.5}$$

where R_f^ϕ and R_f^V are, respectively, the sets of nodes ranked in the top f fraction according to vulnerability and infection probability. $|R_f^\phi| = fN$ is the number of nodes in R_f^ϕ. A higher recognition rate indicates a higher precision of using nodal infection probability to identify the top f fraction most vulnerable nodes.

We define the overall recognition quality ξ as the area under the $r_{\phi v}(f)$ function:

$$\xi = \int_0^1 r_{\phi v}(f) df \tag{4.6}$$

The recognition quality $0 \le \xi \le 1$ measures the overall performance of using infection probability to rank airports in vulnerability. The quality $\xi = \frac{1}{2}$ is obtained by the random ranking, which selects uniformly at random f fraction of nodes as the top f fraction most vulnerable ones. The maximum ξ corresponds to the case when $r_{\phi v}(f) = 1 \; \forall f$, which means that $R^v = R^\phi$.

4.3. Results and Discussion

4.3.1. Experiment description

Our heterogeneous SIS model has three control parameters δ, c and θ. In order to understand the influence of the parameters on the performance of the model, we consider all possible combinations of the parameters. We consider for c all possible values within [0,2] and with step size 0.02. Similarly, θ can be any value within [0,2] and with step size 0.1. The smaller step size of c is motivated by the high sensitivity of the model's performances (especially the recognition quality ξ) on c. This is because the term $\left(\frac{s_i}{s_{max}}\right)^\theta$ in the recovery rate of a node can be small, when θ is large, especially in view of the heterogeneous node strength distribution (see 4.2). Given the underlying network G_1, G_2 or G_3, and given the parameter c and θ, the prevalence in the meta-stable state that can be derived via NIMFA is an increasing function of $1/\delta$. We consider the optimal value of δ, which is denoted as δ_o, as the one that minimizes $(E[\phi] - \frac{1}{N}\sum_{i=1}^{N} v_i)^2$, i.e. when the average nodal infection probability is the closest to the average airport vulnerability. We obtained it via Brent optimization algorithm [38, 39]. For each possible c, θ and the underlying network G_1, G_2 or G_3, which together determine the δ_o, we derive the infection probability for each node via the NIMFA. The performance of the corresponding model is evaluated in comparison with the airport vulnerabilities via the Jensen-Shannon divergence JSD and the recognition quality ξ. We compare the performance of the heterogeneous SIS model on each network with the corresponding homogeneous model. In the baseline homogeneous SIS model on a given network, the infection rate of a link is $\beta_{ij} = w_{ij}$, while the homogeneous recovery rate $\delta(c+1)$ is tuned effectively as one parameter so that the average infection probability is the closest to the average vulnerability.

4.3.2. Performance of the heterogeneous SIS model

The Jensen Shannon divergence JSD evaluates the similarity between nodal infection and vulnerability distribution, whereas the recognition quality ξ assesses the capability of identifying the most vulnerable airports according to their corresponding infection probabilities. In this section we explore the performance of the heterogeneous SIS model in comparison with the baseline homogeneous SIS model. If we aim to develop a model to reproduce the vulnerability distribution alone (to minimize the JSD) or the ranking of nodal vulnerability (to maximize ξ), but not both at the same time, the heterogeneous SIS model evidently outperforms the homogeneous one. As shown in 4.5, the minimal possible JSD and the maximal ξ achieved by the heterogeneous model are far lower and higher respectively than those obtained by the homogeneous model. The minimal JSD and the maximal ξ are not obtained by the heterogeneous model at the same time, i.e. via the same parameter set θ and c.

Furthermore, the data points on the top-left panel in each sub-figure of 4.5 correspond to the parameter sets with which the heterogeneous model outperforms the homogeneous one in reproducing both the vulnerability distribution and ranking the airports in vulnerability. Among those points, those that lead to an evidently high recognition quality are within the parameter range $\theta > 1$ and $c = 0.02$, when the recovery rate is highly heterogeneous. The heterogeneous SIS model on the unweighted network G_1 could possibly achieve slightly better recognition quality than the model on G_2 and G_3. The homogeneous model on network G_1 however, performs worse than that on G_2 and G_3 in recognition quality. The network G_1, which contains less information than the other two networks, is sufficient for the heterogeneous model to perform well. When $c = 0.02$, the heterogeneous model achieves the best performance in ξ. This suggests that a fine tuning of the c within the range $(0, 0.02)$ may further improve the performance of the model. The parameter sets that we have considered are sufficient for us to illustrate that the heterogeneous SIS model could perform better than the homogeneous one.

4.3.3. THE INFECTION PROBABILITY VERSUS THE NODE STRENGTH OF A NODE

Identifying the most vulnerable airports is crucial for operations. In this section, we aim to understand why the heterogeneous SIS model better recognizes vulnerable airports, i.e. is higher in recognition quality than the homogeneous model. In the homogeneous SIS model, a node with a large strength tends to have a high infection probability. In the heterogeneous SIS model, a node with a large strength has high rates of getting infected by its neighbors, contributing to a high infection probability. On the other hand, a node with a large strength could have a large recovery rate when $\theta > 0$. These two factors imply that a node with a large node strength does not necessarily have a high infection probability. In this section, we explore whether the better performance of our heterogeneous SIS model in recognition quality corresponds to its better capability to reproducing the relationship between the vulnerability and strength of a node if compared to the homogeneous SIS model.

We take network G_1 as an example. The heterogeneous SIS model on G_1 achieves the highest recognition quality ξ when $c = 0.02$ and $\theta = 1.5$. We consider the SIS model when $c = 0.02$ whereas θ varies and when $\theta = 1.5$ whereas c varies. We plot the vulnerability ϕ and the meta-stable infection probability v (derived by the heterogeneous SIS model or the homogeneous SIS baseline model) of a node versus the strength of the node in 4.6a. When $\theta < 1$, and $c = 0.02$, the infection probability increases monotonically with the strength of a node (see 4.6a). When $\theta > 1$, the new phenomena unfolds: high nodal infection probability is obtained by nodes with an intermediate strength, but not those having a small nor large strength. A large θ attributes to the heterogeneity of the recovery rates, allowing nodes with a large strength to have a small infection probability. Given the $\theta = 1.5$, 4.6b shows that the nodal infection probability increases monotonically with the node strength when c is large, e.g. $c > 1$. A large c reduces the heterogeneity of the recovery rate. When c is small, the

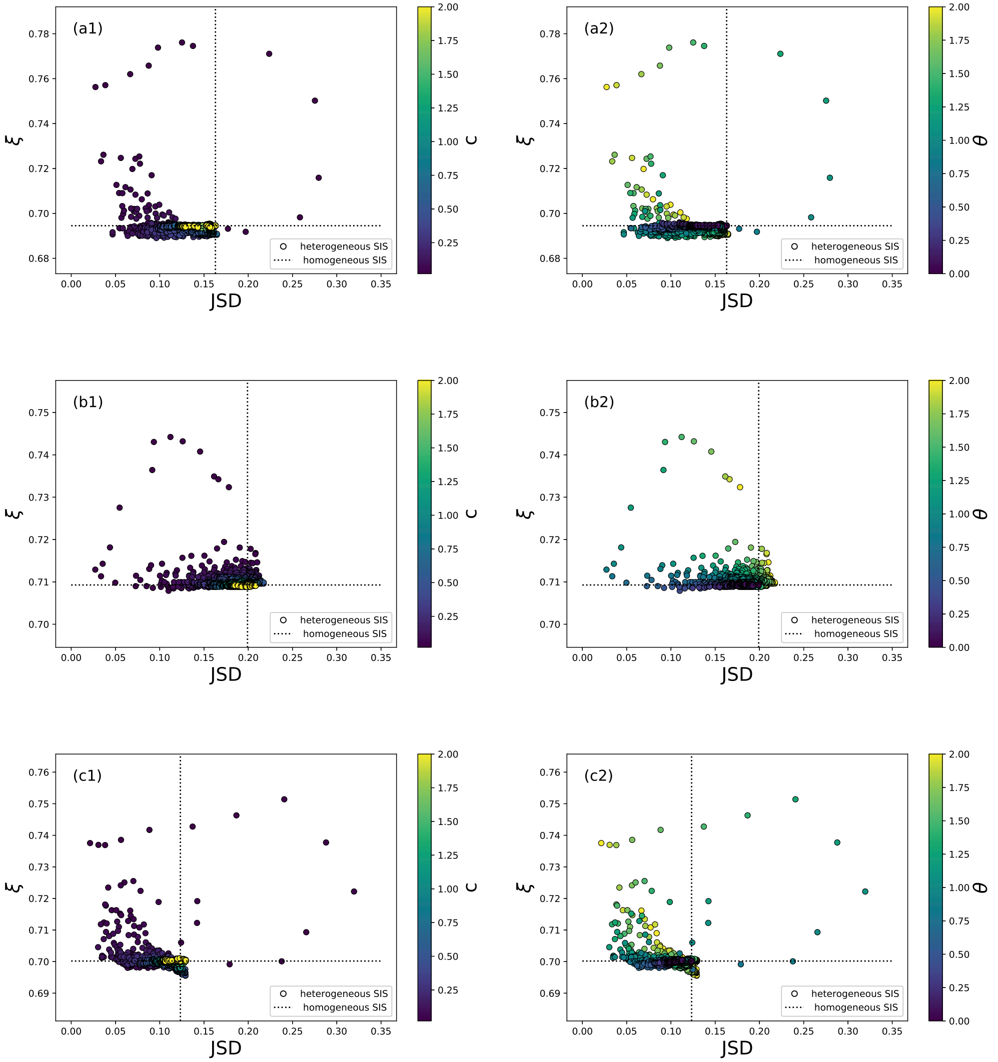

Fig. 4.5: **The scatter plot of the recognition quality ξ versus Jensen-Shannon divergence JSD for both heterogeneous and homogeneous SIS model with diverse parameter sets**. The scatter plot is obtained in network G_1 (figure a1, a2), G_2 (b1, b2) and G_3 (c1, c2). Points correspond to the heterogeneous model, where $\theta \in [0,2]$ with step size 0.1 and $c \in (0,2]$ with step size 0.02. The points are colored according to parameter c in a1, b1, c1 and according to θ in a2, b2, c2. The dash lines correspond to the baseline homogeneous model.

maximal vulnerability has also been obtained by nodes with an intermediate node strength.

The node strength that leads to the maximal infection probability increases as c increases because a larger c makes the recovery rate more homogeneous. In the extreme case, the most heterogeneous case, when $\theta > 1$ and $c = 0$, v decreases monotonically as the node strength increases, which can be seen in 4.6a. In this special case, a larger $\theta > 1$ corresponds to a steeper decrease. The relative magnitude of the constant term c with respect to the node strength dependant term $(\frac{s_i}{s_{max}})^{\theta}$ of δ_i decides when the phenomena occurs that the infection probability increases first and decreases afterward as the node strength increases. Figure 4.7 illustrates the cumulative distribution $Pr[(\frac{S}{s_{max}})^{\theta} \leq x]$ of the term $(\frac{s_i}{s_{max}})^{\theta}$. The model on G_1 that maximizes the recognition quality is obtained when $\theta = 1.5$ and $c = 0.02$ (observed within the range we have searched for). In this case, the constant c is larger than the term $(\frac{s_i}{s_{max}})^{\theta}$ of δ_i in less than 70% of the nodes. The model where $(\frac{s_i}{s_{max}})^{\theta} \leq c$ in most nodes (e.g. when $c = 0.1$ and $\theta = 1.5$) is not optimal. These observations motivate that we may identify the optimal parameter set more efficiently by better choosing the search space.

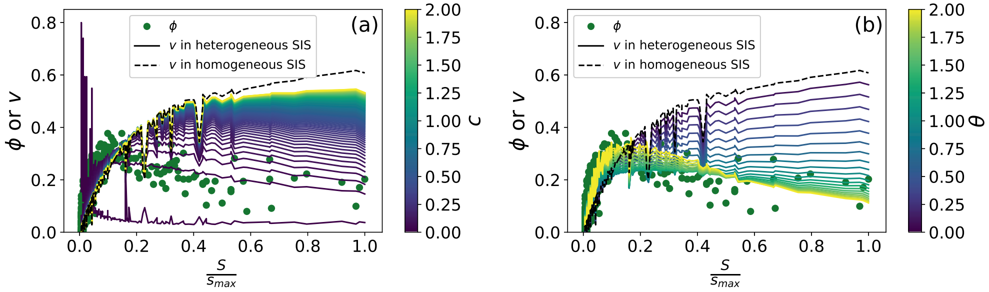

Fig. 4.6: **Airport vulnerability and nodal infection probability versus normalized node strength.** The scatter plot of the vulnerability ϕ (points) and infection probability v (lines) of a node versus the normalized node strength $\frac{s_i}{s_{max}}$ of the node on the underlying network topology G_1. The black dashed line corresponds to the baseline homogeneous model ($\theta = 0$). Solid lines correspond to the heterogeneous model with $\theta = 1.50$, colored according to the parameter c (figure a) or with $c = 0.02$, colored according to the parameter θ (figure b).

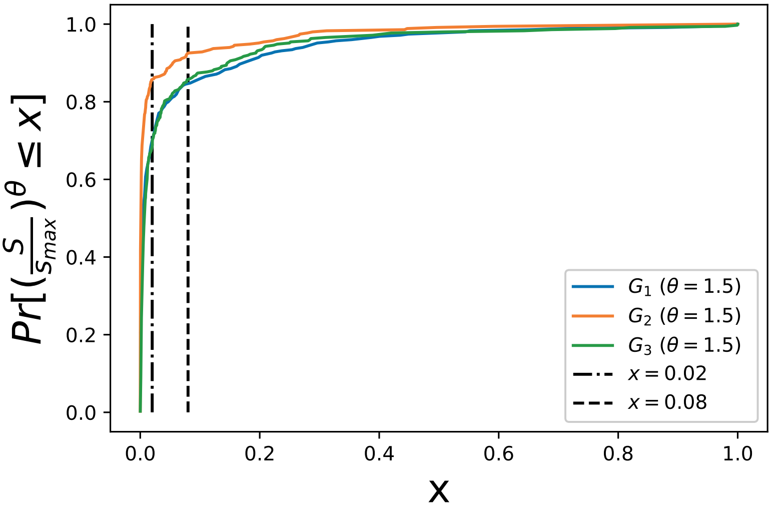

Fig. 4.7: **Cumulative distribution of** $(\frac{S}{s_{max}})^{\theta}$ **in each network when** $\theta = 1.5$.

4.4. Conclusion

We model airport traffic congestion contagion as a heterogeneous SIS spreading process on an airport transportation network, aiming to identify airport's vulnerability, i.e. probability of being congested, using nodal infection probabilities derived from our model. Three airline networks are constructed to capture diverse information e.g. flight frequency and duration and the infection rate of each link is assumed to be proportional to its link weight. Per node, we introduce an heterogeneous recovery rate which is a function of its node strength. The model is evaluated via its capability to reproduce the distribution of nodal vulnerability and to rank airports in vulnerability. Our model evidently outperforms the SIS model with a homogeneous recovery rate in ranking airports from both perspectives. One explanation of the better performance of our heterogeneous model in reproducing the ranking of airports in vulnerability is that: the phenomena that the vulnerability is the largest at airports whose strength in the airline network is neither too large nor too small can be only captured by the heterogeneous model. In particular, a node with a large strength has high rates (link weights) of getting infected by its neighbors, whereas its large recovery rate could reduce its infection probability. Finally, the simplest airline network that represents which airports have direct flight(s) in between already allows the heterogeneous model to evidently outperform the homogeneous one.

The identification of vulnerable airports is crucial for airport operations. Beyond, our model may facilitate the development and evaluation of optimization strategies. The optimization problem can be, e.g. which airports should be invested in improving their capacity thus reducing their vulnerability or in improving their recovery rates in order to minimize the global vulnerability. The derived model that describes how congestion at one airport spreads to other airports could be used to evaluate optimization solutions as a starting point. Such questions require as well further improvement and validation of the model, accounting for e.g. other operational factors and the time varying nature of airport vulnerability. The definition of airport vulnerability can also be generalized by considering e.g. the extent of congestion at an airport.

BIBLIOGRAPHY

[1] Réka Albert and Albert-László Barabási. “Statistical mechanics of complex networks”. In: *Reviews of Modern Physics* 74.1 (2002), p. 47.

[2] Mark EJ Newman. “The structure and function of complex networks”. In: *SIAM review* 45.2 (2003), pp. 167–256.

[3] Stefano Boccaletti et al. “Complex networks: Structure and dynamics”. In: *Physics Reports* 424.4-5 (2006), pp. 175–308.

[4] István Z Kiss, Joel C Miller, Péter L Simon, et al. “Mathematics of epidemics on networks”. In: *Cham: Springer* 598 (2017). URL: `%5Curl%7Bdoi:10.1007/978-3-319-50806-1%7D`.

[5] Massimiliano Zanin and Fabrizio Lillo. “Modelling the air transport with complex networks: A short review”. In: *Eur. Phys. J. Spec. Top.* 215.1 (2013), pp. 5–21.

[6] Romualdo Pastor-Satorras et al. “Epidemic processes in complex networks”. In: *Rev. Mod. Phys.* 87.3 (2015), p. 925.

[7] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge University Press, 2008.

[8] Daqing Li et al. “Epidemics on interconnected lattices”. In: *EPL (Europhysics Letters)* 105.6 (2014), p. 68004.

[9] Piet Van Mieghem, Jasmina Omic, and Robert Kooij. “Virus spread in networks”. In: *IEEE/ACM Transactions On Networking* 17.1 (2008), pp. 1–14.

[10] Bo Qu and Huiijuan Wang. “SIS epidemic spreading with correlated heterogeneous infection rates”. In: *J. Phys. A* 472 (2017), pp. 13–24.

[11] Bo Qu and Huijuan Wang. “SIS epidemic spreading with heterogeneous infection rates”. In: *IEEE TNSE* 4.3 (2017), pp. 177–186.

[12] Cong Li, Ruud van de Bovenkamp, and Piet Van Mieghem. “Susceptible-infected-susceptible model: A comparison of N-intertwined and heterogeneous mean-field approximations”. In: *Phys. Rev. E* 86.2 (2012), p. 026116.

[13] Piet Van Mieghem. “The N-intertwined SIS epidemic network model”. In: *Computing* 93.2-4 (2011), pp. 147–169.

[14] Cong Li, Huijuan Wang, and Piet Van Mieghem. “Epidemic threshold in directed networks”. In: *Phys. Rev. E* 88.6 (2013), p. 062802.

[15] Zimo Yang and Tao Zhou. “Epidemic spreading in weighted networks: An edge-based mean-field solution”. In: *Phys. Rev. E* 85.5 (2012), p. 056106.

[16] Dan Lu et al. “Resilience of epidemics for SIS model on networks”. In: *Chaos Interdiscip. J. Nonlinear Sci.* 27.8 (2017), p. 083105.

[17] Bo Qu et al. “Ranking of nodal infection probability in susceptible-infected-susceptible epidemic”. In: *Scientific Reports* 7.1 (2017), pp. 1–10.

[18] Alain Barrat et al. “The architecture of complex weighted networks”. In: *Proceedings of the National Academy of Sciences* 101.11 (2004), pp. 3747–3752.

[19] Roger Guimera et al. “The worldwide air transportation network: Anomalous centrality, community structure, and cities’ global roles”. In: *Proceedings of the National Academy of Sciences* 102.22 (2005), pp. 7794–7799.

[20] Aura Reggiani et al. “Network measures in civil air transport: a case study of Lufthansa”. In: *Networks, Topology and Dynamics.* Springer, 2009, pp. 257–282.

[21] Ding-Ding Han, Jiang-Hai Qian, and Jin-Gao Liu. “Network topology and correlation features affiliated with European airline companies”. In: *Phys. A* 388.1 (2009), pp. 71–81.

[22] LP Chi and X Cai. “Structural changes caused by error and attack tolerance in US airport network”. In: *Int. J. Mod. Phys. B* 18.17n19 (2004), pp. 2394–2400.

[23] Sean M Wilkinson, Sarah Dunn, and Shu Ma. “The vulnerability of the European air traffic network to spatial hazards”. In: *Natural Hazards* 60.3 (2012), pp. 1027–1036.

[24] Pablo Fleurquin, José J Ramasco, and Victor M Eguiluz. “Systemic delay propagation in the US airport network”. In: *Scientific Reports* 3 (2013), p. 1159.

[25] C Ciruelos et al. “Modelling delay propagation trees for scheduled flights”. In: *Proceedings of the 11th USA/EUROPE Air Traffic Management R&D Seminar, Lisbon, Portugal.* 2015, pp. 23–26.

[26] B Baspinar and E Koyuncu. “A data-driven air transportation delay propagation model using epidemic process models”. In: *International Journal of Aerospace Engineering* 2016 (2016).

[27] Seddik Belkoura and Massimiliano Zanin. “Phase changes in delay propagation networks”. In: *arXiv preprint arXiv:1611.00639* (2016).

[28] Massimiliano Zanin, Seddik Belkoura, and Yanbo Zhu. “Network analysis of chinese air transport delay propagation”. In: *Chinese Journal of Aeronautics* 30.2 (2017), pp. 491–499.

[29] Richard De Neufville and Amedeo Odoni. *Airport Systems. Planning, Design and Management.* 2003.

[30] Lucas Lacasa, Miguel Cea, and Massimiliano Zanin. “Jamming transition in air transportation networks”. In: *J. Phys. A* 388.18 (2009), pp. 3948–3954.

[31] Meead Saberi et al. “A simple contagion process describes spreading of traffic jams in urban networks”. In: *Nature Communications* 11.1 (2020), pp. 1–9.

[32] *United States Bureau of Transportation Statistics.* `http://www.transtats.bts.gov`.

[33] Klemens Köstler et al. “Modeling Airport Congestion Contagion by SIS Epidemic Spreading on Airline Networks”. In: *International Conference on Complex Networks and Their Applications.* Springer. 2019, pp. 385–398.

[34] Jukka-Pekka Onnela et al. “Intensity and coherence of motifs in weighted complex networks”. In: *Phys. Rev. E* 71.6 (2005), p. 065103.

[35] Mark EJ Newman. “Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality”. In: *Phys. Rev. E* 64.1 (2001), p. 016132.

[36] Stefania Ottaviano et al. “Community Networks with Equitable Partitions”. In: *Multilevel Strategic Interaction Game Models for Complex Networks*. Springer, 2019, pp. 111–129.

[37] Jianhua Lin. “Divergence measures based on the Shannon entropy”. In: *IEEE Transactions on Information theory* 37.1 (1991), pp. 145–151. URL: `%5Curl%7Bdoi:https://doi.org/10.1109/18.61115%7D`.

[38] Richard P. Brent. “An algorithm with guaranteed convergence for finding a zero of a function”. In: *The computer journal* 14.4 (1971), pp. 422–425.

[39] William H Press et al. *Numerical recipes in C*. Cambridge university press Cambridge, 1992.

5

Reflections and recommendations

IN this thesis, we propose characterization methods for time-varying networks that could take into account the joint topological and temporal properties of connections among nodes. Moreover, we show a methodology to understand to what extent we can identify or approximate an unknown underlying spreading process, given the observation of node activities and the topology of the network on which the process unfolds, in the specific case of congestion contagion among airports. We show our methods, findings and results in three technical chapters. In order to conclude the circular path started in the introduction, in Section 5.1 we answer the research questions and discuss limitations of the work presented in this thesis. In lights of these reflections, we propose then possible future directions in Section 5.2.

5.1. MAIN CONTRIBUTION AND REFLECTIONS

In Chapter 2, we proposed systematic methods to jointly characterize the topological and temporal properties of contacts in a time-evolving network. By applying our methods to real world networks, we identify substantial differences between physical and virtual contacts. First, contacts occurring in short temporal delays tend to occur also close in their topological location, more evidently in virtual contact networks than in physical contact ones. This is supported by higher local temporal correlation of contacts belonging to the ego networks of different links. Such local temporal correlation manifests as long trains of consecutive contacts in a ego network[1]. These consecutive contacts are, in general, activations of many different links belonging to the same ego network. This is particularly evident in virtual contacts and in those physical contact network where the contacts are less constrained in space, such as pupils in primary school or visitors in museum, than in more spatially constrained social settings, such as individuals in a workplace. Such findings may suggest that contacts with a low cost may better facilitate social contagion, i.e. the influence of a node activity on the activity of its neighbors.

These results are limited to traditional pairwise time-evolving networks, where interactions at a given time among a group of nodes composed of $d > 2$ nodes are decomposed in a clique of $\binom{d}{2}$ contacts occurring at the same timestamp. Such group interactions are observed in many networked systems such as brain, social and collaboration networks.

In Chapter 3, we proposed systematic methods to characterize temporal networks without decomposing such group (or higher-order) events in pairwise interactions, i.e. higher order temporal networks. We applied our methods to eight physical contact and five collaboration higher-order temporal networks. Coherently with what observed in Chapter 2, in physical contacts, events relatively close in time tend to occur also close in topology. Moreover, events of different orders usually overlap in component nodes. The occurrence of such local events is also usually correlated in time and supports the observed correlation between temporal and topological distances of events. The temporal and topological correlation of events observed in collaboration networks is instead either weak or absent. In these networks, events of different orders overlap in component nodes but their local temporal correlation almost disappears. These different results seems to reflect the fundamental differences between the two kind of networks. In physical contacts indeed individuals participate in events driven by physical proximity, so that participants of a higher-order event may results in a higher chance for them or a subgroup of them in the near future.

[1] The contacts of the ego networks of a link connecting i to j are the contacts that involve either node i or j.

Differently, the evolution of the higher-order temporal network of scientific collaboration is likely driven more by their content and creation process. Furthermore, the topological overlap of events with different orders in component nodes observed in both physical contacts and collaboration networks, suggested that nodes participating in many events (groups) of a given order tend to interact in many events (groups) of a different order. Hence, nodes are consistent in interactions with respect to frequency and diversity across different orders.

The insights obtained from the first and second chapter of this thesis show that, except the case of collaboration networks, the topological-temporal correlation in pairwise and higher-order temporal networks is supported by the temporal correlation of neighboring (hyper-)links. This seems to suggest that an epidemic spreading process could potentially reproduce properties of temporal networks. A temporal network can indeed be considered as a static network, together with an unknown dynamical process unfolding on it and determining the link activities at each timestamp. Note that, thanks to the line graph transformation, the links of a network can be represented as nodes of the corresponding line graph of the network. The problem of modelling a temporal network then is fundamentally equivalent to that of reproducing an unknown dynamical process unfolding on a static network and determining the activity of nodes/links. Despite some initial evidence, in temporal networks, we do not have sufficient domain knowledge to justify the assumption that the activity of links evolve on a network according to a spreading process.

Thus, in Chapter 4, we propose a methodology to identify the epidemic spreading process that can model the node activity in a simplified case, where the activity of a node clearly influence the state of its neighbors, i.e., the airport congestion contagion mediated by the air transportation network. We propose three airline static weighted networks to capture diverse properties of flight connections, e.g. their frequency or duration. We assume that the infection rate of each link is assumed to be proportional to its link weight. Each node is assigned a heterogeneous recovery rate which is a function of its node strength: this reflects the different capabilities of airports of different size to handle congestion. We evaluate the model via its ability to reproduce two key properties of the airport congestion: the distribution of airport congestion probability and the ranking of airports according to their chance to be congested. Our model evidently outperforms the SIS model with a homogeneous recovery rate from both perspectives. The better performance of our heterogeneous model in reproducing the ranking of airports in vulnerability may be due to the fact that only the heterogeneous model can capture the phenomena that the largest congestion probability is observed in airports whose strength in the airline network is neither too large nor too small. In our model, indeed, a node with a large strength has high chance of getting infected by its neighbors, while its large recovery rate could reduce its infection probability. Finally, the simplest airline topology, where two nodes are connected if the corresponding airports have direct flight in between already allows the heterogeneous model to evidently outperform the homogeneous one.

5.2. FUTURE WORK

In Chapters 2 and 3, dedicated to characterization methods , we proposed topological and temporal distances between contacts/events, we studied the relation between these two quantities in the case of pairwise and higher order temporal networks, and then investigated which temporal or topological properties of contacts/events can (partially) explain such relation. On the other hand, in Chapter 4, we proposed a method to identify (or approximate) the underlying contagion mechanism of airport congestion. We obtained from data the node state (airport congestion) at each time, the network topology over which the process unfolds and we identify key properties of the dynamical process, e.g. the vulnerability of an airport in congestion that our proposed model should reproduce. Then, we proposed a SIS model where the nodal recovery rate is heterogeneous, function of the degree (or strength) of the corresponding node and show that it evidently outperform a SIS model where the recovery rates is the same at each node in reproducing airport congestion vulnerability. In this last section we try to briefly present some of the possible research directions inspired by the work presented in this thesis.

Generalizations of characterization methods. In the first two chapter of this thesis, we have proposed methods that can characterize jointly the temporal and topological properties of contacts/events. Our study of the relation between topological and temporal distance of pairwise/group interactions can be generalized, for example, by considering different types of distances apart from the traditional shortest path, such as the effective resistance or the spatial distance (in case of spatially embedded networks). Another promising generalization of methods of the Chapter 2 seems to include also directed networks. Moreover, the generalized methods proposed in Chapter 3 solve the problem of generalizing the characterization method of the first chapter to higher order temporal networks by considering each interaction order as an independent interaction layer. Thanks to this, we can almost straightforwardly extend these methods to multilayer temporal networks. A final direction deemed as promising is to generalize the definition of higher-order event to other types of temporal motifs. For example, different events occurring at the same time overlapping in some component nodes may be merged in a single event.

Effect of the topological-temporal correlation of contacts/events on dynamical process unfolding on networks. Another promising research direction that can benefit from our proposed characterization methods is the traditional study of how dynamical processes unfolding on the networks are influenced by properties of the underlying networks. Previous studies on temporal networks have shown that specific temporal properties of link activations, e.g., burstiness, clearly affects the dynamics on networks. To our knowledge, no study has investigated how the topological-temporal correlation of contacts influence key properties of epidemic spreading, synchronization process or random walks on temporal networks. More than that, the randomized reference models for higher order temporal networks proposed in Chapter 3 offer tools to investigate how temporal properties of events influence those dynamic processes that explicitly take into account group interactions, such as generalized spreading, synchronization and random-walk processes.

Modelling temporal networks In the first two chapter of this thesis, we proposed characterization methods to investigate the joint topological and temporal properties of contact/events. Thanks to these methods, we discovered substantial differences among different kind of networks, e.g. physical/virtual contacts and scientific collaborations. Such differences are usually supported by a local temporal correlation of contact/events overlapping in component nodes. This seems to suggest that an epidemic spreading model can reproduce key properties of a temporal network. The fourth chapter showed that a modified epidemic spreading models is able to reproduce key properties of congestion dynamics of an airports. The method proposed in Chapter 4 can be used as a starting point to model temporal networks, by using the line graph of the time aggregated topology instead of the airline network, and substituting the link activity to the airport congestion. However, we expect a further generalized epidemic spreading model to be able to reproduce other fundamental properties of temporal networks, e.g. burstiness.

Acknowledgements

I am extremely moved while I am writing these acknowlegments. This is so far the most prestigious event of my early carreer. I hope that these words will be able to give back at least a bit of the universe of good that I received from all people that share this pathway with me in the last years.

If I have to think about the first important event that lead to this thesis, this is what occurred in an evening of talks in Taizè, summer 2014. It was me, **Gabriele** and **Fabio** and the key sentence was *"It seems like you prefer to go to fish with a boat because you are scared by what you have on mainland"*, referencing to the fact that I was looking for any hobbies to distract myself from my wrong choice about which bachelor course to attend. I was almost bachelor graduated in energy engineering but wanted to move to physics. Being here today submitting the thesis for a Ph.D. in Network Science after a Master in Physics of Complex Systems it seems that that sentence not only had a huge impact on me, but also lead to a so far decent direction. I will be always in debt with you guys.

Of course, the second incredible coincidence that lead to today was how this Ph.D. project started. It was September 2018 and I was starting a Ph.D. project in Turin. I sent an application to what I was feeling it was the best project for me. Fundamental problems, pure network science. I loved it. We set three interviews in three days with **Huijuan** and **Alan** and was able to receive their offer right in time to decline the offer for a position at the University of Turin without wasting any of their funding. I will be always endlessly grateful with you for the huge opportunity you gave me at that time. And for how you were able to manage me throughout the immense challenges of moving from Italy for the first time in my life to live four years 1000 km north and then dealing with the perfect psychological storm of pursuing a Ph.D. during COVID-19 period. More than that, **Huijuan**, your esteem for me in many moments of this path was higher than the one that I had for myself, and your ability to deal with any aspect of my character helped immensely, and determined the positive outcome of these five years. You also gave me the chance collaborate with a giant as **Prof. Shlomo Havlin**, undeniably the highest scientific honor I have ever had.

Then **Omar**, **Jay**, **Matteo**, **Li** and **Harlley**, the time at the department would have been completely different without sharing talks and coffe breaks with you. And then of course **Shilun**, **Xiu-Xiu**, **Manel**, **Roger**, **Sandy**, **Andrew**, **Julian**, **Elvin**, **Odette**, **Cynthia**, **Saskia** and all people that share at least a bit of my path at MMC Department, thank you very much. A bit of this thesis belong to any of you.

A particular mention is about the very first person that I met in the Netherlands: **Maurizio**. An older brother, a mentor and an invaluable support, since that rainy day in November, when you invited me to have a beer. It was the very same day in which I had arrived, I was feeling stuck in that Airbnb room almost completely occupied by the bed and it was the first time that I was realizing the big life step I was doing.

Then, my family in Delft. This adventure here in the Netherlands is divided in two parts: before and after meeting you guys. I had no idea that that time that I met **Giulie'** and **Pippo** after a basketball training, and then the rest of the group having lunch together at Aula would have changed completely these five years. In the first 26 years of my life I had always been Albi. Now I am Albi/Segü/Caffetti'/Coffee. It has also downsides, every time someone talks about coffee I turn in that direction. Only people as creative as **Gigi**, **Ste** or **Pippo** could have invented these nicknames. You made me part of this group without apparent reasons: I was not spending that much time with you, I had a different age and even different interests. I will be always grateful to you. Then, many dinners at **Gigi**'s with **la Ceci** and **Vladdi'** and the smart working from library, sharing that period with **Nico**, **Ceci** and **Fra**. The kindness of **Nico**, the spontaneity of **Ceci** and the sympathy of **Fra**. After two years of Covid it was the best healing that I could aim for. When I go to the library, it still feels strange not finding you there. Then the masterpiece represented by a quote on the birthday cake *"A specter is wondering around Delft, it is the specter of Segü"*, only you could ideate something so personal, precise and original on a birthday cake. Thanks to **Matteo**, **Lisa** and **Debe**, not only for the cake, but also because you have always welcomed me at your house, even when I was without the heating system in the night, coming without any early notice. I felt many times a bit as Harry Potter at Weasley's, with family. **Ari**, I was the first person of being introduced to you by **Galgü**[2] and I still feel very happy for this: you were always there when I needed a true friend to talk with. Thank also to **Ste**, the best Lazio supporter one could meet (and go on a surf trip with), and then **Ari**, **Edo** and **Brenti'**: it's like you've always been part of this group. **Kia**, the fact that after a tumultuous Covid lockdown spent together we still meet for beers at PSOR says a lot. I cannot exclude also **Giorgio**, **Matteo**, **Seba** and **Serena**. In those months you have always welcomed me, even in your last travel together in Bologna and Pesaro, even if I was not as close as you were with each other. Thank you again.

All of you have always been supporting, smart, affectionate, unconventional, or simply, really, namely, friends.

Almost at the end of this list, **my uncles, family friends and close friends from Turin, so Claudio, Silvio, Dario and Simo**: I have seen you very few times in the last years, but I have never felt that your affection for me has changed.

Finally, I thank immensely, incredibly and heartfully **my family**. **My granpa**, even only for the effort of trying everytime to understand what is my professional situation, in academia, in the world of 21th century. Your ability of adapting yourself according to the different historical periods that you live will be always something that I hope to inherit from you as a grandson.

My parents, not only for these five years, but for the immense support that I have received since I decided to change path after the bachelor. You supported me when I started that foolish plan of studying theoretical physics courses, even if my grades at the bachelor were definitely not that good, and, as everyone was keep repeating at the time: *"after a degree in physics you are not finding so many jobs, you either become a developer or a researcher"*. The data science revolution had not impacted the professional world yet. Then, you willingly accepted the idea that me, your only child, that had never lived abroad, was separating from you for five (or more?) years. I think that this is definitely not easy, but you managed to accept this and never made me feel guilty for this choice. We had a lot of video calls during these five years and the biggest merit of this result is yours, because you have always helped to see the things in perspective and stabilize my choice, that sometimes could have been much more impulsive.

[2] the Calcutta's concert we went all together at Amsterdam will be the only time in my life in which the singer was at the same time on the stage and among us

I started this thesis with the opening sentence of Boris Pasternak about happiness, not only because it shared how much social interconnections (or links) affect deeply our lives, but also because it is one of my favourite sentences in general. To metaphorically close this thesis, I will conclude this acknowledgement section with another version of the same sentence, a paraphrase written by Christopher Mc Candless few days before dying in the magic bus in Alaska: *"Happiness is real only when is shared"*. My highest satisfaction will be not having the opportunity to defend this thesis, but sharing and then, hopefully, celebrating this event with all of you. Thank you again.

CURRICULUM VITÆ

Alberto CERIA

24-09-1992 Born in Turin, Italy.

EDUCATION

2006–2011 Scientific Lyceum
Liceo Scientifico E. Majorana, Turin, Italy

2011–2014 B.Sc. in Energy Engineering
Politecnico di Torino, Turin, Italy

2014-2017 M.Sc. in Physics of Complex Systems (Cum Laude)
Università di Torino, Turin, Italy
Supervisors: Dr. Andrè Panisson and Prof. Dr. Ciro Cattuto
Thesis: Partially supervised learning for mental illness detection

2018-2023 Ph.D. in Computer Science
Tecnische Universiteit Delft, Delft, The Netherlands
Supervisors: Prof. Dr. Alan Hanjalic and Dr. Ir. Huijuan Wang
Title: Characterization and modeling of time-varying networks

PROFESSIONAL WORK EXPERIENCE

2016-2017 Research Internship
Institute for Scientific Interchange (I.S.I.) Foundation, Turin, Italy

2017-2018 Research Scholarship
Institute for Scientific Interchange (I.S.I.) Foundation, Turin, Italy

List of Publications

6. Jung-Muller, M., **Ceria, A.**, & Wang, H. (2023) *Higher-order temporal network prediction*, under revision.

5. Zou, L., **Ceria, A.**, & Wang, H. (2023) *Short-and long-term temporal network prediction based on network memory*, under revision.

4. **Ceria, A.**, & Wang, H. (2023) *Temporal-topological properties of higher-order evolving networks.* Scientific Reports, 13(1), 5885.

3. **Ceria, A.**, Havlin, S., Hanjalic, A., & Wang, H. (2022). *Topological–temporal properties of evolving networks.* Journal of Complex Networks, **10(5)**, cnac041.

2. **Ceria, A.**, Köstler, K., Gobardhan, R., & Wang, H. (2021). *Modeling airport congestion contagion by heterogeneous SIS epidemic spreading on airline networks.* Plos One, 16(1), e0245043.

1. Köstler, K., Gobardhan, R., **Ceria, A.**, & Wang, H. *Modeling Airport Congestion Contagion by SIS Epidemic Spreading on Airline Networks.* (2019). in International Conference on Complex Networks and Their Applications (2020), 385–398..