Spreading Processes on Networks
Roles of Nodes, Links, and Hyperlinks

Zhang, S.

DOI
10.4233/uuid:f2cc3e96-7d62-4359-a707-6508c44dd4b5

Publication date
2025

Document Version
Final published version

Citation (APA)
Zhang, S. (2025). *Spreading Processes on Networks: Roles of Nodes, Links, and Hyperlinks*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:f2cc3e96-7d62-4359-a707-6508c44dd4b5

Spreading Processes on Networks

Roles of Nodes, Links, and Hyperlinks

Spreading Processes on Networks

Roles of Nodes, Links, and Hyperlinks

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 27 May 2025 om 15:00 uur

door

Shilun Zhang

**Master of Science in Computer Science,
University of Electronic Science and Technology of China,
Sichuan, China,
geboren te Hubei, China.**

Dit proefschrift is goedgekeurd door de

promotor: Prof. dr. A. Hanjalic
promotor: 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. S.T. Gaito	Università degli Studi di Milano
Dr. P. De Meo	University of Messina
Prof. dr. T. Li	Erasmus University Rotterdam
Prof. dr. ir. R.E. Kooij	Technische Universiteit Delft
Prof. dr. O. Cats	Technische Universiteit Delft

Keywords: Temporal Network, Higher-order Network, Spreading Process, Mitigation, Prediction
Printed by: ProefschriftMaken

ISBN 978-94-6518-037-3

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

CONTENTS

SUMMARY

Spreading processes are ubiquitous in nature and society, from the diffusion of information in social platforms to the spread of diseases within populations. Many real-world systems can be represented as networks, where a piece of information or a disease spreads along links connecting nodes. Different nodes and links often differ in their network properties and play distinct roles in a spreading process. Based on network properties of nodes or links, practitioners may be interested in identifying key nodes as the seed nodes to maximally diffuse a piece of information, or removing specific links to mitigate the spreading. In this thesis, we study the roles of a node or a link in a spreading process from three different perspectives and investigate how these roles relate to the properties of nodes and links within the underlying network.

We first explore how the network properties of a node can be used to predict the spreading influence of the node, defined as the average number of nodes that are ultimately infected when this node is the only seed node. Previous studies have shown that combining node properties derived from local and global topological information can better predict nodal influence than using a single metric. In Chapter 2, we investigate whether using relatively local information is sufficient for the prediction. To address this question, we define an iterative metric set by leveraging the iterative process used to derive classical nodal centralities like eigenvector centrality. The iterative metric set progressively incorporates more global information and is used as the feature set in a regression model to predict nodal spreading influence. The iterative metric set is then used as the feature set in a regression model to predict the spreading influence of a node. We find that the model using the iterative metric set that includes relatively local information achieves comparable prediction quality with the method that includes both local and global information, in various networks.

A spreading process can be mitigated by blocking social contacts, i.e., time-specific interactions. In Chapter 3, we investigate how the network properties of a contact are associated with the mitigation effect when the contact is blocked. We develop probabilistic contact blocking strategies, which remove contacts (temporal links) based on their properties in a temporal network, to mitigate the spread of a Susceptible-Infected-Recovered spreading process. The removal probability of a contact depends on a given centrality metric of the corresponding link in the time-aggregated network and the occurring time of the contact. We propose diverse link centrality metrics, and each centrality metric leads to a unique contact blocking strategy. Our results indicate that the spread of the epidemic is most effectively mitigated when contacts between node pairs that have fewer contacts and contacts that occur earlier in time are more likely to be removed.

The role of a link in a spreading process can also be reflected by the extent to which the link is used in the process. Many real-world systems may involve interactions among groups of more than two individuals and can therefore be represented as temporal higher-order networks. Chapter 4 explores the Susceptible-Infected threshold spreading process unfolding on temporal higher-order networks with two objectives: (1) to understand the contribution of each hyperlink to the spreading process, defined as the average number of nodes that are directly infected via the activation of the hyperlink starting from an arbitrary seed node, and (2) to investigate hyperlinks with what network properties tend to contribute more to the spreading process. This understanding is crucial for developing effective strategies to mitigate a spreading process. Given a temporal higher-order network, we propose to construct a weighted higher-order network, the so-called diffusion backbone, where the weight of each hyperlink denotes its contribution to the spreading process. We then systematically design centrality metrics for hyperlinks in a temporal higher-order network, where each centrality metric captures a specific property of the hyperlink within a temporal higher-order network and is used to estimate the ranking of hyperlinks by their weights in the backbone. We find and explain why certain centrality metrics can better estimate the contributions of hyperlinks under different parameters of the spreading process.

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

1

Introduction

1.1. BACKGROUND

Spreading processes are ubiquitous in various systems of nature and society. Daily examples include diffusion of information among individuals and spread of infectious diseases within populations. Advances in information technology have facilitated efficient communication, enabling people to easily share and exchange information online. However, alongside the benefits, we have also witnessed the spread of misinformation on an unprecedented scale. Empirical studies have shown that false information tends to spread faster than true information on social media platforms like Twitter [1]. The spread of online misinformation has been found to influence political elections [2] and harm public health during pandemics [3]. Similarly, while the advances in transportation have improved mobility, they have also facilitated the spread of viruses, increasing the society's susceptibility to infectious diseases [4, 5]. These emerging challenges highlight the importance of understanding how spreading processes unfold in the real-world systems, as this understanding is crucial for not only suppressing the spread of misinformation or disease, but also for maximizing the spread of useful information [6].

Many real-world systems consist of components that may interact with each other. Such interactions between components in a system can be represented as a network, where nodes stand for components and links connecting nodes represent interactions among components. Networks serve as substrates for spreading processes, where a piece of information or the disease spreads along links (interactions) between nodes (individuals). Given that the spread of information (like a news or post) is akin to the disease spread, the spread of information or disease can be modelled by epidemic spreading models [7] such as *susceptible-infected* (SI) model and *susceptible-infected-recovered* (SIR) model. In the SI model, each node is in one of the two possible states at any time: susceptible or infected. A susceptible node can be infected by each of its infected neighboring nodes independently with an infection probability. In the SIR model, in addition to the infection process, each infected node can recover with a recovery probability, and recovered nodes no longer spread the disease. Information diffusion can occur in a similar way. For example, consider the diffusion of a piece of news in a social network. If person X initially knows the news (infected), he/she possibly shares it with his/her friends (susceptible) who have not yet heard it. Those who become informed (infected) may then share the news with their own friends who remain uninformed. Moreover, individuals who already know the news might lose interest (recovered) and stop spreading it for some reasons. Many other spreading processes, e.g., the adoption of behaviour and the spread of failures that can lead to catastrophic system collapse, can be modelled using epidemic models.

A spreading process is intrinsically affected by the underlying network structure . Understanding the role of the network in spreading processes is one of the key challenges in studying spreading processes. Initial studies assumed networks as *static networks*, where nodes and links between nodes remain unchanged over time. Static networks are useful for representing systems that have a constant topology and only involve pairwise interactions. This limits them in their capability to describe real-world systems in a broad sense. Firstly, many systems are rarely static but exhibit time-varying network topologies. A typical example is a human face-to-face contact network, where two nodes (individuals) only contact (interact) with each other at specific time periods. Such networks can be represented as *temporal networks*, where links between pairs of nodes are activated and deactivated over time [8]. Unlike static networks, spreading processes on temporal networks can be intrinsically different. For example, if node X contacts node Y at time t_1 and node Y contacts node Z at time t_2, the spread of information from X to Z is only possible if $t_1 \leq t_2$. Secondly, many real-world systems involve interactions among more than two nodes [9], e.g., individuals often interact in groups, and scientific collaboration may include multiple researchers. Such systems can be represented as a *temporal higher-order network*, where higher-order links, also called hyperlinks, are activated and deactivated over time. In temporal higher-order networks, a hyperlink can involve a group of nodes of an arbitrary size.

Large efforts have been made to explore how the properties of networks influence the spreading process unfolding on them. In static networks, the distribution of node degree (the number of connections a node has) has been shown to affect a spreading process, such as the reproduction number, which reflects the potential for a disease to spread within a population. In temporal networks, the distribution of inter-event times, which is the waiting time between two consecutive link activations, can influence the speed of epidemic spreading. The temporality of temporal higher-order networks was found to impact the onset of endemic state in epidemic processes [10]. These insights into how properties of an entire network influence the spreading process are particularly valuable for network design aimed at specific objectives, such as improving the efficiency of information transmission in communication systems or preventing catastrophic system collapse in infrastructure networks.

1.2. THESIS SCOPE AND CONTRIBUTION

In addition to understanding the impact of network-level properties on a spreading process, it is of practical value to unravel the roles of individual nodes and links. On the one hand, in practical scenarios, there is often a need to facilitate or suppress a spreading process on a given network. For example, to maximize the diffusion of useful content, one might aim to select key nodes as seed nodes for the spreading process. Conversely, to mitigate the spread of diseases or misinformation, the objective may be to immunize specific nodes or block particular links. The central challenge is to strategically select key nodes or links based on their properties in the network to achieve the desired outcome. On the other hand, since nodes and links can differ significantly in their properties, it is crucial to understand based on what properties some nodes or links contribute more in the transmission of information or disease than others. This understanding helps in designing effective strategies to either facilitate or suppress the spreading. The focus of this thesis is to understand these roles of nodes and links in a spreading process and how the roles are associated with the properties of nodes and links in the underlying network.

Firstly, we focus on the prediction of spreading influence of a node based on the properties of the node in the underlying static network. The spreading influence of a node is defined as the average number of nodes that eventually get infected in the SIR process when the node is the only seed node. Initial studies on the node influence prediction problem aimed to identify the most influential nodes among all nodes based on the network topology. These studies proposed to rank nodes by a node centrality metric such as node degree, eigenvector centrality [11], coreness [12, 13], among others [14].

The highest-ranked nodes are then identified as the most influential. A node centrality metric captures a certain topological feature of a node in the network. For example, node degree encodes the local information (1-hop neighborhood) of a node, while other centrality metrics like eigenvector centrality encode the global information of a node in the network. It has been found that no single centrality metric consistently outperforms all other centrality metrics across diverse types of networks. Subsequent studies have proposed methods to integrate local and global centralities or their rankings, these methods usually exhibit better performance than merely using a local or global centrality. Local and global centrality metrics have been shown to complement each other, achieving universally good performance in locating the most influential nodes across various real-world networks [15]. However, global centrality metrics often have high computational complexity, limiting their application to large-scale networks. Moreover, the non-trivial correlation among different centrality metrics [16] makes it difficult to know to what extent global nodal properties are needed to estimate nodal spreading influence. In **Chapter 2**, we explore the following fundamental question: *Can the spreading influence of a node be effectively estimated using relatively local information, i.e., topological information derived from the neighborhood within a small hop count from the target node?*

In **Chapter 3**, we focus on a more realistic problem on how to mitigate the epidemic spreading process on a temporal network by strategically selecting contacts to block. The motivation is that the spread of epidemics or information can be mitigated via reducing physical contacts in reality. For instance, measures implemented during the Covid-19 pandemic—such as curfews, remote work, and social distancing—aim to limit physical interactions. Recent work [17] proposed link removal strategies based on link properties in the time-aggregated network of the temporal network to suppress epidemic outbreaks in the SI process on a temporal network. Each of the proposed strategies blocks a portion of links with a certain property in the aggregated network, thus all contacts associated with the selected links are removed from the temporal network. With such a framework, it was revealed that links with different properties in the aggregated network play different roles in mitigating the spreading process. However, this study has several limitations. Firstly, it is often impractical to block all contacts associated with selected links in real-world scenarios. Secondly, the properties based on the aggregated network ignore the temporal information of links in the temporal network, which has a crucial impact on the spreading process on temporal networks. Thirdly, the study only used the outbreak size as measure of the mitigation effect. These limitations motivate us to investigate how to suppress the epidemic spreading via blocking contacts. In **Chapter 3**, we broaden our investigation by systematically exploring the following question: *Given a temporal network, how can we effectively mitigate the spreading process by selecting the contacts to block?*

The role of a link in a spreading process can also be reflected by how likely the link is used in the spreading process. Understanding such a role of links is crucial to addressing challenging optimization problems, such as determining which node pairs or temporal contacts should be stimulated to maximize the prevalence of information spreading. Zhan et al. [18] studied the probability of a link appearing in the spreading trajectories of a spreading process on a pairwise temporal network, and investigated the relationship between this probability of a link and properties of the link within the temporal network.

They found that links that activate for a large number of times and activate early in time tend to be have a high probability to appear in spreading trajectories. Contreras et al. [19] considered spreading processes on higher-order networks and found that the parameters of the spreading process affect the probability that a node is directly infected by another node. However, their study is limited to higher-order static networks. The understanding of the role of a hyperlink in the spreading process on temporal higher-order networks is still unknown. This understanding is essential for the design of strategies to mitigate a spreading process via incentivizing the activity of critical groups. In **Chapter 4**, we investigate the contribution of a hyperlink, defined as the number of nodes directly infected through the activations of the hyperlink, in a spreading process on temporal higher-order networks, and explore the question: *which kind of hyperlinks, i.e., hyperlinks with what specific properties in a temporal higher-order network, contribute more to the spreading process?*

1.3. PUBLICATION RELATED TO THIS THESIS

The following papers are completed by the author of this thesis while pursuing the Ph.D. degree at the Delft University of Technology.

1. **S. Zhang**, A. Hanjalic, and H. Wang, Predicting nodal influence via local iterative metrics, Scientific Reports 14, 4929 (2024). [**Chapter 2**]

2. **S. Zhang**, X. Zhao, and H. Wang, Mitigate SIR epidemic spreading via contact blocking in temporal networks, Applied Network Science 7, 2 (2022). [**Chapter 3**]

3. **S. Zhang**, A. Ceria ,and H. Wang, Diffusion backbone of temporal higher-order networks, Communication Physics (submitted) [**Chapter 4**].

1.4. HOW TO READ THIS THESIS

Chapters 2, 3, and 4 of this thesis comprise original publications. The corresponding publication references can be found in the footnote at the beginning of each chapter. Each chapter is a stand-alone work and can be read without reference to previous chapters. The content length and depth may vary across the chapters as we published the corresponding papers in different scientific journals or conferences.

REFERENCES

1. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359,** 1146–1151 (2018).
2. Bovet, A. & Makse, H. A. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* **10,** 7 (2019).
3. Roozenbeek, J. *et al.* Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science* **7,** 201199 (2020).
4. Kraemer, M. U. *et al.* The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368,** 493–497 (2020).
5. Hazarie, S., Soriano-Paños, D., Arenas, A., Gómez-Gardeñes, J. & Ghoshal, G. Interplay between population density and mobility in determining the spread of epidemics in cities. *Communications Physics* **4,** 191 (2021).
6. Morone, F. & Makse, H. A. Influence maximization in complex networks through optimal percolation. *Nature* **524,** 65–68 (2015).
7. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Reviews of Modern Physics* **87,** 925–979 (2015).
8. Holme, P. Modern temporal network theory: a colloquium. *The European Physical Journal B* **88,** 1–30 (2015).
9. Battiston, F. *et al.* Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports* **874,** 1–92 (2020).
10. Chowdhary, S., Kumar, A., Cencetti, G., Iacopini, I. & Battiston, F. Simplicial contagion in temporal higher-order networks. *Journal of Physics: Complexity* **2,** 035019 (2021).
11. Klemm, K., Serrano, M. Á., Eguíluz, V. M. & Miguel, M. S. A measure of individual role in collective dynamics. *Scientific reports* **2,** 292 (2012).
12. Kitsak, M. *et al.* Identification of influential spreaders in complex networks. *Nature Physics* **6,** 888–893 (2010).
13. Lü, L., Zhou, T., Zhang, Q.-M. & Stanley, H. E. The H-index of a network node and its relation to degree and coreness. *Nature Communications* **7,** 1–7 (2016).
14. Lü, L. *et al.* Vital nodes identification in complex networks. *Physics Reports* **650,** 1–63 (2016).
15. Bucur, D. Top influencers can be identified universally by combining classical centralities. *Scientific Reports* **10,** 1–14 (2020).
16. Wang, H., Hernandez, J. M. & Van Mieghem, P. Betweenness centrality in a weighted network. *Physical Review E* **77,** 046105 (2008).
17. Zhan, X.-X., Hanjalic, A. & Wang, H. *Suppressing information diffusion via link blocking in temporal networks* in *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8* (2020), 448–458.

18. Zhan, X.-X., Hanjalic, A. & Wang, H. Information diffusion backbones in temporal networks. *Scientific Reports* **9,** 1–12 (2019).

19. Contreras, D. A., Cencetti, G. & Barrat, A. Infection patterns in simple and complex contagion processes on networks. *PLOS Computational Biology* **20,** e1012206 (2024).

2

Predicting Nodal Influence via Local Iterative Metrics

Nodal spreading influence is the capability of a node to activate the rest of the network when it is the seed of spreading. Combining nodal properties (centrality metrics) derived from local and global topological information respectively has been shown to better predict nodal influence than using a single metric. In this chapter, we investigate to what extent local and global topological information around a node contributes to the prediction of nodal influence and whether relatively local information is sufficient for the prediction. We show that by leveraging the iterative process used to derive a classical nodal centrality such as eigenvector centrality, we can define an iterative metric set that progressively incorporates more global information around the node. We propose to predict nodal influence using an iterative metric set that consists of an iterative metric from order 1 to K produced in an iterative process, encoding gradually more global information as K increases. Three iterative metrics are considered, which converge to three classical node centrality metrics, respectively. In various real-world networks and synthetic networks with community structures, we find that the prediction quality of each iterative based model converges to its optimal when the metric of relatively low orders ($K \sim 4$) are included and increases only marginally when further increasing K. This fast convergence of prediction quality with K is further explained by analyzing the correlation between the iterative metric and nodal influence, the convergence rate of each iterative process, and network properties. The prediction quality of the best performing iterative metric set with $K = 4$ is comparable with the benchmark method that combines seven centrality metrics: their prediction quality ratio is within the range $[91\%, 106\%]$ across all three quality measures and networks. In two spatially embedded networks with an extremely large diameter, however, iterative metric of higher orders, thus a large K, is needed to achieve comparable prediction quality to the benchmark.

2.1. Introduction

Spreading processes are ubiquitous in various systems of nature and society. Examples include the spreading of epidemics, the propagation of information, and cascade of failures. Complex networks, usually considered as the underlying structure of such systems, provide the substrate upon which the spreading process unfolds via links connecting nodes. The spreading influence of a node represents the extent to which the node, where the spread originates, can eventually activate other nodes in the network. For a given spreading process, the spreading influence of a node is defined as the expected outbreak size when the spreading process starts from the node, also called the seed node. Due to the topological heterogeneity of nodes in many real networks [1], some nodes may have significantly higher spreading influence and are evidently more influential than the other nodes [2–4]. Identifying these influential nodes and predicting their spreading influence is crucial for controlling the spread of epidemics [5, 6] or rumors [7, 8], promoting strategic marketing [9–11], quantifying the impact of researchers and publications [12], and more [13–15].

Two generic influence prediction problems have been addressed in prior research. The first involves identifying the most influential nodes among all nodes based on the given network topology. To solve this problem, previous studies have proposed to rank nodes by a single nodal topological metric, so-called centrality metric [16–18], which encodes either local [19, 20] or global [16, 21] topological information around a given node.

The highest-ranked nodes are then identified as the most influential ones. Nonetheless, these prior work suggests that no single centrality metric can outperform all other centralities for different epidemic parameters and in diverse types of networks, since a centrality metric only captures a certain topological feature of a node. It has been shown that nodal degree, i.e., number of 1-hop neighbors, is more (less) predictive than eigenvector centrality [22] when the spreading rate is small (large) [6, 23]. The coreness better predicts the top spreaders than nodal degree in Susceptible-Infected-Recovered model below epidemic threshold. Further studies put forward methods to integrate local and global centralities or their rankings. Zhe Li et al. [24] used the sum of normalized degree, eigenvector centrality, and coreness as the mass of a node in a gravity model to derive a new nodal metric. Andrea Madotto et al. [25] aggregated the ranking lists by local and global node centralities to produce a new ranking list based on the correlations between the rankings. These methods usually exhibit better performance than merely using a local or global centrality.

In many practical scenarios, it is possible to observe or derive the spreading influences of a small fraction of nodes. For example, the average number of retweets of content posted by a node can be used as an approximation of the spreading influence of the node [6, 26]. This motivates the second influence prediction problem: identify the most influential nodes given the network topology and the influence of a small fraction of nodes. Bucur [27] recently proposed to train a statistical model on the set of nodes whose spreading influences are known to classify the rest of nodes into binary classes, representing whether a node is among the top (e.g., top 10%) influential ones or not. The statistical model maps the relation between the class of a node in spreading influence and centrality metrics including both local centrality metrics like degree and global centrality metrics like betweenness [28] and eigenvector centrality. These centrality metrics were shown to be able to complement each other to achieve universally good performance in locating the most influential nodes across various real-world networks. However, global centrality metrics have a high computational complexity, which limits their application to large-scale networks. Moreover, the non-trivial correlation among different metrics makes it difficult to interpret to what extent global nodal properties are needed to estimate nodal spreading influence.

To bridge this gap, we will systematically explore two foundational questions: how local and global topological information around a node contribute to the prediction of the spreading influence of this node, and whether relatively local information, i.e., topological information derived from the neighborhood within a small hopcount from a target node, can predict its nodal spreading influence effectively. The general prediction task is considered: given the topology of a network and the spreading influences of a fraction of nodes, how to predict the spreading influences of the other nodes in the network, beyond their ranking. To solve the prediction task, a node-level regression model is trained on the set of nodes whose spreading influences are known and used to predict the influences of the remaining nodes. To understand how local and global topological information contribute to the prediction, we design the input of the regression model based on nodal properties as follows. We show that by leveraging the iterative process used to derive a classical node centrality such as eigenvector centrality, we can define an iterative metric that gradually encodes more global information as the order grows.

Then, an iterative metric set that consists of an iterative metric from order 1 to order K is used as input features of the regression model. For example, the number of k-hop walks originate from a node, which is determined by the k-hop neighborhood of the node, can be derived in an iterative process starting from $k = 1$. The resultant iterative metric set is composed of the iterative metric (the number k-hop walks) with order $k \in [1, K]$ after K iterations. The benefits of using an iterative metric set to predict nodal influence are as following. Firstly, it allows us to explore to what extent global network information is needed to estimate the nodal influence, i.e., is K necessarily large for accurate prediction? Second, it enables us to identify the prediction method with low computational cost, that is, the regression model with an iterative metric set of small K. Moreover, in practical applications, one has the flexibility to choose an appropriate K to achieve a well-balanced trade-off between prediction accuracy and computational efficiency. The intuition is illustrated in Figure 2.1, which shows a network example of 1000 nodes with community structure generated by Lancichinetti–Fortunato–Radicchi model [29]. The red-colored nodes are the top 10% nodes when nodes are ranked by spreading influence (top left), eigenvector centrality (EC, top middle, which corresponds the component of the eigenvector corresponds to the largest eigenvalue of the adjacency matrix), degree (DC, top right), number of 2-hop (bottom left), 3-hop (bottom middle) 4-hop (bottom right) walks originating from a node, respectively. The example suggests that the number of 2-, 3- and 4-hop walks possibly reflect nodal spreading influence better than the global metric (eigenvector centrality). Furthermore, it has been observed and partially proved in previous work that a centrality metric like betweenness with a high computational complexity is correlated with local metrics derived from a low order neighborhood [18, 30]. Hence, global network information, i.e., large K, is not necessarily needed in nodal influence prediction.

In this work, we consider three iterative metrics, which converge, respectively to three global node centrality metrics: eigenvector centrality, PageRank centrality [31], and H index of a node [32]. The computation of each iterative metric set can be done in $\mathcal{O}(K \cdot |E|)$ time, where $|E|$ is the number of links in the network. Based on each iterative metric set, a statistical regression model is built and trained to predict nodal influence. We evaluate the prediction quality of the corresponding three regression models, in comparison with a benchmark [27], i.e., the regression model that uses 7 nodal centrality metrics, in both real-world networks and synthetic networks with community structure. We find that in almost all networks, an iterative metric set with $K \sim 4$ is able to accurately predict nodal spreading influence, and the prediction quality increases marginally when more global metrics are included as K grows. This suggests the low computational complexity of our iterative metric based prediction methods. Additionally, the best performing iterative metric based model with $K \sim 4$ performs as well as the benchmark model, which has higher computational cost due to the computation of global centrality metrics. An exception holds for two infrastructure networks, i.e., US power grid and Chicago regional road network, which are spatially embedded networks and have an extremely large diameter (> 40). In these two networks, nearly optimal prediction quality is achieved only when using the iterative metric set that includes metrics of large orders, thus when K is large. Hence, the proposed iterative metric method utilizing relatively local network information could predict nodal influence as well as the benchmark in networks with the small-world property and has a lower computational complexity.

This chapter is organized as follows. In Section 2.2, we introduce the definition of nodal spreading influence and iterative metrics, and regression models to predict nodal influence. Section 2.3 evaluates the performance of the proposed influence predication methods in both real-world networks and synthetic networks with community structure. Section 2.4 summarizes our findings and discusses limitations and potential extensions of our work.

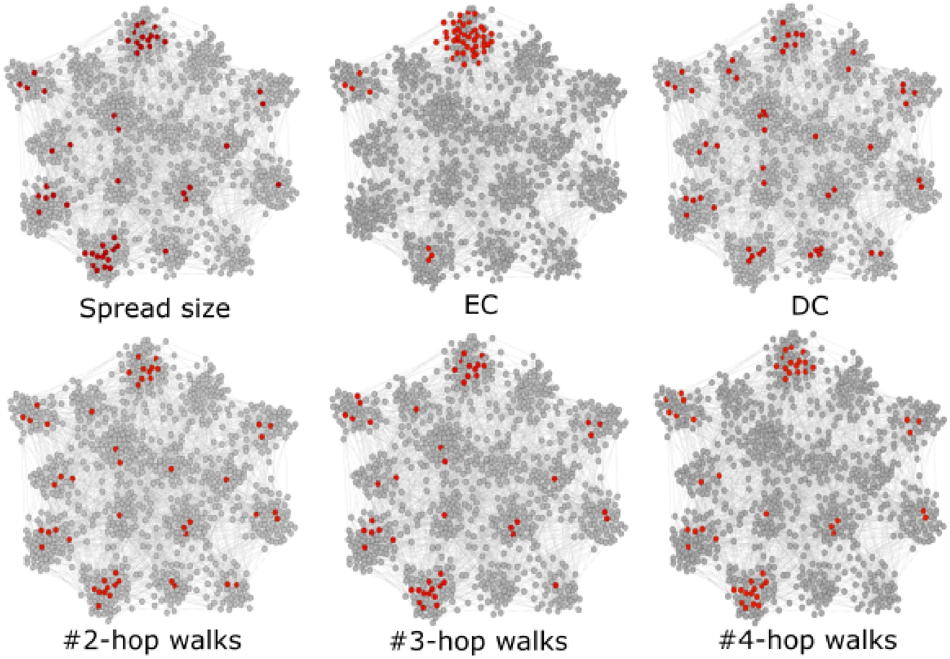

Figure 2.1: Location of top ranked nodes in a network generated by LFR model. The red-colored nodes are the top 10% nodes when nodes are ranked by spread size (top left), eigenvector centrality (EC, top middle), degree centrality (DC, top right), 2-hop walk counts (bottom left), 3-hop walk counts (bottom middle), and 4-hop walk counts (bottom right), respectively.

2.2. METHODS

In this section, we present the definition of nodal spreading influence (Section 2.2.1), followed by the definition of iterative metrics (Section 2.2.2). We then describe the regression model that uses an iterative metric set to predict nodal spreading influence (Section 2.2.3).

2.2.1. NODAL SPREADING INFLUENCE

We consider the continuous-time Susceptible-Infected-Recovered (SIR) spreading process on a static network [3, 33]. At any time, each node can be in one of three possible states: susceptible, infected, or recovered. In the beginning, one seed node gets infected, while the rest are susceptible. A susceptible node gets infected by each of its infected neighbors at an infection rate β, and each infected node recovers at a recovery rate γ. Both the infection and recovery processes are independent Poisson processes. In the steady state, all nodes are either susceptible or recovered. The ratio $\lambda = \beta/\gamma$ is called the effective infection rate. Without loss of generality, we assume the recovery rate $\gamma = 1$, thus $\lambda = \beta$. For a given network, an epidemic threshold λ_c exists. When $\lambda > \lambda_c$, a non-zero fraction of recovered nodes exist in the stable state. When $\lambda < \lambda_c$, the epidemic dies out. The number of recovered nodes in the steady state, or equivalently, the number of nodes that have ever been infected is called the outbreak size.

The spreading influence of a node is defined as the average outbreak size when the node is chosen as the seed node. We derive the influence of a node as the average outbreak size over $r = 10^4$ realizations of the SIR spreading process on a given network. When the effective infection rate $\lambda \ll \lambda_c$ or when $\lambda \gg \lambda_c$, nodes tend to have similar influence. We focus on predicting influence when the effective infection rate is around the epidemic threshold, e.g., $\lambda = 0.5\lambda_c, \lambda_c 1.5\lambda_c, 2\lambda_c$. This is when nodes differ evidently in influence, and influence prediction is crucial. We estimate the epidemic threshold λ_c using the numerical approach introduced in [34]. Specifically, referring to ρ as a random variable denoting the influence of a random node in the network, we consider the variability $\sqrt{\langle\rho^2\rangle - \langle\rho\rangle^2}/\langle\rho\rangle$ as a function of λ. The epidemic threshold λ_c is then the value of λ that maximizes the variability.

2.2.2. Iterative metrics

Given an undirected network $G = (V, E)$, where V is the set of nodes and E is the set of links between nodes in V, the network can be represented by the adjacency matrix A, whose element $A_{ij} = 1$ if there is a link between node i and j, otherwise $A_{ij} = 0$. Various node centrality metrics have been proposed to measure the topological importance of a node, such as eigenvector centrality, PageRank, and coreness [32]. For a given centrality metric, the centralities of all nodes can be denoted by a vector $\mathscr{M}$, where the entry $\mathscr{M}_i$ represents the centrality of node i. The iterative process used to derive the corresponding iterative metric set starts with an initial metric vector $\mathscr{M}^{(0)}$ and updates the metric vector based on a specific rule $\mathscr{M}^{(k)} = f(\mathscr{M}^{(k-1)})$. Eventually, this process converges to the target centrality metric $\mathscr{M}$. We refer to the derived metric vectors $\{\mathscr{M}^{(k)}, k = 1, 2, ...K\}$ as the iterative metric set.

In this work, we consider three iterative processes that converge to three global centrality metrics: eigenvector centrality, PageRank centrality, and coreness of a node, respectively. Three different iterative metrics are derived using these processes.

- **Normalized Walk Count (NWC)**. We adopt the power iteration process for the computation of eigenvector centrality to derive the NWC iterative metric. The centrality vector is initialized as the normalized all-one vector $w^{(0)} = u/\sqrt{N}$, where u is the all-one vector, and is updated iteratively following the updating equation $w^{(k)} = Aw^{(k-1)}/||Aw^{(k-1)}||$. The k-th order NWC follows $w^{(k)} = A^k u/||A^k u||$. Its element $w_i^{(k)}$ represents the normalized number of distinct k-hop walks starting from node i and can be derived from the neighborhood within k hops of the node i. As k increases, $w^{(k)}$ converges to the eigenvector centrality w. The rate of convergence is determined by the ratio of the largest eigenvalue $\lambda_1(A)$ and the second largest eigenvalue $\lambda_2(A)$ of the adjacency matrix A of the network. The convergence rate is higher when $\frac{|\lambda_2(A)|}{|\lambda_1(A)|}$ is smaller [35].

- **Visiting Probability (VP)** is derived using the iteration process for the computation of PageRank centrality [31]. The metric vector is initiated as the normalized all-one vector, $p^{(0)} = u/N$, and updated iteratively as $p_i^{(k)} = \alpha \sum_{j=1}^{N} A_{ji} p_j^{(k-1)}/d_j + (1-\alpha)/N$, where d_j is the degree of node j and the teleportation parameter α is set to 0.85, which is a common choice for calculating the PageRank centrality [36]. As k increases, $p_i^{(k)}$ converges to PageRank centrality. The updating equation can be formulated in matrix form: $p^{(k)} = Gp^{(k-1)}$, where $G = \alpha A^T D^{-1} + \frac{1-\alpha}{N} uu^T$, matrix D is a diagonal matrix with $D_{ii} = \sum_j A_{ij}$. Since matrix G is a stochastic matrix, the largest eigenvalue $\lambda_1(G) = 1$. The rate of convergence is determined by the second largest eigenvalue $\lambda_2(G)$ of the matrix G. The smaller $|\lambda_2(G)|$ is, the faster the convergence is [35]. The iterative process can be interpreted as a random walk: the walker starts at a randomly selected node. At each time step, with probability α it moves to a random neighbor of the current visiting node, and with probability $1-\alpha$ it jumps to a node that is randomly selected from the network. The k-th order iterative metric $p_i^{(k)}$ of a node i is the probability that node i is visited by the random walker at the k-th hop. Since the information of neighbors' degree is needed in each iteration step, $p_i^{(1)}$ actually encodes 2-hop neighbors' information. Similarly, the $(k+1)$-hop neighborhood information of a node i is needed to derive $p_i^{(k)}$.

- **H index (HI)** [32]. The 1-st order H index is defined as the degree of a node, i.e. $h_i^{(1)} = d_i$. The k-th order H index of node i can be derived as $h_i^{(k)} = \mathscr{H}[h_{j_1}^{(k-1)}, h_{j_2}^{(k-1)}, ..., h_{j_{d_i}}^{(k-1)}]$, where $j_1, ..., j_{d_i}$ are neighbors of node i and $\mathscr{H}$ is an operator that returns an integer. Specifically, $h_i^{(k)}$ is the maximum integer such that at least $h_i^{(k)}$ elements of $[h_{j_1}^{(k-1)}, h_{j_2}^{(k-1)}, ..., h_{j_{d_i}}^{(k-1)}]$ are no less than $HI_i^{(k)}$. It has been proved that $h^{(k)}$ will converge to the coreness [16, 37] as k increases.

The iterative rules f in the three iterative processes only involve operations among a node's 1-hop neighbors. As a result, the metric vector $\mathscr{M}^{(k)}$ after one step iteration encodes information about the neighborhood one hop further than $\mathscr{M}^{(k-1)}$ (see Section 2.5.1 for a more detailed explanation). Given an iterative process, the obtained metric set $\{\mathscr{M}_i^{(1)}, \mathscr{M}_i^{(2)}, ..., \mathscr{M}_i^{(K)}\}$ will be used to predict the influence of node i using the regression model described in Nodal influence prediction method subsection. The parameter K controls the scope of information around a node encoded in the iterative metric set $\{\mathscr{M}_i^{(1)}, \mathscr{M}_i^{(2)}, ..., \mathscr{M}_i^{(K)}\}$.

2.2.3. NODAL INFLUENCE PREDICTION METHOD

We assume two key types of information are given to predict nodal influence. Firstly, the network topology is known. Secondly, the influences of a small fraction of nodes are available. In practical scenarios, these influences can often be estimated from real-world diffusion data within social media networks. Our objective is to predict the influences of the remaining nodes in the network. We approach the prediction of nodal influence as a node-level regression problem. Specifically, given a static network $G = (V, E)$ represented by its adjacency matrix A and the spreading influences of a fraction q of nodes, which is randomly selected and denoted as S_q, we aim to predict spreading influences of the remaining $1-q$ nodes, referred to as S_{1-q}.

We choose $q = 10\%$ assuming only the influences of a small fraction of nodes are known. We train a statistical regression model, which maps the nodal features into the influence of a node, on the training node set S_q, and evaluate it on the remaining test node set S_{1-q}. For each of the three iterative metrics, the iterative metric set $\{\mathcal{M}_i^{(1)}, \mathcal{M}_i^{(2)}, ..., \mathcal{M}_i^{(K)}\}$ is used as nodal features in the regression model to predict nodal influence. As a benchmark model, we consider a regression model that uses the same set of 7 classic centrality metrics as in Bucur's classification model [27] as nodal features. These 7 centrality metrics include both local and global centrality metrics and are able to complement each other in improving the performance in the node classification task. Finally, we evaluate the prediction quality of the regression models based on 50 realizations of the random sampling of the training node set S_q and the training of the regression model.

We choose the Random Forest Regression model (RFR), a classic model that captures the nonlinear relationship between input features and the outcome variable, i.e., nodal influence, in our case. We also considered the Ridge regression, a linear regression model with L2 regularization, and obtained qualitatively similar observations (see Appendix) as the Random Forest Regression.

2.3. Results

We evaluate the performance of the regression models based on each of the three iterative metrics and the benchmark model based on classic centrality metrics, first in real-world networks in Performance analysis in real-world networks subsection, and afterwards in synthetic networks with community structures in Prediction on networks with communities subsection. Finally, we explore the performance of these models in relation to parameters of the spreading process in Prediction of nodal spreading influence near epidemic threshold subsection.

2.3.1. Networks and measures to evaluate prediction quality

We consider 9 real-world networks that differ in network properties such as size and and diameter (i.e. the largest shortest path length between a node pair among all possible node pairs), including four online social networks (advogato, facebook, deezerEU, github), a scientific collaboration networks (Arxiv Astro), a file sharing network (Gnutella04), two infrastructure networks (US power grid, ChicagoRegional road network), and an email communication network (Email Enron). All the datasets are obtained from the repository of KONECT project [38, 39]. We treat all networks as simple, undirected and unweighted. Basic properties of these networks are listed in Table 2.1. Notably, the two infrastructure networks, US powergrid and ChicagoRegional, have significantly larger diameters, higher modularity, and lower average degree than the other networks.

We evaluate the prediction quality of the proposed regression models using the fol-

Dataset	$\|N\|$	$\|E\|$	$\langle d\rangle$	Diameter	Q	λ_c
advogato	5042	41791	16.577	9	0.408	0.020
Arxiv-astrophics (astroph)	17903	196972	22.004	14	0.626	0.015
enron	33696	180811	10.732	13	0.608	0.013
facebook	63392	816886	25.773	15	0.632	0.010
Gnutella04 (gnu04)	10876	39994	7.355	10	0.386	0.080
github	37700	289003	15.332	11	0.453	0.011
Deezer EU (deezereu)	28281	92752	6.559	21	0.683	0.070
US power grid (uspower)	4941	6594	2.669	46	0.935	0.870
ChicagoRegional (Chicago)	12979	20627	3.179	106	0.931	1.230

Table 2.1: Basic properties of each real-world network considered: Number of nodes $|N|$, number of links $|E|$, average node degree $\langle d\rangle$, network diameter, the modularity Q [1], and epidemic threshold λ_C of the SIR process on the network.

lowing 3 classic measures:

The **coefficient of determination** r^2 measures the proportion of the variance in the dependent variable (nodal influence) that is predictable from the input features in the regression model. r^2 is defined as:

$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{2.1}$$

Here, y_i and $\hat{y}_i$ are the ground truth and the predicted nodal influence of node i given by the regression model, respectively. $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the mean value of y_i.

Kendall's correlation coefficient $\tau(\hat{s}, s)$ measures the similarity of the two ranking lists of nodes based on the predicted nodal influence $\hat{s}$ and the ranking based on the actual nodal influence obtained by SIR simulation. A value of 1 for $\tau(\hat{s}, s)$ indicates that the predicted nodal influence gives the same node ranking as the ground truth, while a value of -1 indicates that the two rankings are reverse. Kendall's correlation coefficient [40] $\tau(\hat{s}, s)$ is defined as follows:

$$\tau(\hat{s}, s) = \frac{n_c - n_d}{\sqrt{(n_c + n_d + T) * (n_c + n_d + U)}} \tag{2.2}$$

where n_c and n_d are the total number of node pairs that are concordant and discordant respectively, based on the influence s and the predicted influence $\hat{s}$. For example, node pair (i, j) is concordant if $(\hat{s}_i - \hat{s}_j)(s_i - s_j) > 0$, and is discordant if $(\hat{s}_i - \hat{s}_j)(s_i - s_j) < 0$. T is the number of node pairs that have the same influence but different predicted influence, i.e., $s_i = s_j, \hat{s}_i \neq \hat{s}_j$ and U is the number of node pairs that have the same predicted influence but different influence, i.e., $\hat{s}_i = \hat{s}_j, s_i \neq s_j$.

Recognition rate of top-f**%** measures the performance of a regression model in identifying the most influential f% nodes in the test set S_{1-q}. It is calculated as the fraction of nodes that are present in the top f% of both the ranking by predicted nodal influence $\hat{s}$ and the ranking by actual nodal influence s. A higher recognition rate of top-f% implies better performance of the regression model in identifying the most influential nodes.

2.3.2. PERFORMANCE ANALYSIS IN REAL-WORLD NETWORKS

We focus on the prediction of spreading influence when the effective infection rate of the SIR spreading process is $\lambda = \lambda_c$, where the epidemic threshold λ_c of each network is identified using the method described in Method section. The values of λ_c of each real-world network are shown in Table 2.1. Later in this section, we will discuss how the choice of the effective infection rate around the epidemic threshold impacts the performance of influence prediction methods.

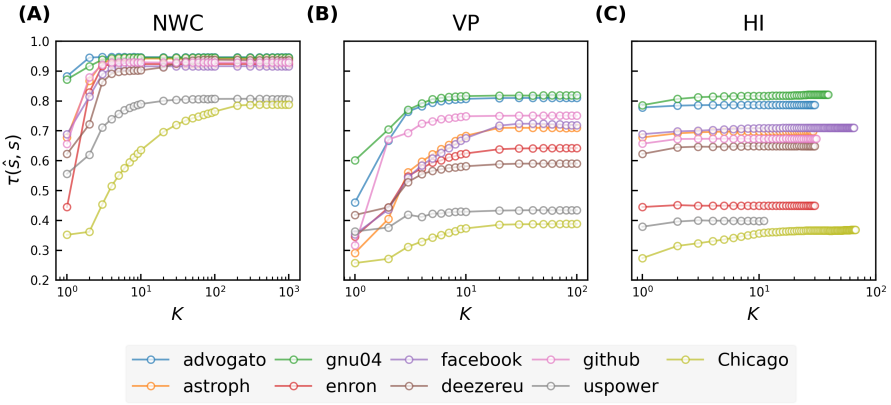

Figure 2.2: Kendall correlation between the actual nodal spreading influence s and the influence $\hat{s}$ predicted by a regression model based on NWC (panel A), VP (panel B), and H index (panel C) respectively. Results are averaged over 50 realizations of training set sampling and model training.

We predict nodal influence in real-world networks using the iterative metric based regression models. Each model uses an iterative metric set $\{\mathcal{M}_i^{(1)}, \mathcal{M}_i^{(2)}, ..., \mathcal{M}_i^{(K)}\}$ as input features. Thus, topological information of the K-hop ($K+1$-hop for VP) neighborhood of each node is used by the regression model for influence prediction. These regression models are evaluated using the evaluation metrics introduced in Section 2.2. In Figure 2.2, we show the Kendall correlations $\tau(\hat{s}, s)$ between the actual nodal influence s and the influence $\hat{s}$ predicted by a regression model as a function of K in real-world networks. As K grows, higher order iterative metrics are included, and the prediction quality increases. For all three iterative metrics, the prediction quality converges relatively fast as K increases. As shown in Figure 2.2 (A), the prediction quality of the NWC based model is already close to the highest at a small K ($K \sim 4$) and only increases marginally by choosing a $K > 4$. For example, the prediction quality when $K = 4$ reaches at least 95% of the highest prediction quality of the NWC based model. This suggests that a regression model using relatively local topological information could already achieve comparably good prediction quality as the one using more global information. This finding does not hold for the two infrastructure networks with an extremely large diameter, for which an iterative metric of higher orders (i.e., $K > 4$) is needed to achieve optimal prediction quality.

To understand why an iterative metric method achieves nearly its optimal prediction quality with a small $K \sim 4$ in all networks except for the two networks without the small-world property, we first explore the correlation $\tau(\mathscr{M}^{(k)}, s)$ between the k-th order iterative metric $\mathscr{M}^{(k)}$ and the spreading influence s. As shown in Figure 2.3 (A-C), each iterative metric $\mathscr{M}^{(k)}$ exhibits positive correlation with spreading influence for any order k, indicating that each iterative metric has certain predictive power. As k increases, the correlation $\tau(\mathscr{M}^{(k)}, s)$ increases when k is small and achieves nearly the highest correlation around $k \sim 4$, implying the high predictive power of iterative metrics of up to order 4 in those small-world networks.

Secondly, we study the convergence of the iterative metric $\mathscr{M}^{(k)}$ as the order k grows. As k increases, each centrality metric $\mathscr{M}^{(k)}$ converges to the global centrality metric $\mathscr{M}^*$. The three iterative metrics converge to three global metrics: eigenvector centrality, PageRank centrality, coreness, respectively. Figure 2.3 (D-F) shows the Kendall's correlation $\tau(\mathscr{M}^{(k)}, \mathscr{M}^*)$ between the k-th order metric $\mathscr{M}^{(k)}$ and the global metric $\mathscr{M}^*$ as a function of k for each iterative metric. For each iterative metric, $\mathscr{M}^{(k)}$ converges to $\mathscr{M}^*$ with different convergence rates in different networks. Importantly, $\mathscr{M}^{(k)}$ exhibits relatively high correlation with $\mathscr{M}^*$ at $k \sim 4$ in most networks. Hence, the predictive power of an high-order iterative metric could be inherited by a low-order iterative metric. This explains why the corresponding regression model improves in prediction quality only marginally as K increases when $K \geq 4$. Furthermore, the large correlation $\tau(h^{(k)}, h^*)$ for any k, as shown in Figure 2.3 (F), explains why the prediction quality of the regression model based on HI hardly improves when K grows, as observed in Figure 2.2 (C).

In the two infrastructure networks with a large diameter and strong community structure, iterative metrics converge relatively slowly, indicating the possibility that a large K or high-order iterative metric is needed for better prediction quality. Still, the convergence of the prediction quality $\tau(\mathscr{M}^{(k)}, s)$ is faster than that of the metric NWC $\tau(\mathscr{M}^{(k)}, \mathscr{M}^*)$. This is likely because the higher-order metric is less predictive, thus possibly less needed for the prediction, as shown in the decreasing trend of the correlation $\tau(\mathscr{M}^{(k)}, s)$ with an increasing k when k is large. The different performance of the iterative metric based model in the two infrastructure networks from the other networks as well as the weakness of using a single classical centrality to predict influence precisely in networks with community structure [41, 42] motivate us to investigate the impact of the strength of community structure on nodal influence prediction in the next section.

To gain insight into why each iterative metric $\mathscr{M}^{(k)}$ exhibits relatively high correlation with $\mathscr{M}^*$ at $k \sim 4$ in most networks, we investigate the average size of the k-hop neighborhood, i.e., the fraction of nodes that is reachable (covered) from a random node in k hops. This indicates the proportion of nodes whose information is considered in the metric $\mathscr{M}^{(k)}$. Figure 2.3 (G) shows that in most real-world networks, more than half of nodes are reachable from a random node within 4 hops. Hence, the 4-th order iterative metric possibly captures the topological information of a significant amount of nodes, supporting why $\tau(\mathscr{M}^{(k)}, \mathscr{M}^*)$ is high when $k \sim 4$. The 4-hop coverage of network *deezer EU* and the two infrastructure networks is lower than in the other networks, which is likely due to their community structure or large diameter. Correspondingly, $\tau(\mathscr{M}^{(k)}, \mathscr{M}^*)$ when $k \sim 4$ for NMC is relatively lower in these three networks.

Among all three iterative metrics, NWC achieves evidently the highest prediction quality when $K \sim 4$. This is supported by the higher correlation $\tau(s, w^{(k)})$ between the NWC centrality $w^{(k)}$ and the spreading influence s at each order k, as shown in Figure

2.3 (A-C).

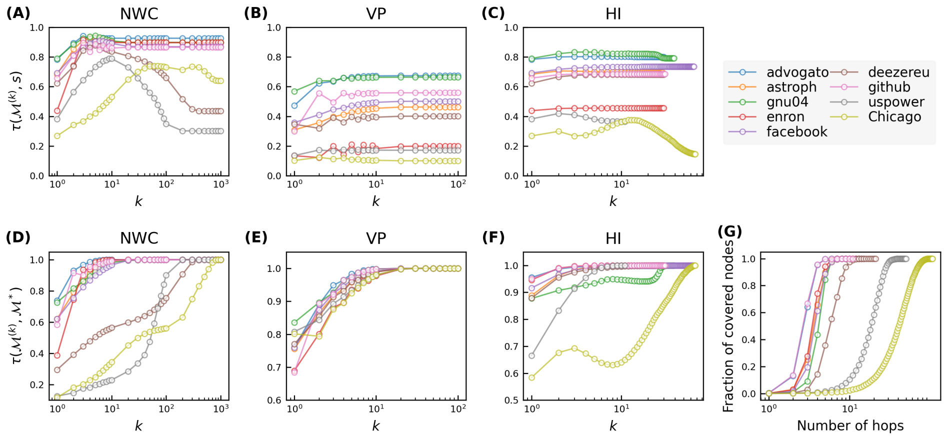

Figure 2.3: Kendall correlation between nodal spreading influence s and different orders of NWC ($w^{(k)}$, panel A), VP ($p^{(k)}$, panel B), and H index ($h^{(k)}$, panel C), and the convergence of NWC (D), VP (E), HI (F), measured by the Kendall's correlation between the iterative metric after k iterations and the corresponding global centrality metrics, as a function of iteration number k in 9 real-world networks. (G) shows the coverage, i.e. the average fraction of nodes covered by hopping step out from a node, as a function of the number of hops.

It has been found that combining local and global node centrality metrics can more accurately identify top influencers than using either local or global centralities alone [27]. Hence, we build a benchmark regression model that uses the same 7 centrality metrics (local ones like degree and global ones like closeness) as in the classification model in [27] as input features. Now, we compare the prediction quality of the proposed iterative metric based models with the benchmark model. We choose $K = 4$ for iterative metric based models. The choice of $K = 4$ corresponds the case where the iterative metric based model only uses relatively local information, which ensures the computational efficiency and reasonably good prediction quality in most networks.

Figure 2.4 shows three evaluation measures of the regression models: r^2 (left panel), Kendall correlation between the actual nodal spreading influence s and the predicted influence $\hat{s}$ of the node by a regression model (middle panel), and the recognition rate of top 10% nodes (right panel). Across all real-world networks, the prediction quality of NWC based model is evidently better than the other two iterative metric based models. In all networks except for the two infrastructure networks, NWC based model achieves prediction quality comparable to the benchmark model. The prediction quality ratio between NWC based model and the benchmark model is within the range $[91\%, 101\%]$ for any of the three evaluation measures. In those two infrastructure networks uspower and Chicago, the NWC based model with $K = 4$ performs worse than the benchmark, whereas NWC based model with a large K performs as well as the benchmark, achieving 96% to 105% of the prediction quality of the benchmark model.

Moreover, the computational complexity of NWC based model with $K = 4$ is lower

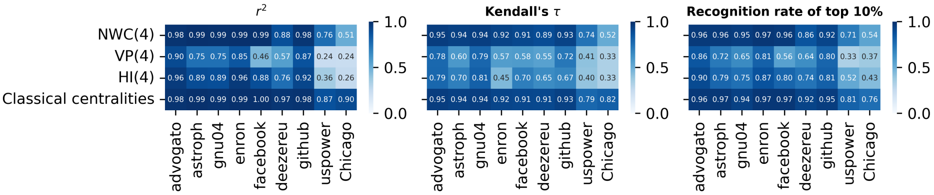

Figure 2.4: Comparison of prediction quality across different empirical networks (horizontal axis) of four prediction models based on different metrics (vertical axis): Normalized Walk Count when $K = 4$ (NWC(4)), Visiting Probability when $K = 4$ (VP(4)), H index when $K = 4$ (HI(4)), and classic centrality metrics [27]. Three panels correspond to different evaluation measures of prediction quality: r^2 (left panel), Kendall's τ (middle panel), and recognition rate of top 10% nodes (right panel), respectively. Results are averaged over 50 realizations of Random Forest Model training process.

than that of the benchmark model, which requires the computation of global centrality metrics. We summarize in Table 2.2 the computational complexity of an iterative metric of orders up to K and the 7 classical centrality metrics used in the benchmark model for all nodes. In each iteration of an iterative process, the iterative metric of each node is updated via aggregating the metrics of its 1-hop neighbors derived in the previous iteration. Thus, updating the metric for all nodes in each iteration requires $2|E|$ basic operations. The computational complexity of an iterative metric set $\{\mathcal{M}_i^{(1)}, \mathcal{M}_i^{(2)}, ..., \mathcal{M}_i^{(K)}\}$ for all nodes equals that of $\mathcal{M}_i^{(K)}$ for all nodes, which is $\mathcal{O}(K \cdot |E|)$. Hence, a relatively small K facilitates the application of iterative metric based method in large networks. In contrast, the global metrics used in the benchmark model, such as closeness centrality, have a higher complexity.

Iterative metric	degree, neighborhood,two-hop neighborhood	coreness	eigenvector, PageRank	closeness
$\mathcal{O}(K \cdot \lvert E \rvert)$	$\mathcal{O}(\lvert E \rvert)$	$\mathcal{O}(\lvert E \rvert)$	$\mathcal{O}(K^* \cdot \lvert E \rvert)$	$\mathcal{O}(\lvert V \rvert \lvert E \rvert)$

Table 2.2: Comparison of the computational complexity of different nodal metrics for all nodes in a network: an iterative metric set $\{\mathcal{M}^{(1)}, ..., \mathcal{M}^{(K)}\}$ and classical centrality metrics used in the benchmark model. Neighborhood stands for the sum of degrees of direct neighbors, and two-hop neighborhood are the sum of degrees of nodes that are two hops away. K^* is the number of iterations at which the iterative process to compute the centrality metric converges.

2.3.3. PREDICTION ON NETWORKS WITH COMMUNITIES

Community structure has been observed in many real-world networks [43], where nodes within a community are densely connected while nodes from different communities have fewer connections. The existence of communities affects significantly the spreading process unfolding on a network [44, 45] and has been ignored in most centrality metrics used to predict nodal influence [42, 46]. Here we evaluate the performance of our influence prediction methods in networks with community structures and investigate how community structure affects the prediction quality. To this end, we adopt the Lancichinetti–Fortunato–Radicchi (LFR) model [29] to generate networks with power-law degree distribution and community size distribution, as observed in real-world networks. One advantage of LFR model is that the strength of the community structure in the generated networks can be changed via tuning its parameters. We use LFR model to generate networks with the following properties: network size $N = 1000$ and $N = 10000$ respectively, the exponent of the power-law degree distribution $\tau_1 = 2$, and exponent of the power-law community size distribution $\tau_2 = 3$, the average degree $\langle k \rangle = 10$, the maximum degree $d_{max} = \sqrt{10N}/2$, the range of community sizes $[50, \sqrt{10N}]$. The mixing parameter μ represents the fraction of inter-community links of a node. When $\mu = 0$, the generated networks have the strongest community structure, with communities being disjoint from each other. The model with $\mu = 1$ generates networks where all links fall between different clusters. When $\mu > 0.5$, the community structure is not evident anymore [29]. We set $\mu = [0.02, 0.05, 0.1, 0.2, 0.3, 0.4]$, thus six networks with different strength of communities are generated. We will focus on the results for $N = 1000$, since results for $N = 10000$ (as shown in the Appendix) lead to the same observation. The generated networks vary in network properties such as diameter and modularity, as shown in Table 2.3 and Table 2.4.

μ	Diameter	Q	λ_c
0.02	10	0.924	0.090
0.05	6	0.872	0.080
0.1	5	0.608	0.070
0.2	5	0.632	0.070
0.3	5	0.386	0.070
0.4	5	0.453	0.070

Table 2.3: Basic properties of networks generated by LFR model with different mixing parameter μ and $N = 1000$: network diameter, the modularity Q, epidemic threshold λ_C of the SIR process on the network.

We first evaluate our iterative metric based models in predicting nodal influence in LFR networks when the effective infection rate of the SIR model is around epidemic threshold, i.e., $\lambda = \lambda_c$. Figure 2.5 (A-C) show Kendall correlations $\tau(\hat{s}, s)$ between the nodal spreading influence s and the prediction $\hat{s}$ by a regression model based on an iterative metric set $\{\mathscr{M}^{(1)}, \mathscr{M}^{(2)}, ..., \mathscr{M}^{(K)}\}$, as a function of K in LFR networks. Like what we observed in real-world networks, the prediction quality increases as K increases. Notably, the prediction quality only improves marginally when choosing a $K > 4$. This can be understood by the correlation $\tau(\mathscr{M}^{(k)}, s)$ between $\mathscr{M}^{(k)}$ and nodal influence s, which is shown in Figure 2.6 (A-C). As k increases up to $k \sim 4$, the correlation $\tau(\mathscr{M}^{(k)}, s)$ increases. As k increases further, the correlation tends to decrease. This decreasing trend is more evident in networks with more evident community structure, but not observed in real-world networks that have a relatively small diameter and modularity as shown in Figure 2.3. This suggests that high-order ($k > 4$) iterative metrics are less predictive than an iterative metric of an order around $k \sim 4$, thus less needed to predict nodal influence in networks with a higher modularity. Furthermore, we explore the convergence of an iterative metric $\mathcal{M}^{(k)}$ as k increases. Figure 2.6 (D-F) show the Kendall's correlation $\tau(\mathcal{M}^{(k)}, \mathcal{M}^*)$ as a function of k for the three iterative metrics, respectively. For NWC, the correlation tends to be lower when $k \sim 4$ as the mixing parameter μ gets smaller or equivalently in network with more evident community structure. In networks with strong community structure, NWC converges relatively slowly. Still, the prediction quality of the regression models in these networks is close to optimal when $K \sim 4$, since the higher order metric is less predictive. This is also in line with the intuition that in networks with strong community structure and when the infection rate is around the critical epidemic threshold, nodal influence is supposed to be mainly determined by nodal property derived within or around the community that the node belongs to.

Figure 2.6 (G) shows the average fraction of nodes that are reachable (covered) from a randomly chosen node within k hops, i.e., the so called coverage, as a function of k. In networks with strong community structure (small μ), the coverage and $\tau(\mathcal{M}^{(k)}, \mathcal{M}^*)$ when $k \sim 4$ tend to be small. In such networks, an order $k \sim 4$ iterative metric encodes topological information of a small fraction of nodes, which explains partially the weak correlation $\tau(\mathcal{M}^{(k)}, \mathcal{M}^*)$ when $k \sim 4$.

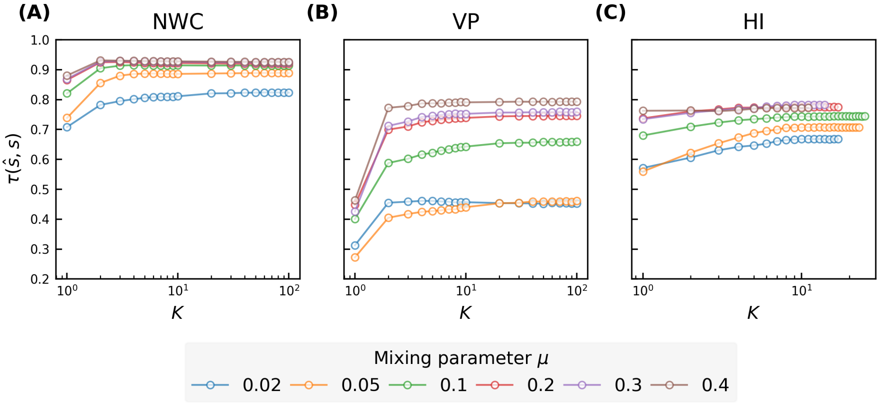

Figure 2.5: Kendall correlation between nodal spreading influence $\hat{s}$ predicted by different numbers of iterative metrics as features and nodal spreading influence given by SIR simulations of NWC (A), VP (B), and H index (C). Results are averaged over 50 realizations of training set sampling and model training.

Now we compare the prediction quality of iterative metric based models (when $K = 4$) with the benchmark model in LFR networks via the same three evaluation measures as in real-world networks. Figure 2.7 shows that NWC based model with $K = 4$ performs comparably as (mostly slightly better than) the benchmark model, the prediction quality ratio between NWC based model and the benchmark model ranges from 95% to 106%. Among the three iterative metric based models, NWC based model performs the best whereas VP based model performs the worst. As the strength of community structure grows, all models perform worse. This can be explained by the small (large) correlation $\tau(\mathcal{M}^{(k)}, s)$ in networks with a strong (weak) community structure, as shown in Figure 2.6

(A-C). The same has also been observed in real-world networks. As shown in Figure 2.4, both the NWC based model and the benchmark model perform the worst in the two infrastructure networks that have the stronger community structure than the other considered real-world networks. In the two infrastructure networks, the correlation $\tau(\mathcal{M}^{(k)}, s)$ is also weaker (see Figure 2.3).

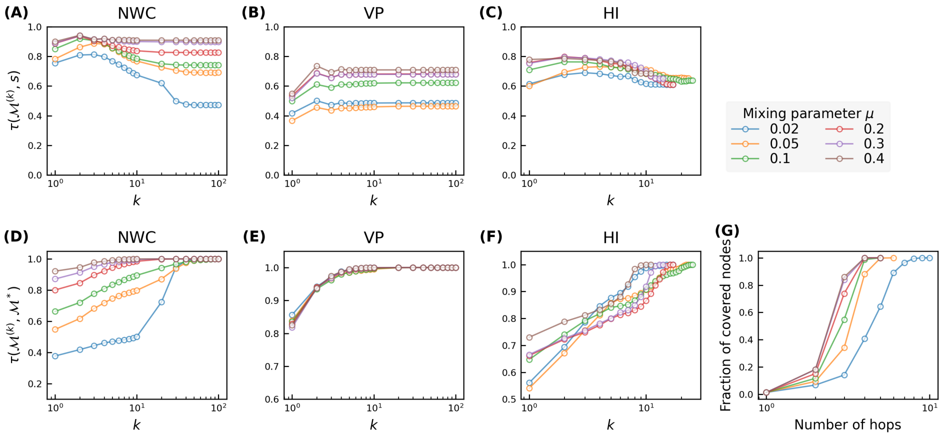

Figure 2.6: Kendall correlation between nodal spreading influence s and different orders of NWC ($w^{(k)}$, panel A), VP ($p^{(k)}$, panel B), and H index ($h^{(j)}$, panel C), and the convergence of NWC (D), VP (E), HI (F), measured by the Kendall's correlation between the iterative metric after k iterations and the corresponding global centrality metrics, as a function of iteration number k in Lancichinetti–Fortunato–Radicchi (LFR) networks with different $\mu = 0.02, 0.05, 0.1, 0.2, 0.3, 0.4$. (G) shows the coverage, i.e. the average fraction of nodes covered by hopping step out from a node, as a function of the number of hops.

2.3.4. Prediction of nodal spreading influence near epidemic threshold

So far, we have focused on the influence prediction problem, where the influence is defined for the SIR epidemic spreading process with $\lambda = \lambda_c$. It has been shown that the change of parameters in the epidemic spreading can lead to different rankings of nodes according to their influences [23, 47, 48]. Hence, we evaluate the average prediction quality of a regression model over all the networks except for the two infrastructure networks, at various effective infection rates around the epidemic threshold λ_c. Figure 2.8 (top panel) shows that NWC outperforms VP and HI, as λ varies from $0.5 \cdot \lambda_c$ to $2.0 \cdot \lambda_c$. NWC based model with $K = 4$ and the benchmark model show comparable prediction quality. Their prediction quality is less sensitive to the effective infection rate λ. In the two infrastructure networks (Figure 2.8 bottom panel), the NWC based model with $K = 4$ exhibits lower prediction quality than the benchmark at different effective infection rates except that they perform similarly at $\lambda = 0.5 \cdot \lambda_c$, when the SIR spreading is relatively local.

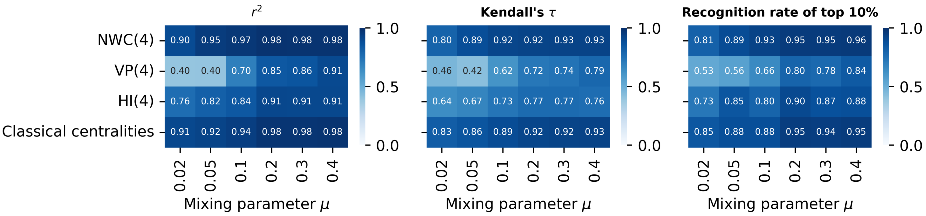

Figure 2.7: Prediction performance on model networks generated with LFR model with varying mixing parameter μ (horizontal axis) of five sets of metrics (vertical axis): Normalized Walk Count when $K = 4$ (NWC(4)), Visiting Probability when $K = 4$ (VP(4)), H index when $K = 4$ (HI(4)), and classical centralities. Three panels correspond to different evaluation measures of predictive models: r^2 (left panel), Kendall τ (middle panel) and recognition rate of top 10% nodes (right panel), respectively. Results are averaged over 50 realizations of training process of Random Forest Model.

2.4. DISCUSSION AND FUTURE WORK

In summary, we explore to what extent local and global topological information of a node is needed for the prediction of nodal spreading influence and whether relatively local topological information around a node is sufficient for the prediction. We propose to predict nodal influence by an iterative metric set derived from an iterative process. Three iterative metrics are considered: Normalized Walk Counts (NWC), Visiting Probability (VP), and H index (HI), which converge to eigenvector centrality, PageRank, and H index, respectively. The regression model using an iterative metric set as input features is trained on a fraction of nodes whose influence is known and is used to predict the nodal influence of the remaining nodes. We evaluate and interpret the performance of these three iterative metric based models in predicting nodal influence in SIR spreading processes with diverse effective infection rates around the epidemic threshold, on both real-world networks and synthetic networks with different strength of community structure. We find that the prediction quality of each iterative metric based model converges to its optimal when the iterative metric set of relatively low orders (up to order 4) are included and increases only marginally when further increasing K. This is explained via the correlation between an iterative metric of order k and nodal influence and the fast convergence of each iterative metric. The prediction quality of the best performing iterative metric set (NWC) with $K = 4$ is comparable with the benchmark method that combines seven centrality metrics. In two spatially embedded networks with an extremely large diameter and modularity, however, iterative metric of higher orders, thus a large K, is needed to achieve comparable predict quality as the benchmark. These findings suggest that the NWC metric of relatively low orders contain sufficient information to predict nodal influence reasonably well in networks with the small-world property, whereas its computation complexity is lower than that of the global centrality metrics needed by the benchmark model. In these networks, the NWC metric has almost the highest correlation with nodal influence when $k \approx 4$ in most networks, indicating that a node with more distinct 4-hop walks starting from the node tends to be more influential. However, the interpretability of the iterative metric-based regression model is limited by

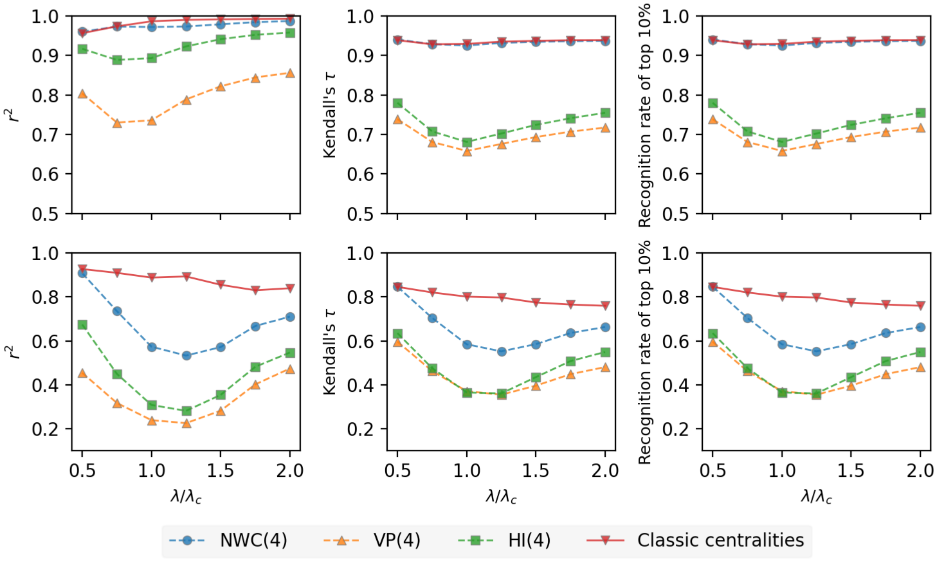

Figure 2.8: Average prediction quality over all considered real-world networks (shown in Table 2.1) excluding the two spatially embedded networks (top panels) and over the two spatially embedded networks (bottom panels) as a function of λ/λ_c of 4 different metric sets: Normalized Walk Counts (NWC), Visiting Probability (VP), H index (HI), and 7 node centralities [27]. Three columns correspond to different evaluation measures of predictive models: r^2 (left panel), kendall's τ (middle panel), and recognition rate of top 10% nodes (right panel), respectively. Results are averaged over 50 realizations of training process of Random Forest Model.

the strong correlation among the iterative metric of different orders. Nodes with what kind of combination of low order the iterative metrics are more influential remains an interesting question.

This study has several limitations that call for further exploration. Firstly, we observe the trend that a larger K is needed for the iterative metric based method to perform close to its optimal in networks with a significant large diameter. It is interesting to explore the minimal K needed for the NWC based model to perform at least, for example, 95% of the optimal performance of the model in relation to the diameter of the network. Secondly, the diameter and strength of community structure are possibly correlated in real-world networks and network models. We have observed the influence of community structure or diameter on the prediction quality of NWC based model and the benchmark model. An open question is how the diameter influences the prediction quality while the community strength is fixed. For both objectives, network models with a controllable diameter and more real-world networks, especially those without the small-world property are needed. Thirdly, we confine ourselves to the SIR spreading process on a static network. However, in many scenarios, both the spreading process and the underlying topology can be more complicated. Our proposed method can be extended to explore its capability of predicting nodal influence defined in such more complex context using local network information.

2.5. Appendix

2.5.1. Computation of an iterative metric set and complexity analysis

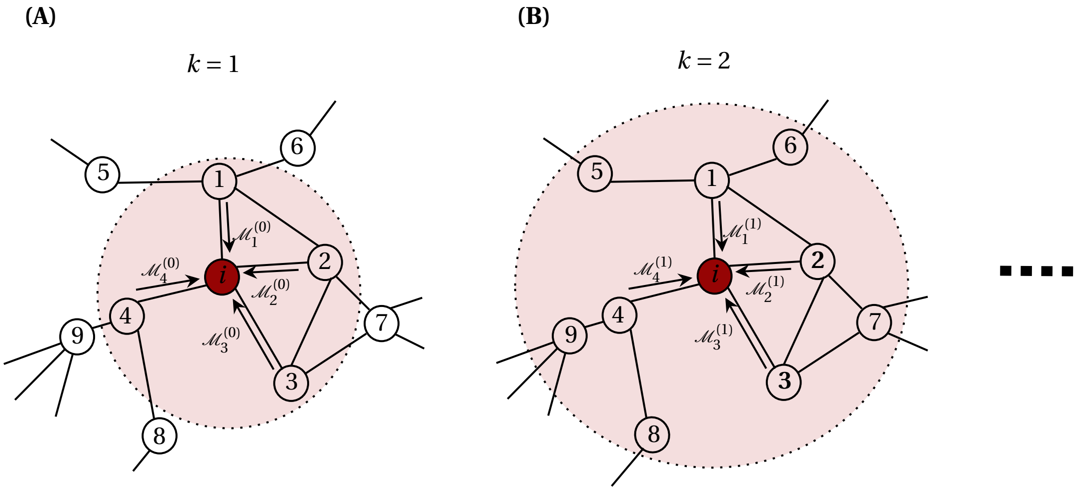

Figure 2.9: An illustration of the neighborhood considered to compute an iterative metric of order k at any node i.

Here, we exemplify how gradually more network information around a node is incorporated in the computation an iterative metric of a higher order. In each iteration k of an iterative process, the computation of an iterative metric $\mathscr{M}_i^{(k)}$ of node i is derived via aggregating the metrics of its 1-hop neighbors derived in the previous iteration $k-1$. Take the example of NWC computation in the network shown in Figure 2.9. The iterative process of NWC is governed by $w^{(k)} = Aw^{(k-1)}/|Aw^{(k-1)}|$. Initially, $\mathscr{M}_i^{(0)} = 1/\sqrt{N}$ for any node i, which encode no topological information. In iteration $k = 1$, the metric $\mathscr{M}_i^{(1)}$ is derived as the sum of the metrics of its neighbors from iteration 0, i.e., $\mathscr{M}_i^{(1)} = \sum_{j \in \mathscr{N}(i)} \mathscr{M}_j^{(0)}$, where node i's 1-hop neighbor set $\mathscr{N}(i) = \{1,2,3,4\}$. Thus, $\mathscr{M}_i^{(1)}$ encodes the information of 1-hop neighborhood of node i as illustrated in the pink area in Figure 2.9 (A). In iteration $k = 2$, the metric $\mathscr{M}_i^{(2)}$ is derived as $\mathscr{M}_i^{(2)} = \sum_{j \in \mathscr{N}(i)} \mathscr{M}_j^{(1)} = 4\mathscr{M}_i^{(0)} + \mathscr{M}_5^{(0)} + \mathscr{M}_6^{(0)} + 2\mathscr{M}_7^{(0)} + \mathscr{M}_8^{(0)} + \mathscr{M}_9^{(0)}$, which encodes up to 2-hop neighborhood information of node i as illustrated in the pink area in Figure 2.9 (B). Consequently, for any iteration k, the metric $\mathscr{M}_i^{(k)}$ of node i encodes information of up to k-hop neighborhood.

2.5.2. Training of regression models

Given a network G, we train a Random Forest regression (RFR) model based on the $q = 10\%$ of nodes whose nodal influences are known and use the trained model to predict the influence of the rest nodes. We adopt `RandomForestRegressor` in the Python library scikit-learn to implement our regression models. The hyperparameters are determined via 5-fold cross-validation on the training set, where the number of trees considered in RFR, n_estimators, is tuned in the range $[50, 1000]$. The final RFR with the identified hyperparameters is trained on the training set.

Performance of all iterative based models in nodal influence prediction is the average over 50 realizations of training set sampling and model training.

2.5.3. RESULTS FOR PERFORMANCE OF ITERATIVE METRIC BASED MODELS

Figure 2.10 and Figure 2.11 show the Mean Squared Errors, r^2, and Recognition rate of top 10% nodes of the three iterative metric based RFR models in all considered real-world networks and LFR networks, respectively. Figure 2.12 and Figure 2.13 shows the results of iterative metric based Ridge regression models in all considered real-world networks.

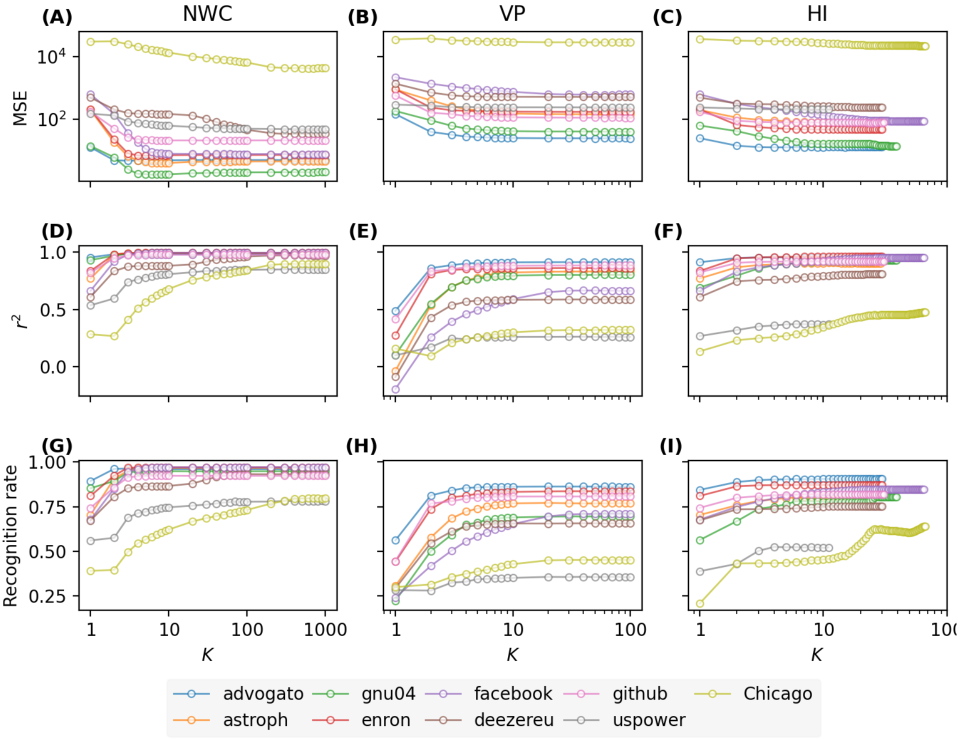

Figure 2.10: Mean Squared Errors, r^2, and Recognition rate of top 10% nodes of the Normalized Walk Count, Visiting Probability, and H index based models, respectively, in nodal influence prediction as a function of the size K of an iterative metric set in each of the 9 real-world networks.

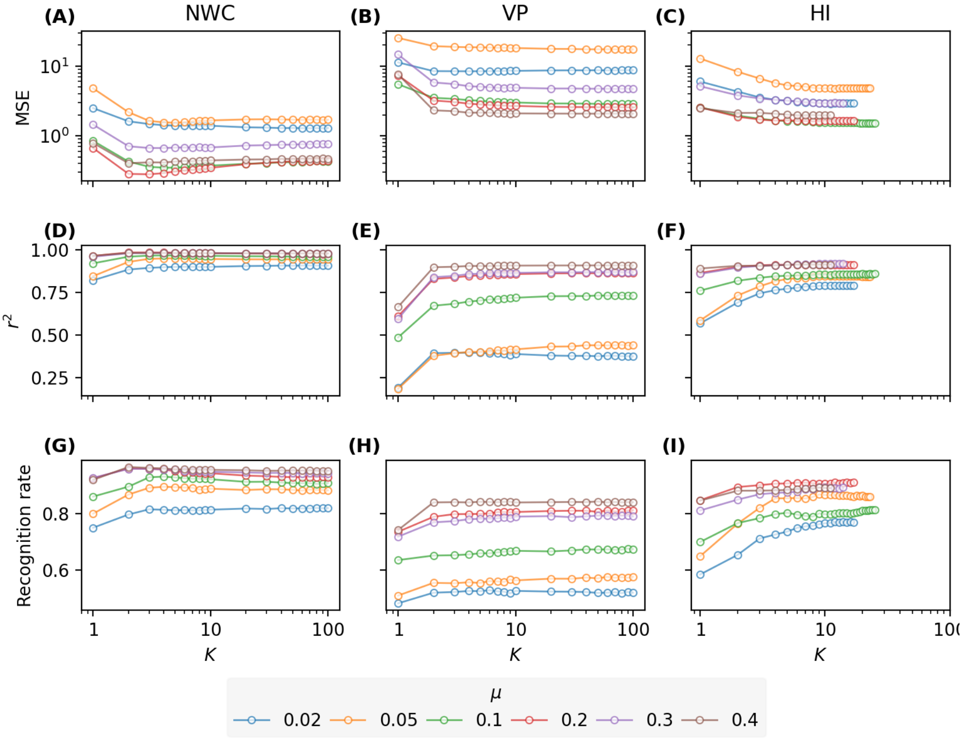

Figure 2.11: Mean Squared Errors, r^2, and Recognition rate of top 10% nodes of the Normalized Walk Count, Visiting Probability, and H index based models, respectively, in nodal influence prediction as a function of the size K of an iterative metric set in LFR networks with 1000 nodes, where the mixing parameter $\mu = [0.02, 0.05, 0.1, 0.2, 0.3, 0.4]$.

μ	Diameter	Q	λ_c
0.02	11	0.971	0.060
0.05	8	0.933	0.050
0.1	7	0.859	0.050
0.2	6	0.728	0.050
0.3	6	0.614	0.040
0.4	6	0.497	0.040

Table 2.4: Basic properties of networks generated by LFR model with 10000 nodes using different mixing parameter μ: network diameter, the modularity Q, epidemic threshold λ_c of the SIR process on the network.

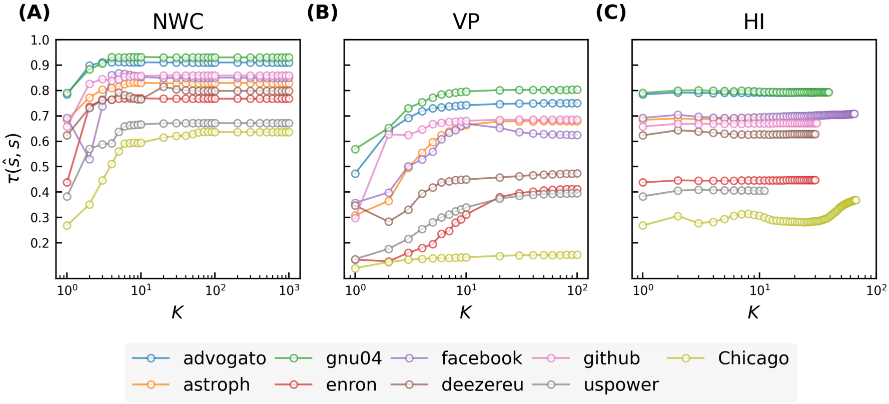

Figure 2.12: Kendall correlation between the actualy nodal influence s and nodal spreading influence $\hat{s}$ predicted by each iterative metric (NWC, VP or HI) based Ridge regression model as a function of the size K of an iterative metric set in each of the 9 real-world networks.

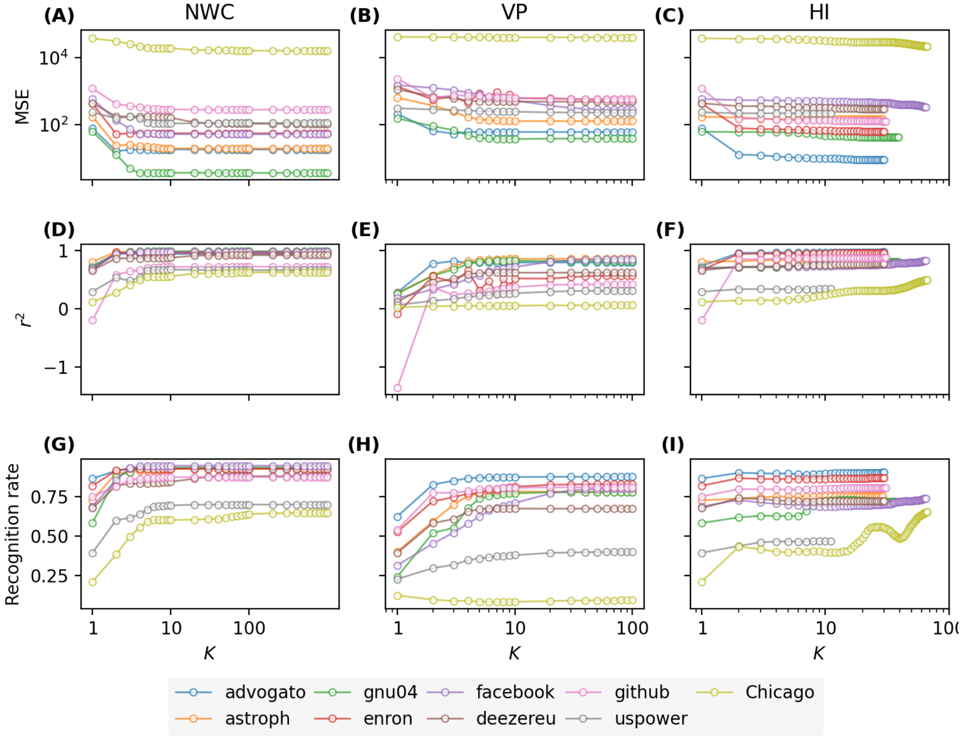

Figure 2.13: Mean Squared Errors, r^2, and Recognition rate of top 10% nodes of the Normalized Walk Count, Visiting Probability, and H index based Ridge regression models, respectively, in nodal influence prediction as a function of the size K of an iterative metric set in each of the 9 real-world networks.

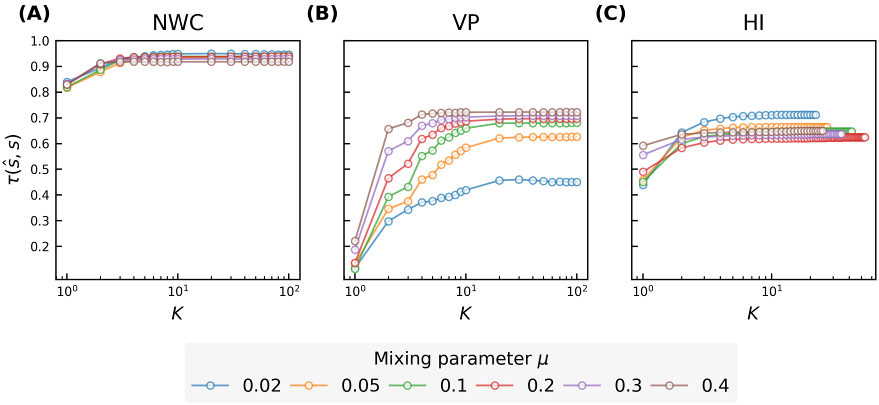

Figure 2.14: Kendall correlation between the actualy nodal influence s and nodal spreading influence $\hat{s}$ predicted by each iterative metric (NWC, VP or HI) based Random Forest regression model as a function of the size K of an iterative metric set in LFR network with 10000 nodes.

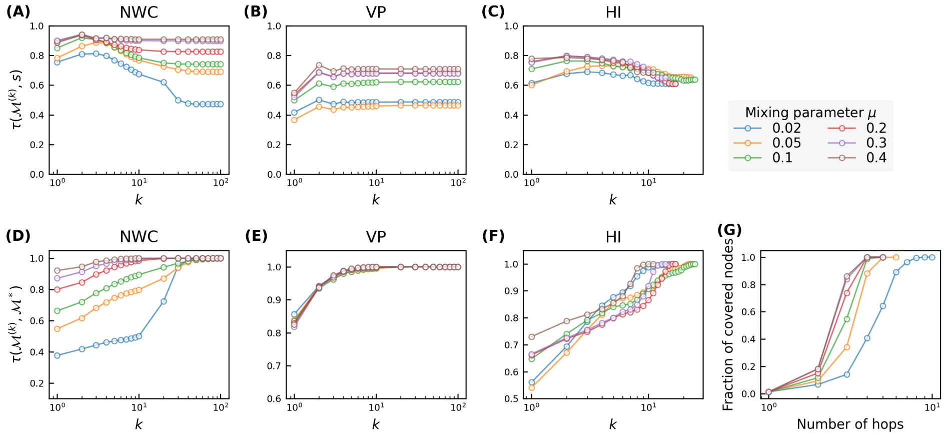

Figure 2.15: Kendall correlation between nodal spreading influence s and different orders of NWC ($w^{(k)}$, panel A), VP ($p^{(k)}$, panel B), and H index ($h^{(j)}$, panel C), and the convergence of NWC (D), VP (E), HI (F), measured by the Kendall's correlation between the iterative metric after k iterations and the corresponding global centrality metrics, as a function of iteration number k in LFR networks with 10000 nodes and different $\mu = 0.02, 0.05, 0.1, 0.2, 0.3, 0.4$. (G) shows the coverage, i.e. the average fraction of nodes covered by hopping step out from a node, as a function of the number of hops.

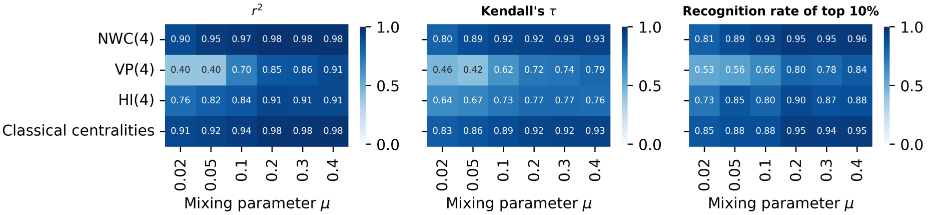

Figure 2.16: Prediction performance comparison in LFR networks with 10000 nodes of four sets of metrics (vertical axis): Normalized Walk Count when $K = 4$ (NWC(4)), Visiting Probability when $K = 4$ (VP(4)), H index when $K = 4$ (HI(4)), and seven centralities. Three panels correspond to different evaluation measures of predictive models: r^2 (left panel), Kendall's τ (middle panel), and recognition rate of top 10% nodes (right panel), respectively. The prediction quality ratio between NWC based model and the benchmark model ranges from 95% to 106% for all evaluation measures. Results are averaged over 50 realizations of Random Forest Model training process.

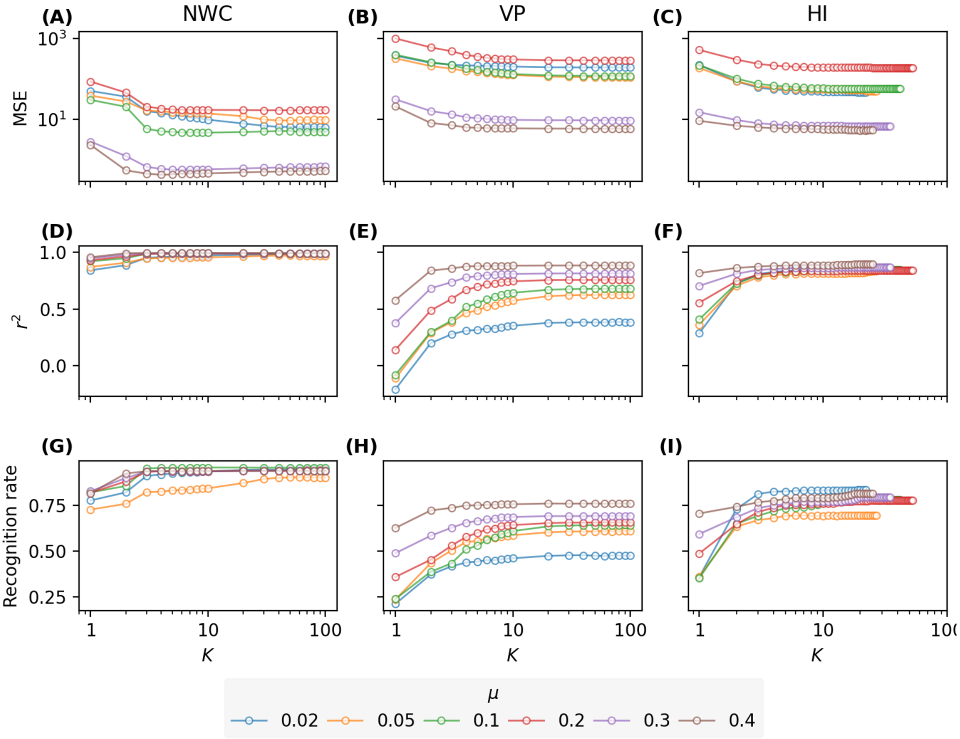

Figure 2.17: Mean Squared Errors, r^2, and Recognition rate of top 10% nodes of the Normalized Walk Count, Visiting Probability, and H index based models, respectively, in nodal influence prediction as a function of the size K of an iterative metric set in LFR networks with 10000 nodes, where the mixing parameter $\mu = [0.02, 0.05, 0.1, 0.2, 0.3, 0.4]$.

References

1. Newman, M. *Networks* (Oxford university press, 2018).
2. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438,** 355–359 (2005).
3. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Reviews of Modern Physics* **87,** 925–979 (2015).
4. Hu, Y. *et al.* Local structure can identify and quantify influential global spreaders in large scale social networks. *Proceedings of the National Academy of Sciences* **115,** 7468–7472 (2018).
5. Woolhouse, M. E. *et al.* Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proceedings of the National Academy of Sciences* **94,** 338–342 (1997).
6. Pei, S. & Makse, H. A. Spreading dynamics in complex networks. *Journal of Statistical Mechanics: Theory and Experiment* **2013,** P12002 (2013).
7. Chen, X. & Wang, N. Rumor spreading model considering rumor credibility, correlation and crowd classification based on personality. *Scientific Reports* **10,** 1–15 (2020).
8. Bovet, A. & Makse, H. A. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* **10,** 7 (2019).
9. Watts, D. J. & Dodds, P. S. Influentials, networks, and public opinion formation. *Journal of Consumer Research* **34,** 441–458 (2007).
10. Leskovec, J., Adamic, L. A. & Huberman, B. A. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)* **1,** 5–es (2007).
11. Kempe, D., Kleinberg, J. & Tardos, É. *Influential nodes in a diffusion model for social networks* in *International Colloquium on Automata, Languages, and Programming* (2005), 1127–1138.
12. Zhou, Y.-B., Lü, L. & Li, M. Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity. *New Journal of Physics* **14,** 033033 (2012).
13. Zhan, X.-X., Li, Z., Masuda, N., Holme, P. & Wang, H. Susceptible-infected-spreading-based network embedding in static and temporal networks. *EPJ Data Science* **9,** 30 (2020).
14. Wang, J., Xu, S., Mariani, M. S. & Lü, L. The local structure of citation networks uncovers expert-selected milestone papers. *Journal of Informetrics* **15,** 101220 (2021).
15. Zhang, S., Medo, M., Lü, L. & Mariani, M. S. The long-term impact of ranking algorithms in growing networks. *Information Sciences* **488,** 257–271 (2019).
16. Kitsak, M. *et al.* Identification of influential spreaders in complex networks. *Nature Physics* **6,** 888–893 (2010).
17. Lü, L. *et al.* Vital nodes identification in complex networks. *Physics Reports* **650,** 1–63 (2016).

18. Li, C., Li, Q., Van Mieghem, P., Stanley, H. E. & Wang, H. Correlation between centrality metrics and their application to the opinion model. *The European Physical Journal B* **88,** 1–13 (2015).

19. Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C. & Zhou, T. Identifying influential nodes in complex networks. *Physica A: Statistical mechanics and its applications* **391,** 1777–1787 (2012).

20. Lawyer, G. Understanding the influence of all nodes in a network. *Scientific Reports* **5,** 1–9 (2015).

21. Klemm, K., Serrano, M. Á., Eguíluz, V. M. & Miguel, M. S. A measure of individual role in collective dynamics. *Scientific reports* **2,** 292 (2012).

22. Maharani, W., Gozali, A. A., *et al. Degree centrality and eigenvector centrality in twitter* in *2014 8th International Conference on Telecommunication Systems Services and Applications (TSSA)* (2014), 1–5.

23. Liu, J.-G., Lin, J.-H., Guo, Q. & Zhou, T. Locating influential nodes via dynamics-sensitive centrality. *Scientific Reports* **6,** 1–8 (2016).

24. Li, Z. & Huang, X. Identifying influential spreaders by gravity model considering multi-characteristics of nodes. *Scientific Reports* **12,** 9879 (2022).

25. Madotto, A. & Liu, J. Super-spreader identification using meta-centrality. *Scientific Reports* **6,** 38994 (2016).

26. Pei, S., Muchnik, L., Andrade Jr, J. S., Zheng, Z. & Makse, H. A. Searching for superspreaders of information in real-world social media. *Scientific Reports* **4,** 1–12 (2014).

27. Bucur, D. Top influencers can be identified universally by combining classical centralities. *Scientific Reports* **10,** 1–14 (2020).

28. Wang, H., Hernandez, J. M. & Van Mieghem, P. Betweenness centrality in a weighted network. *Physical Review E* **77,** 046105 (2008).

29. Lancichinetti, A., Fortunato, S. & Radicchi, F. Benchmark graphs for testing community detection algorithms. *Physical Review E* **78,** 046110 (2008).

30. Bartolucci, S., Caccioli, F., Caravelli, F. & Vivo, P. Ranking influential nodes in networks from aggregate local information. *Physical Review Research* **5,** 033123 (2023).

31. Page, L., Brin, S., Motwani, R. & Winograd, T. *The PageRank citation ranking: Bringing order to the web.* tech. rep. (Stanford InfoLab, 1999).

32. Lü, L., Zhou, T., Zhang, Q.-M. & Stanley, H. E. The H-index of a network node and its relation to degree and coreness. *Nature Communications* **7,** 1–7 (2016).

33. Kiss, I. Z., Miller, J. C., Simon, P. L., *et al.* Mathematics of epidemics on networks. *Cham: Springer* **598** (2017).

34. Shu, P., Wang, W., Tang, M. & Do, Y. Numerical identification of epidemic thresholds for susceptible-infected-recovered model on finite-size networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **25,** 063104 (2015).

35. Björck, Å. *et al. Numerical methods in matrix computations* (Springer, 2015).

36. Gleich, D. F. PageRank beyond the Web. *SIAM Review* **57,** 321–363 (2015).

37. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. K-core organization of complex networks. *Physical Review Letters* **96,** 040601 (2006).

38. Kunegis, J. The koblenz network collection. *URL: http://konect. uni-koblenz. de/(accessed 16.04. 23)* (2020).

39. Kunegis, J. *Konect: the koblenz network collection* in *Proceedings of the 22nd International Conference on World Wide Web* (2013), 1343–1350.

40. Kendall, M. G. The treatment of ties in ranking problems. *Biometrika* **33,** 239–251 (1945).

41. Ghalmane, Z., Cherifi, C., Cherifi, H. & Hassouni, M. E. Centrality in complex networks with overlapping community structure. *Scientific Reports* **9,** 1–29 (2019).

42. Rajeh, S., Savonnet, M., Leclercq, E. & Cherifi, H. Characterizing the interactions between classical and community-aware centrality measures in complex networks. *Scientific Reports* **11,** 10088 (2021).

43. Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Physics Reports* **659,** 1–44 (2016).

44. Saxena, R., Kaur, S. & Bhatnagar, V. Social centrality using network hierarchy and community structure. *Data Mining and Knowledge Discovery* **32,** 1421–1443 (2018).

45. Kumar, M., Singh, A. & Cherifi, H. *An efficient immunization strategy using overlapping nodes and its neighborhoods* in *Companion Proceedings of the The Web Conference 2018* (2018), 1269–1275.

46. Costantini, L., Sciarra, C., Ridolfi, L. & Laio, F. Measuring node centrality when local and global measures overlap. *Physical Review E* **105,** 044317 (2022).

47. Šikić, M., Lančić, A., Antulov-Fantulin, N. & Štefančić, H. Epidemic centrality—is there an underestimated epidemic impact of network peripheral nodes? *The European Physical Journal B* **86,** 1–13 (2013).

48. Qu, B., Li, C., Van Mieghem, P. & Wang, H. Ranking of nodal infection probability in susceptible-infected-susceptible epidemic. *Scientific Reports* **7,** 9233 (2017).

3

Mitigate SIR Epidemic Spreading via Contact Blocking in Temporal Networks

Progress has been made in how to suppress epidemic spreading on temporal networks via blocking all contacts of targeted nodes or node pairs. In this chapter, we develop contact blocking strategies that remove a fraction of contacts from a temporal (time evolving) human contact network to mitigate the spread of a Susceptible-Infected-Recovered (SIR) epidemic. We define the probability that a contact $c(i, j, t)$ is removed as a function of a given centrality metric of the corresponding link $l(i, j)$ in the aggregated network and the time t of the contact. The aggregated network captures the number of contacts between each node pair. A set of 12 link centrality metrics have been proposed and each centrality metric leads to a unique contact removal strategy. These strategies together with a baseline strategy (random removal) are evaluated in empirical contact networks via the average prevalence, the peak prevalence and the time to reach the peak prevalence. We find that the epidemic spreading can be mitigated the best when contacts between node pairs that have fewer contacts and early contacts are more likely to be removed. A strategy tends to perform better when the average number contacts removed from each node pair varies less. The aggregated pruned network resulted from the best contact removal strategy tends to have a large largest eigenvalue, a large modularity and probably a small largest connected component size.

3.1. INTRODUCTION

Networks, such as physical contact networks and online social networks, facilitate the spread of epidemics and information. The study of epidemic spreading first assumed the topology of networks to be static [1, 2], while many real-world networks are not static as nodes and links can appear and disappear over time, thus can be better represented as temporal networks [3]. For example, human contact networks such as face-to-face contact networks [4] are temporal networks, which can be described by a sequence of contacts (or temporal links) between pairs of individuals occurring at discrete time steps. The increasing availability of network data with temporal information has fostered research on how the temporal aspect of networks can affect dynamic processes such as the spreading of epidemics [5, 6] and information [7] on temporal networks. Epidemic/information spreading can be mitigated via reducing physical contacts. Covid-19 measures like curfew, working at home, social distancing all aim to block physical contacts. These measures treat at least a subgroup of the population in the same way. In this work, we address the further question of how to mitigate the epidemic spreading more effectively via selecting the contacts to block heterogeneously and strategically. We propose to develop contact removal strategies utilizing the network properties of contacts.

We consider real-world physical contact networks, where only the connection between nodes evolves (appears when there is a contact and disappears) over time whereas the nature/type of nodes and contacts do not change . In this case, a temporal network observed within a time window $[0, T]$ can be represented by $\mathscr{G} = (\mathscr{N}, \mathscr{C})$, where $\mathscr{N}$ is the node set observed within $[0, T]$, size $N = |\mathscr{N}|$ is the number of nodes in the network, $\mathscr{C} = \{c(i, j, t), t \in [0, T], i, j \in \mathscr{N}\}$ is the set of contacts between pairs of nodes in $\mathscr{N}$, with contact (i, j, t) representing the interaction between node i and node j at time step t. A contact $c(i, j, t)$, also called a temporal link, describes interaction/connection between node i and j at a specific time t. A node without any contact at time t can be regarded as inactive or not observed at that time step. We confine ourselves to the Susceptible-

Infected-Recovered (SIR) epidemic spreading model [1] on a temporal network instead of more realistic spreading processes: Initially at $t = 0$, a seed node is selected to be infected whereas all the other nodes are susceptible; When a contact happens between an infected node and a susceptible node at any time step, the susceptible node becomes infected with a probability β; Each infected node becomes recovered with a probability γ at each time step. A recovered node will neither be infected nor infect any other node. The contacts to block will be selected based on the (time) aggregated network $\mathscr{G}_{\mathscr{W}}$ of the temporal network $\mathscr{G}$. Aggregated network represented as $\mathscr{G}_{\mathscr{W}} = (\mathscr{N}, \mathscr{L})$ is a weighted network with the same node set $\mathscr{N}$ as temporal network $\mathscr{G}$, $\mathscr{L}$ is the set of weighted links, two nodes i and j in $\mathscr{G}_{\mathscr{W}}$ are connected by a link $l(i,j)$ if they have at least one contact in temporal network $\mathscr{G}$ and link $l(i,j)$ is associated with a weight recording the number of contacts in $\mathscr{G}$ between the two nodes. In the rest of this chapter, links refer to the links in the aggregated network, and contacts will not be called temporal links anymore to avoid confusion. Contacts between two nodes i and j can be regarded as the corresponding link $l(i,j)$ in the aggregated network activated at specific time steps.

The objective is to mitigate the epidemic spreading via blocking a given percentage ϕ of contacts, selected based on the aggregated network. The fraction ϕ of contacts removed corresponds to the cost of the mitigation. To launch a contact removal intervention during the time window $[0, T]$, the information of the aggregated network of the temporal network $\mathscr{G}$ observed in $[0, T]$ needs to be known at $T = 0$. Such aggregated network is assumed to be given in our work, whereas in practice, it can be estimated based on the temporal network observed before $T = 0$. Predicting the aggregated network is more feasible compared to predicting the temporal network in $[0, T]$. The latter, i.e. long-term prediction of time specific and possibly noisy contacts challenges machine learning approaches that target at short-term predictions. Hence, we focus on the development of contact removal strategies based on the aggregated network, instead of the complete temporal network information which is difficult to obtain.

We propose probabilistic contact removal strategies. Specifically, the probability that a contact $c(i,j,t)$ is removed is a generic function of a centrality metric [8] of link $l(i,j)$ in the aggregated network and the time t of the contact. Each centrality metric leads to a unique mitigation strategy in contact removal. The impact of an SIR epidemic spreading can be evaluated via the following performance measures, which will be used to evaluate the mitigation strategies: the average prevalence over time, where the prevalence at a time step is the number of infected nodes; the maximal prevalence, so called peak height, which suggests the maximal demand in e.g. hospital resources; the time to reach the peak prevalence, so called peak time, which indicates the time to prepare the medical resources for the peak demand.

The mitigation strategies that we have proposed are evaluated in 6 real-world temporal networks. We find that the mitigation effect is better when contacts between node pairs that have fewer contacts are removed with a higher probability. Removing contacts that occur earlier in time could further enhance the mitigation effect. A strategy tends to better mitigate the epidemic spreading if the average number of contacts removed varies less among node pairs. Furthermore, we analyze properties of the aggregated pruned network resulted from each contact blocking strategy. We find that the optimal strategy tends to lead to an aggregated pruned network with a large largest eigenvalue, a large modularity and a possibly a small largest connected component. Networks with a large modularity and a small largest connected component are difficult for an epidemic to spread. Static networks with a small largest eigenvalue have been shown to be robust against epidemic spreading i.e. have a high epidemic threshold for Susceptible-Infected-Susceptible epidemic. The resultant aggregated pruned network after contact removal, however, may lead to a low prevalence if its largest eigenvalue is large. This suggests that the temporal information of contacts, may lead to new phenomena that can not be captured by static network studied.

Recent work has been devoted to understand the influence of temporal networks on dynamic processes and especially the mitigation of epidemic spreading. A first line of reseach has studied the mitigation of epidemic spreading via node-level approaches. Génois et. al have shown that vaccination of individuals who act as bridges between communities in time-aggregated network can efficiently prevent epidemic outbreaks [9]. Gemmetto et. al have investigated the epidemic mitigation via excluding a sub-group of nodes in a temporal network in school environments [10]. Another line of research has focused on link-based approaches to suppress epidemic outbreaks. Link removal strategies based on link centrality metrics in the aggregated network has been studied in [11]. These strategies select the links in the aggregated network to block, thus removing all contacts associated with the selected links. In this work, we investigate in-depth at contact level, i.e. how to select a given number of contacts to remove to suppress epidemic spreading. To the best of our knowledge, few works have studied contact-level approaches to suppress epidemic spreading. Our previous work [12] has addressed the same question, however, was confined to Susceptible-Infected (SI) model, which is a special case of SIR model. In this work, we consider the SIR model, broaden and deepen our investigation towards a more comprehensive evaluation of mitigation effect and a more systematic analysis of the properties of the pruned network to explain the performance of the strategies. In view of the uncertainty of realistic temporal network data, we further check the robustness of our finding in the relative effectiveness of proposed mitigation strategies when the temporal networks are under the perturbation, i.e. when the time (ordering) of contacts is uncertain.

3.2. Methods

We will firstly propose our contract removal strategies. Afterwards, we will introduce the real-world temporal networks and simulations that will be used to simulate the epidemic spreading process and further to evaluate the effect of the mitigation strategies.

3.2.1. Contact blocking strategies

We select the contacts to block based on a given centrality metric in the aggregated network and the time of each contact. Specifically, the probability that a contact $c(i,j,t)$ is removed is defined as a function of the given centrality metric of the corresponding link $l(i,j)$ in the aggregated network $\mathscr{G}_{\mathscr{W}}$ and the time t of the contact. This function also ensures that a fraction ϕ of contacts are removed on average.

LINK CENTRALITY METRICS

We propose a set of link centrality metrics based on node centrality metrics for the aggregated network $\mathcal{G}_\mathcal{W}$. The aggregated network $\mathcal{G}_\mathcal{W}$ is a weighted network constructed from a temporal network $\mathcal{G}$. The weight of each link in the aggregated network represents the number of contacts between the two corresponding nodes in the temporal network. Each centrality metric below will lead afterwards to a unique mitigation strategy:

- *Degree product* of a link $l(i,j)$ refers to $d(i)\cdot d(j)$, where $d(i)$ is the degree of node i defined as the number of links incident to node i in the aggregated network.
- *Strength product* of a link $l(i,j)$ refers to $s(i)\cdot s(j)$, where $s(i)$ is the strength of node i defined as the total weights of all the links incident to node i in aggregated network. The strength of a node tells the total number of contacts the node has.
- *Betweenness* is the number of shortest paths that traverse the link between all possibly node pairs in the unweighted aggregated network [13].
- *Link weight* of a link $l(i,j)$ in aggregated network refers to the total number of contacts between node i and j in the corresponding temporal network.
- *Weighted eigenvector component product* is the product of the principal eigenvector components of the link's two end nodes. The principal eigenvector is the eigenvector corresponds to the largest eigenvalue of the weighted aggregated network.
- *Unweighted eigenvector component product* is the product of the principal eigenvector components of the link's two end nodes. The principal eigenvector is the eigenvector corresponds to the largest eigenvalue of the unweighted aggregated network.

Besides the proposed strategies based on the aforementioned link centrality metrics, we introduce a baseline strategy called *Random removal*. In the *Random removal* strategy, the probability for each contact $c(i,j,t)$ to be removed is independent of the centrality of $l(i,j)$. Or equivalently, *Random removal* sets the centrality value as 1 for all links.

CONTACT REMOVAL PROBABILITY

Given a link centrality metric m, we can derive the centrality m_{ij} for each link $l(i,j)$ in the aggregated network. Consider the simple case where the probability that a contact $c(i,j,t)$ between i and j is removed is independent of the time t and we first propose the removal preference p_{ij}:

$$p_{ij} = m_{ij}\frac{\phi\sum_{lk} w_{lk}}{\sum_{lk}(w_{lk}m_{lk})} \tag{3.1}$$

where w_{ij} is the weight of link $l(i,j)$ in the aggregated network or equivalently the number of contacts between i and j, ϕ is the expected fraction of contacts to be removed, thus we have $\sum_{ij} p_{ij} w_{ij} = \phi \sum_{lk} w_{lk}$, the expected number of contacts to be removed. The removal preference p_{ij} of a contact between any node pair i and j is proportional to the centrality m_{ij} of the corresponding link $l(i,j)$.

We cannot use the removal preference p_{ij} directly as the removal probability of a contact between node i and j in view of the following. Some centrality metrics could be highly heterogeneous. The removal preference p_{ij} is possibly larger than 1 if the centrality measure m_{ij} of the link $l(i,j)$ is large. To deal with this issue, we propose an iterative process to derive the contact removal probability by re-normalizing p_{ij}, where $i,j \in \mathcal{N}$: we assign removal probabilities 1 to those contacts whose removal preference p_{ij} according to (3.1) is larger than one, and re-normalize p_{ij} among the contacts with $p_{ij} \leq 1$ to satisfy $\sum_{ij} p_{ij} w_{ij} = \phi \sum_{ij} w_{ij}$. We repeat this normalization process until the removal preference p_{ij} of all contacts are between 0 and 1, while the actual average fraction of contacts blocked is ϕ. Now we define $\tilde{p}_{ij}$ as the re-normalized p_{ij} via the proposed iterative process, and $\tilde{p}_{ij}$ is used as the removal probability of each contact between node i and node j.

We further generalize the definition of the contact removal preference p_{ij} as

$$p^*_{ij} = m^{\alpha}_{ij} \frac{\phi \sum_{lk} w_{lk}}{\sum_{lk}(w_{lk} m^{\alpha}_{lk})} \tag{3.2}$$

The removal preference of a contact $c(i,j,t)$ is proportional to a polynomial function of m_{ij}. The definition (3.1) of p_{ij} is a special case when $\alpha = 1$ of definition (3.2). The random strategy, i.e. all contacts have the same probability of being removed, corresponds to the case when $\alpha = 0$. Consider (3.1) where the reciprocal metric $\frac{1}{m_{ij}}$ is taken as a new centrality metric. The corresponding strategy is equivalent to the general definition (3.2) where metric m_{ij} is considered and $\alpha = -1$.

In this work, we consider the definition (3.1) of p_{ij} using the aforementioned list of centrality metrics and their reciprocals as well as the random strategy, which correspond to the general definition of (3.2) where $\alpha = 1, -1, 0$, respectively.

Finally, we generalize our strategy by considering the timestamps of the contacts. This is motivated by the intuition that early intervention, e.g. blocking early contacts, could be possibly more effective. We propose a time-dependent contract removal preference $p_{ij}(t)$:

$$p_{ij}(t) = m_{ij} f(t) \frac{\phi \sum_{lk} w_{lk}}{\sum_{lk}(w_{lk} m_{lk} f(t))} \tag{3.3}$$

where $f(t)$ describes the preference to remove contacts at specific period. The preference that $c(i,j,t)$ is removed is proportional to $m_{ij} \cdot f(t)$. The same aforementioned normalization process is applied to this generalized contact removal preference to derive the removal probability of each contact.

As a start, we consider $f(t) = 4 \cdot 1_{t \leq T/2} + 1_{t > T/2}$, $f(t) = 1_{t \leq T/2} + 4 \cdot 1_{t > T/2}$ and $f(t) = 1$, where the indicator function 1_y is one if the condition y is true, and otherwise it is 0. They correspond to the preference of removing contacts happening early in $[1, T/2]$, late in $(T/2, T]$ and no preference for the timestamps of the contacts, respectively.

3.2.2. Datasets

The following real-world physical contact networks will be considered:

- HighSchool11&12 record the physical contacts between students in a high school in Marseilles, France. [14]. The two datasets consider two different groups of students.

Datasets	Nodes	Links	Contacts	Duration
HighSchool11 (HS11)	126	1709	28561	3.15
HighSchool12 (HS12)	180	2220	45047	8.44
WorkPlace13 (WP13)	92	755	9827	11.43
WorkPlace15 (WP15)	217	4274	78249	11.50
MIT1	74	355	29107	6.99
MIT2	45	200	22714	6.99
MIT	96	5078	1086404	232.30

Table 3.1: Basic properties of real-world networks: the number of nodes, links (in the aggregated network) and contacts, respectively. The duration refers to the duration T of the observation window [1,T] in the units of days.

- WorkPlace13&15 capture the contacts between individuals in an office building in France [9]. The two datasets are measured from different groups of individuals respectively.

- MIT are human contact network among students of the Massachusetts Institute of Technology [15, 16]. The MIT dataset has been measured for about 8 months.

All networks are undirected. Their properties are given in Table 3.1. The duration of each time step is 1 second in all the networks. For the MIT dataset, we choose randomly two observation period, each of about one-week time. The temporal networks corresponding to these two periods are called MIT1 and MIT2. In this way, all the six temporal networks (HighSchool11&12, WorkPlace13&15, MIT1&2) are comparable in observation window. They will be used to study the impact of the mitigation strategies on the average prevalence over time, the focus of this work.

However, most networks have a short duration of the observation window, within 12 days, besides MIT. In order to observe the peak (increase and afterwards decrease of) prevalence in the SIR process, the observation window of a temporal network needs to be long in duration. When we study the performance measure like peak height/prevalence and peak time, we repeat each of the temporal network HighSchool11&12, WorkPlace13&15 respectively for 10 times. The constructed networks, *HighSchool11&12, *WorkPlace13&15 which repeats one temporal network periodically are also called periodic networks [5]. Each constructed network has a duration ten times as large as the original network . We consider the 4 constructed network *HighSchool11&12, *WorkPlace13&15 and the MIT dataset to study the performance of the strategies in terms of peak prevalence and peak time.

3.2.3. SIMULATION

In this subsection, we will introduce the simulation of the SIR spreading process and the choice of parameters. The performance measures to evaluate the mitigation strategies will be discussed in the next section.

We consider the following discrete time SIR spreading process: a seed node is chosen to be infected at $t = 0$, while the other nodes are susceptible at $t = 0$. Each contact between an infected node and a susceptible node could lead to an infection with probability β. At each time step, each infected node recovers with a recovery probability γ. We consider infection probability $\beta = 0.01$ as an example. In this case and when $\gamma = 0$, the prevalence at T is around the order of 10% in the first six temporal networks. Furthermore, we consider the recover probability per time step $\gamma = 1.22 \cdot 10^{-6}$ or $\gamma = 0$. The former, $\gamma = 1.22 \cdot 10^{-6}$ leads approximately to a recovery probability 10% per day.

In the simulation, we simulate the exact infection and recovery process except the following approximation in the recovery process. If there is no contact in the whole network for the period $t_0, t_0 + t$, we update the state of each node only at the end of this time window $t_0 + t$ instead of at each of the t time steps. In the datasets we have considered, the longest gap that no contact happens is around one day. Correspondingly, the average prevalence is the number of infected nodes over the time steps when at least one contact happens in the network.

Given a temporal network and a centrality metric, we compute the contact removal preference (3.1) for each contact based on the aggregated network of the temporal network and derive further the removal probability of each contact via the normalization process of the contact removal preference. We select each node as a possible seed node and iterate the following for five times per seed node: the fraction ϕ of contacts to be removed are selected according to contact removal probabilities; The SIR process starting from the given seed is performed on the pruned temporal network resulted from the removal of the selected contacts; the prevalence ρ is recorded at each time step when there is a contact in the network. Given a network and a centrality measure, we obtain the prevalence at a time step as the average over the five iterations per every seed node. The average prevalence over all time steps when there is at least one contact is used as the key performance to evaluate the contact removal strategies. The fraction ϕ of contacts to be removed is a control parameter and $\phi = 10\%$ and $\phi = 30\%$ are considered. Simulations are performed in the same way when the time factor $f(t)$ are taken into account via the contact removal probability (3.3).

3.3. Results

In this section, we evaluate our contact removal strategies via three performance measures: the average prevalence and the peak height (the maximal number of infected at a time step) and the peak time (the time to reach the peak height/prevalence).

3.3.1. Performance evaluation

Average prevalence

Firstly, we evaluate the strategies as defined in (3.1) where the probability that a contact $c(i, j, t)$ is removed is independent of the time t of the contact but do depend a centrality metric of the link $l(i, j)$ in the aggregated network. In total, 13 strategies are considered that correspond to the aforementioned centrality metrics and their reciprocals. Figure 3.1 exemplifies the prevalence $\rho(t)$ over time in two periodic networks *HighSchool12 and *WorkPlace15 when each of the 13 strategies is performed and 10% contacts are removed. The ordering of the prevalence $\rho(t)$ at each time step for the 13 strategies are relatively stable over time. The relative performance of the mitigation strategies in terms of average prevalence over time seems not sensitive to duration of the observation time window.

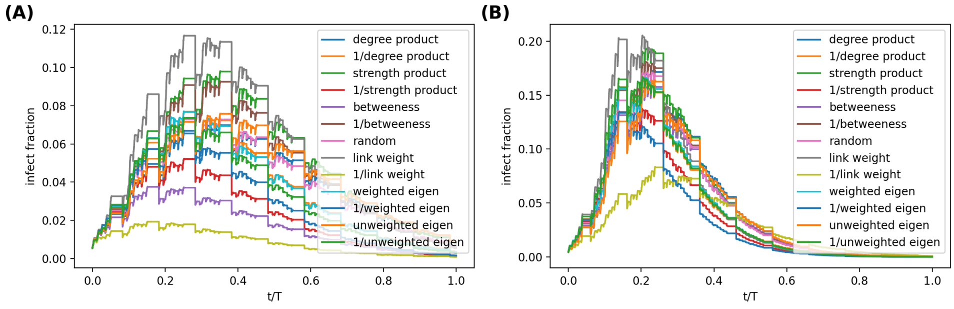

Figure 3.1: The prevalence ρ of the SIR model over time in periodic network *HighSchool12 (A) and *WorkPlace15 (B), when mitigated via 13 contact blocking strategies defined by (3.1) respectively. The infection rate is $\beta = 0.01$ per time step, the recovery rate is $\gamma = 1.22 \cdot 10^{-6}$ per time step, approximately 10% per day and 30% of the contacts are removed.

We use the original network HighSchool11&12, WorkPlace13&15, MIT1&2 to evaluate the blocking strategies with respect to the average prevalence. These networks are comparable in duration of the observation time window, i.e. within 12 days. The performance of each strategy in each network is evaluated via the the average prevalence $E[\rho]$, i.e., the average fraction of infected nodes over the time steps when there is at least one contact in the network. We start with the simple case when the recovery rate $\gamma = 0$. In this case, the SIR model is equal to the Susceptible-Infected (SI) model. The average prevalence when contacts are removed according to each strategy are shown in Table 3.2 and

Metrics	HS11	HS12	WP13	WP15	MIT1	MIT2
degree product	0.043	0.038	0.027	0.102	0.106	0.193
1/degree product	0.044	0.041	0.028	0.107	0.097	0.183
strength product	0.049	0.042	0.028	0.106	0.110	0.193
1/strength product	0.046	0.040	0.027	0.108	0.098	0.164
betweeness	0.046	0.037	0.027	0.106	0.097	0.178
1/betweeness	0.047	0.041	0.028	0.109	0.112	0.189
random	0.045	0.040	0.028	0.106	0.109	0.202
link weight	0.052	0.042	0.028	0.122	0.111	0.189
1/link weight	**0.038**	**0.032**	**0.025**	**0.084**	**0.084**	0.159
weighted eigen	0.050	0.041	0.028	0.108	0.121	0.197
1/weighted eigen	0.048	0.040	0.027	0.107	0.095	**0.158**
unweighted eigen	0.041	0.040	0.027	0.100	0.104	0.196
1/unweighted eigen	0.046	0.040	0.029	0.107	0.099	0.187

Table 3.2: The average prevalence $E[\rho]$ when the recovery rate is $\gamma = 0\%$ per step, and $\phi = 10\%$ of the contacts are removed from each temporal network using removal probability (3.1) based on each centrality metric.

3.3, where $\phi = 10\%$ and $\phi = 30\%$ contacts are removed respectively. In most networks, the 1/link weight performs the best among all 13 strategies. The same has been observed when the recovery rate is $\gamma = 1.22 \cdot 10^{-6}$ per step, approximately 10% per day, as shown in Table 3.4 and 3.5. These observations suggest that removing contacts between nodes that have few contacts tends to be the most effective in reducing the average prevalence.

Furthermore, we consider the time dependent contact removal strategies where the contact removal probability $p_{ij}(t)$ is defined in (3.3). When $f(t) = 4 \cdot 1_{t \leq T/2} + 1_{t>T/2}$, a contact happening early in time i.e. $t < T/2$ is 4 times more likely to be removed than a contact occurring late $t > T/2$. When $f(t) = 1_{t \leq T/2} + 4 \cdot 1_{t>T/2}$, contacts happening late i.e. $t > T/2$ are more likely to be removed. Contact removal strategies based on each of these two $f(t)$ examples and each centrality metric are evaluated via the average prevalence. Their performance when $\gamma = 0$, $\phi = 10\%$ is shown in Table 3.6 and 3.7, where early and later contacts are more likely removed respectively. Comparing these results and the time-independent strategies (Table 3.3) or equivalently when $f(t) = 1$, we find that removing earlier contacts better suppresses the epidemic spreading. The same has been observed when $\gamma = 1.22 \cdot 10^{-6}$, $\phi = 10\%$ (see Table 3.8 and 3.9). Moreover, metric 1/link weight tends to have the best performance independent of the choice of $f(t)$. Therefore, the epidemic spreading can be better mitigated when contacts between node pairs that have few contacts and happening early are more probable to be removed.

Metrics	HS11	HS12	WP13	WP15	MIT1	MIT2
degree product	0.026	0.026	0.021	0.057	0.084	0.184
1/degree product	0.037	0.031	0.024	0.072	0.073	0.142
strength product	0.038	0.029	0.024	0.068	0.099	0.184
1/strength product	0.030	0.028	0.022	0.063	0.063	0.109
betweeness	0.032	0.026	0.022	0.059	0.074	0.151
1/betweeness	0.032	0.030	0.023	0.068	0.102	0.164
random	0.032	0.027	0.022	0.064	0.088	0.168
link weight	0.043	0.034	0.024	0.088	0.107	0.183
1/link weight	**0.020**	**0.018**	**0.020**	**0.038**	**0.055**	0.119
weighted eigen	0.032	0.031	0.023	0.070	0.101	0.177
1/weighted eigen	0.043	0.030	0.024	0.070	0.064	**0.099**
unweighted eigen	0.026	0.027	0.022	0.056	0.092	0.167
1/unweighted eigen	0.040	0.030	0.023	0.075	0.080	0.141

Table 3.3: The average prevalence $E[\rho]$ when the recovery rate is 0% per step, and $\phi = 30\%$ of the contacts are removed from each temporal network using removal probability (3.1) based on each centrality metric.

Metrics	HS11	HS12	WP13	WP15	MIT1	MIT2
degree product	0.034	0.023	0.014	0.051	0.074	0.132
1/degree product	0.038	0.023	0.014	0.052	0.071	0.124
strength product	0.038	0.024	0.014	0.051	0.069	0.131
1/strength product	0.037	0.023	0.013	0.049	0.061	**0.110**
betweeness	0.037	0.024	0.013	0.050	0.064	0.130
1/betweeness	0.038	0.023	0.015	0.050	0.072	0.130
random	0.036	0.024	0.014	0.047	0.075	0.126
link weight	0.043	0.024	0.015	0.058	0.078	0.139
1/link weight	**0.031**	**0.020**	**0.013**	**0.040**	**0.061**	0.111
weighted eigen	0.038	0.024	0.014	0.051	0.078	0.133
1/weighted eigen	0.039	0.024	0.014	0.055	0.072	0.122
unweighted eigen	0.033	0.022	0.013	0.045	0.076	0.138
1/unweighted eigen	0.039	0.024	0.013	0.050	0.068	0.127

Table 3.4: The average prevalence $E[\rho]$ when the recovery rate is $\gamma = 1.22 \cdot 10^{-6}$ per step, approximately 10% per day, and $\phi = 10\%$ of the contacts are removed from each temporal network using removal probability (3.1) based on each centrality metric.

Metrics	HS11	HS12	WP13	WP15	MIT1	MIT2
degree product	0.021	0.015	0.011	0.026	0.060	0.118
1/degree product	0.032	0.018	0.012	0.033	0.051	0.093
strength product	0.031	0.017	0.011	0.031	0.069	0.121
1/strength product	0.024	0.016	0.011	0.030	0.044	0.075
betweeness	0.025	0.015	0.012	0.027	0.050	0.108
1/betweeness	0.026	0.018	0.011	0.030	0.070	0.110
random	0.026	0.016	0.012	0.031	0.062	0.105
link weight	0.036	0.021	0.012	0.040	0.068	0.121
1/link weight	**0.017**	**0.011**	**0.010**	**0.017**	**0.039**	0.081
weighted eigen	0.027	0.018	0.012	0.032	0.062	0.124
1/weighted eigen	0.037	0.018	0.012	0.034	0.042	**0.068**
unweighted eigen	0.021	0.016	0.011	0.026	0.063	0.119
1/unweighted eigen	0.032	0.018	0.012	0.034	0.056	0.094

Table 3.5: The average prevalence $E[\rho]$ when the recovery rate is $\gamma = 1.22 \cdot 10^{-6}$ per step, approximately 10% per day, and $\phi = 30\%$ of the contacts are removed from each temporal network using removal probability (3.1) based on each centrality metric.

Metrics	HS11	HS12	WP13	WP15	MIT1	MIT2
degree product	0.040	0.037	0.027	0.101	0.109	0.191
1/degree product	0.044	0.038	0.028	0.106	0.097	0.171
strength product	0.045	0.039	0.027	0.109	0.107	0.184
1/strength product	0.044	0.039	0.026	0.100	0.091	0.159
betweeness	0.041	0.034	0.027	0.098	0.099	0.165
1/betweeness	0.044	0.040	0.027	0.102	0.107	0.185
random	0.041	0.037	0.028	0.101	0.102	0.184
link weight	0.049	0.040	0.028	0.122	0.118	0.188
1/link weight	**0.035**	**0.030**	**0.026**	**0.080**	**0.081**	0.160
weighted eigen	0.045	0.040	0.028	0.108	0.096	0.192
1/weighted eigen	0.047	0.041	0.028	0.102	0.098	**0.159**
unweighted eigen	0.038	0.038	0.027	0.097	0.104	0.197
1/unweighted eigen	0.050	0.041	0.029	0.107	0.103	0.170

Table 3.6: The average prevalence $E[\rho]$ when the recovery rate is $\gamma = 0\%$ per step and $\phi = 10\%$ of the contacts are removed from each temporal network using removal probability (3.3) based on each centrality metric and $f(t) = 4 \cdot 1_{t \le T/2} + 1_{t>T/2}$. Contacts occurring early in time i.e. $t < T/2$ are more likely to be removed.

Metrics	HS11	HS12	WP13	WP15	MIT1	MIT2
degree product	0.043	0.040	0.027	0.106	0.109	0.193
1/degree product	0.047	0.042	0.028	0.110	0.102	0.186
strength product	0.051	0.042	0.027	0.109	0.111	0.200
1/strength product	0.045	0.040	0.028	0.105	0.095	0.172
betweeness	0.046	0.038	0.027	0.107	0.101	0.191
1/betweeness	0.046	0.042	0.027	0.111	0.115	0.193
random	0.048	0.041	0.028	0.107	0.108	0.200
link weight	0.051	0.045	0.029	0.114	0.114	0.191
1/link weight	**0.041**	**0.035**	**0.026**	**0.086**	**0.089**	**0.161**
weighted eigen	0.048	0.041	0.028	0.112	0.108	0.191
1/weighted eigen	0.050	0.041	0.028	0.112	0.097	0.166
unweighted eigen	0.046	0.040	0.027	0.107	0.109	0.200
1/unweighted eigen	0.050	0.043	0.027	0.108	0.103	0.191

Table 3.7: The average prevalence $E[\rho]$ when the recovery rate is $\gamma = 0\%$ per step and $\phi = 10\%$ of the contacts are removed from each temporal network using removal probability (3.3) based on each centrality metric and $f(t) = 1_{t \le T/2} + 4 \cdot 1_{t>T/2}$. Contacts occurring late in time i.e. $t > T/2$ are more likely to be removed.

Metrics	HS11	HS12	WP13	WP15	MIT1	MIT2
degree product	0.032	0.023	0.013	0.046	0.070	0.127
1/degree product	0.039	0.025	0.014	0.049	0.065	0.113
strength product	0.038	0.024	0.014	0.051	0.072	0.133
1/strength product	0.037	0.023	0.013	0.048	0.062	0.108
betweeness	0.034	0.021	0.014	0.047	0.062	0.114
1/betweeness	0.033	0.023	0.014	0.048	0.069	0.134
random	0.033	0.023	0.013	0.048	0.070	0.131
link weight	0.037	0.024	0.013	0.054	0.079	0.131
1/link weight	**0.029**	**0.019**	**0.012**	**0.038**	**0.054**	**0.104**
weighted eigen	0.036	0.025	0.013	0.050	0.067	0.128
1/weighted eigen	0.038	0.025	0.014	0.050	0.064	0.111
unweighted eigen	0.030	0.023	0.013	0.044	0.073	0.126
1/unweighted eigen	0.038	0.024	0.014	0.051	0.068	0.113

Table 3.8: The average prevalence $E[\rho]$ when the recovery rate is $\gamma = 1.22 \cdot 10^{-6}$ per step, and $\phi = 10\%$ of the contacts are removed from each temporal network using removal probability (3.3) based on each centrality metric and $f(t) = 4 \cdot 1_{t \le T/2} + 1_{t > T/2}$. Contacts occurring early in time i.e. $t < T/2$ are more likely to be removed.

Metrics	HS11	HS12	WP13	WP15	MIT1	MIT2
degree product	0.036	0.025	0.014	0.052	0.074	0.134
1/degree product	0.039	0.024	0.014	0.053	0.071	0.124
strength product	0.039	0.025	0.014	0.052	0.078	0.139
1/strength product	0.037	0.024	0.014	0.052	0.070	0.117
betweeness	0.037	0.024	0.014	0.051	0.065	0.126
1/betweeness	0.036	0.025	0.014	0.051	0.077	0.133
random	0.037	0.025	0.014	0.051	0.077	0.138
link weight	0.041	0.026	0.015	0.055	0.075	0.129
1/link weight	**0.033**	**0.021**	0.014	**0.042**	**0.059**	**0.110**
weighted eigen	0.040	0.025	0.014	0.051	0.082	0.141
1/weighted eigen	0.041	0.025	0.014	0.054	0.067	0.115
unweighted eigen	0.036	0.025	**0.013**	0.050	0.078	0.136
1/unweighted eigen	0.039	0.024	0.014	0.056	0.067	0.123

Table 3.9: The average prevalence $E[\rho]$ when the recovery rate is $\gamma = 1.22 \cdot 10^{-6}$ per step and $\phi = 10\%$ of the contacts are removed from each temporal network using removal probability (3.3) based on each centrality metric and $f(t) = 1_{t \le T/2} + 4 \cdot 1_{t > T/2}$. Contacts occurring late in time i.e. $t > T/2$ are more likely to be removed.

Properties of the pruned network

The pruned network is the resultant temporal network after contacts being removed according to a strategy. In this section, we explore the relation between the properties of the pruned network and the average prevalence, resulted from a contact removal strategy. This could help us understand what kind of pruned networks may lead to a low prevalence. We focus on time-independent contact removal strategies to illustrate our method.

The average number of contacts removed between any node pair i and j or link $l(i,j)$ in the aggregated network is $p_{ij}w_{ij}$, where w_{ij} is the number of contacts between i and j and p_{ij} is the probability that a contact between i and j is removed. The average number of contacts removed by strategy 1/link weight is the same for all links in the aggregated network[1]. We explore whether a strategy that removes a similar number of contacts per node pair (link) may better mitigates the epidemic spreading. Figure 3.2 (B) demonstrates the scatter plot of the average prevalence $E[\rho]$ versus $\sqrt{Var[p_{ij}w_{ij}]}$ for each strategy when $\phi = 10\%$ contacts are removed and the recovery rate is $\gamma = 0$ per step. We find that, in each network, a strategy tends to reduce the average prevalence $E[\rho]$ more if $\sqrt{Var[p_{ij}w_{ij}]}$ is small. The same can be observed when the recovery rate γ and removal fraction ϕ vary (see (B) of Figure 3.3, 3.4, 3.5).

Each pruned network is a temporal network. We investigate three properties of the aggregated network W^* of the pruned network. Each element W^*_{ij} in the weighted adjacency matrix W^* of the aggregated pruned network tells the number of contacts between i and j in the pruned network.

We explore firstly the largest eigenvalue $\lambda_1(W^*)$ of the aggregated pruned network in relation the corresponding average prevalence resulted from each strategy. Consider the Susceptible-Infected-Susceptible SIS epidemic spreading process on a static network. It has been shown that the largest eigenvalue of the network suggests the robustness of the network subject to epidemic spreading [2, 17–19]. When the effective infection rate, i.e. infection rate divided by the recovery rate, is above (below) the threshold $\tau_c \sim \frac{1}{\lambda_1(W^*)}$, a none-zero (zero) fraction of the population is infected in the meta-stable state. A static network whose largest eigenvalue is small has a large epidemic threshold, thus is robust against epidemic spreading.

Would a pruned network with a small $\lambda_1(W^*)$ lead to a low prevalence according to the findings of SIS model on static networks? Figure 3.2(A), 3.3(A), 3.4(A), 3.5(A) respectively show the scatter plot of the average prevalence $E[\rho]$ versus $\lambda_1(W^*)$ of the aggregated pruned network [2] for each strategy in each network. We observe the opposite: the best strategy with the lowest prevalence tends to lead to a pruned network with a large largest eigenvalue. Such inconsistency can be possibly explained as follows. First, a network that is robust against SIS epidemic spreading is not necessarily robust against SIR epidemic spreading. Each link in the aggregated pruned network can transmit the epidemic maximally once in SIR model whereas possibly multiple times in SIS models.

[1] In the simulation, the average number of contacts removed per link by strategy 1/link weight may differ slightly among the links. When the removal probability $p_{ij} > 1$, we set $p_{ij} = 1$ and re-normalize the removal probabilities of the other links to ensure that a fraction ϕ of contacts are removed.

[2] Given a temporal network and a contact removal strategy, we have simulated per seed node 5 realizations of contact removal and a SIR spreading process on each resultant pruned network. The $\lambda_1(W^*)$ in the scatter plot is the average over the $5N$ realizations of the pruned network.

That is why removing many contacts from links whose end nodes have a high strength may better reduce the largest eigenvalue and better suppress the SIS epidemic but not the SIR epidemic spreading. Second, a network with a low epidemic threshold does not implies a high prevalence when the effective infection rate is above the epidemic threshold. Finally, the aggregated pruned network can not capture the temporal information of contacts, which influence the spread of an epidemic.

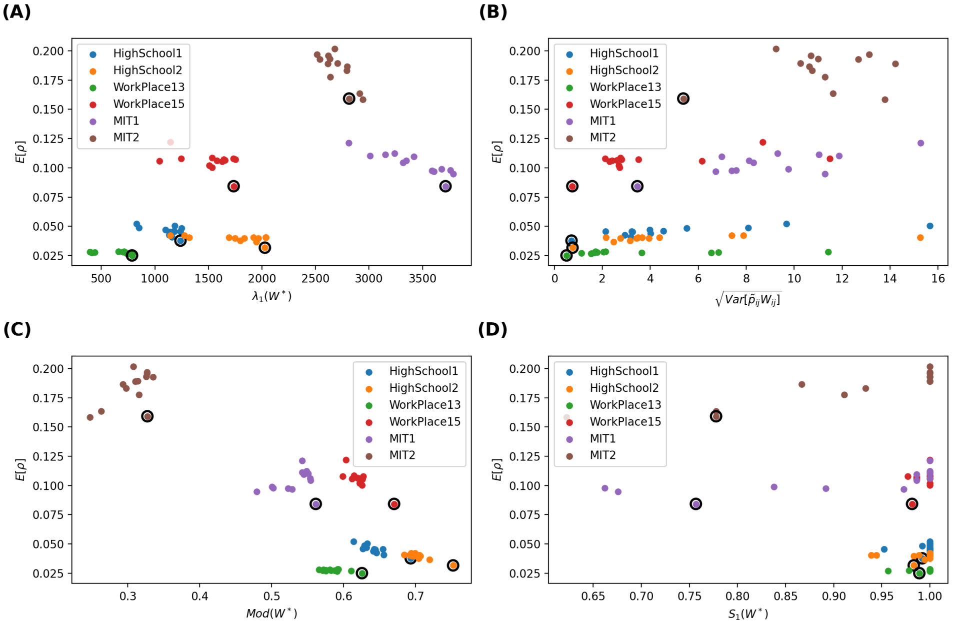

Figure 3.2: Scatter plot of the average prevalence $E[\rho]$ versus the largest eigenvalue $\lambda_1(W^*)$ of the aggregated pruned network (A), the standard deviation $\sqrt{Var[\tilde{p}_{ij}w_{ij}]}$ of the average number of contacts removed from a node pair (B) the modularity $Mod(W^*)$ (C) and the relative size of the largest connected component of the aggregated pruned network (D), respectively. A fraction $\phi = 10\%$ of the contacts are removed. The recovery rate is $\gamma = 0$ per step. The results obtained with 1/link weight strategy are circled.

Furthermore, we consider the modularity $Mod(W^*)$ of the aggregated pruned network. Given a weighted network and a given partition of all the nodes into non-overlapping communities, the quality of this community partition can be measured by the modularity [20, 21] $\frac{1}{2L}\sum_{i,j=1}^{N}(W^*_{ij}-\frac{s_is_j}{2L})\delta_{C_iC_j}$, where s_i is the strength of node i, C_i is the label of the community to which node i belongs to, the Kronecker delta function $\delta_{C_iC_j}=1$ if $C_i = C_j$ or else $\delta_{C_iC_j}=0$. The modularity of a partition describes the extent to which that more link weights are within each community than link weights between communities. The modularity $Mod(W^*)\in[0,1]$ of a network is the maximal modularity that could be obtained via network/node partition. We compute the modularity of an aggregated pruned network via the Louvain method [22]. The scatter plot in Figure 3.2(C), 3.3(C), 3.4(C), 3.5(C) shows that the optimal contact removal strategy that obtains the minimal

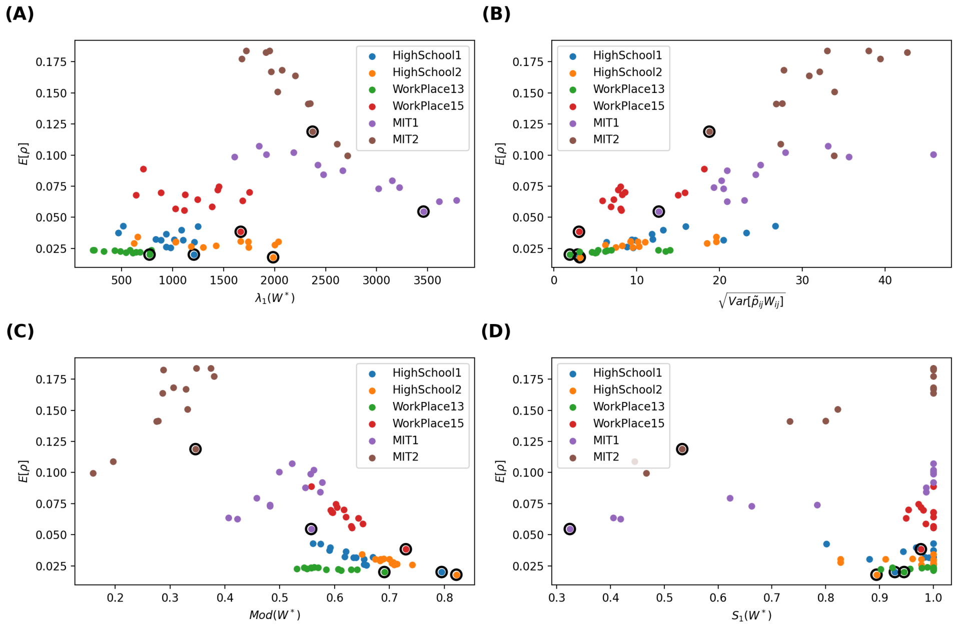

Figure 3.3: Scatter plot of the average prevalence $E[\rho]$ versus the largest eigenvalue $\lambda_1(W^*)$ of the aggregated pruned network (A), the standard deviation $\sqrt{Var[\tilde{p}_{ij} w_{ij}]}$ of the average number of contacts removed from a node pair (B) the modularity $Mod(W^*)$ (C) and the relative size of the largest connected component of the aggregated pruned network (D), respectively. A fraction $\phi = 30\%$ of the contacts are removed. The recovery rate is $\gamma = 0$ per step. The results obtained with 1/link weight strategy are circled.

average prevalence tends to result in a pruned network that has a large modularity. A network with a large modularity is more robust against epidemic spreading.

Finally, we explore the relative size $S_1(W^*)$ of the largest connected component of the aggregated pruned network. We wonder whether the optimal strategy reduced the prevalence via disconnecting the network. As shown in the bottom-right figure of Figure 3.2(D), 3.3(D), 3.4(D), 3.5(D), most pruned networks still have a relative large component $S_1(W^*) \sim 1$. Exceptions are observed for in MIT1 and MIT2, where strategies may evidently disconnect the aggregated pruned network. In such cases, the optimal strategy tends to lead to a relatively small largest component size $S_1(W^*)$. This is in line with the finding that efficient immunization strategy should keep the largest connected component size small [23].

In summary, the optimal mitigation strategy tends to lead to an aggregated pruned network with a large largest eigenvalue, a large modularity and possibly a small largest connected component (in case contact removal strategies evidently disconnect the pruned network). Moreover, a strategy seems to better reduce the prevalence if it removes a similar number of contacts from the links. These observations together further support our previous explanation why the optimal strategy could result in an aggregated pruned net-

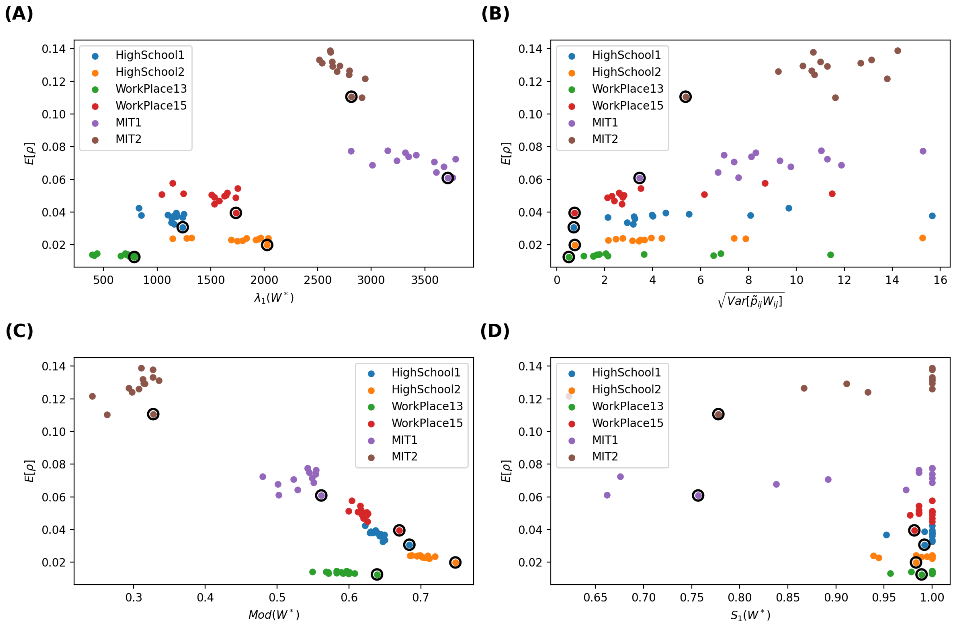

Figure 3.4: Scatter plot of the average prevalence $E[\rho]$ versus the largest eigenvalue $\lambda_1(W^*)$ of the aggregated pruned network (A), the standard deviation $\sqrt{Var[\tilde{p}_{ij}w_{ij}]}$ of the average number of contacts removed from a node pair (B) the modularity $Mod(W^*)$ (C) and the relative size of the largest connected component of the aggregated pruned network (D), respectively. A fraction $\phi = 10\%$ of the contacts are removed. The recovery rate is $\gamma = 1.22 \cdot 10^{-6}$ per step, approximately 10% per day. The results obtained with 1/link weight strategy are circled.

work with a large largest eigenvalue: the optimal strategy tends to remove a similar number of contacts from links, keeping the hubs, i.e. nodes with a large node strength. Such hubs contribute to a large largest eigenvalue and thus a low epidemic threshold for SIS epidemic spreading. However, the modular structure of the pruned network limits the prevalence of an epidemic, which can not be captured directly by the largest eigenvalue.

PEAK HEIGHT AND PEAK TIME

The peak height/prevalence and peak time suggest the maximal demand in e.g., healthcare resources and the time to prepare for the highest demand in resources, respectively. We consider the 4 constructed network *HighSchool11&12, *WorkPlace13&15 and the MIT dataset to study the performance of the strategies in terms of peak prevalence and peak time.

For each centrality metric or strategy, we simulate the SIR spreading process five times for every possible seed node. The peak height is found as the maximum prevalence in each spreading process. Table 3.10 shows the average peak height over all $5 \cdot N$ realizations of the spreading processes. We find that the strategy 1/link weight results in the smallest peak height. The average peak height shown in Table 3.10 differs from the

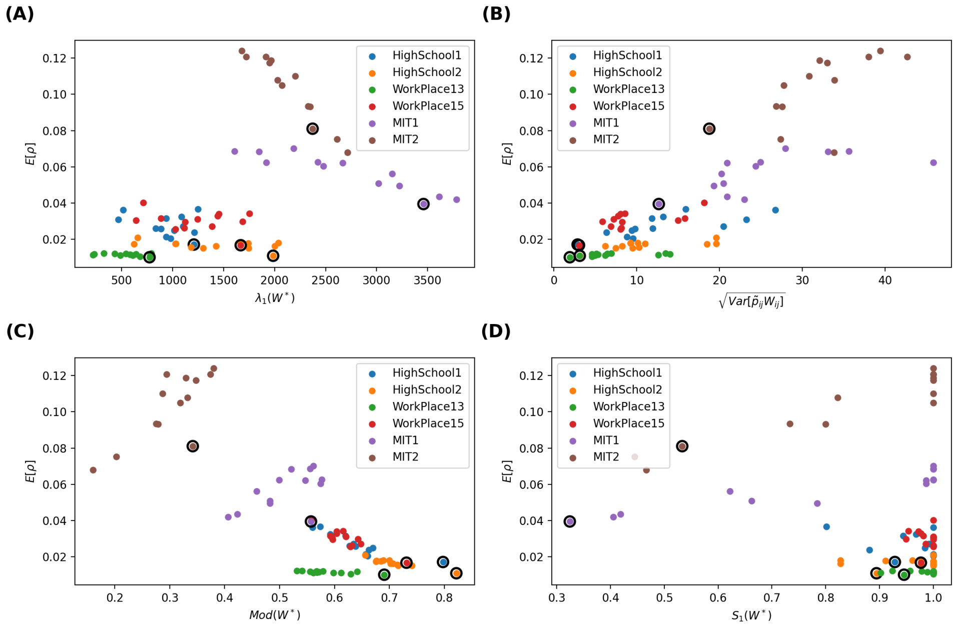

Figure 3.5: Scatter plot of the average prevalence $E[\rho]$ versus the largest eigenvalue $\lambda_1(W^*)$ of the aggregated pruned network (A), the standard deviation $\sqrt{Var[\tilde{p}_{ij}w_{ij}]}$ of the average number of contacts removed from a node pair (B) the modularity $Mod(W^*)$ (C) and the relative size of the largest connected component of the aggregated pruned network (D), respectively. A fraction $\phi = 30\%$ of the contacts are removed. The recovery rate is $\gamma = 1.22 \cdot 10^{-6}$ per step, approximately 10% per day. The results obtained with 1/link weight strategy are circled.

maximal prevalence in Figure 3.1, which corresponds to the maximum of the average prevalence over the $5 \cdot N$ realizations.

Similarly, the average peak time, i.e. time to reach the maximal prevalence over all spreading processes started at every possible seed node is derived and given in Table 3.11. Interestingly, the peak time for strategy 1/link weight is not always the smallest. Strategy 1/link weight leads to the lowest peak height and possibly a longer peak time.

Robustness

Temporal networks measured in real-world scenarios possibly contain noise, e.g., uncertainty of the ordering of contacts or occurring time of contacts. We would like to explore whether our findings in the relative effectiveness of the proposed mitigation strategies still holds when the temporal networks measured are subject to such type of uncertainty.

We assume the temporal networks that we have so far analyzed are measured relatively precisely. For each of these temporal networks, we apply two approaches, respectively, to generate the corresponding temporal networks perturbed by the aforementioned uncertainty. The duration of one time step in the original temporal networks is either 1 second or 20 seconds. We split the observation period $[0, T]$ of a temporal network into non-overlapping bins, whose duration is $\Delta = 60$ seconds to further perturb the networks. We first adopt the uncertainty model I used in [24], which randomly reshuffles the timestamps of the contacts within each bin of $\Delta = 60$ seconds. This model encapsulates the uncertainty of the ordering of contacts that happen at similar time. Given the uncertainty model I (one network realization as an example) of each original temporal network, we evaluate the contact blocking strategies in the same way as in the original network and their performance is given in Table 3.12 and Table 3.13. We find that the ranking of the strategies does not change in model I compared to that in the original temporal networks and the 1/link weight remains the best strategy. Our finding seems to be robust against minor uncertainty in the ordering of contacts.

To capture the uncertainty of the exact occurring time of contacts, we use our uncertainty model II, where each contact's occurring time is measured in the time resolution of $\Delta = 60$ seconds instead of second. In other words, the number of contacts between each pair of nodes in each bin of $\Delta = 60$ seconds is known in model II. However, the exact occurring time of the contacts happening within each bin in precision of seconds is unknown. For each snapshot/bin of $\Delta = 60$ seconds, model II constructs a weighted network, where the weight between two nodes counts the number of contacts between them that occur within the bin of $\Delta = 60$ seconds. Each weighted network is thus an aggregated network of the original temporal network over 60 seconds. The performance of each blocking strategy on model II are shown in Table 3.14 and Table 3.15, demonstrating that strategy 1/link weight outperforms the others, the same as observed in the original temporal networks. Hence, our evaluation of the strategies is robust against network perturbations that models the uncertainty of temporal network data.

Metrics	*HS11	*HS12	*WP13	*WP15	MIT
degree product	0.355	0.121	0.018	0.235	0.163
1/degree product	0.346	0.119	0.020	0.239	0.156
strength product	0.356	0.130	0.020	0.246	0.182
1/strength product	0.301	0.109	0.017	0.225	0.130
betweeness	0.314	0.099	0.018	0.230	0.171
1/betweeness	0.354	0.120	0.020	0.242	0.164
random	0.346	0.119	0.018	0.236	0.167
link weight	0.384	0.132	0.020	0.261	0.170
1/link weight	**0.279**	**0.077**	**0.016**	**0.193**	**0.110**
weighted eigen	0.371	0.128	0.020	0.237	0.182
1/weighted eigen	0.298	0.115	0.021	0.227	0.144
unweighted eigen	0.347	0.117	0.019	0.235	0.169
1/unweighted eigen	0.343	0.112	0.018	0.241	0.164

Table 3.10: The peak height i.e. the highest prevalence over time, when the recovery rate is 10% per day, and $\phi = 10\%$ of the contacts are removed from each temporal network using removal probability (3.1).

Metrics	*HS11	*HS12	*WP13	*WP15	MIT
degree product	3.714	2.518	0.877	1.381	0.429
1/degree product	3.514	2.519	1.695	1.380	0.429
strength product	3.317	2.520	1.678	1.381	0.429
1/strength product	3.321	2.522	1.258	1.382	0.430
betweeness	3.717	2.778	1.208	1.381	0.429
1/betweeness	3.320	2.521	1.078	**1.380**	0.429
random	3.120	2.521	1.145	1.381	0.429
link weight	3.119	**2.516**	1.297	1.381	**0.429**
1/link weight	4.119	2.719	**0.617**	1.904	0.589
weighted eigen	3.120	2.519	1.111	1.381	0.430
1/weighted eigen	**3.117**	2.519	1.314	1.381	0.429
unweighted eigen	3.915	2.520	1.079	1.381	0.482
1/unweighted eigen	3.120	2.519	1.043	1.381	0.482

Table 3.11: The peak time in units of t/T before the maximum prevalence is achieved. The recovery rate is 10% per day, and $\phi = 10\%$ of the contacts are removed from each temporal network using removal probability (3.1).

Metrics	HS11	HS12	WP13	WP15	MIT1	MIT2
degree product	0.028	0.027	0.022	0.059	0.089	0.180
1/degree product	0.038	0.031	0.025	0.072	0.075	0.142
strength product	0.038	0.029	0.023	0.069	0.101	0.186
1/strength product	0.030	0.027	0.022	0.065	0.061	0.107
betweeness	0.031	0.025	0.022	0.061	0.073	0.154
1/betweeness	0.036	0.030	0.023	0.064	0.100	0.175
random	0.033	0.027	0.023	0.063	0.092	0.165
link weight	0.042	0.033	0.024	0.087	0.105	0.185
1/link weight	**0.021**	**0.018**	**0.020**	**0.037**	**0.056**	0.124
weighted eigen	0.034	0.032	0.023	0.068	0.104	0.187
1/weighted eigen	0.047	0.030	0.024	0.071	0.060	**0.101**
unweighted eigen	0.026	0.028	0.022	0.053	0.097	0.176
1/unweighted eigen	0.041	0.030	0.024	0.074	0.077	0.142

Table 3.12: The average prevalence $E[\rho]$ in uncertainty model I when the recovery rate is $\gamma = 0$ per step, and $\phi = 30\%$ of the contacts are removed from each temporal network using removal probability (3.1) based on each centrality metric.

Metrics	HS11	HS12	WP13	WP15	MIT1	MIT2
degree product	0.022	0.016	0.013	0.026	0.060	0.112
1/degree product	0.031	0.018	0.012	0.033	0.052	0.098
strength product	0.030	0.018	0.012	0.030	0.068	0.120
1/strength product	0.025	0.017	0.012	0.031	0.042	0.081
betweeness	0.024	0.016	0.011	0.027	0.048	0.095
1/betweeness	0.027	0.018	0.012	0.031	0.068	0.115
random	0.026	0.017	0.011	0.030	0.055	0.112
link weight	0.036	0.021	0.013	0.041	0.064	0.125
1/link weight	**0.017**	**0.011**	**0.010**	**0.017**	**0.035**	0.082
weighted eigen	0.027	0.019	0.012	0.033	0.066	0.120
1/weighted eigen	0.035	0.019	0.013	0.032	0.042	**0.067**
unweighted eigen	0.021	0.017	0.011	0.025	0.065	0.121
1/unweighted eigen	0.033	0.018	0.012	0.034	0.051	0.096

Table 3.13: The average prevalence $E[\rho]$ in uncertainty model I when the recovery rate is $\gamma = 1.22 \cdot 10^{-6}$ per step, approximately 10% per day, and $\phi = 30\%$ of the contacts are removed from each temporal network using removal probability (3.1) based on each centrality metric.

Metrics	HS11	HS12	WP13	WP15	MIT1	MIT2
degree product	0.022	0.016	0.011	0.025	0.061	0.114
1/degree product	0.031	0.018	0.013	0.033	0.047	0.093
strength product	0.029	0.017	0.012	0.031	0.065	0.122
1/strength product	0.026	0.017	0.012	0.032	0.041	0.077
betweeness	0.025	0.015	0.012	0.029	0.051	0.103
1/betweeness	0.027	0.017	0.013	0.030	0.068	0.113
random	0.028	0.017	0.012	0.030	0.061	0.109
link weight	0.035	0.022	0.013	0.043	0.073	0.128
1/link weight	**0.017**	**0.011**	**0.011**	**0.018**	**0.038**	0.086
weighted eigen	0.028	0.019	0.012	0.033	0.070	0.130
1/weighted eigen	0.036	0.018	0.012	0.034	0.043	**0.068**
unweighted eigen	0.019	0.015	0.011	0.025	0.066	0.118
1/unweighted eigen	0.032	0.018	0.011	0.036	0.053	0.096

Table 3.14: The average prevalence $E[\rho]$ in uncertainty model II when the recovery rate is $\gamma = 0$ per step, and $\phi = 30\%$ of the contacts are removed from each temporal network using removal probability (3.1) based on each centrality metric.

Metrics	HS11	HS12	WP13	WP15	MIT1	MIT2
degree product	0.021	0.016	0.012	0.022	0.055	0.114
1/degree product	0.026	0.017	0.012	0.030	0.051	0.099
strength product	0.027	0.017	0.012	0.026	0.069	0.110
1/strength product	0.023	0.016	0.011	0.029	0.039	0.078
betweeness	0.023	0.015	0.012	0.025	0.051	0.095
1/betweeness	0.024	0.017	0.012	0.028	0.060	0.109
random	0.022	0.016	0.011	0.028	0.058	0.103
link weight	0.030	0.020	0.013	0.036	0.066	0.117
1/link weight	**0.016**	**0.011**	**0.011**	**0.017**	**0.036**	0.081
weighted eigen	0.025	0.017	0.013	0.029	0.060	0.123
1/weighted eigen	0.029	0.018	0.011	0.031	0.041	**0.066**
unweighted eigen	0.020	0.015	0.011	0.024	0.061	0.108
1/unweighted eigen	0.028	0.017	0.012	0.032	0.051	0.094

Table 3.15: The average prevalence $E[\rho]$ in uncertainty model II when the recovery rate is $\gamma = 1.22 \cdot 10^{-6}$ per step, and $\phi = 30\%$ of the contacts are removed from each temporal network using removal probability (3.1) based on each centrality metric.

3.4. CONCLUSIONS

In this chapter, we have developed and evaluated contact blocking strategies in order to mitigate SIR epidemic spreading on a temporal network. The probability that a contact $c(i, j, t)$ is removed is defined as a generic function of a given centrality metric of the corresponding link $l(i, j)$ in the corresponding aggregated network and time t. In total 12 centrality metrics or strategies and a baseline strategy (random removal) have been considered. The strategy (1/link weight) that tends to remove contacts between node pairs with few contacts and removes early contacts seems to mitigate the epidemic spreading the best, with respect to the average prevalence, the peak prevalence and the time needed to reach the peak prevalence. This suggests that the removal of contacts along weak social ties in an early phase tends better suppress the epidemic spreading. Removing a large number of contacts from few node pairs is likely too costly to be effective. We demonstrate further that our finding, i.e., the 1/link weight strategy tends to outperform, still holds when uncertainty is introduced into original temporal networks via reshuffling the ordering of contacts and enlarging the temporal resolution, respectively.

Characterization of the pruned network resulted from the contact removal of a given strategy provides insights why some strategies outperform the others: an optimal strategy (1/link weight) leads to an aggregated pruned network with a large largest eigenvalue, a large modularity and a possibly small largest connected component size. A strategy tends to perform better when a similar number of contacts are removed from links. These findings are in line with our understanding that a network with a small largest connected component, a large modularity prohibits epidemic spreading. However, the large largest eigenvalue achieved by the optimal strategy seems to contradict our understanding that a static network with a large largest eigenvalue tends to facilitate SIS epidemic spreading with respect to its small epidemic threshold. We explain this seemingly inconsistency with respect to the difference between SIR and SIS models, between epidemic threshold and prevalence, and the complexity introduced by the temporal contacts that cannot be captured by the aggregated network.

A few limitations of our work should be noticed and could be explored in future work. First, we have confined ourselves to the SIR model with limited choice of parameters and a few real-world networks. SIR model is a simplified model of the epidemic spreading process, whereas real-world epidemic spreading can be more complicated. Hence, our conclusion regarding the effectiveness of the mitigation strategies cannot be generalized directly to real-world epidemic mitigation. It is essential to explore further generalized choice of more realistic epidemic spreading model. Second, the dependency of removal preference $p_{ij}(t)$ on time, i.e., $f(t)$, that we have we chosen is one of the simplest forms. Other forms of time-dependent function $f(t)$ could be further explored, especially those that are feasible for policy makers. The contact removal strategies proposed is based on the knowledge of the aggregated network over the observation window, the period when we intervene the spreading process. One challenging question is how to estimate or predict this aggregated network based on the observation of the aggregated network in the past. Beyond the aggregated network, contact removal strategies can also be based on temporal and topological properties of contacts.

REFERENCES

1. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Reviews of Modern Physics* **87,** 925–979 (2015).

2. Wang, H. *et al.* Effect of the interconnected network structure on the epidemic threshold. *Physical Review E* **88,** 022801 (2013).

3. Holme, P. & Saramäki, J. Temporal networks. *Physics Reports* **519,** 97–125 (2012).

4. Zhao, K., Stehlé, J., Bianconi, G. & Barrat, A. Social network dynamics of face-to-face interactions. *Physical Review E* **83,** 056109 (2011).

5. Zhang, Y.-Q., Li, X. & Vasilakos, A. V. Spectral analysis of epidemic thresholds of temporal networks. *IEEE Transactions on Cybernetics* **50,** 1965–1977 (2017).

6. Karsai, M. *et al.* Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E* **83,** 025102 (2011).

7. Scholtes, I. *et al.* Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nature Communications* **5,** 1–9 (2014).

8. Newman, M. *Networks* (Oxford university press, 2018).

9. Génois, M. *et al.* Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science* **3,** 326–347 (2015).

10. Gemmetto, V., Barrat, A. & Cattuto, C. Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC Infectious Diseases* **14,** 1–10 (2014).

11. Zhan, X.-X., Hanjalic, A. & Wang, H. *Suppressing information diffusion via link blocking in temporal networks* in *International Conference on Complex Networks and Their Applications* (2019), 448–458.

12. Zhao, X. & Wang, H. *Suppressing Epidemic Spreading via Contact Blocking in Temporal Networks* in *International Conference on Complex Networks and Their Applications* (2020), 444–454.

13. Wang, H., Hernandez, J. M. & Van Mieghem, P. Betweenness centrality in a weighted network. *Physical Review E* **77,** 046105 (2008).

14. Fournet, J. & Barrat, A. Contact patterns among high school students. *PLOS One* **9,** e107878 (2014).

15. Kunegis, J. *Konect: the koblenz network collection* in *Proceedings of the 22nd International Conference on World Wide Web* (2013), 1343–1350.

16. Eagle, N. & Pentland, A. S. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing* **10,** 255–268 (2006).

17. Van Mieghem, P., Omic, J. & Kooij, R. Virus spread in networks. *IEEE/ACM Transactions On Networking* **17,** 1–14 (2008).

18. Ottaviano, S., De Pellegrini, F., Bonaccorsi, S., Mugnolo, D. & Van Mieghem, P. in *Multilevel Strategic Interaction Game Models for Complex Networks* 111–129 (Springer, 2019).

19. Qu, B. & Wang, H. Sis epidemic spreading with heterogeneous infection rates. *IEEE Transactions on Network Science and Engineering* **4,** 177–186 (2017).

20. Newman, M. E. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103,** 8577–8582 (2006).

21. Ge, X. & Wang, H. Community overlays upon real-world complex networks. *The European Physical Journal B* **85,** 1–10 (2012).

22. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P10008 (2008).

23. Schneider, C. M., Mihaljev, T., Havlin, S. & Herrmann, H. J. Suppressing epidemics with a limited amount of immunization units. *Physical Review E* **84,** 061911 (2011).

24. Antulov-Fantulin, N., Lančić, A., Šmuc, T., Štefančić, H. & Šikić, M. Identification of patient zero in static and temporal networks: Robustness and limitations. *Physical Review Letters* **114,** 248701 (2015).

4

Diffusion Backbone of Temporal Higher-order Networks

Temporal higher-order networks, where each hyperlink involving a group of nodes are activated or deactivated over time, are recently used to represent complex systems such as social contacts, interactions or collaborations that occur at specific times. Such networks are substrates for social contagion processes such as the diffusion of information and opinions. In this chapter, we consider eight temporal higher-order networks derived from human face-to-face interactions in various contexts and the Susceptible-Infected threshold process on each of these networks: whenever a hyperlink is active and the number of infected nodes in the hyperlink exceeds a threshold Θ, each susceptible node in the hyperlink is infected independently with probability β. The objective is to understand (1) the contribution of each hyperlink to the diffusion process, namely, the average number of nodes that are infected directly via the activation of the hyperlink when the diffusion starts from an arbitrary seed node, and (2) hyperlinks with what network properties tend to contribute more. Such understanding is crucial for further development of strategies to mitigate a diffusion process.

We first propose to construct the information diffusion backbone. The backbone is a weighted higher-order network, where the weight of each hyperlink denotes the contribution of the hyperlink to a given diffusion process. Secondly, we find that the backbone, or the contribution of hyperlinks, is dependent on the parameters β and Θ of the diffusion process, which is also supported by our theoretical analysis of the backbone when $\beta \to 0$. Thirdly, we systematically design centrality metrics, i.e., network properties, for hyperlinks in a temporal higher-order network and use each centrality metric to estimate the ranking of hyperlinks by the weight in the backbone. Finally, we find and explain why different centrality metrics can better estimate the contributions of hyperlinks for different parameters of the diffusion process.

4.1. Introduction

Complex networks serve as substrates for the diffusion of information, where a piece of information propagates along links connecting couples of nodes. Complex systems in nature and society are rarely static but exhibit time-varying network topologies [1–3]. Traditionally, such systems can be represented as temporal networks, where links between pairs of nodes are activated and deactivated over time. For example, a human physical contact network is usually experimentally recorded as a collection of time-resolved contacts, where a contact denotes an interaction between a pair of nodes at a specific timestamp.

Prior works have revealed that properties of temporal networks such as the inter-event time distribution can affect the dynamics of processes unfolding on the temporal network, e.g., impact the speed of epidemic spreading or diffusion [4–8], or impede the random walk explorations [9]. The properties of nodes, node pairs, and subgraphs [10] in temporal networks have been explored in order to identify which kind of nodes, node pairs, or subgraphs to activate (or block) to maximize (or to minimize) the spread of information (or epidemics) [11–14]. It was found that that vaccinating nodes with certain properties in a temporal network leads to a lower outbreak size when mitigating an epidemic spreading process on temporal networks [11, 12]. Ciaperoni et al. [10] proposed to identify the subgraphs of a temporal network using a generalized k-core decomposition method and showed that the removal of temporal links belong to these subgraphs leads to large decrease in the final outbreak size of a spreading process.

Despite advances made in the past decade, studies of temporal networks have mostly focused on pairwise interactions, which fall short of representing a wide variety of real-world systems [15–18]. For example, individuals [19, 20] or animals [21, 22] may interact in a group of a size larger than two and the collaboration in a scientific paper may involve more than two researchers [23, 24]. Such higher-order (group) interactions can be represented as temporal higher-order networks, where a group or hyperlink is activated (when it has an interaction or contact) and deactivated over time. Previous studies have found shared properties of real-world temporal higher-order networks representing human physical interactions, such as the temporal correlation between hyperlinks in activity [19, 25–27]. It was also demonstrated that the properties of temporal higher-order networks influence the behavior of dynamical processes. For instance, the duration of hyperlink activations was found to affect the onset of endemic state in epidemic spreading processes [28], while the time ordering of hyperlink activations impacts the consensus reached in nonlinear consensus dynamics [29]. This body of research has mainly revealed how global properties of the entire temporal higher-order network influence the behavior of dynamical processes.

However, the role or contribution of a hyperlink in a spreading process, e.g., the number of nodes that are infected directly via the activation of the hyperlink, on a temporal higher-order network remains unexplored. Recently, Zhan et al. [13] studied the contribution of pairwise links in a diffusion process on a temporal pairwise network, finding that links that activate frequently earlier in time tend to contribute more when the infection probability is large. Contreras et al. [30] investigated spreading processes on static higher-order network and found that the parameters of the spreading process affect the probability of a node being directly infected by another node. To understand which kind of hyperlinks contribute more to the diffusion process unfolding on temporal higher-order networks, new methods are required due to the new dynamics of the diffusion process.

In this work, we aim to understand the role of hyperlinks and investigate which kind of hyperlinks, or hyperlinks with what properties, tend to contribute more to a spreading process on the temporal higher-order network. Empirical evidence has shown that in social phenomena, such as the diffusion of rumors or the adoption of norms and behaviors, contact with a single active neighbor is often insufficient to trigger adoption by an individual [31–33]. Moreover, effects such as peer-pressure can occur as a consequence of the simultaneous exposition to many active members in a group gathering. As a result, a number of generalized models on higher-order networks have been proposed recently to study social contagion, which is also referred as a spreading or diffusion process [34–36]. In this work, we model the social contagion process by generalizing the Susceptible-Infected threshold spreading process, which is originally defined on a static higher-order network [35], to a spreading process unfolding on a temporal higher-order network: initially, one seed node is infected while the other nodes are susceptible; when a hyperlink is active at any time, if the number of infected nodes within the hyperlink exceeds a threshold Θ, each susceptible node in the hyperlink is independently infected with a probability β. The threshold Θ reflects how many exposures to infected nodes in a group interaction are required to trigger the infection of each susceptible node within the group. We propose to represent the contribution of each hyperlink to a diffusion process by constructing the diffusion backbone. The diffusion backbone is a static higher-order network that is the union of all hyperlinks that appear in at least one diffusion trajectory of the spreading process, starting from an arbitrary seed node. Each hyperlink in the backbone is assigned a weight that reflects the average number of nodes directly infected through the activation of this hyperlink, over different choices of the seed node and realizations of the spreading process.

We construct the diffusion backbones for real-world temporal higher-order networks underlying SI threshold processes with diverse parameters, β and Θ. Eight temporal higher-order networks derived from human face-to-face interactions in various contexts are considered. Firstly, we investigate how the infection probability β and threshold Θ influence the constructed diffusion backbone. The backbone is shown to be dependent on the parameters of the spreading process via both experiments and theoretical analysis of the backbone when $\beta \to 0$. Secondly, we explore which properties of a hyperlink tend to results in a large weight of the hyperlink in the diffusion backbone thus a significant contribution to the diffusion process. We propose different centrality metrics or properties for hyperlinks in the underlying temporal higher-order network and use each centrality metric to estimate the ranking of hyperlinks in their contribution to the spreading process. Each proposed metric is based only on the partial temporal higher-order network observed at the hyperlinks itself and its neighboring hyperlinks. This allows efficient identification of hyperlinks that contribute more to the diffusion. For different ranges of the process parameters, different (parameter-free) metrics perform the best, approaching the optimal performance of complex centrality metrics with control parameters. This is further explained through physical and theoretical interpretations.

Our findings elucidate the contributions of hyperlinks to a spreading process process. These findings could be insightful for the design of strategies to facilitate (or suppress) the information spreading on temporal higher-order networks via e.g., incentivizing the activation of selected groups.

4.2. Methods

4.2.1. Temporal higher-order networks

A temporal (pairwise) network can be represented by $\mathscr{G} = (\mathscr{N}, \mathscr{C})$, where $\mathscr{N}$ is the set of nodes (or individuals), and $\mathscr{C} = \{(l(u,v),t) | u,v \in \mathscr{N}, t \in [1,T]\}$ is the set of events. Each event $(l(u,v),t) \in \mathscr{C}$ represents the pairwise interaction between node u and v occurring at discrete time t. A temporal network $\mathscr{G}$ can be aggregated along the time dimension, giving the (time) aggregated network, denoted as $G = (\mathscr{N}, C)$. A link between the node pair (u,v) exists in aggregated network G, i.e., $l(u,v) \in C$, if and only if there is at least one interaction in the temporal network between u and v during $[1,T]$. Each link $l(u,v) \in C$ in the aggregated network $G(\mathscr{N}, C)$ is indexed with an integer j, and the j-th link l_j is associated with a weight w_j, which is the number of times that link l_j has been activated in the temporal network.

However, people often gather in larger groups where more than two individuals interact simultaneously. The classic pairwise representation of temporal networks is limited in describing such group interactions, requiring the formalism of temporal higher-order networks. A temporal higher-order network (or temporal hypergraph) is represented as $\mathscr{H} = (\mathscr{N}, \mathscr{E})$, where $\mathscr{E} = \{(h(u_1,...,u_d), t) | u_1,...,u_d \in \mathscr{N}, t \in [1,T]\}$ is the set of higher-order interactions (events) involving an arbitrary number of nodes from the node set $\mathscr{N}$. Each higher-order event $(h(u_1,...,u_d), t) \in \mathscr{E}$ denotes a group interaction among the d individual nodes at time t, where $h(u_1,...,u_d) = \{u_1,...,u_d\}$ denotes an order-d hyperlink among a set of the corresponding d individuals. The size d of the group is also referred to as the order of hyperlink $h(u_1,...,u_d)$. For example, the order of a dyadic and triadic hyperlink are 2 and 3 respectively. The higher-order aggregated network of a temporal higher-order network $\mathscr{H}$ is denoted as $H = (\mathscr{N}, E)$, where E is the set of hyperlinks. Hyperlink $h(u_1,...,u_d)$ exists in E if and only if hyperlink $h(u_1,...,u_d)$ has been activated at least once in the temporal higher-order network. Each hyperlink in the aggregated network H is indexed with an integer j and associated with a weight w_j, where $j \in [1,|E|]$ and $|E|$ is the total number of hyperlinks in H. The weight w_j represents the number of times that hyperlink h_j has been activated during $[1,T]$. We denote the activation of link h_j in the temporal higher-order network by a time series $x_j(t)$, where $t \in [1,T]$ and $x_j(t) = 1$ if hyperlink h_j is activated at time t, otherwise $x_j(t) = 0$. A temporal higher-order network $\mathscr{H}$ can thus be equivalently represented by its aggregated network H with each link h_j in H further associated with its activity time series $x_j(t)$.

4.2.2. DIFFUSION PROCESS ON TEMPORAL HIGHER-ORDER NETWORKS

We consider a social contagion process on a temporal higher-order network, where each node is in one of two states at any time: infectious or susceptible. Initially, one seed node is infected while the other nodes are susceptible. Susceptible nodes can be infected through interactions with other infected nodes: when a hyperlink is active at any time, if the number of infected nodes within the hyperlink exceeds a threshold Θ, each susceptible node within the hyperlink gets infected independently with a probability β; otherwise, it remains susceptible. The traditional Susceptible-Infected (SI) process on a temporal pairwise network can be regards as a special case of the above diffusion process when all hyperlinks in $\mathscr{H}(\mathscr{N}, \mathscr{E})$ are dyadic (order-2), and the threshold is $\Theta = 1$. When $\Theta = 1$ and $\beta = 1$, the diffusion process on a higher order temporal network becomes equivalent to the traditional SI process on the corresponding temporal pairwise network, where each higher-order interaction in the temporal higher-order network is treated as interactions between each pair of nodes within the hyperlink. We consider two cases in this work: $\Theta = 1$ and $\Theta = d-1$. When $\Theta = d-1$, the threshold for a hyperlink is dependent on the order d of the hyperlink. Hence, hyperlinks of a higher order have a higher threshold.

Furthermore, the aforementioned diffusion model can be considered as a adjusted version of the model studied in Ref. [35], with the following distinctions. We consider a discrete time Susceptible-Infectious (SI) diffusion process on temporal higher-order networks. In contrast, a continuous time Susceptible-Infectious-Susceptible (SIS) process on static higher-order networks has been studied in Ref. [35]. Our choice of discrete time process is because the underlying temporal higher-order networks evolve at discrete time. We start with the simple SI process instead of SIS process, which requires both network data of longer periods and better method to understand and identifies the steady state.

4.2.3. Empirical datasets

We apply our analysis to 8 real-world human physical contact datasets from *SocioPatterns*. These datasets contain collections of face-to-face interactions among individuals in various social contexts, including hospital (Hospital), primary school (Primaryschool2013), high school (Highschool2012, Highschool2013), workplace (Workplace2015), museum (Infectious), and conferences (HT2009, SFHH). The face-to-face interactions are recorded as pairwise contacts, where an interaction is stored when two individuals face each other at a distance of approximately $\lesssim 1.5$ meters over a 20-second interval. Each original dataset naturally records only pairwise interactions, from which we deduce the corresponding temporal higher-order network via the common method already used in the literature [19, 25–27]. Specifically, at any time t, if there are $d(d-1)/2$ pairwise interactions between each nodes pair of a set of d nodes, thus forming a clique, we promote these $d(d-1)/2$ pairwise interactions to an interaction of order d. For example, three temporal links $(l(a,b),t)$, $(l(b,c),t)$ and $(l(a,c),t)$ in the temporal network $\mathcal{G}(\mathcal{N},\mathcal{C})$ at time t are considered as a single temporal hyperlink $(h(a,b,c),t)$ in the corresponding temporal higher-order network $\mathcal{H}(\mathcal{N},\mathcal{E})$. Since a clique of order d contains all its sub-cliques of orders $d' < d$, only the maximal clique is promoted to a higher-order event. Furthermore, we preprocess the datasets by removing time steps without any interaction in the whole network and also excluding nodes that are not in the largest connected component of the higher-order aggregated network H. The basic statistics of each preprocessed dataset are presented in Table 4.1.

Network	$\lvert\mathcal{N}\rvert$	$\lvert E\rvert$	$\lvert\mathcal{E}\rvert$	T
infectious	410	3350	14725	1393
primaryschool	242	12704	106879	3101
highschool2012	180	2645	42105	11274
highschool2013	327	7818	172035	7376
hospital	75	1825	27835	9454
ht09	113	2434	19037	5247
workplace15	217	4903	73823	18489
SFHH	403	10541	54306	3510

Table 4.1: Statistics of real-world temporal higher-order networks after data processing. The number of nodes $\lvert\mathcal{N}\rvert$, the number of hyperlinks $\lvert E\rvert$, the number of higher-order events $\lvert\mathcal{E}\rvert$, and the number of time steps T are shown.

4.2.4. Diffusion backbone

Given a temporal higher-order network $\mathcal{H}(\mathcal{N},\mathcal{E})$ and a diffusion process unfolding on $\mathcal{H}$, we quantify the contribution of a hyperlink (a group of nodes) to the diffusion process as the average number of nodes that are directly infected via the activation of the hyperlink. Specifically, we propose the following construction of a diffusion backbone to represent the contribution of each hyperlink to the diffusion process. The diffusion backbone of $\mathcal{H}(\mathcal{N},\mathcal{E})$ is a higher-order (static) network denoted as $B(\mathcal{N},E_B)$. At time step $t = 0$, one seed node is infected while all the other nodes are susceptible. For each seed node i, we construct a diffusion trajectory $\mathscr{T}_i(\Theta,\beta)$ as the union of the hyperlinks through which at least one susceptible node gets infected directly, during time period $[1,T]$. Each hyperlink h_j in the trajectory $\mathscr{T}_i$ is associated with a weight $w_j^{\mathscr{T}_i}$ denoting the number of nodes in h_j that are infected through the activation of hyperlink h_j directly. The diffusion backbone is then defined as the union of all diffusion trajectories starting from each node in $\mathscr{N}$, i.e., $B(\Theta,\beta)=\cup_{i\in\mathscr{N}}\mathscr{T}_i(\Theta,\beta)$. Each hyperlink h_j in $B(\Theta,\beta)$ is associated with a weight w_j^B, which is the average weights of the same hyperlink h_j over all $|\mathscr{N}|$ diffusion trajectories, i.e., $w_j^B=\frac{1}{|\mathscr{N}|}\sum_{i\in\mathscr{N}}w_j^{\mathscr{T}_i}$. An illustration of the construction of the diffusion backbone in case of $\beta=1$ is shown in Figure 4.1. In case of $0<\beta<1$, the diffusion process is stochastic, the diffusion backbone can be obtained by averaging over multiple independent realizations. In this work, we choose the number of realizations as $5\cdot 10^4$. The convergence of the backbone as the number of realizations increases is discussed in Appendix 4.5.2. The backbone $B(\mathscr{N},E_B)$ encodes how many nodes on average are infected directly through each hyperlink when the seed node node is randomly selected. By definition, the total weights of all hyperlinks in the backbone $B(\beta,\Theta)$ plus one is equal to the average number $|\mathscr{N}|\rho$ of infected nodes during time $[1,T]$ per seed node, i.e., $\sum_{j\in E_B}w_j^B+1=|\mathscr{N}|\rho$, since each infected node except the seed node leads to an increment of 1 in the weight of a hyperlink .

Now, we derive the backbone analytically for the limiting case when $\beta\to 0$. When $\Theta=1$, the diffusion backbone $B(\beta\to 0,\Theta=1)$ approaches the higher-order aggregated network $H(\mathscr{N},E)$ in topology, which can be explained as follows. Firstly, consider an arbitrary node i as the seed node and one of its 1-hop neighbors v in the higher-order aggregated network, i.e., i and v have at least a (group) interaction. The probability that the information diffuses from node i to v through an interaction that involves both nodes is β. Similarly, the total probability that i infects v via a hyperlink h_j is βw_j, where w_j is the weight of hyperlink h_j in the aggregated higher-order network, or equivalently, the total number of activations of h_j in the temporal higher-order network. The total probability that the information diffuses from node i to v is the total weight of all hyperlinks that include both i and v in the aggregated higher-order network times β, thus of order β. Furthermore, consider a 2-hop neighbor u of seed node i in the aggregated higher-order network H, i.e., node u has interactions with at least one 1-hop neighbor of i but has no interaction with node i. The probability that the information diffuses from node i to a one hop neighbor of i, which spreads the information further to node u is proportional to β^2, which is negligibly small compared to the probability for i to infect a first-hop neighbor. Hence, the diffusion trajectory $\mathscr{T}_i$ starting from any seed node i approaches the ego network of node i in the aggregated higher-order network H, which comprises node i, its 1-hop neighbors and any hyperlink that includes i and at least one of its 1-hop neighbors. Hence, the diffusion backbone $B(\beta\to 0,\Theta=1)$, which is the union of all $|\mathscr{N}|$ diffusion trajectories, has the same topology as the aggregated higher-order network H. The weight of an arbitrary hyperlink h_j in the backbone is $w_j^B=\beta|h_j|(|h_j|-1)\cdot w_j\cdot\frac{1}{|\mathscr{N}|}$, because the activation of h_j could spread the information only if a component node of h_j is the seed and then each of the other $|h_j|-1$ component nodes could possible get infected.

Consider the backbone when $\Theta=d-1$ and $\beta\to 0$. The same analysis applies to the weight w_j^B of a dyadic hyperlink, thus the weight w_j^B of a dyadic hyperlink is the same as that in the backbone $B(\beta \rightarrow 0, \Theta = 1)$, i.e., $w_j^B = 2\beta w_j \cdot \frac{1}{|\mathcal{N}|}$. The weight of an order-3 hyperlink h_j in $B(\beta \rightarrow 0, \Theta = d-1)$ approaches $\frac{2}{|\mathcal{N}|}\beta^2 \sum_{t=1}^{T} x_j(t) \cdot \left(\sum_{l \in \mathscr{L}^{sub}(j)} \sum_{\iota < t} x_l(\iota)\right)$ where the function $x_j(t)$ indicates whether the target hyperlink h_j is activated at time t (i.e., $x_j(t) = 1$) or not (i.e., $x_j(t) = 0$) and $\mathscr{L}^{sub}(j)$ includes all dyadic hyperlinks that share two common nodes with h_j. This second-order estimation can be explained as follows. Only when the seed node is a component node of h_j, could the activation of h_j at any time t infect a node with a probability of order β^2: the probability for a second node in h_j to get infected before t is β times the total number of dyadic interactions the seed node has with the component nodes in h_j before t. Taking into account the possibility that each component node of h_j could be the seed node, the total probability that a second component node is infected before t is 2β times the total number of activations $\sum_{l \in \mathscr{L}^{sub}(j)} \sum_{\iota < t} x_l(\iota)$ of all sub-links $\mathscr{L}^{sub}(j)$ of h_j before t. If a second component node gets infected before t, the activation of h_j could infect a third component node with probability β. If the seed node is outside h_j, the probability for the hyperlink to infect another node is negligibly small compared to that when the seed is a component node of h_j.

4.2.5. OBSERVATION TIME WINDOWS

Each real-world temporal higher-order network has a unique observation window $[1, T]$, determined by its measurement. To better understand the relationship between a hyperlink's network properties and its contribution to the diffusion process, should we consider the full observation window or only a portion of it as the dataset? To address this, we examine the evolution of the average prevalence $\rho(t)$ over time, where $1 \leq t \leq T$ and T is the duration of original observation window, for each of the eight empirical temporal higher-order networks. This helps us understand events occurring at which time period may contribute to the spreading, thus is relevant for the construction of the diffusion backbone. In case of $\Theta = 1$ and $\beta = 1.0$, the evolution of average prevalence ρ at each timestamp is shown in Figure 4.2. In some networks like *highschool2013*, the average prevalence increases rapidly in a short time and hardly anymore afterward, while in other networks like *infectious*, the average prevalence increases continuously over time. This implies that in *highschool2013*, the diffusion backbone $B(\Theta = 1, \beta = 1.0)$ is mainly determined by the interactions that occur in the early period, while in *infectious*, all interactions during $[1, T]$ could contribute to the diffusion process. The above observation motivates us to consider various observation time windows from each data set when exploring the relation between properties of a hyperlink and its contribution in the diffusion process. Specifically, for each temporal higher-order network, we consider three time windows of different lengths, which are derived as follows. We examine the average prevalence $\rho(t)$ when $\Theta = 1$ and $\beta = 1.0$, where $1 \leq t \leq T$. We consider the following three observation windows, i.e., $[1, T_{p\%}]$, where $T_{p\%}$ is the earliest time when the average prevalence ρ reaches $p\% \cdot \rho(T)$, i.e., $T_{p\%} = \min\{t : \rho(t) \geq p\% \cdot \rho(T)\}$, and $p \in \{30, 60, 90\}$. For each time window, we construct a temporal higher-order network that includes all the higher-order events occur during $[1, T_{p\%}]$. Eventually, a total of three temporal higher-order networks are derived from each dataset. In the following, we will

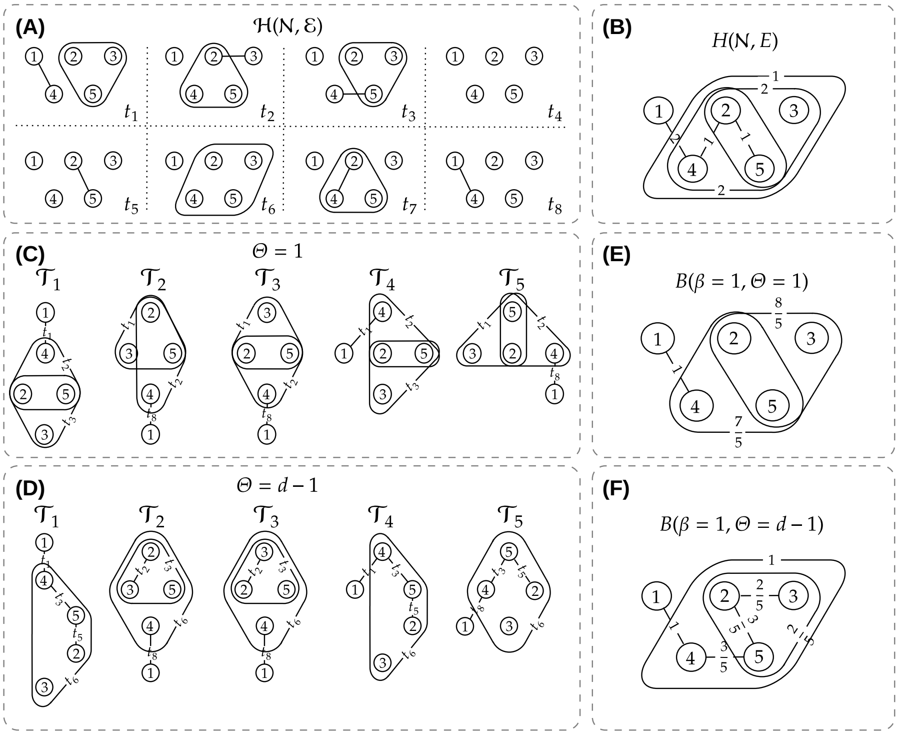

Figure 4.1: An illustration of the construction of the diffusion backbone of a temporal higher-order network when the infection probability of the diffusion model is $\beta = 1.0$. (A) A temporal higher-order network $\mathcal{H}$ with 5 nodes and $T = 8$ time steps. (B) The time-aggregated higher-order network H, with a weight w_j associated with each hyperlink j. (C-D) Diffusion trajectory $\mathcal{T}_i(\beta = 1)$ when $\Theta = 1$ (panel C) and $\Theta = d - 1$ (panel D) for each possible seed node i. (E-F) The diffusion backbone B when $\Theta = 1$(panel E) and when $\Theta = d - 1$ (panel F).

present the results for the time window with the largest length, $[1, T_{90\%}]$, and we refer to the Appendix 4.5.3 for results related to other time windows, which lead to qualitatively similar findings.

4.3. RESULTS

In this section, we will construct the diffusion backbone for each of the real-world temporal higher-order networks listed in Table 4.1 with its corresponding time window, e.g., $[1, T_{90\%}]$. We focus on two extreme cases for the threshold of the diffusion model: $\Theta = 1$ and $\Theta = d - 1$, with the infection probability β ranging from 0.001 to 1.0 on a logarithmic scale, i.e., $\beta \in [0.001, 1.0]$, to construct the backbones. To understand which kind of hyperlinks are used more in the diffusion process thereby acquire higher weights in the backbone under different parameters Θ and β, we will conduct the following analysis. Firstly, we will explore whether and how diffusion backbone changes with the two

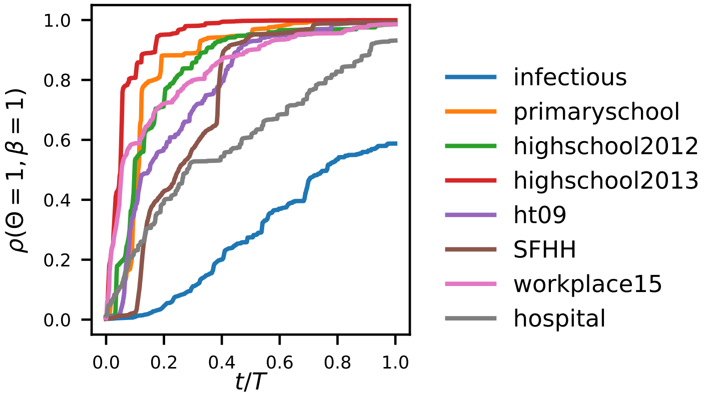

Figure 4.2: The average prevalence ρ of the spreading process with threshold $\Theta = 1$ and infection probability $\beta = 1.0$ on each original temporal network over time t. The time steps are normalized by the total number of time steps T of each dataset.

parameters β and Θ of the diffusion model. This analysis will help understand whether different properties of hyperlinks should be used to estimate the ranking of hyperlinks by their weight in the backbone when the parameters of the diffusion process vary. Secondly, we will propose a set of hyperlink centrality metrics that capture diverse properties of a hyperlink in the temporal higher-order network. By examining the correlation between each proposed metric and the weight of a hyperlink in the backbone, we will uncover hyperlinks with what properties acquire higher weights in the diffusion backbone under different process parameters. In the following, we will present the results for *SFHH* dataset with time window $[1, T_{90\%}]$, as the results for other datasets and different time windows show qualitatively similar trends (see Appendix).

4.3.1. Influence of process parameters on the backbone

In Section 4.2.4, we have analytically derived the weighted backbone in the limit case of $\beta \rightarrow 0$. The backbone is shown to differ when Θ varies. Now, we explore how the backbone $B(\beta, \Theta)$ changes as the infection probability β and threshold Θ vary.

Total weight and number of links and in the backbone

We first examine the prevalence of the diffusion process, which is equivalent to the total weights of all hyperlinks in the backbone plus one. In Figure 4.3 (A), we show the prevalence as a function of the infection probability β. The wide range of prevalence under different β and Θ shows that our constructed backbones span a broad dynamical space. Given the threshold Θ, the prevalence increases with the infection probability β, suggesting that hyperlinks contribute more to the diffusion. The prevalence gets suppressed when Θ changes from 1 to $d-1$, because for the larger threshold $\Theta = d-1$, the diffusion through hyperlinks of higher orders ($d > 2$) is impeded, which motivates us to explore the contribution of hyperlinks of different orders separately. The relative contribution c_d^B of order-d hyperlinks to a diffusion process is reflected by the total weight

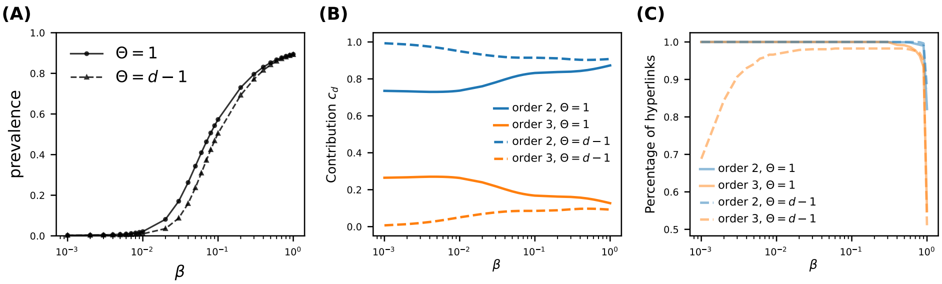

Figure 4.3: (A): The prevalence as a function of β when $\Theta = 1$ and $\Theta = d-1$ respectively. (B): The contribution of order-2 and order-2 hyperlinks as a function of β when $\Theta = 1$ and $\Theta = d-1$, respectively. (C): The percentage of order-2 and order-3 hyperlinks that appear in in backbone B as a function of β when $\Theta = 1$ and $\Theta = d-1$. Results are shown for the *SFHH* dataset.

of all order-d hyperlinks in the backbone normalized by the total weight of all hyperlinks in the backbone. The relative contribution c_d^B indicates the percentage of infected nodes that are infected through the activations of order-d hyperlinks. Hence, $\sum_d c_d^B = 1$. Since the number of hyperlinks of orders $d > 3$ in each temporal higher-order network is small, we focus on hyperlinks of order 2 and order 3. Figure 4.3 (B) shows the contribution c_d^B of order 2 and order 3, respectively, under two distinct thresholds: $\Theta = 1$ and $\Theta = d-1$, as β increases. Generally, dyadic (order-2) hyperlinks exhibit the highest contribution, indicating that the majority of node infections occur through dyadic interactions. This is because order-2 interactions comprise the largest proportion of interactions in all the temporal higher-order networks and their threshold to spread information $\Theta = d-1 = 1$ is relatively low. Given the same β, the relative contribution of order-3 links when $\Theta = d-1$ is smaller than it is when $\Theta = 1$, as expected. Let us consider the case when $\Theta = d-1$. When $\beta \to 0$, $c_2^B \to 1$ and the contribution of higher orders ($d > 2$) approaches zero, which is consistent with our theoretical analysis (Section 4.2). In this case, only few nodes get infected such that the number of infected nodes in a hyperlink of order $d > 2$ can hardly reach $d-1$. As β grows, more nodes get infected, which increases the chance that the information diffuses through triadic (higher-order) interactions. As a consequence, the contribution of triadic (dyadic) hyperlinks increases (decreases). When $\Theta = 1$, the contribution of order-3 hyperlinks decreases as β increases, because nodes within a triadic link could possibly get infected via interactions of dyadic links (large in number) before the activation of the triadic link, reducing the probability for a triadic link to diffuse the information.

Figure 4.3 (C) shows the percentage of hyperlinks of a specific order d in the aggregated higher-order network H that appear in the backbone $B(\beta, \Theta)$. When $\Theta = 1$ and $0 < \beta < 1$, each hyperlink in the temporal higher-order network $\mathcal{H}$ has a nonzero probability to appear in diffusion trajectories starting from every possible seed node. Thus, the diffusion backbone $B(\Theta = 1, 0 < \beta < 1)$ contains all hyperlinks in the aggregated network H. While in the deterministic case of $\beta = 1$, the diffusion from a source node to a target node follows the time-respecting paths that arrives at the target node the earliest in time. Hence, only hyperlinks that are present in such paths from one node to any other node are included in the diffusion backbone. This explains the drop in the fraction of hyperlinks that appears in the backbone as $\beta \to 1$, independent of Θ. When $\Theta = d-1$ and $0 < \beta < 1$, each dyadic (order-2) hyperlink in the aggregated network appears in the backbone with a non-zero probability. However, not all order-3 hyperlinks necessarily appear in the backbone, because the condition $\Theta = d-1$ for an order-3 hyperlink to diffuse the information is possibly difficult to meet, especially when β is small.

Correlation between backbones under different process parameters

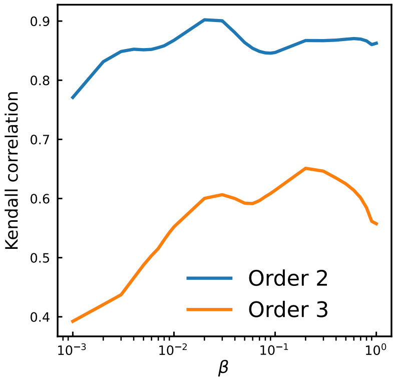

Figure 4.4: The Kendall correlation between $w_j^B(\Theta = 1)$ and $w_j^B(\Theta = d-1)$ for order-2 hyperlinks and order-3 hyperlinks, in the *SFHH* dataset.

In this work, we aim to understand if a certain property of hyperlinks in the temporal higher-order network can be used to estimate the rankings of hyperlinks of the same order d by the weight w_j^B in the backbone. In the following, we study how the rankings of hyperlinks of a given order d by the weight w_j^B change with process parameters β and Θ. If the backbone changes with process parameters, different properties of hyperlinks maybe needed to identify hyperlinks with the highest backbone weight.

Consider hyperlinks of order d in the aggregated higher-order network H. We investigate the Kendall rank correlation between their weights in backbone $B(\beta, \Theta = 1)$ and backbone $B(\beta, \Theta = d-1)$. Figure 4.4 shows that the Kendall correlation for order-2 hyperlinks is in general higher than for order-3 hyperlinks, independent of β. This suggests that different properties of hyperlinks may be needed to estimate the weights of order-3 hyperlinks in $B(\beta, \Theta = 1)$ and in $B(\beta, \Theta = d-1)$, respectively. Furthermore, we examine the Kendall correlation between the weights of these hyperlinks in the backbones constructed with two different infection probabilities and the same threshold , which is shown in Figure 4.5. We find that when the difference between the two infection probabilities is small (large), the correlation is large (small). This suggests that backbones constructed under different infection probabilities differ in their link weights and different properties of hyperlinks may be required to estimate the ranking of hyperlinks in

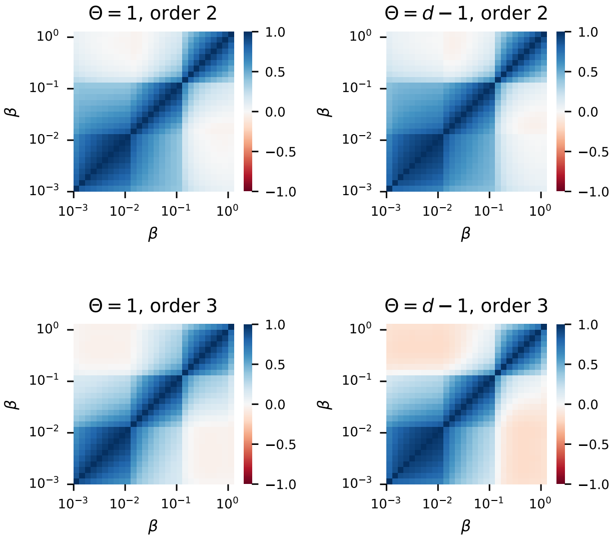

Figure 4.5: The Kendall correlation between the weight $w_j^B(\beta,\Theta)$ of an order-d hyperlink obtained for two different infection probabilities, β, in the *SFHH* dataset. Two left (right) panels correspond to the case of $\Theta = 1$ ($\Theta = d-1$). Two upper (lower) panels show the Kendall correlation for hyperlinks of order $d = 2$ (order $d = 3$).

their backbone weight as β varies.

4.3.2. CENTRALITY METRICS FOR HYPERLINKS BASED ON LOCAL TEMPORAL HIGHER-ORDER NETWORKS

Given a temporal higher-order network $\mathcal{H}$, which network property of a hyperlink is correlated with the weight of the hyperlink in the backbone $B(\beta,\Theta)$ under each parameter set (β,Θ)? We will design different centrality metrics for a hyperlink, each reflecting a specific property of a hyperlink within its local temporal higher-order network. In the next subsection, we will investigate how well each metric can be used to estimate the ranking of hyperlinks based on their contribution to the diffusion process.

The motivation for using local network information is threefold. Firstly, the information of local connections is more accessible than the global network information. Secondly, computing centrality metrics based on local connections is more efficient than those based on a larger set of connections. Thirdly, local connections of a hyperlink could be more relevant than the rest of the temporal higher-order network, thus sufficient in estimating the weight of the hyperlink in the backbone, particularly when infection probability β is relatively low. As discussed in Section 4.2.4, when $\beta \to 0$ and $\Theta = 1$, only 1-hop neighbors of the seed node could possibly get infected. Hence, the backbone weight of a hyperlink depends only on the number of times the hyperlink has been activated and the order of the hyperlink. As β increases, the nodes that are further than 1-

hop from the seed node may also get infected with a nonzero probability. In this case, the activity (temporal connection) of neighboring hyperlinks of a target hyperlink could influence the weight of the target hyperlink in the backbone. For example, neighboring hyperlinks with a large number of activations may infect the component nodes of the target hyperlink, enabling the target hyperlink to meet its threshold condition and contribute to the spreading process.

We design local centrality metrics based on the activity of the target hyperlink itself and the activity of its neighboring hyperlinks. Since dyadic hyperlinks are more abundant than higher-order hyperlinks in each of the considered temporal higher-order networks, for an arbitrary target hyperlink h_j, we consider two different sets of neighboring dyadic hyperlinks respectively: all *adjacent dyadic hyperlinks* that share at least one common node with the target h_j, denoted as set $\mathscr{L}^{adj}(j)$, and the *dyadic sub-hyperlinks* that include all dyadic hyperlinks that share two common nodes with h_j, denoted as set $\mathscr{L}^{sub}(j)$. Any dyadic sub-hyperlink is also an adjacent dyadic hyperlink, so $\mathscr{L}^{sub}(j) \subseteq \mathscr{L}^{adj}(j)$. We propose firstly four centrality metrics capturing static or temporal properties of these two types of local neighborhoods, respectively. We refer to Table 4.2 for a summary of the notation used in this chapter.

Static adjacent hyperlink based metric ξ_j^{adj} of hyperlink h_j is defined as:

$$\xi_j^{adj}(\alpha) = w_j \cdot \left(1 + \sum_{l \in \mathscr{L}^{adj}(j)} w_l\right)^{\alpha}, \tag{4.1}$$

Static sub-hyperlink based metric ξ_j^{sub} is defined similarly as equation (4.1), except that the set of sub-hyperlinks $\mathscr{L}^{sub}(j)$ is considered instead of adjacent hyperlinks$\mathscr{L}^{adj}(j)$:

$$\xi_j^{sub}(\alpha) = w_j \cdot \left(1 + \sum_{l \in \mathscr{L}^{sub}(j)} w_l\right)^{\alpha}. \tag{4.2}$$

The static metrics ξ_j^{adj} and ξ_j^{sub} are determined by the number of activations of the target hyperlink, w_j, as well as the total number of activations of neighboring hyperlinks in set $\mathscr{L}^{adj}(j)$ and $\mathscr{L}^{sub}(j)$, respectively. The parameter α is the scaling parameter, which is a real constant and determines the contribution of neighboring hyperlinks to the metrics. Each proposed centrality metric is used to estimate the ranking of hyperlinks in backbone weight. Hence, using the logarithm of the metric e.g., $\log(\xi^{adj}) = \log w_j + \alpha \log\left(1 + \sum_{l \in \mathscr{L}^{adj}(j)} w_l\right)$, to predict the ranking is the same as using ξ_j^{adj}. This reveals also how α controls the relative contribution of the adjacent links and the target link itself. These two metrics are motivated by the possibility that a hyperlink that has many activations in its neighborhood and in itself may contribute more to a diffusion process.

Temporal adjacent hyperlink based metric Ξ_j^{adj} further considers the time ordering between the activations of the target hyperlink and the activations of hyperlinks in

$\mathscr{L}^{adj}(j)$. We define the metric Ξ_j^{adj} of a hyperlink h_j as:

$$\Xi_j^{adj}(\alpha) = \sum_{t=1}^{T} x_j(t) \cdot \left(1 + \sum_{l \in \mathscr{L}^{adj}(j)} \sum_{\iota < t} x_l(\iota)\right)^{\alpha}, \tag{4.3}$$

Temporal sub-hyperlink based metric Ξ_j^{sub} is defined in the same way as Ξ^{adj}, except that the contribution of neighboring hyperlinks in set $\mathscr{L}^{sub}(j)$ is considered. The metric Ξ_j^{sub} of a hyperlink h_j is defined as:

$$\Xi_j^{sub}(\alpha) = \sum_{t=1}^{T} x_j(t) \cdot \left(1 + \sum_{l \in \mathscr{L}^{sub}(j)} \sum_{\iota < t} x_l(\iota)\right)^{\alpha}. \tag{4.4}$$

In equation (4.3) and (4.4), the function $x_j(t)$ indicates whether the target hyperlink h_j is activated at time t (i.e., $x_j(t) = 1$) or not (i.e., $x_j(t) = 0$). Taking the metric Ξ_j^{adj} as an example, it assigns a weight to each activation of the target hyperlink h_j. The weight for an activation at time t, given by $(1+\sum_{l\in\mathscr{L}^{adj}(j)}\sum_{\iota<t} x_l(\iota))^{\alpha}$, depends on the total number of activations of neighboring hyperlinks in set $\mathscr{L}^{adj}(j)$ that occurred before time t. Metric Ξ_j^{adj} tends to be large if hyperlink h_j has a large number of events (activations), and prior to each of its events, a large number of events have already occurred in adjacent hyperlinks. Consider an order-3 target hyperlink, when $\Theta = d-1$ and β is small, the target hyperlink could possibly infect its component nodes only when it is active (has an event) and two component nodes have already been infected at that time, which is more likely to happen when the neighboring hyperlinks have a large number of activations before. When β is large, the large number of events occurring in the adjacent links before the activation of the target link could lead to the infection of all the component nodes of the target link, reducing the chance for the target hyperlink to infect its component nodes. Both scenarios motivate us to examine the number of events occurring in adjacent links before each event at the target hyperlink. Actually, the design of Ξ_j^{sub} is directly motivated by theoretical weight of an order-3 hyperlink in $B(\beta \to 0, \Theta = d-1)$ derived in Section 4.2.4, i.e., $\frac{2}{|\mathscr{N}|}\beta^2\sum_{t=1}^{T} x_j(t)\cdot\left(\sum_{l\in\mathscr{L}^{sub}(j)}\sum_{\iota<t} x_l(\iota)\right)$. Hence, Ξ_j^{sub} with $\alpha = 1$ is supposed to estimate well the ranking of order-3 links in their weights in $B(\beta \to 0, \Theta = d-1)$.

When $\alpha = 0$, all metrics are equal to the weight w_j of the target hyperlink in the higher-order aggregated network, i.e., $\xi_j^{adj}(0) = \xi_j^{sub}(0) = \Xi_j^{adj}(0) = \Xi_j^{sub}(0) = w_j$. When $\alpha > 0$ ($\alpha < 0$), the events in the neighborhood either $\mathscr{L}^{adj}(j)$ or $\mathscr{L}^{sub}(j)$, contribute positively (negatively) to each centrality metric. Order 2 hyperlinks have no neighboring sub-hyperlinks, i.e., the set $\mathscr{L}^{sub}(j)$ is empty, which means $\xi_j^{sub}(\alpha) = \Xi_j^{sub}(\alpha) = w_j$.

Furthermore, we consider a metric that has been proposed in [13] and used to estimate the diffusion backbone for temporal pairwise networks when $\beta = 1$.

Time-scaled weight Ω_j of hyperlink h_j is defined as:

$$\Omega_j(\mu) = \sum_{t=1}^{T} x_j(t) \cdot t^{\mu}, \tag{4.5}$$

where μ is the scaling parameter. When $\mu = 0$, the metric Ω_j equals the weight w_j of hyperlink h_j in the aggregated network. When $\mu < 0$ ($\mu > 0$), the metric Ω_j assigns higher (smaller) weights to activations that occur earlier in time. This metric is motivated by the possibility that a hyperlink that has activations that are large in number and early in time may contribute more to a diffusion process, especially when β is relatively large.

Finally, we combine the proposed temporal metrics, which capture the relatively time ordering of the activations of neighboring hyperlinks and the activations of the target hyperlink, with the metric time-scaled weight that captures the times of the activations of the target hyperlink itself.

Combined metrics Φ_j^{adj} and Φ_j^{sub} combine the time-scaled weight Ω_j with metric Ξ_j^{adj} and Ξ_j^{sub}, respectively.

$$\Phi_j^{adj}(\mu,\alpha) = \sum_{t=1}^{T} x_j(t)\cdot t^{\mu}\cdot\left(1+\sum_{l\in\mathscr{L}^{adj}(j)}\sum_{\iota<t} x_l(\iota)\right)^{\alpha}, \tag{4.6}$$

$$\Phi_j^{sub}(\mu,\alpha) = \sum_{t=1}^{T} x_j(t)\cdot t^{\mu}\cdot\left(1+\sum_{l\in\mathscr{L}^{sub}(j)}\sum_{\iota<t} x_l(\iota)\right)^{\alpha}, \tag{4.7}$$

where the real constants μ_j and α are two scaling parameters. We have $\Phi_j^{adj}(\mu=0,\alpha) = \Xi_j^{adj}$, $\Phi_j^{sub}(\mu=0,\alpha) = \Xi_j^{sub}$, $\Phi_j^{adj}(\mu,\alpha=0) = \Phi_j^{sub}(\mu,\alpha=0) = \Omega_j(\mu)$.

4.3.3. ESTIMATING HYPERLINK WEIGHT w_j^B USING LOCAL CENTRALITY METRICS

NEW METRICS WITH A SINGLE PARAMETER α

We examine the performance of each centrality metric with the scaling parameters $\alpha \in [-3,3]$ and $\mu \in [-3,3]$ in predicting the ranking of hyperlinks of a specific order d by the weights w_j^B. It is challenging to pre-select the parameter(s) of a metric when performing such a prediction task, especially when the number of parameters is large. Hence, we first evaluate the performance of the four new metrics with a single parameter α: $\xi_j^{adj}(\alpha)$, $\xi_j^{sub}(\alpha)$, $\Xi_j^{adj}(\alpha)$, $\Xi_j^{sub}(\alpha)$, our main focus. The objectives are to (1) understand how the scaling parameter α affects their performance and whether certain values of the scaling parameter lead to the highest performance of a metric across all networks, and (2) compare the performance of these four metrics. Later in Section 4.3.4, we will compare the performance of these four metrics with the other three metrics we have introduced in Section 4.3.2, time-scaled weight from literature and two combined metrics. The objective is to explore whether these four metrics with one parameter or with evident choice of the parameter could perform close to the combined metrics with two scaling parameters.

	Variable	Description
Higher-order network	$\mathcal{H}$	Temporal higher-order network
	H	Time-aggregated higher-order network of $\mathcal{H}$
	$\mathcal{N}$	The set of nodes in $\mathcal{H}$
	$\mathcal{E}$	The set of higher-order events in $\mathcal{H}$
	E	The set of hyperlinks in H
	T	Total number of time steps in $\mathcal{H}$
	h_j	Hyperlink in H
	w_j	Weight of hyperlink h_j in H
Diffusion model and backbone	β	Infection probability
	Θ	Threshold
	$\mathcal{T}_i$	Diffusion trajectory when node i is the seed
	B	Diffusion backbone
	w_j^B	Weight (contribution) of h_j in B
	c_d^B	Contribution of order d in B
Centrality metrics	$\xi_j^{adj}(\alpha)$	Static adjacent hyperlink based metric
	$\xi_j^{sub}(\alpha)$	Static sub-hyperlink based metric
	$\Xi_j^{adj}(\alpha)$	Temporal adjacent hyperlink based metric
	$\Xi_j^{sub}(\alpha)$	Temporal sub-hyperlink based metric
	$\Omega_j(\mu)$	Time-scaled weight proposed in ref. [13]
	Φ_j^{adj}	Combined metric that combines Ω_j and Ξ_j^{adj}
	Φ_j^{sub}	Combined metric that combines Ω_j and Ξ_j^{sub}

Table 4.2: Notation used in this chapter.

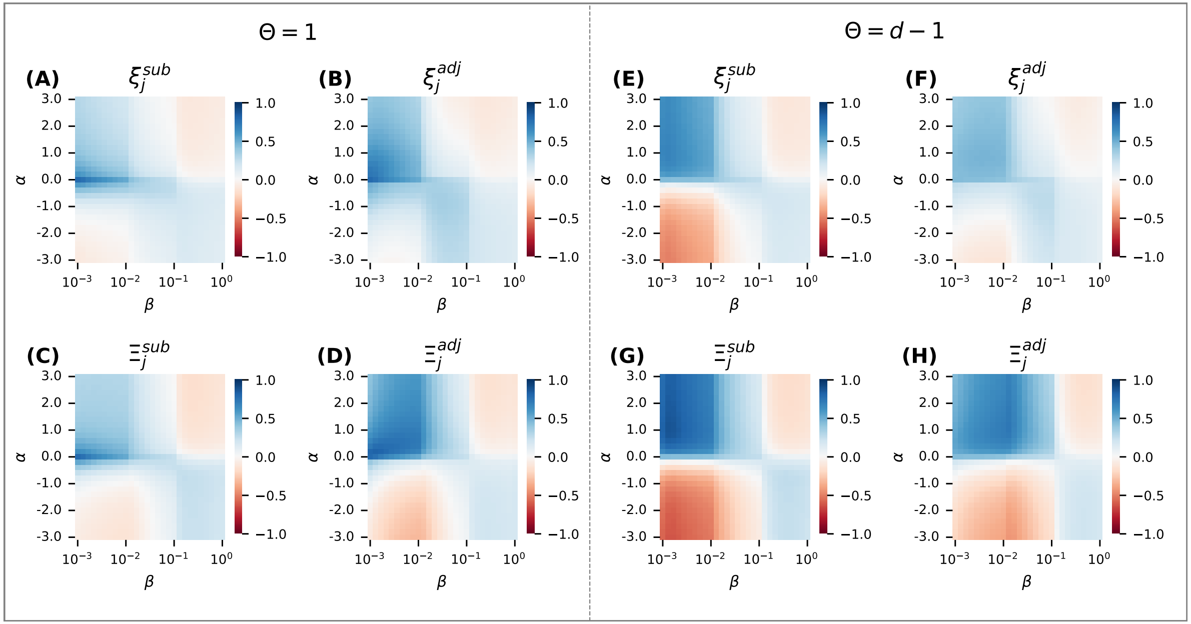

Figure 4.6: The Kendall correlation between centrality metric $\xi_j^{adj}(\alpha)$, $\xi_j^{sub}(\alpha)$, $\Xi_j^{adj}(\alpha)$ or $\Xi_j^{sub}(\alpha)$ and the weight w_j^B of the triadic (order-3) hyperlink in the backbone B, as a function of the infection probability β and the scaling parameter α, for $\Theta = 1$ (left panels, A-D) and $\Theta = d - 1$ (right panels, E-H). The results are shown for the *SFHH* dataset.

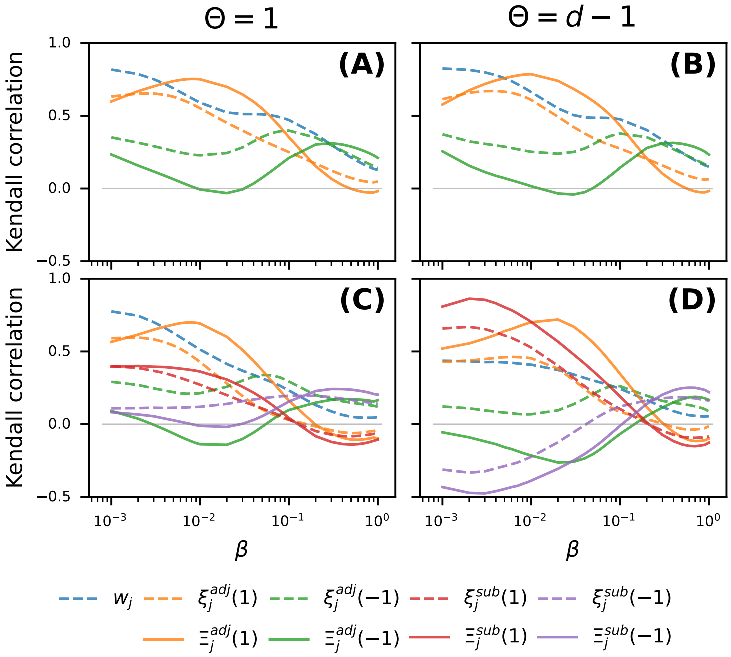

Figure 4.7: The Kendall correlation between a local metric with the scaling parameter $\alpha \in \{0, 1, -1\}$ and the weight w_j^B of the hyperlink in the backbone B, as a function of infection probability β, in the *SFHH* dataset. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

Figure 4.6 (A-H) show the Kendall rank correlation between each of the four new metrics and the weight w_j^B of an order-3 hyperlink as a function of the scaling parameter α and the infection probability β, when $\Theta = 1$ (panel (A-D)) and $\Theta = d-1$ (panel (E-H)). A general observation is that when β is relatively small (large), each metric with a positive (negative) α of a hyperlink is positively correlated with the weight of the hyperlink in the backbone. This is also observed for order-2 hyperlinks. Indeed, when β is small, a large number of activations of neighboring hyperlinks could increase the chance that nodes in the target hyperlink get infected, which increases the chance that activations of the target hyperlink afterwards could infect other nodes. However, when β is sufficiently large, the chance that all component nodes in the target hyperlink get infected through the activations of neighboring hyperlinks may increase, which suppresses the chance that the activation of the target hyperlink afterwards could infect other nodes.

Given the parameters β and Θ, the best prediction performance of each metric is roughly achieved when $\alpha = 0, 1$ or -1 in each real-world network, which is further shown by the comparison of the performance of each metric with $\alpha = 0, 1, -1$ and its best performance across $\alpha \in [-3, 3]$ in Figure 4.15 in Appendix. Hence, we will focus metrics w_j, $\xi_j^{adj}(1)$, $\xi_j^{adj}(-1)$, $\Xi_j^{adj}(1)$, $\Xi_j^{adj}(-1)$, $\xi_j^{sub}(1)$, $\xi_j^{sub}(-1)$, $\Xi_j^{sub}(1)$, $\Xi_j^{sub}(-1)$, where the parameter has been selected. Since order-2 hyperlinks have no sub-hyperlinks, which makes $\xi_j^{sub}(\alpha) = \Xi_j^{sub}(\alpha) = w_j$ according to the definitions, only metrics, w_j, $\xi_j^{adj}(1)$, $\xi_j^{adj}(-1)$, $\Xi_j^{adj}(1)$, $\Xi_j^{adj}(-1)$, will be considered for order-2 hyperlinks.

Figure 4.7 shows the Kendall correlation between each metric with $\alpha \in \{0, 1, -1\}$ with the weight w_j^B of a hyperlink as a function of β, for order-2 and order-3 hyperlinks respectively. Generally, the best-performing centrality metric varies depending on the process parameters Θ and β, because the backbone is dependent on the process parameters. Consider the region of $\beta \to 0$. Metric $\Xi_j^{sub}(1)$ exhibits the highest correlation with the weight w_j^B of order-3 hyperlinks in backbone $B(\beta \to 0, \Theta = d-1)$, as shown in Figure 4.6 (D), while metric w_j is the best-performing metric to estimate the weight w_j^B of order-2 hyperlinks in backbone $B(\beta \to 0, \Theta = d-1)$ (Figure 4.7 (B)) and to estimate the weight w_j^B of hyperlinks of both order 2 (Figure 4.7 (A)) and order 3 (Figure 4.7 (C)) in backbone $B(\beta \to 0, \Theta = 1)$. These two observations are in line with the backbone $B(\beta \to 0, \Theta)$ derived analytically in Section 4.2.4.

As β increases but is still small ($< 10^{-1}$), metric $\Xi_j^{adj}(1)$ outperforms the other metrics. In this range of β, nodes that are more than 1 hop away from the seed could be infected with a non-zero probability. A larger number of activations of the adjacent hyperlinks could lead to a higher probability of infection of the component nodes of the target hyperlink, which makes the target hyperlink more likely to meet the threshold to infect other component node(s) when activated.

When β is large ($> 10^{-1}$), our temporal metrics with $\alpha = -1$ tend to perform the best. However, their performance is still relatively low, i.e., the correlation with backbone weight is evidently lower than 0.5. This suggests that when β is large, the diffusion becomes global, local network information alone may be insufficient to well predict the contribution of a hyperlink to the diffusion process.

In general, when the infection probability β is not large ($\beta < 10^{-1}$), either $\Xi_j^{adj}(1)$,

$\Xi_j^{sub}(1)$ or w_j performs best, depending on the process parameters and the order of hyperlinks. These observations are qualitatively similar in other datasets and different observation time windows, though the specific range of β for the a metric to perform the best may vary (see Appendix).

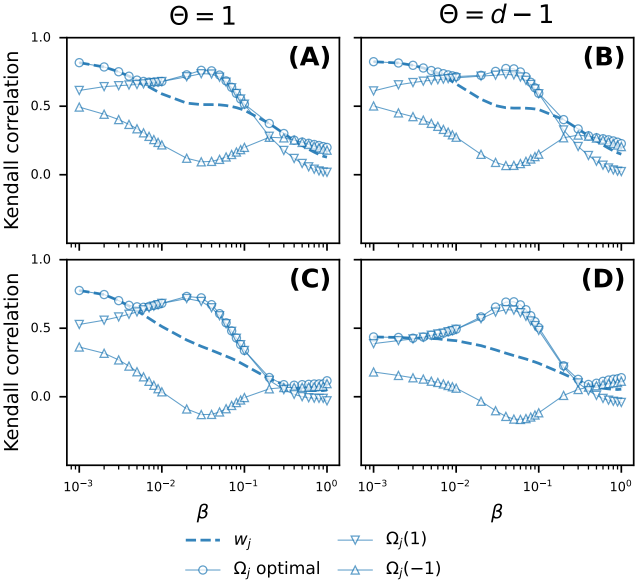

Figure 4.8: The Kendall correlation between the time-scaled weight $\Omega_j(\mu)$ with $\mu \in \{0,1,-1\}$ and the weight $w_j^B(\beta,\Theta)$ of a hyperlink in the backbone, in *SFHH* dataset. This is compared with the optimal Kendall correlation (circles) of $\Omega_j(\mu)$ achieved all possible choices of $\mu \in [-3,3]$. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d-1$, respectively.

4.3.4. COMPARISON WITH OTHER TEMPORAL CENTRALITY METRICS

We further compare the performance of $\Xi_j^{adj}(1)$, $\Xi_j^{sub}(1)$ and w_j with the other three metrics: the time-scaled weight $\Omega_j(\mu)$ and the two combined metrics $\Phi_j^{adj}(\mu,\alpha)$, $\Phi_j^{sub}(\mu,\alpha)$, in predicting the ranking of hyperlinks of a given order by their weights in the backbone $B(\beta,\Theta)$.

We first investigate whether the metric $\Omega_j(\mu)$ with a specific choice of the scaling parameter μ tends to lead to the highest prediction performance across all networks. As shown in Figure 4.8, when β is small ($\beta < 10^{-2}$), the optimal/highest Kendall correlation between metric $\Omega_j(\mu)$ and the weight w_j^B, is achieved approximately by $w_j = \Omega_j(0)$. When β increases but is still smaller than 10^{-1}, the performance of $\Omega_j(1)$ is close to the optimal Kendall correlation, indicating that hyperlinks that activate frequently later in time tend to contribute more. As β increases further ($\beta > 10^{-1}$), the optimal Kendall correlation is approximately achieved by metric $\Omega_j(-1)$, suggesting that hyperlinks that

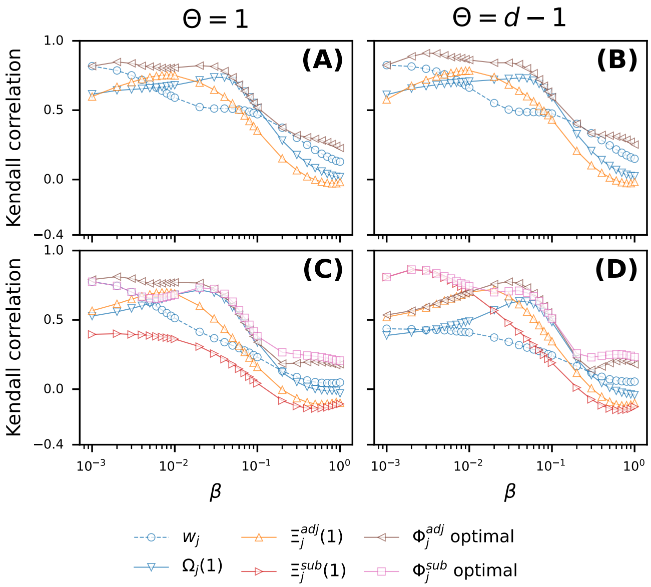

Figure 4.9: The Kendall rank correlation between the four metrics w_j, $\Omega_j(1)$, Ξ_j^{adj} and Ξ_j^{sub} and the weight w_j^B of a hyperlink in the backbone, in comparison to the optimal Kendall correlation between each combine metric and the weight w_j^B, in *SFHH* dataset. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

activate earlier in time tend to contribute more to the diffusion process when β is large. Next, we will compare so far the best performing metrics $\Xi_j^{adj}(1)$, $\Xi_j^{sub}(1)$, w_j and $\Omega_j(1)$, which do not require parameter calibration, with the two combined metrics in estimating the ranking of hyperlinks of a given order by their weights in the backbone $B(\beta, \Theta)$.

Given the infection probability β and the threshold Θ, we perform a grid search for the scaling parameters μ and α within the range $[-3.0, 3.0]$ to find the highest Kendall correlation of each combined metric when estimating the ranking of order-2 and order-3 hyperlinks, respectively. The optimal performance of a combined metric, say Φ_j^{adj}, is the upper bound for the performance of Ξ_j^{adj}, w_j and the time-scaled weight Ω_j, which are special cases of the combined metric. As shown in Figure 4.9, the best performance achieved by metrics $\Xi_j^{adj}(1)$, $\Xi_j^{sub}(1)$, w_j and $\Omega_j(1)$ is close to the optimal performance of the two combined metrics when $\beta < 10^{-1}$. We compare further the performance of the four metrics without the need for parameter calibration: $\Xi_j^{adj}(1)$, $\Xi_j^{sub}(1)$, w_j and $\Omega_j(1)$. When $\beta \to 0$, either metric w_j or $\Xi_j^{sub}(1)$ performs the best depending on Θ and the order of hyperlinks under estimation, which is in line with our analytical analysis. As β increases ($\beta \approx 10^{-2}$), $\Xi_j^{adj}(1)$ outperforms. A large number of activations of the adjacent hyperlinks could cause the infection of component nodes in a target hyperlink thus fulfilling the threshold condition for the target to infect other nodes. As β increases further but still within the range of $\beta < 10^{-1}$, $\Omega_j(1)$ performs the best. A hyperlink with a large number of activations that occur relatively late in time tend to contribute more the spreading process. Activations that occur late, when more nodes are infected thus the threshold condition is likely met, tend to contribute effectively to the spreading process. In summary, for different ranges of the infection rate β, different parameter free centrality metrics estimate the backbone weights the best, close to the optimal performance of combined metrics achieved by parameter searching.

4.4. CONCLUSION AND FUTURE WORK

CONCLUSION AND FUTURE WORK

In this chapter, the contribution of a hyperlink in a temporal higher-order network to a spreading process is defined as the average number of nodes infected via the activation of the hyperlink. We explored which properties of a hyperlink in the temporal higher-order network lead to a high contribution. A generalized Susceptible-Infected threshold process with infection probability β and threshold Θ is considered on eight real-world temporal higher-order networks derived from human face-to-face contacts in various contexts. Firstly, we proposed the construction of the diffusion backbone B where the weight of each hyperlink equals the the contribution of the hyperlink to the diffusion process starting from a random seed node. In the limiting case when $\beta \to 0$, the backbone weight of a hyperlink was derived analytically based on local temporal connections around the hyperlink. Secondly, we illustrated the dependency of the backbone $B(\beta, \Theta)$ on the two process parameters β and Θ. This is also evidenced by the different backbones derived when Θ varies and $\beta \to 0$. To explore which properties of hyperlinks are associated with high contributions to the diffusion process, we designed centrality metrics for a hyperlink to estimate the ranking of hyperlinks by their weights in backbone $B(\beta, \Theta)$. Each proposed metric is defined based on the activity of the target hyperlink and the activity of its neighboring hyperlinks in the temporal higher-order network. Different metrics are shown to predict the best for different parameter sets (β, Θ) of the process, approaching the optimal performance of combined metrics achieved via parameter searching. The reason why certain properties of a hyperlink result in a high contribution to the process is further explained.

There are several limitations in this work that call for further exploration. Firstly, we only considered the Susceptible-Infected threshold process as the diffusion model on temporal higher-order networks and focused only on two extreme cases for the threshold Θ. A more comprehensive investigation of diffusion models is needed. Secondly, it is interesting to explore other types of temporal higher-order networks, such as scientific collaborations, which may have different properties than the human interaction networks considered in this work. Thirdly, the backbone constructed captures the contribution of each hyperlink. This definition can be extended to study the contribution of each node. Finally, it is promising to investigate how to use the backbone or our proposed centrality metrics that well estimate the backbone to mitigate the dynamic process on the network via the blocking of selected hyperlinks.

4.5. APPENDIX

4.5.1. STATISTICS OF EMPIRICAL DATASETS

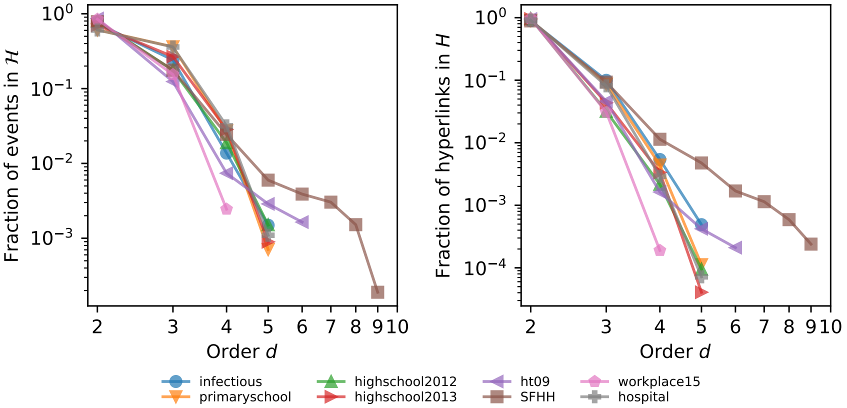

Figure 4.10: The fraction of events (hyperlinks) that are of order d in each real-world temporal higher-order network with the original observation time window $[1, T]$ is shown in the left (right) panel.

Dataset	$T_{30\%}$	$T_{60\%}$	$T_{90\%}$
infectious	519	784	1104
primaryschool	287	359	994
highschool2012	916	1477	3664
highschool2013	195	395	1252
hospital	1266	3942	7702
ht09	437	1154	2361
workplace	574	2133	8773
SFHH	483	1124	1421

Table 4.3: The number of time steps in the observation time windows $[1, T_{p\%}]$ that we choose for each empirical dataset.

4.5.2. CONVERGENCE OF DIFFUSION BACKBONE

We explore whether $R = 50000$ realizations of the diffusion process starting from each seed node is sufficient to obtain a representative backbone for $0 < \beta < 1$. Given the infection probability β and the threshold Θ, we construct the diffusion backbones $B(\beta, \Theta)$ resulting from different numbers R of independent realizations. Then, we measure the Kendall correlation between the backbones in weight resulting from R realizations and 50000 realizations, for order-2 and order-3 hyperlinks, respectively.

Figure 4.11 and Figure 4.12 show that ss the number of realizations R grows, the Kendall correlation $\tau(R)$ increases quickly and approaches one, suggesting that the ranking of hyperlinks by their weights in the backbone gradually becomes stable. This supports our choice of $R = 50000$ realizations. Similar trends are observed when changing the observation time window, as shown in Figure 4.13 and Figure 4.14.

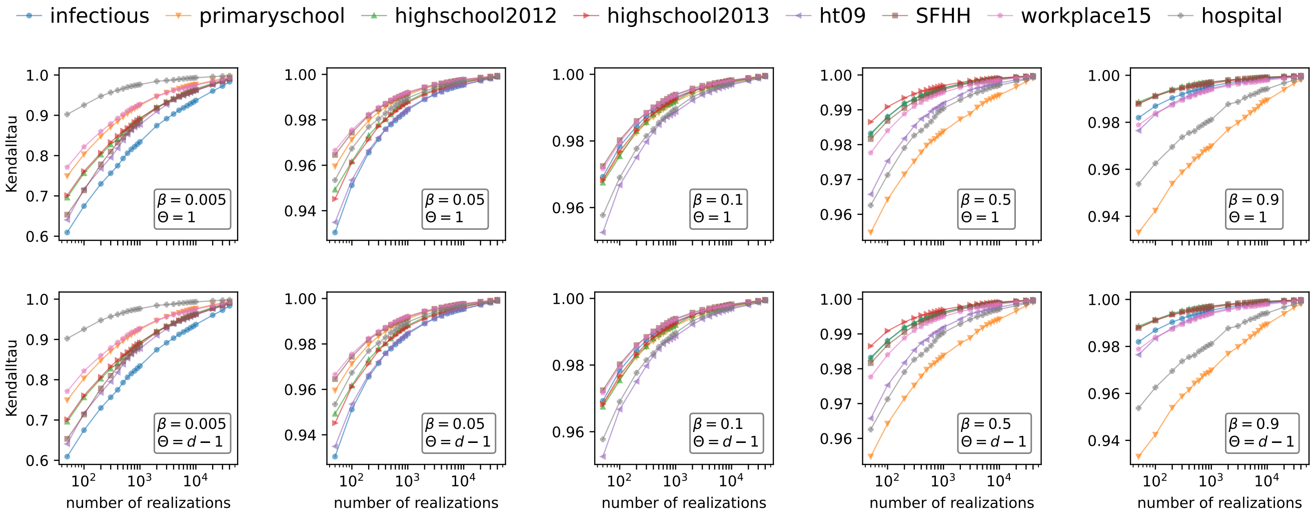

Figure 4.11: The Kendall correlation between the weights in the backbones resulting from R realizations and 50000 realizations, as a function of R for order-2 hyperlinks. This is shown for each empirical dataset with the observation time window $[1, T_{90\%}]$.

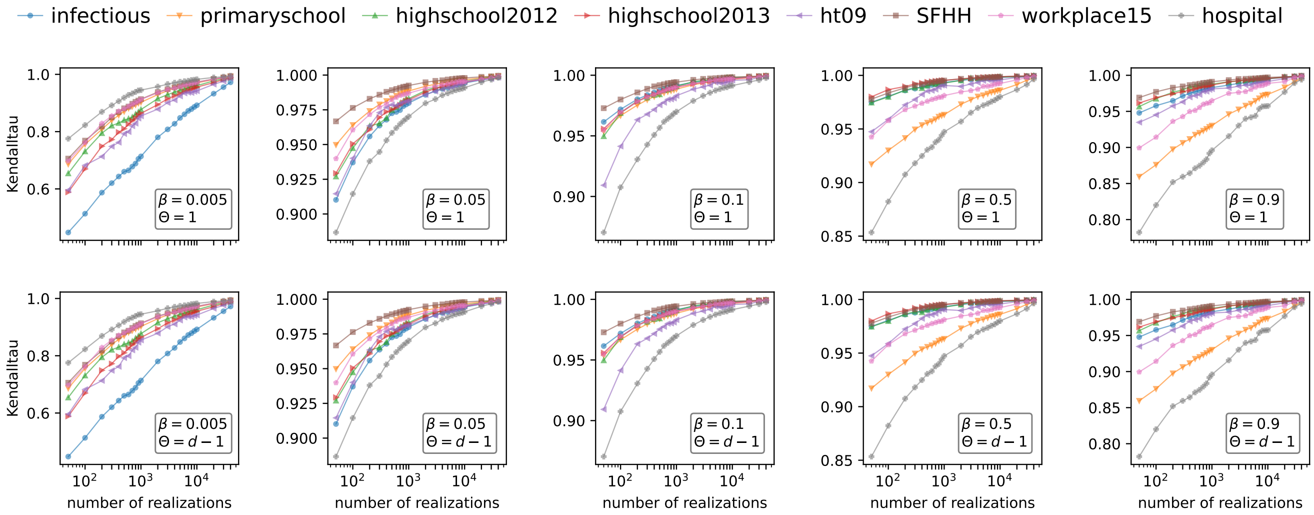

Figure 4.12: The Kendall correlation between the weights in the backbones resulting from R realizations and 50000 realizations, as a function of R for order-3 hyperlinks. This is shown for each empirical dataset with the observation time window $[1, T_{90\%}]$.

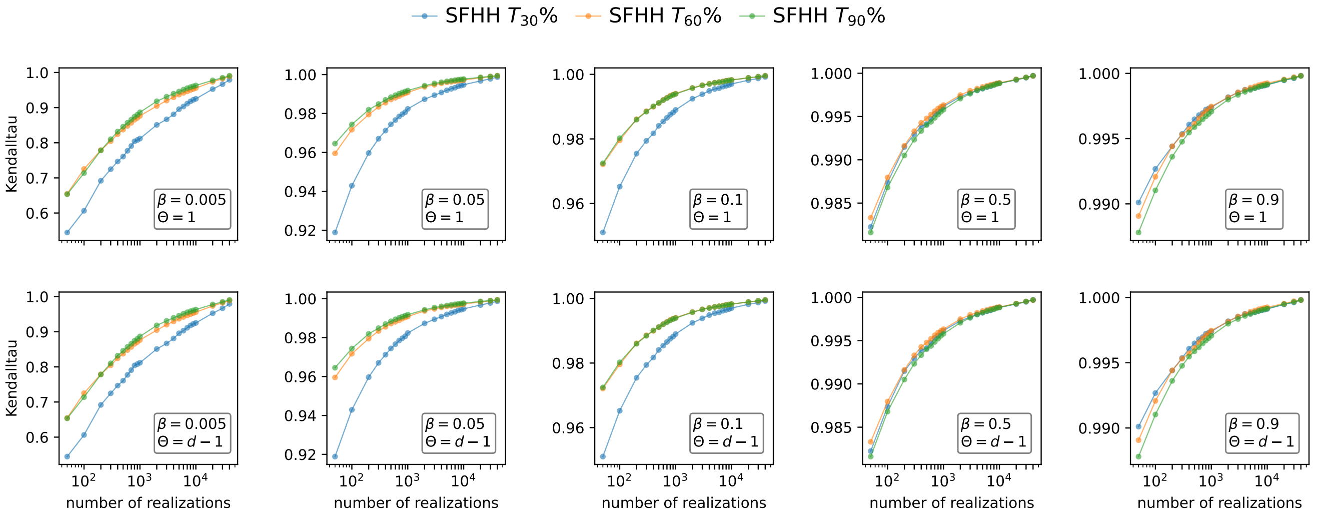

Figure 4.13: The Kendall correlation between the weights in the backbones resulting from R realizations and 50000 realizations, as a function of R for order-2 hyperlinks, in *SFHH* dataset with different observation time windows.

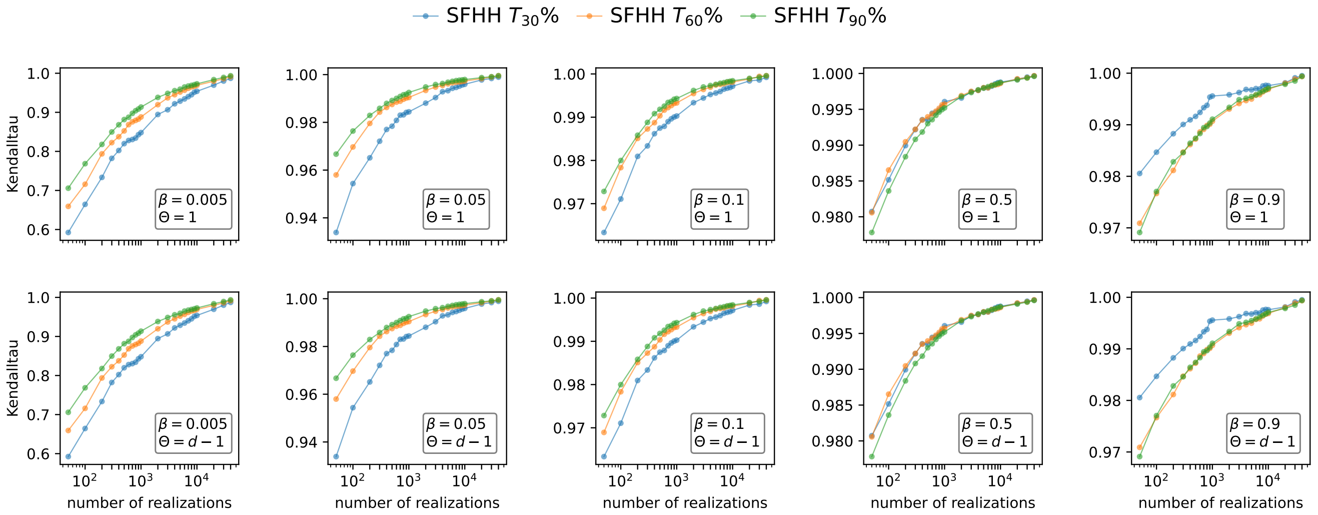

Figure 4.14: The Kendall correlation between the weights in the backbones resulting from R realizations and 50000 realizations, as a function of R for order-3 hyperlinks, in *SFHH* dataset with different observation time windows.

4.5.3. RESULTS FOR *SFHH* DATASET

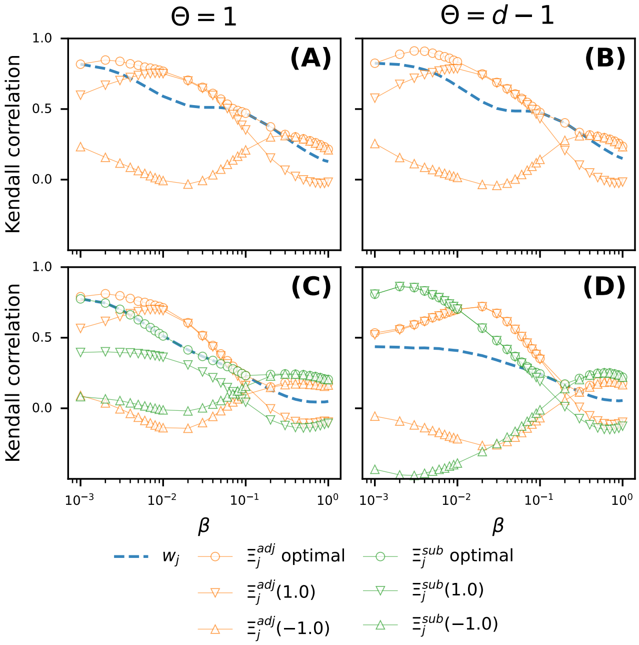

Figure 4.15: The *SFHH* dataset with the observation time window $[0, T_{90\%})$. The Kendall rank correlation between the metrics w_j, $\Xi_j^{adj}(1)$, $\Xi_j^{adj}(-1)$, $\Xi_j^{sub}(1)$ and $\Xi_j^{sub}(-1)$, and the weight w_j^B of a hyperlink, in comparison to the optimal Kendall correlation between metrics, Ξ_j^{adj}, Ξ_j^{sub}, and the weight w_j^B. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

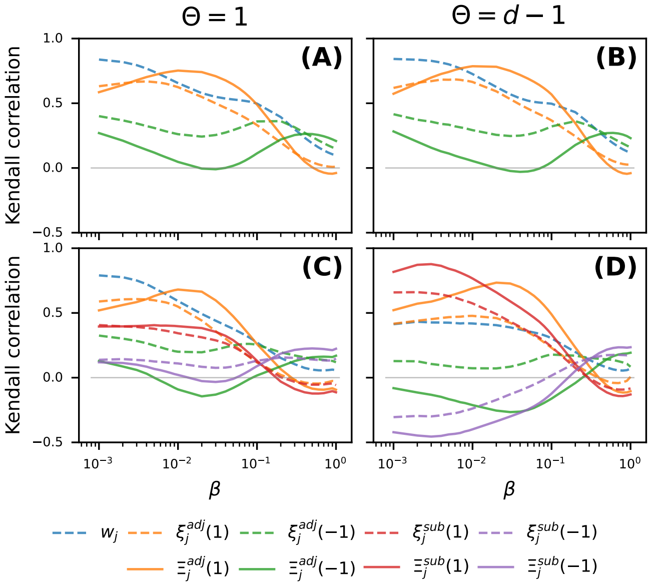

Figure 4.16: The Kendall correlation between a centrality metric with the scaling parameter $\alpha \in \{0, 1, -1\}$ and the weight w_j^B of a hyperlink in the backbone B, as a function of infection probability β, in *SFHH* dataset with the time observation time window $[0, T_{60\%}]$. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

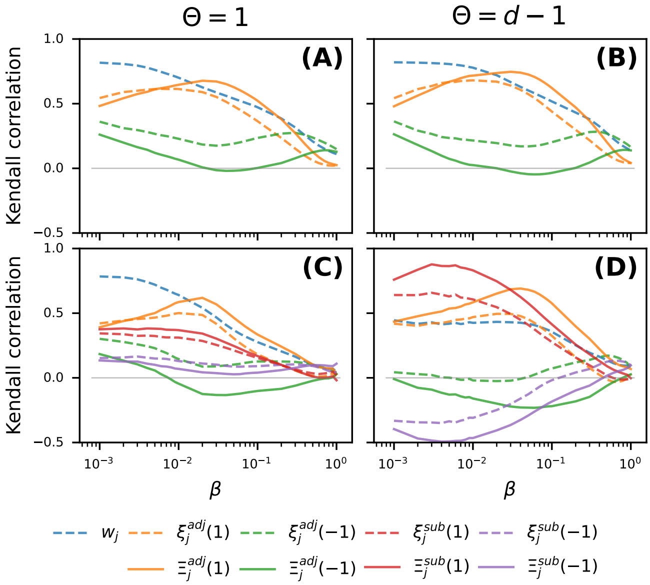

Figure 4.17: The Kendall correlation between a centrality metric with the scaling parameter $\alpha \in \{0, 1, -1\}$ and the weight w_j^B of a hyperlink in the backbone B, as a function of infection probability β, in *SFHH* dataset with the time observation time window $[0, T_{30\%}]$. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

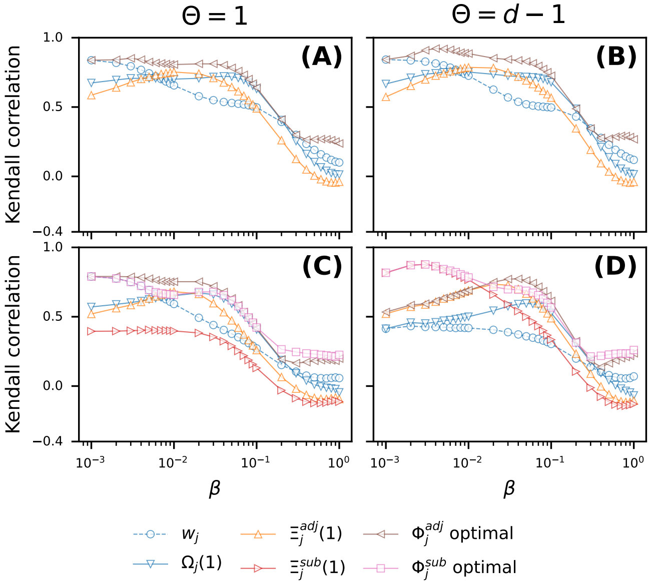

Figure 4.18: The Kendall rank correlation between each of the four metrics w_j, $\Omega_j(1)$, Ξ_j^{adj} and Ξ_j^{sub} and the weight w_j^B of a hyperlink in the backbone, in comparison to the optimal Kendall correlation between each combined metric and the weight w_j^B, in *SFHH* dataset with the time observation time window $[0, T_{60\%}]$. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

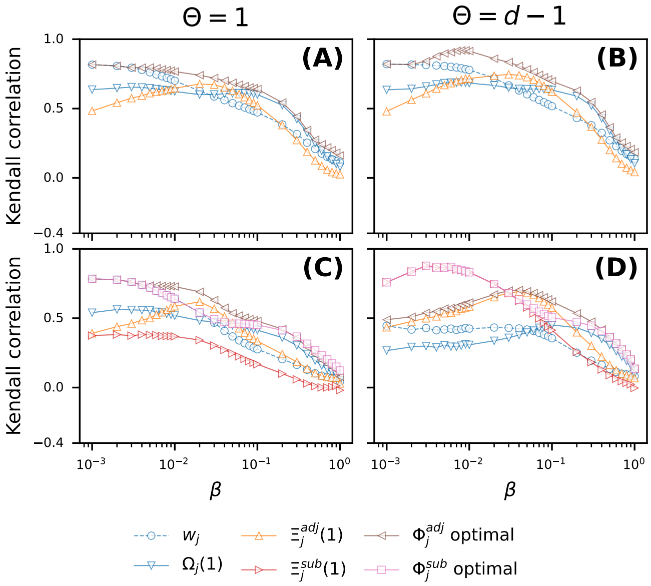

Figure 4.19: The Kendall rank correlation between each of the four metrics w_j, $\Omega_j(1)$, Ξ_j^{adj} and Ξ_j^{sub} and the weight w_j^B of a hyperlink in the backbone, in comparison to the optimal Kendall correlation between each combined metric and the weight w_j^B, in *SFHH* dataset with the time observation time window $[0, T_{30\%}]$. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d-1$, respectively.

4.5.4. RESULTS FOR OTHER DATASETS

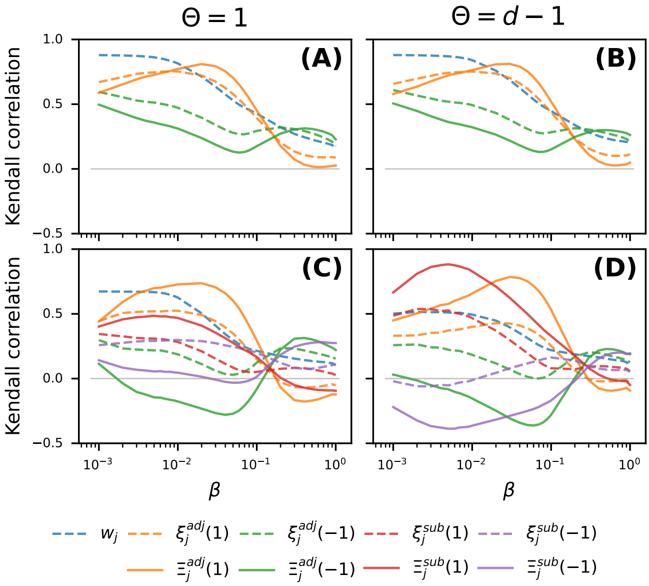

Figure 4.20: The *infectious* dataset. The Kendall correlation between a centrality metric with the scaling parameter $\alpha \in \{0, 1, -1\}$ and the weight w_j^B of a hyperlink in the backbone B, as a function of infection probability β. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

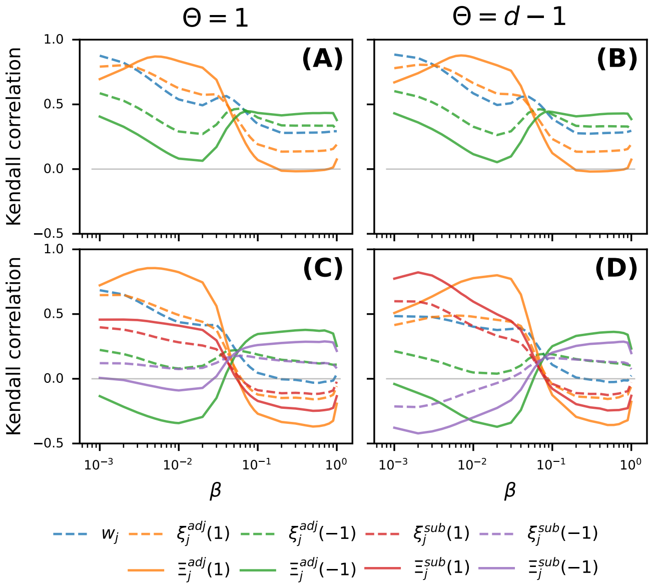

Figure 4.21: The *primaryschool* dataset. The Kendall correlation between a centrality metric with the scaling parameter $\alpha \in \{0, 1, -1\}$ and the weight w_j^B of a hyperlink in the backbone B, as a function of infection probability β. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

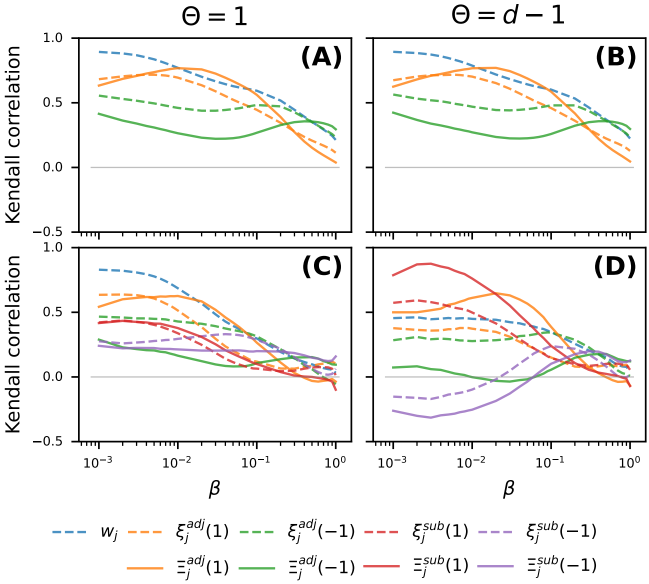

Figure 4.22: The *highschool2012* dataset. The Kendall correlation between a centrality metric with the scaling parameter $\alpha \in \{0, 1, -1\}$ and the weight w_j^B of a hyperlink in the backbone B, as a function of infection probability β. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

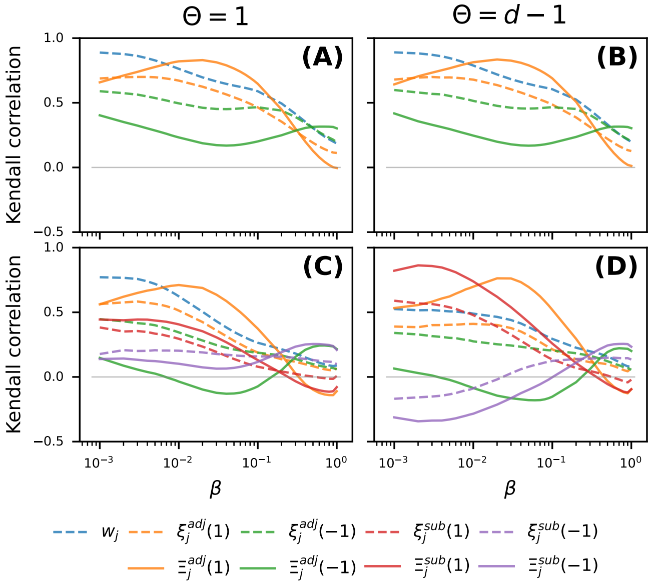

Figure 4.23: The *highschool2013* dataset. The Kendall correlation between a centrality metric with the scaling parameter $\alpha \in \{0, 1, -1\}$ and the weight w_j^B of a hyperlink in the backbone B, as a function of infection probability β. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

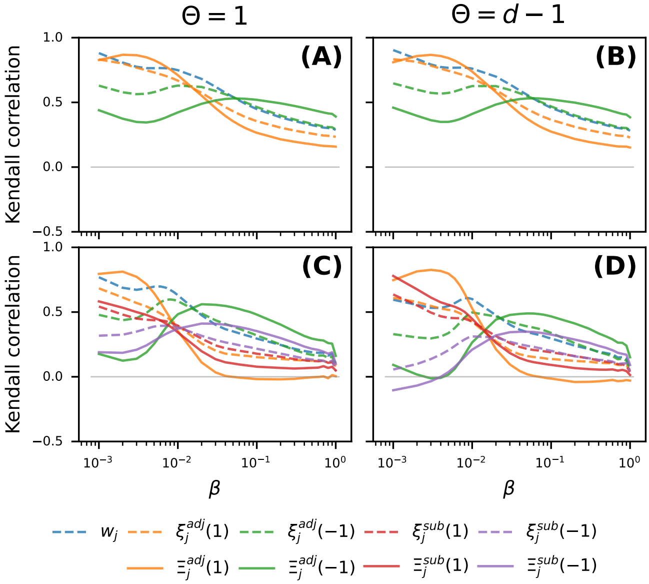

Figure 4.24: The *hospital* dataset. The Kendall correlation between a centrality metric with the scaling parameter $\alpha \in \{0, 1, -1\}$ and the weight w_j^B of a hyperlink in the backbone B, as a function of infection probability β. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

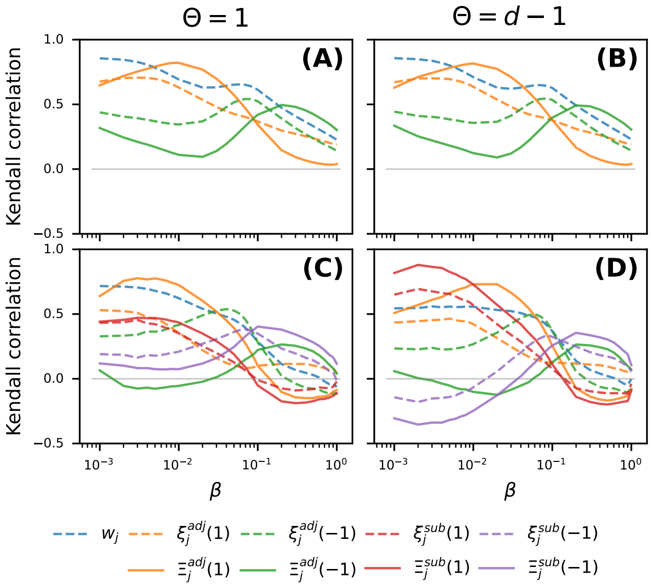

Figure 4.25: The *ht09* dataset. The Kendall correlation between a centrality metric with the scaling parameter $\alpha \in \{0,1,-1\}$ and the weight w_j^B of a hyperlink in the backbone B, as a function of infection probability β. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

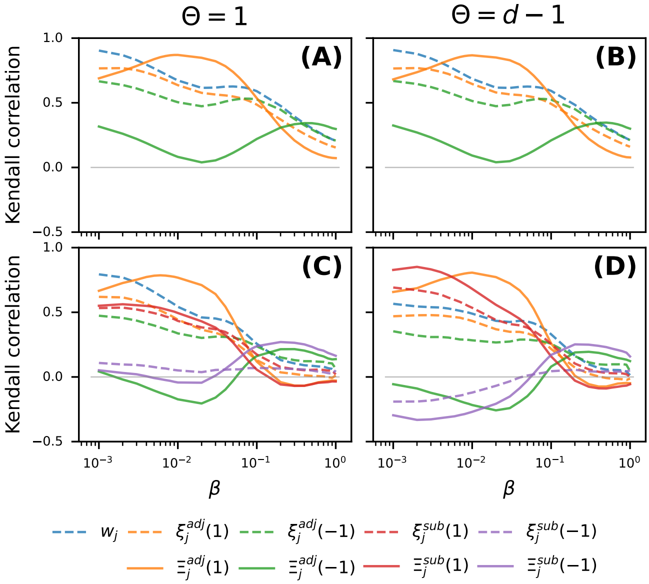

Figure 4.26: The *workplace15* dataset. The Kendall correlation between a centrality metric with the scaling parameter $\alpha \in \{0, 1, -1\}$ and the weight w_j^B of a hyperlink in the backbone B, as a function of infection probability β. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

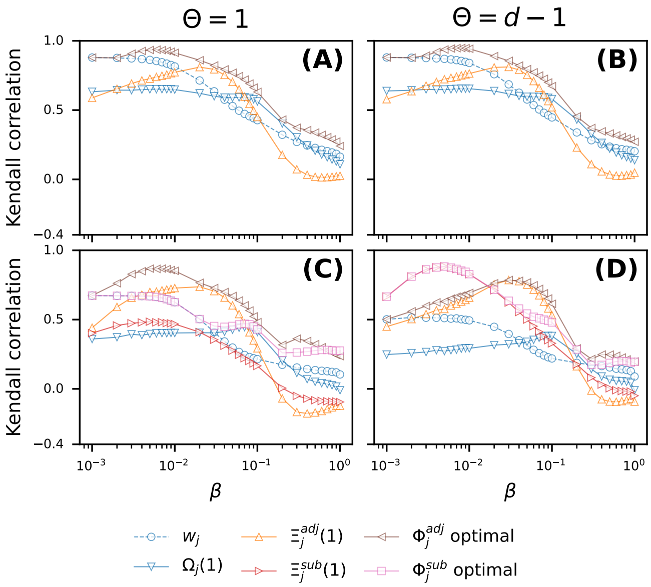

Figure 4.27: The *infectious* dataset. The Kendall rank correlation between each of the four metrics $w_j, \Omega_j(1), \Xi_j^{adj}$ and Ξ_j^{sub} and the weight w_j^B of a hyperlink in the backbone, in comparison to the optimal Kendall correlation between each combined metric and the weight w_j^B. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

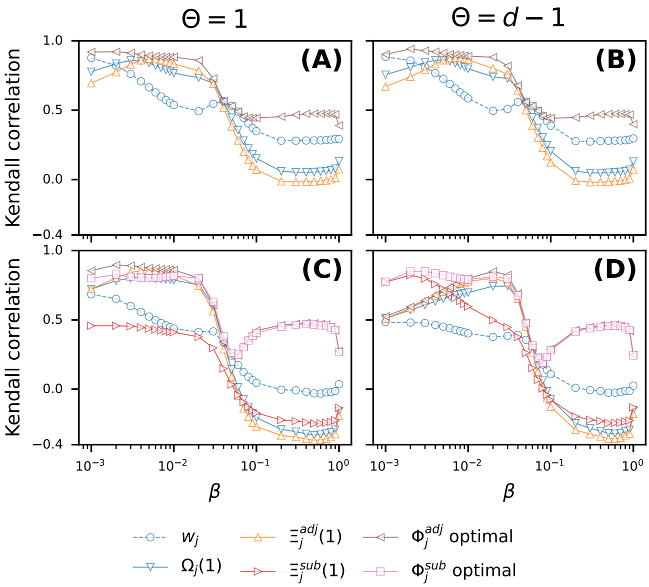

Figure 4.28: The *primaryschool* dataset. The Kendall rank correlation between each of the four metrics w_j, $\Omega_j(1)$, Ξ_j^{adj} and Ξ_j^{sub} and the weight w_j^B of a hyperlink in the backbone, in comparison to the optimal Kendall correlation between each combined metric and the weight w_j^B. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

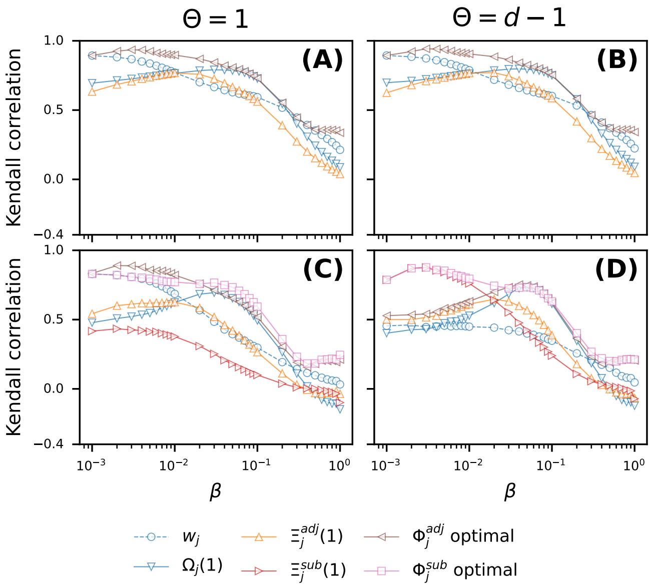

Figure 4.29: The *highschool2012* dataset. The Kendall rank correlation between each of the four metrics w_j, $\Omega_j(1)$, Ξ_j^{adj} and Ξ_j^{sub} and the weight w_j^B of a hyperlink in the backbone, in comparison to the optimal Kendall correlation between each combined metric and the weight w_j^B. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

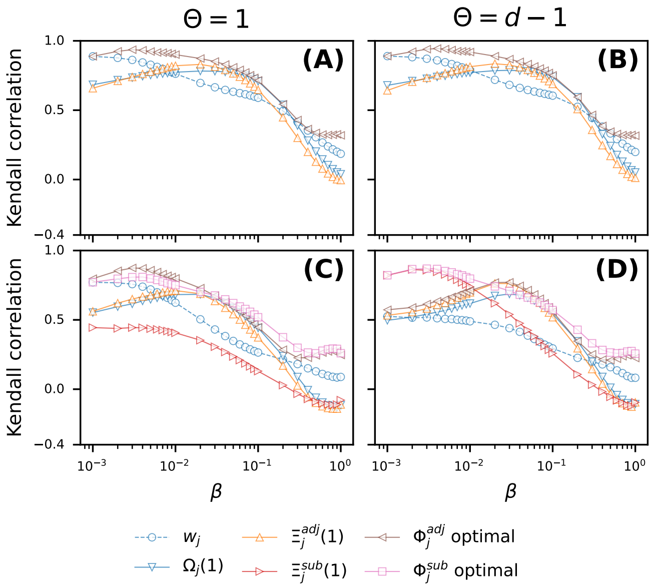

Figure 4.30: The *highschool2013* dataset. The Kendall rank correlation between each of the four metrics w_j, $\Omega_j(1)$, Ξ_j^{adj} and Ξ_j^{sub} and the weight w_j^B of a hyperlink in the backbone, in comparison to the optimal Kendall correlation between each combined metric and the weight w_j^B. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

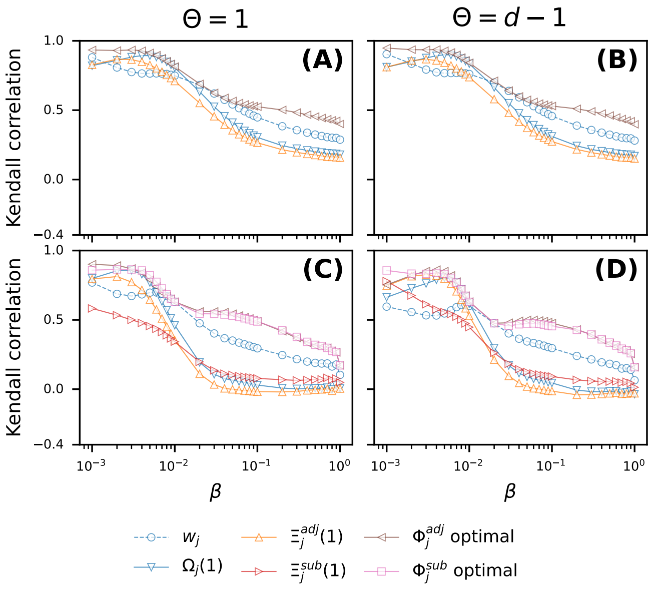

Figure 4.31: The *hospital* dataset. The Kendall rank correlation between each of the four metrics w_j, $\Omega_j(1)$, Ξ_j^{adj} and Ξ_j^{sub} and the weight w_j^B of a hyperlink in the backbone, in comparison to the optimal Kendall correlation between each combined metric and the weight w_j^B. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

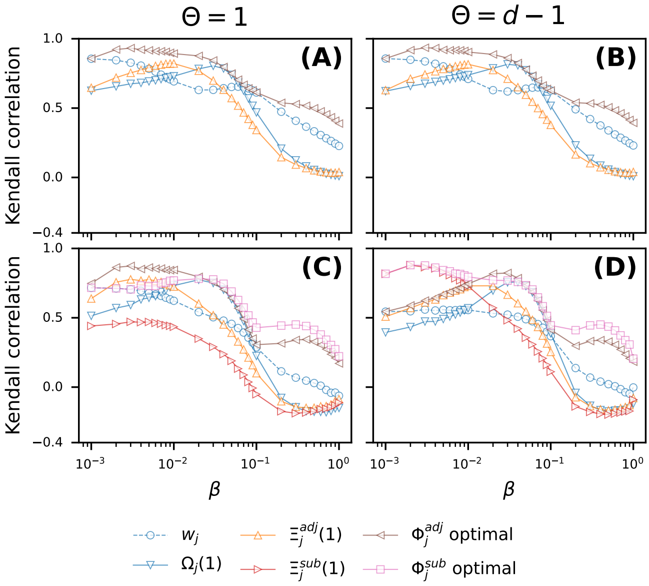

Figure 4.32: The *ht09* dataset. The Kendall rank correlation between each of the four metrics w_j, $\Omega_j(1)$, Ξ_j^{adj} and Ξ_j^{sub} and the weight w_j^B of a hyperlink in the backbone, in comparison to the optimal Kendall correlation between each combined metric and the weight w_j^B. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

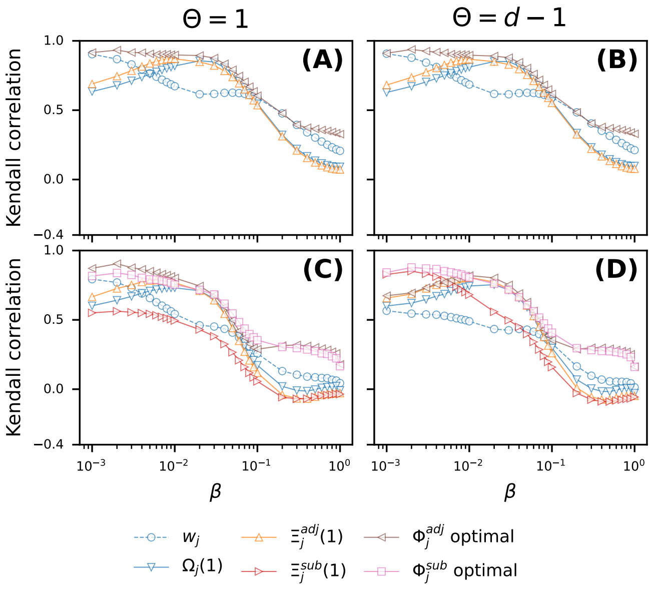

Figure 4.33: The *workplace15* dataset. The Kendall rank correlation between each of the four metrics w_j, $\Omega_j(1)$, Ξ_j^{adj} and Ξ_j^{sub} and the weight w_j^B of a hyperlink in the backbone, in comparison to the optimal Kendall correlation between each combined metric and the weight w_j^B. Results are shown for dyadic hyperlinks (A-B) and triadic hyperlinks (C-D). Two columns corresponds to $\Theta = 1$ and $\Theta = d - 1$, respectively.

REFERENCES

1. Holme, P. & Saramäki, J. Temporal networks. *Physics Reports* **519,** 97–125 (2012).
2. Holme, P. Modern temporal network theory: a colloquium. *The European Physical Journal B* **88,** 1–30 (2015).
3. Masuda, N. & Lambiotte, R. *A guide to temporal networks* (World Scientific, 2016).
4. Karsai, M. *et al.* Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E* **83,** 025102 (2011).
5. Lambiotte, R., Tabourier, L. & Delvenne, J.-C. Burstiness and spreading on temporal networks. *The European Physical Journal B* **86,** 1–4 (2013).
6. Masuda, N., Klemm, K. & Eguíluz, V. M. Temporal networks: slowing down diffusion by long lasting interactions. *Physical Review Letters* **111,** 188701 (2013).
7. Rocha, L. E. & Blondel, V. D. Bursts of vertex activation and epidemics in evolving networks. *PLoS Computational Biology* **9,** e1002974 (2013).
8. Unicomb, S., Iñiguez, G., Gleeson, J. P. & Karsai, M. Dynamics of cascades on burstiness-controlled temporal networks. *Nature Communications* **12,** 133 (2021).
9. Starnini, M., Baronchelli, A., Barrat, A. & Pastor-Satorras, R. Random walks on temporal networks. *Physical Review E* **85,** 056115 (2012).
10. Ciaperoni, M. *et al.* Relevance of temporal cores for epidemic spread in temporal networks. *Scientific Reports* **10,** 12529 (2020).
11. Lee, S., Rocha, L. E., Liljeros, F. & Holme, P. Exploiting temporal network structures of human interaction to effectively immunize populations. *PloS One* **7,** e36439 (2012).
12. Starnini, M., Machens, A., Cattuto, C., Barrat, A. & Pastor-Satorras, R. Immunization strategies for epidemic processes in time-varying contact networks. *Journal of Theoretical Biology* **337,** 89–100 (2013).
13. Zhan, X.-X., Hanjalic, A. & Wang, H. Information diffusion backbones in temporal networks. *Scientific Reports* **9,** 1–12 (2019).
14. Zhang, S., Zhao, X. & Wang, H. Mitigate SIR epidemic spreading via contact blocking in temporal networks. *Applied Network Science* **7,** 2 (2022).
15. Battiston, F. *et al.* Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports* **874,** 1–92 (2020).
16. Battiston, F. *et al.* The physics of higher-order interactions in complex systems. *Nature Physics* **17,** 1093–1098 (2021).
17. Boccaletti, S. *et al.* The structure and dynamics of networks with higher order interactions. *Physics Reports* **1018,** 1–64 (2023).
18. Wang, W. *et al.* Epidemic spreading on higher-order networks. *Physics Reports* **1056,** 1–70 (2024).
19. Cencetti, G., Battiston, F., Lepri, B. & Karsai, M. Temporal properties of higher-order interactions in social networks. *Scientific Reports* **11,** 7028 (2021).

20. Sekara, V., Stopczynski, A. & Lehmann, S. Fundamental structures of dynamic social networks. *Proceedings of the National Academy of Sciences* **113,** 9977–9982 (2016).

21. Iacopini, I., Foote, J. R., Fefferman, N. H., Derryberry, E. P. & Silk, M. J. Not your private tête-à-tête: leveraging the power of higher-order networks to study animal communication. *Philosophical Transactions B* **379,** 20230190 (2024).

22. Musciotto, F., Papageorgiou, D., Battiston, F. & Farine, D. R. Beyond the dyad: uncovering higher-order structure within cohesive animal groups. *BioRxiv,* 2022–05 (2022).

23. Patania, A., Petri, G. & Vaccarino, F. The shape of collaborations. *EPJ Data Science* **6,** 1–16 (2017).

24. Abella, D. *et al.* Unraveling higher-order dynamics in collaboration networks. *arXiv preprint arXiv:2306.17521* (2023).

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

26. Iacopini, I., Karsai, M. & Barrat, A. The temporal dynamics of group interactions in higher-order social networks. *arXiv preprint arXiv:2306.09967* (2023).

27. Gallo, L., Lacasa, L., Latora, V. & Battiston, F. Higher-order correlations reveal complex memory in temporal hypergraphs. *Nature Communications* **15,** 4754 (2024).

28. Chowdhary, S., Kumar, A., Cencetti, G., Iacopini, I. & Battiston, F. Simplicial contagion in temporal higher-order networks. *Journal of Physics: Complexity* **2,** 035019 (2021).

29. Neuhäuser, L., Lambiotte, R. & Schaub, M. T. Consensus dynamics on temporal hypergraphs. *Physical Review E* **104,** 064305 (2021).

30. Contreras, D. A., Cencetti, G. & Barrat, A. Infection patterns in simple and complex contagion processes on networks. *PLOS Computational Biology* **20,** e1012206 (2024).

31. Centola, D. The spread of behavior in an online social network experiment. *science* **329,** 1194–1197 (2010).

32. Karsai, M., Iniguez, G., Kaski, K. & Kertész, J. Complex contagion process in spreading of online innovation. *Journal of The Royal Society Interface* **11,** 20140694 (2014).

33. Mønsted, B., Sapieżyński, P., Ferrara, E. & Lehmann, S. Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PloS one* **12,** e0184148 (2017).

34. Iacopini, I., Petri, G., Barrat, A. & Latora, V. Simplicial models of social contagion. *Nature Communications* **10,** 2485 (2019).

35. De Arruda, G. F., Petri, G. & Moreno, Y. Social contagion models on hypergraphs. *Physical Review Research* **2,** 023032 (2020).

36. Iacopini, I., Petri, G., Baronchelli, A. & Barrat, A. Group interactions modulate critical mass dynamics in social convention. *Communications Physics* **5,** 64 (2022).

5
Conclusion

In this thesis, we focus on understanding the roles of nodes and links in a spreading process. We investigate the roles of nodes and links from three different angles, each covered in a separate chapter. The general objective is to analyze how these roles are associated with the properties of nodes and links, as captured by centrality metrics, within the underlying network.

We first investigate, in Chapter 2, the spreading influence of a node, which indicates the average number of nodes that are eventually infected when it serves as the seed node. Chapter 3 addresses a more practical problem: how to select contacts (temporal links) to block in order to mitigate the epidemic spreading on temporal networks. Here, we investigate the roles of temporal links in mitigating a spreading process. Finally, in Chapter 4, we explore the roles of hyperlinks in a spreading process unfolding on a temporal higher-order network (temporal hypergraphs). We study how many nodes are directly infected through the activation of each hyperlink, which indicates the contribution of a hyperlink to the spreading process.

In this chapter, we summarize the contributions of this thesis and reflect on its limitations in Section 5.1, then discuss the potential directions for future research in Section 5.2.

5.1. Main contribution and reflections

In Chapter 2, we explore to what extent local and global topological information of a node is needed to predict its spreading influence and whether relatively local topological information around a node is sufficient for the prediction. To this end, we propose to predict nodal influence using an iterative metric set derived from an iterative process. The iterative metric set consists of an iterative metric from order 1 to K, encoding progressively more global information as K increases. We consider three iterative metrics: Normalized Walk Counts (NWC), Visiting Probability (VP), and H index (HI), each converging to a corresponding global metric in the iterative process. A regression model using an iterative metric set is trained on a fraction of nodes whose influence is known to predict the influence of the remaining nodes. We evaluate the performance of these three iterative metrics in predicting nodal influence in SIR spreading processes across various effective infection rates around the epidemic threshold, on both real-world networks and synthetic networks with different strength of community structure. Our findings show that the prediction quality of each iterative metric set converges to its optimal when relatively low orders (up to order 4) are included, with only marginal improvement upon adding higher orders. This is explained by the correlation between the iterative metric of each order and nodal influence, and the convergence rate of each iterative process. The best-performing iterative metric set, NWC, achieves prediction quality comparable to the benchmark method, which combines both local and global centrality metrics. In spatially embedded networks with extremely large diameter and modularity, however, the iterative metric of higher orders (thus a large K) are needed to achieve comparable prediction quality as the benchmark. These findings suggest that NWC metrics of relatively low orders (up to order 4) contain sufficient information for predicting nodal influence reasonably well in networks with the small-world property, while being computationally less complex than global centrality metrics in the benchmark model. In most networks, the NWC metric of order $k \approx 4$ has near-maximal correlation with nodal influence, indicating that nodes with a large number of distinct 4-hop walks originating from them tend to be more influential. However, the interpretability of the iterative metric-based regression model is limited because of the strong correlation among an iterative metric of different orders.

In Chapter 3, we investigate how to strategically select contacts to block to mitigate the epidemic spreading on temporal networks. We propose probabilistic contact blocking strategies based on the properties of contacts within the temporal network. Specifically, we define the probability of blocking a contact $c(i,j,t)$ as a function of a given centrality metric of the corresponding link $l(i,j)$ in the time-aggregated network and time t. In total, we consider 12 centrality metrics, each defining a unique strategy, along with a baseline strategy (random removal). We find that the strategy that prioritizes the removal of early contacts between node pairs with fewer contacts achieves the most effective epidemic mitigation in terms of reducing the average prevalence, the peak prevalence, and the time needed to reach the peak prevalence. This suggests that removing contacts associated to weak social ties (i.e., links activated less frequently) in the early phase tends to better suppress the epidemic spreading. Our findings still hold when uncertainty is introduced in the original temporal networks, either by reshuffling the ordering of contacts or by enlarging the temporal resolution.

We further analyze the properties of the pruned network resulting from contact removal according to each strategy. Our analysis shows that the strategy that best mitigate the spreading produces an aggregated pruned network with a high largest eigenvalue, a large modularity, and a possibly a small size of the largest connected component. A strategy tends to perform better when the number of removed contacts of each link is similar. These findings are in line with our understanding that a network with a small largest connected component and high modularity hinders epidemic spreading. However, the high largest eigenvalue in the aggregated pruned network resulting from the best strategy seems to contradict our understanding that a static network with a large largest eigenvalue tends to facilitate SIS epidemic spreading with respect to its small epidemic threshold. This is possibly because of differences between SIR and SIS models, distinctions between epidemic threshold and prevalence, and the complexity introduced by the temporal nature of the network, which cannot be fully captured by the aggregated network.

Chapter 4 focuses on the spreading process unfolding on a temporal higher-order network. In this chapter, we study the contribution of a hyperlink in a temporal higher-order network to the spreading process, defined as the average number of nodes directly infected via the activation of the hyperlink. We explore which properties of a hyperlink in the temporal higher-order network lead to a high contribution to the spreading process. A generalized Susceptible-Infected threshold process with two parameters, infection probability β and threshold Θ, is considered on eight real-world temporal higher-order networks, which are derived from human face-to-face contacts in various contexts. We first proposed the construction of the diffusion backbone, where the weight of each hyperlink represents its contribution to the spreading process starting from a random seed node. In the limiting case where $\beta \to 0$, the weight of a hyperlink in the backbone is derived analytically based on the local temporal connections around the hyperlink. We then illustrate the dependency of the backbone on the two process parameters. This is also evidenced by the backbones analytically derived as $\beta \to 0$ when Θ varies. To explore which properties of hyperlinks are associated with high contributions to the spreading process, we designed centrality metrics for a hyperlink to estimate the ranking of hyperlinks by their weights in the backbone. Each proposed metric is defined based on the activity of the target hyperlink and its neighboring hyperlinks in the temporal higher-order network. Different metrics are shown to achieve the best prediction performance for different parameter sets (β, Θ) of the process, and the best performance approaches the optimal performance of the combined metrics.

5.2. FUTURE WORK

Based on the methods and results in this thesis, we raise some promising future directions.

The impact of network properties on the prediction of nodal influence. In Chapter 2, we observe the trend that an iterative metric of high orders, which encode more global network information, are needed for the iterative metric-based method to perform close to its optimal in networks with larger diameters. It would be valuable to identify the minimal order of an iterative metric needed to achieve, for example, at least 95% of the optimal performance of the iterative metric-based method in relation to properties of the underlying network, such as diameter. Furthermore, the diameter and strength of community structure are possibly correlated in real-world networks and network models. We have observed the influence of community structure and diameter on the prediction quality of the NWC-based method and the benchmark model. An open question is how diameter influences the prediction quality while the community strength is fixed. To address both objectives, network models with a controllable diameter and more real-world networks, especially those without the small-world property, are needed.

Systematic exploration of mitigation of epidemic spreading. In Chapter 3, we have confined ourselves to the SIR model with limited choice of parameters and a few real-world networks. While the SIR model is a simplified representation of epidemic spreading, real-world epidemics are often more complicated. Hence, our conclusion regarding the effectiveness of mitigation strategies cannot be generalized directly to real-world epidemic mitigation. It is essential to explore more realistic epidemic spreading models. Additionally, in our time-dependent contact removal strategies, the time dependence is defined using one of the simplest functional forms. Exploring other time-dependent functions, especially those that are practical for policymakers, would be valuable.

Backbones of other types of higher-order interactions and under different dynamics. In Chapter 4, we consider only the Susceptible-Infected threshold process as the diffusion model on temporal higher-order networks and focus on two extreme cases for the threshold. Indeed, the proposed method can also be generalized to other dynamical processes unfolding on temporal higher-order networks to study the relation between the contribution of each hyperlink and the process dynamics. Moreover, it may be used to investigate how different temporal correlations observed in real temporal higher-order networks can affect the hyperlink contribution. Furthermore, it would be interesting to explore other types of temporal higher-order networks, such as scientific collaborations networks, which may have different properties than the human interaction networks considered in Chapter 4. Finally, it would be promising to investigate how to design strategies to mitigate the dynamic process on the network via blocking selected hyperlinks, based on the backbone or our proposed centrality metrics.