SCIENTIFIC REPORTS

OPEN

Information diffusion backbones in temporal networks

Xiu-Xiu Zhan, Alan Hanjalic & Huijuan Wang

Received: 19 November 2018
Accepted: 11 April 2019
Published online: 01 May 2019

Progress has been made in understanding how temporal network features affect the percentage of nodes reached by an information diffusion process. In this work, we explore further: which node pairs are likely to contribute to the actual diffusion of information, i.e., appear in a diffusion trajectory? How is this likelihood related to the local temporal connection features of the node pair? Such deep understanding of the role of node pairs is crucial to tackle challenging optimization problems such as which kind of node pairs or temporal contacts should be stimulated in order to maximize the prevalence of information spreading. We start by using Susceptible-Infected (*SI*) model, in which an infected (information possessing) node could spread the information to a susceptible node with a given infection probability β whenever a contact happens between the two nodes, as the information diffusion process. We consider a large number of real-world temporal networks. First, we propose the construction of an *information diffusion backbone* $G_B(\beta)$ for a *SI* spreading process with an infection probability β on a temporal network. The backbone is a weighted network where the weight of each node pair indicates how likely the node pair appears in a diffusion trajectory starting from an arbitrary node. Second, we investigate the relation between the backbones with different infection probabilities on a temporal network. We find that the backbone topology obtained for low and high infection probabilities approach the backbone $G_B(\beta \to 0)$ and $G_B(\beta = 1)$, respectively. The backbone $G_B(\beta \to 0)$ equals the integrated weighted network, where the weight of a node pair counts the total number of contacts in between. Finally, we explore node pairs with what local connection features tend to appear in $G_B(\beta = 1)$, thus actually contribute to the global information diffusion. We discover that a local connection feature among many other features we proposed, could well identify the (high-weight) links in $G_B(\beta = 1)$. This local feature encodes the time that each contact occurs, pointing out the importance of temporal features in determining the role of node pairs in a dynamic process.

Both online social networks like Facebook, Twitter and LinkedIn and physical contact networks facilitate the diffusion of information where a piece of information is transmitted from one individual to another through their online or physical contacts or interactions. Information diffusion processes have been modeled by, e.g., independent cascade models[1], threshold models[2] and epidemic spreading models[3–7]. Social networks have been first considered to be static where nodes represent the individuals and links indicate the relation between nodes such as whether they have ever contacted or not[8]. Information is assumed to propagate through the static links according to the aforementioned models. Recently, the temporal nature of contact networks has been taken into account in the spreading processes, i.e., the contacts between a node pair occur at specific time stamps (the link between nodes is time dependent) and information could possibly propagate only through contacts (or temporal links)[9–13]. Consider the *SI* (Susceptible-Infected) spreading process on a temporal network[3,5]. Each individual can be in one of the two states: susceptible (*S*) or infected (*I*). A node in the infected (susceptible) state means that it has (does not have) the information. A susceptible node could get infected with an infection probability β via each contact with an infected node. An infected individual remains infected forever.

Progress has been made in the exploration of how temporal network features[14–18] and the choice of the source node[19,20] influence a diffusion process especially its diffusion size, i.e., the number of nodes reached. However, we lack foundational understanding of which kind of node pairs are likely to contribute to an actual information diffusion process, i.e., appear in an information diffusion trajectory. Such understanding is essential to explain and control the prevalence of information spread (e.g., which node pairs should be stimulated to contact at what time in order to maximize the prevalence?). The contact frequency between nodes, as typically used in static networks,

Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Mekelweg 4, Delft, 2628 CD, The Netherlands. Correspondence and requests for materials should be addressed to H.W. (email: H.Wang@tudelft.nl)

is not the only factor that would affect the appearance of a node pair in an information diffusion trajectory, as we need to consider the time stamps of the contacts as well[21–24]. For instance, the node pairs with a lot of contacts that only happen before the information starts to diffuse are of no importance for the diffusion process.

In this paper, we address the question of which kind of node pairs are likely to contribute to the diffusion of information, considering the *SI* diffusion process as a start. Specifically, we explore how the probability that a node pair appears in a diffusion trajectory is related to local temporal connection features of the two nodes. First, we propose the construction of an *information diffusion backbone* $G_B(\beta)$ for a *SI* spreading process with an infection probability β on a given temporal network. The construction is based on a large number of information diffusion trajectories. The resultant backbone is a weighted network where the weight of each node pair indicates how likely the node pair contributes to a diffusion process that starts from an arbitrary node. We consider a large number of empirical temporal networks. For each network, we construct diffusion backbones for diverse infection probabilities and study the relationship between these backbones. We find that backbone topology varies from $G_B(\beta = 0) \triangleq G_B(\beta \to 0)$ (which equals the integrated weighted network) when the spreading probability β is small to $G_B(\beta = 1)$ when the infection probability is large. The difference between the two extreme backbones $G_B(\beta = 0)$ and $G_B(\beta = 1)$, suggests the extent to which the backbones with diverse infection rates may vary. Finally, we investigate further which local connection feature of a node pair may suggest its high weight in the backbone $G_B(\beta = 1)$. One of the features that we proposed incorporates only the time stamps when contacts occur between a node pair. It outperforms other classic features of a node pair including those derived from the integrated network, which points out the importance of temporal information in determining the role of a node pair in a diffusion process. The computational complexity of $G_B(\beta = 1)$ is high. Our finding of the relation between local temporal features of a node pair and its global contribution to information diffusion allows the approximation of the information backbone by computing a local temporal feature that is of low computational complexity.

The paper is organized as follows. In Section Materials and Methods, we first introduce how to represent a temporal network and then explain the process of constructing the information diffusion backbone for a *SI* diffusion process on a temporal network. Finally, we illustrate a set of empirical temporal networks that will be used in the following experiments. In Section Results, we present our comparative analysis of the constructed backbones for different infection probabilities and for different networks. At the end of this section, we evaluate which local connection features of a node pair, including the measures we proposed, can identify whether the node pair will be connected in the backbone $G_B(\beta = 1)$ and with a high weight or not. A discussion concludes the paper in Section Discussion.

Materials and Methods

Representation of a Temporal Network. A temporal network can be measured by observing the contacts between each node pair at each time step within a given time window $[0, T]$ and represented as $\mathcal{G} = (\mathcal{N}, \mathcal{L})$. Here, $\mathcal{N}$ is the node set, with the size $N = |\mathcal{N}|$ representing the number of nodes in the network, and $\mathcal{L} = \{l(j, k, t), t \in [0, T], j, k \in \mathcal{N}\}$ is the contact set, where the element $l(j, k, t)$ indicates that the nodes j and k have a contact at time step t. A temporal network can also be described by a three-dimensional binary adjacency matrix $\mathcal{A}_{N\times N\times T}$, where the elements $\mathcal{A}(j, k, t) = 1$ and $\mathcal{A}(j, k, t) = 0$ represent, respectively, that there is a contact or no contact between the nodes j and k at time step t.

An integrated weighted network $G_W = (\mathcal{N}, \mathcal{L}_W)$ can be derived from a temporal network $\mathcal{G}$ by aggregating the contacts between nodes over the entire observation time window T. In other words, two nodes are connected in G_W if there is at least one contact between them in $\mathcal{G}$. Each link $l(j, k)$ in $\mathcal{L}_W$ is associated with a weight w_{jk} counting the total number of contacts between node j and k in $\mathcal{G}$. The integrated weighted network G_W can therefore be described by a weighted adjacency matrix $A_{N\times N}$, with its element

$$A(j, k) = \sum_{t=1}^{T} \mathcal{A}(j, k, t) \tag{1}$$

counting the number of contacts between a node pair. An example of a temporal network $\mathcal{G}$ and its integrated weighted network G_W are given in Fig. 1(a) and (b), respectively.

Information Diffusion Backbone. We propose to characterize how node pairs are involved in diffusion processes by constructing information diffusion backbones. We will construct a backbone for the *SI* diffusion process with a given infection probability β on a temporal network defined above. We start with the simplest case when $\beta = 1$. At time step $t = 0$, the seed node i is infected and all the other nodes are susceptible. The trajectory of the *SI* diffusion on $\mathcal{G}$ can be recorded by a *diffusion path tree* $\mathcal{T}_i(\beta)$. The diffusion path tree $\mathcal{T}_i(\beta)$ records the union of contacts, via which information diffuses. We define the diffusion backbone $G_B(\beta) = (\mathcal{N}, \mathcal{L}_B(\beta))$ as the union of all diffusion path trees, i.e., $\bigcup_{i=1}^{N} \mathcal{T}_i(\beta)$, that start at each node as the seed node. The node set of $G_B(\beta)$ is $\mathcal{N}$, and nodes are connected in $G_B(\beta)$ if they are connected in any diffusion path tree. Each link in $\mathcal{L}_B(\beta)$ is associated with a weight w^B_{jk}, which denotes the number of times node pair (j, k) appears in all diffusion path trees. An example of how to construct the diffusion backbone is given in Fig. 1(c) and (d) for $\beta = 1$. The ratio $\frac{w^B_{jk}}{N}$ indicates the probability that the node pair (j, k) appears in a diffusion trajectory starting from an arbitrary seed node.

When $0 < \beta < 1$, the diffusion process is stochastic. In this case, the backbone can be obtained as the average of a number of realizations of the backbones. Per realization, we run the *SI* process starting from each node serving as the seed for information diffusion, obtain the diffusion path trees and construct one realization of the diffusion backbone. The weight w^B_{jk} of a link in $G_B(\beta)$ is the average weight of this link over the h realizations. The

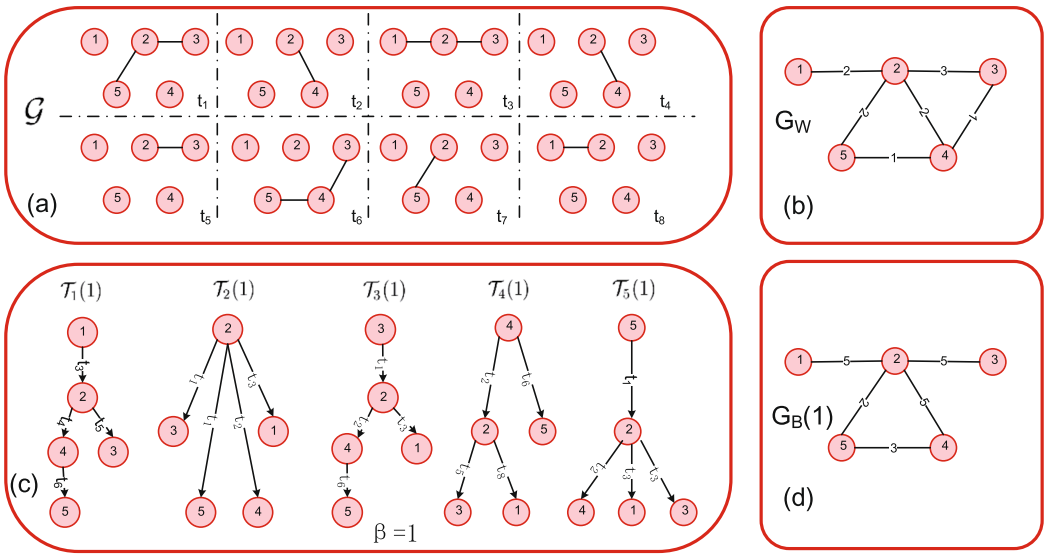

Figure 1. (**a**) A temporal network $\mathcal{G}$ with $N=5$ nodes and $T=8$ time steps. (**b**) The integrated weighted network G_W, in which a link exists between a node pair in G_W as long as there is at least one contact between them in $\mathcal{G}$. The weight of a link in G_W is the number of contacts between the two nodes in $\mathcal{G}$. (**c**) Diffusion path tree $\mathcal{T}_i(\beta)$, where node i is the seed and infection rate is $\beta=1$. (**d**) Diffusion backbone $G_B(1)$, where the infection probability $\beta=1$ in the *SI* diffusion process. The weight on the node pair represents the number of times it appears in all the diffusion path trees.

computational complexity of constructing $G_B(\beta)$ is $\mathcal{O}(N^3Th)$, where T is the length of the observation time window of the temporal network.

Empirical Networks. *Description and basic features.* For the construction and analysis of diffusion backbones, we consider a large number of temporal networks that capture two types of contacts, i.e., physical and virtual contacts. We collect the datasets *Reality mining*[25,26], *Hypertext* 2009[27,28], *High School* 2011[29], *High School* 2012[29], *High School* 2013[30], *Primary School*[31], *Workplace*[32], *Haggle*[33,34] and *Infectious*[28] that record the face-to-face physical contacts of individuals at MIT, ACM Hypertext 2009 conference, a high school, a primary school, a workplace and the Science Gallery, respectively. We also consider virtual contact datasets recording the mailing and message behavior, including *Manufacturing Email*[35,36], *Email Eu*[37], *DNC Email*[38] and *Collegemsg*[39]. The list of the datasets used and their detailed statistics are given in Table 1. We consider only the temporal network topologies measured at discrete time steps in these datasets, whereas the duration of a time step differs among these datasets. We have removed the time steps without any contact in order to consider the steps that are relevant for information diffusion and to avoid the periods that have no contact due to technical errors in measurements.

Observation time windows. We aim to understand which node pair is likely to be connected in the backbone, thus to contribute to a diffusion process and how such connection in the backbone is related to this node pair's temporal connection features. However, real-world temporal networks are measured for different lengths T of time windows as shown in Table 1. If a diffusion process has a relatively high spreading probability or the temporal network has a relatively long observation time window, almost all the nodes can be reached within a short time. The temporal contacts happened afterwards will not contribute to the diffusion process. Hence, we will select the time windows such that all contacts within each selected time window could possibly contribute, or equivalently, are relevant to a diffusion process. On the other hand, we will consider several time windows for each measured temporal network. This will allow us to understand how the time window of a temporal network may influence the relation between the backbones of different spreading probabilities and relation between a node pair's local connection features and its connection in a backbone. We select the observation time windows for each measured temporal network within its original time window $[0, T]$ as follows. On each measured temporal network with its original observation time window $[0, T]$, we conduct the *SI* diffusion process with $\beta=1$ by setting each node as the seed of the information diffusion process and plot the average prevalence ρ at each time step, as illustrated in Fig. 2. The time steps are normalized by the original length of observation window T. The average prevalence at the end of the observation $t/T=1$ is recorded as $\rho(t=T)$. The time to reach the steady state varies significantly across the temporal networks. For networks like *RM*, *HT*2009, the diffusion finishes or stops earlier and contacts happened afterwards are not relevant for the diffusion process. However, the prevalence curves ρ of the last four networks (i.e., *Haggle*, *Infectious*, *DNC* and *Collegemsg*) increase slowly and continuously over the whole period. Actually, we observe these four networks are more heterogeneous than the other networks in terms of the degree distribution of the integrated static network, which are shown in Fig. 3.

For each real-world temporal network with its original length of observation time window T, we consider the following lengths of observation time windows: the time $T_{p\%}$ when the average prevalence reaches p%, where

Network	N	T	$\|\mathcal{C}\|$	$\|\mathcal{L}_W\|$	*Contact Type*
Reality Mining (RM)	96	33,452	1,086,404	2,539	Physical
Hypertext 2009 (HT2009)	113	5,246	20,818	2,196	Physical
High School 2011 (HS2011)	126	5,609	28,561	1,710	Physical
High School 2012 (HS2012)	180	11,273	45,047	2,220	Physical
High School 2013 (HS2013)	327	7,375	188,508	5,818	Physical
Primary School (PS)	242	3,100	125,773	8,317	Physical
Workplace (WP)	92	7,104	9,827	755	Physical
Manufacturing Email (ME)	167	57,791	82,876	3,250	Virtual
Email Eu (EEU)	986	207,880	332,334	16,064	Virtual
Haggle	274	15,662	28,244	2,124	Physical
Infectious	410	1,392	17,298	2,765	Physical
DNC Email (DNC)	1866	1,8682	37,421	4,384	Virtual
Collegemsg	1899	5,8911	59,835	13,838	Virtual

Table 1. Basic features of the empirical networks. The number of nodes (N), the original length of the observation time window (T is number of time steps), the total number of contacts ($|\mathcal{C}|$), the number of links in G_W ($|\mathcal{L}_W|$) and contact type are shown.

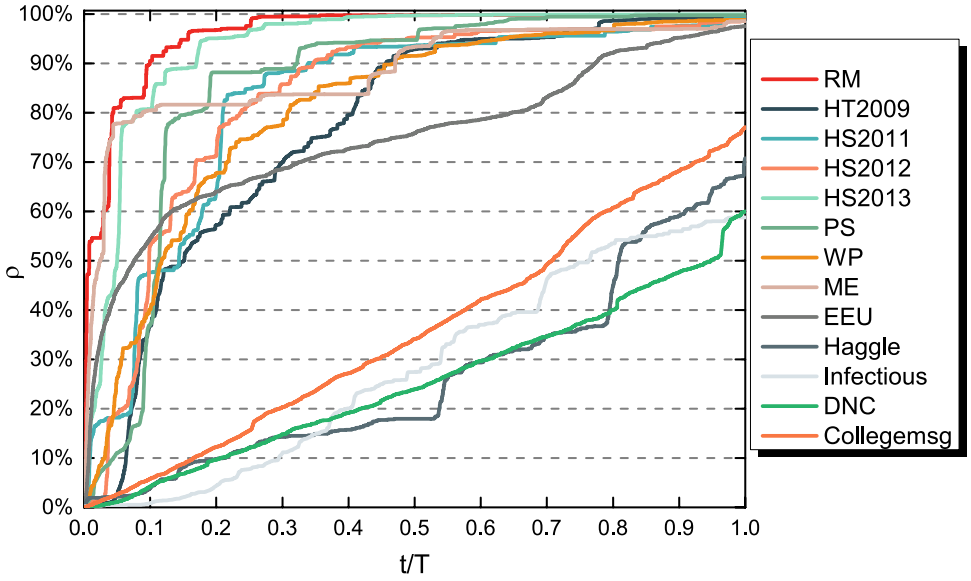

Figure 2. Average prevalence ρ of the *SI* spreading process with $\beta = 1$ on each original empirical temporal network over time. The time steps are normalized by the corresponding observation time window T of each network.

$p \in \{10, 20, \ldots, 90\}$ and $p\% < \rho(t = T)$. For a given measured temporal network $\mathcal{G} = (\mathcal{N}, \mathcal{L})$, we consider maximally 9 observation time windows. For each length $T_{p\%}$, we construct a sub-temporal network, $\mathcal{G}_{p\%} = (\mathcal{N}, \mathcal{L}_{p\%})$, in which $\mathcal{L}_{p\%}$ includes contacts in $\mathcal{L}$ that occur earlier than $T_{p\%}$. The lengths of observation time window $T_{p\%}$ for the empirical networks are shown in Table S1 in the APPENDIX A. For a network like *RM*, we can get 9 sub-networks and for network like *Infectious*, we can only obtain 5 sub-networks. In total, 106 sub-networks are obtained. Contacts in all these sub-networks are relevant for SI diffusion processes with any spreading probability β. Without loss of generality, we will consider all these sub-networks with diverse lengths of observation time windows and temporal network features to study the relationship between diffusion backbones and temporal connection features.

Results

Relationship between Diffusion Backbones. We explore the relationships among the backbones $G_B(\beta)$ with different spreading probabilities $\beta \in [0, 1]$ on the same temporal network. When the infection probability $\beta \to 0$, the backbone $G_B(\beta \to 0)$ approaches the integrated weighted network G_W if the network is finite regarding to its size and the number of contacts. This can be understood as follows. When an arbitrary node i is the seed node, the probability that the information diffuses to any other node j within a given observation time window of length T is $1 - (1 - \beta)^{w_{ij}} = 1 - e^{w_{ij}\log(1-\beta)} \sim 1 - e^{-w_{ij}\beta} \sim w_{ij}\beta$, where w_{ij} is the number of contacts between

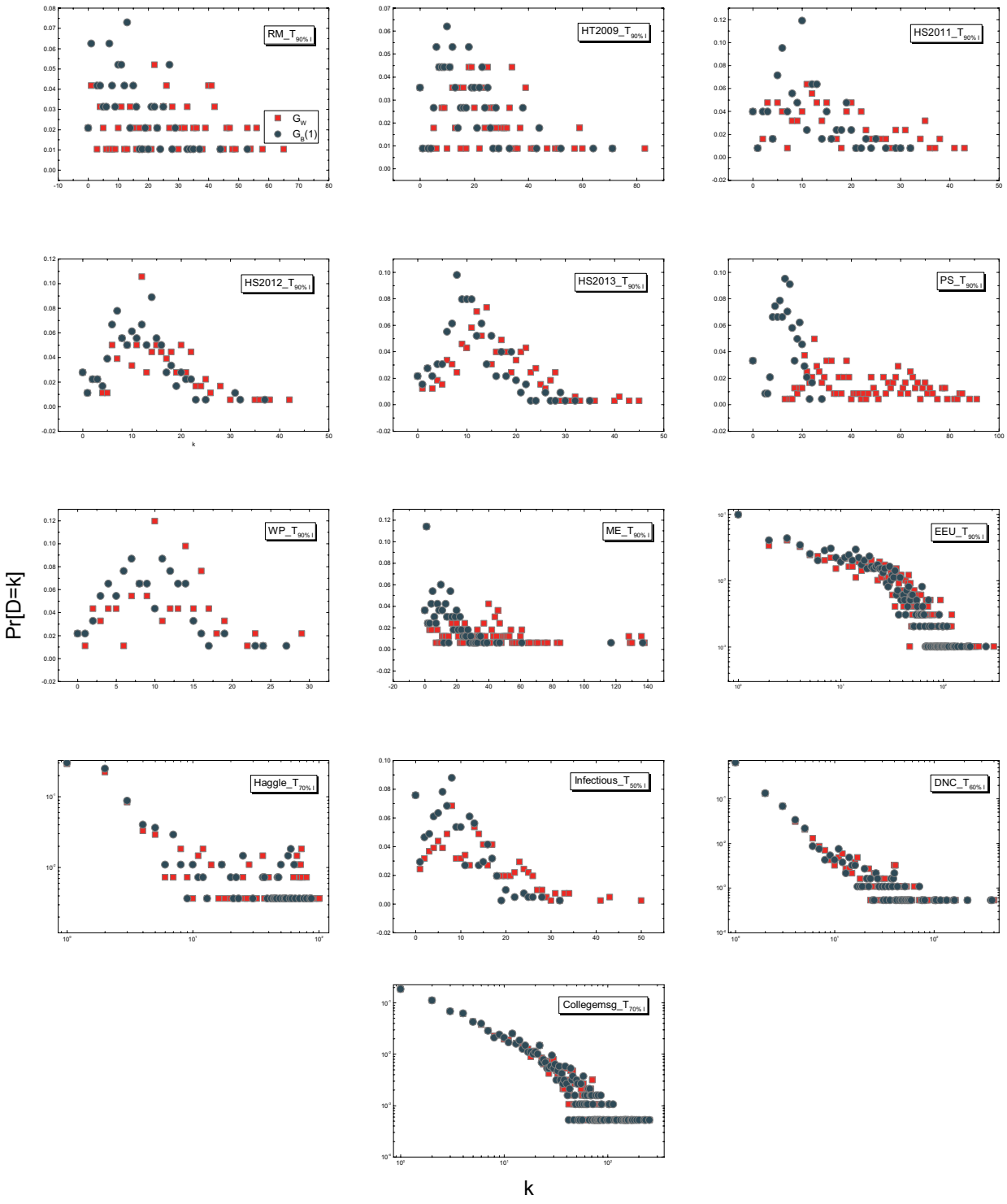

Figure 3. Degree distribution of G_W and $G_B(1)$ for empirical networks with longest observation window.

nodes i and j within the observation time window. Assume that i and j have contact(s), i.e., $w_{ij} > 0$, and node k has no contact with the seed i but has contact(s) with node j. The probability that the information initiated by the seed i diffuses further from j to k is smaller than $w_{ij}w_{jk}\beta^2 \ll w_{ij}\beta$. In other words, the probability that the information diffuses via a second hop node pair (j, k) relative to the seed i (from the view of the integrated network) is negligibly small compared to the first hop node pair (i, j). Hence, the information diffusion tree approaches a tree whose root is the seed node and the leaves are the nodes that have contacts with the seed. The information diffusion backbone, which is the union of the diffusion trees rooted at each node, has the same topology as the integrated network. The weight w_{ij}^B of each link in the backbone is $w_{ij}^B \sim 2w_{ij}\beta$. When the network is infinite in size or

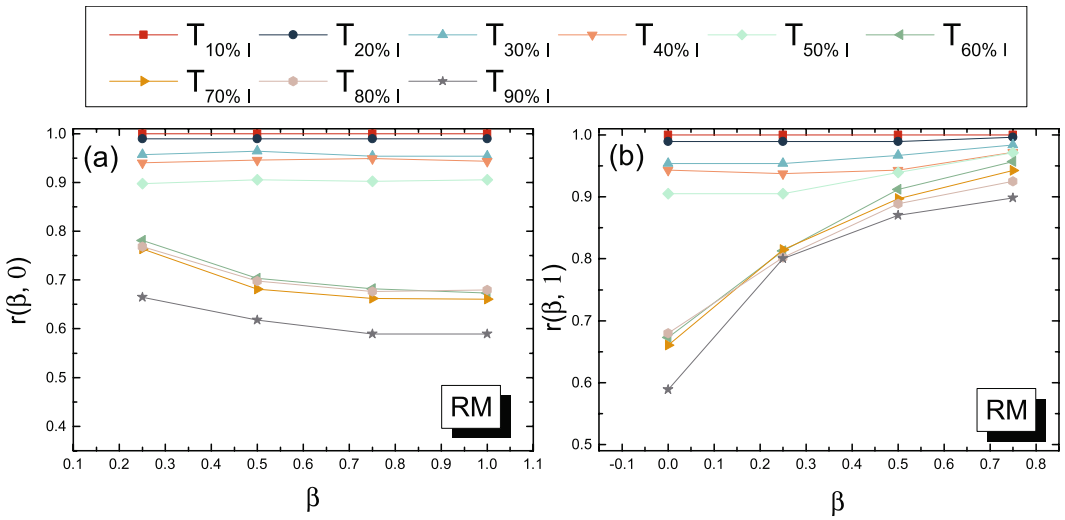

Figure 4. (**a**) Overlap $r(\beta, 0)$ between $G_B(\beta)$ and $G_B(0)$ as a function of β in (sub)networks derived from dataset *RM*; (**b**) Overlap $r(\beta, 1)$ between $G_B(\beta)$ and $G_B(1)$ as a function of β in (sub)networks derived from dataset *RM*. Diffusion backbones ($0 < \beta < 1$) are obtained over 100 iterations.

the number of contacts, $G_B(\beta \to 0) \sim G_W$ is not necessarily true also because a node pair can be a second hop pair relative to many seed nodes.

We denote $G_B(\beta = 0) \triangleq G_B(\beta \to 0) = G_W$ except that the weight of each node pair in the two networks is scaled. When the infection probability β is small, node pairs with more contacts are more likely to appear in the backbone. The backbone $G_B(\beta)$ varies from $G_B(0) = G_W$ when $\beta \to 0$ to $G_B(1)$ when $\beta = 1$.

Overlap in Links between Backbones. We investigate first how different these backbones with different spreading probabilities $\beta \in [0, 1]$ are and whether $G_B(\beta)$ with a small and large β can be well approximated by G_W and $G_B(1)$ respectively.

When $0 < \beta < 1$, every contact has a none zero probability to diffuse the information, especially taking into account the fact that every node could be the seed of the information. Therefore, the topology of $G_B(\beta)$ without considering the link weight is the same as that of G_W and we have $|\mathcal{L}_B(\beta)| = |\mathcal{L}_W|$ when $0 < \beta < 1$. However, note that the observed topology of $G_B(\beta)$ obtained from the simulation which is composed of a limited number of iterations of the spreading process can be a sub-graph of the topology of G_W. We illustrate how the number of iterations affects the ratio of links in the observed $G_B(\beta)$ to $|\mathcal{L}_W|$ in Figure S1(d–f) in the APPENDIX B. It shows that with the increased number h of iterations, $|\mathcal{L}_B(\beta)|$ is getting close to $|\mathcal{L}_W|$ for networks with a large observation time window. For networks with a small observation time window like $RM\_T_{10\%}$, $|\mathcal{L}_B(\beta)|$ tends to approach $|\mathcal{L}_W|$ at a small number of iterations h. For $G_B(1)$, we have $|\mathcal{L}_B(1)| <= |\mathcal{L}_W|$, which is reflected in Fig. 5 (a) where the number of links in $G_B(0)$ and $G_B(1)$ are compared.

The similarity between two backbones or two weighted networks in general can be measured by their overlap in links or node pairs with a high weight. For each backbone $G_B(\beta)$, links in $\mathcal{L}_B(\beta)$ are ordered according to their weights in the backbone in a descending order. Thus the links in the relatively top positions are more likely to be used in the diffusion process. Therefore, for any backbone with $\beta \in [0, 1]$, we consider the top $|\mathcal{L}_B(1)|$ links from $\mathcal{L}_B(\beta)$, which are denoted as $\mathcal{L}_B^*(\beta)$. The similarity or overlap between two backbones like $G_B(\beta)$ and $G_B(\beta = 0)$ can be measured by the overlap between $\mathcal{L}_B^*(\beta)$ and $\mathcal{L}_B^*(0)$, defined as

$$r(\beta, 0) = r(\mathcal{L}_B^*(\beta), \mathcal{L}_B^*(0)) = \frac{|\mathcal{L}_B^*(\beta) \cap \mathcal{L}_B^*(0)|}{|\mathcal{L}_B^*(\beta)|}, \tag{2}$$

For each temporal network, we construct each backbone $G_B(\beta)$, where $\beta = 0.25, 0.5, 0.75, 1$, as the average of $h = 100$ iterations of the *SI* spreading processes starting from each node as the seed, based on the method illustrated in Section Materials and Methods (The validation that 100 iterations are enough to get a stable backbone is given in Figure S1 in the APPENDIX B). The backbone $G_B(\beta = 0)$ equals G_W. The overlap between backbones for dataset *RM* are shown in Fig. 4 as an example. More examples are given in Figure S2 in the APPENDIX C). The overlap $r(\beta, 0)$ tends to decrease with the increase of β and $G_B(\beta = 0)$ well approximates the backbones with a small β. Similarly, $G_B(1)$ well approximates the backbones with a large β. When the observation time window of a temporal network is small, the backbones with different β are relatively similar in topology. In this case, a diffusion path tree tends to have a smaller average depth (The average depth of a tree is the average number of links in the shortest path from the root to another random node in the tree) and a node pair with a large number of contacts is likely to appear or connect in the backbone, which explains why G_W approximates all the backbones including $G_B(1)$. These observations motivate us to explore the two extreme backbones $G_B(0)$ and $G_B(1)$ regarding to how much they differ from or relate to each other.

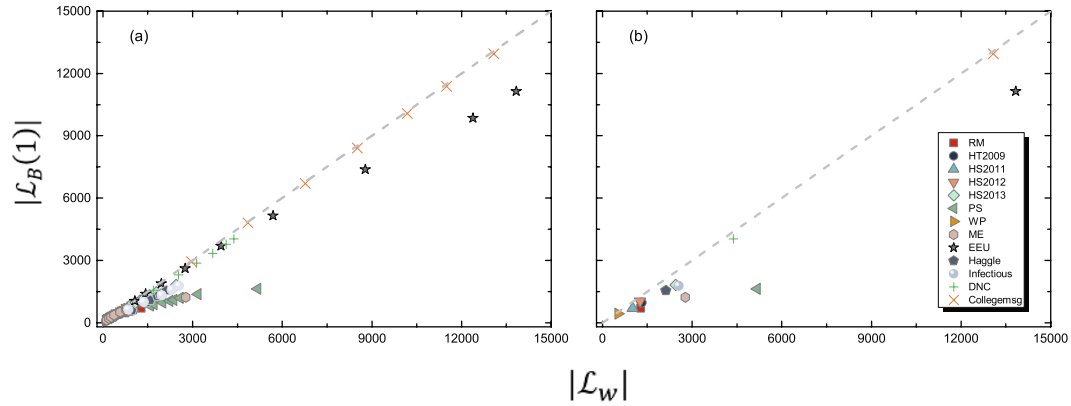

Figure 5. The relationship between the number of links in G_W and $G_B(1)$ for (**a**) all the networks with observation windows given in Table S1 in APPENDIX A; (**b**) the networks with the longest observation windows in each dataset.

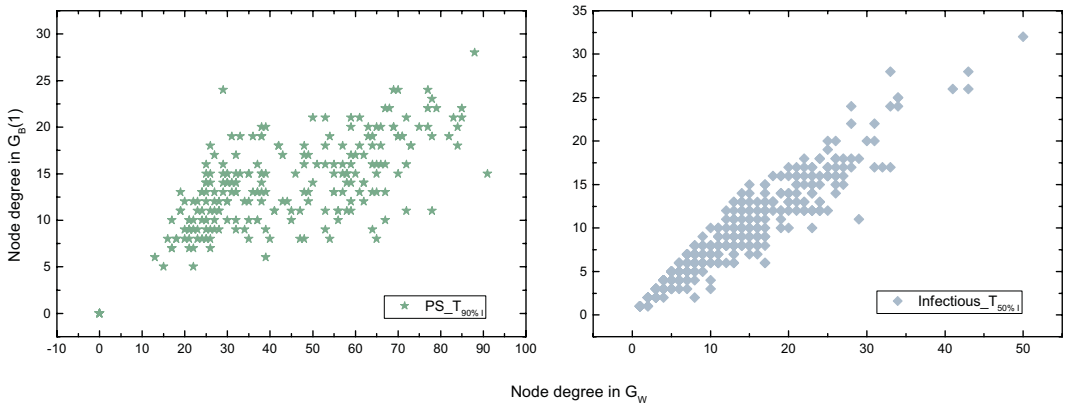

Figure 6. Degree correlation between G_W and $G_B(1)$ for networks *PS* and *Infectious* with the longest observation window respectively.

Degree of a Node in Different Backbones. From now on, we focus on the two extreme backbones $G_B(0) = G_W$ and $G_B(1)$. A node pair that has contact(s) may not necessarily contribute to a diffusion process. Hence, the degree of a node in $G_B(0)$ is larger or equal to its degree in $G_B(1)$. The comparison of the number of links in $G_B(0)$ and $G_B(1)$ in Fig. 5 shows that $G_B(1)$ indeed has less links than $G_B(0)$, especially when the observation time window is large. As explained earlier, $G_B(1)$ and $G_B(0)$ are similar to each other in topology when the observation time window is small.

Furthermore, we explore the degree of a node in $G_W = G_B(0)$ and $G_B(1)$ respectively. Interestingly, a universal finding is that the degree of a node in these two backbones tends to be linearly and positively correlated in all the empirical networks. Table S2 in the APPENDIX E provides the Pearson correlation coefficient between the degree of a node in G_W and in $G_B(1)$, which is above 0.7 for all the networks. Since the topology of $G_B(1)$ is a subgraph of G_W, the degrees of a node in these two networks tend to be linearly correlated if these two networks have a similar number of links. This explains the high degree correlation when the temporal networks have a short observation window. Figure 6 shows the scatter plot of the degree of each node in G_W and $G_B(1)$ respectively for the network with the longest observation window when their backbones G_W and $G_B(1)$ differ much in the number of links derived from two datasets respectively. The strong degree correlation in all these cases suggests that a node with a high degree in G_W tends to have a high degree in $G_B(1)$. A node that has contacts with many others tends to be able to propagate the information directly to many others.

Is this because the degree distribution in G_W is highly heterogeneous that overrules the temporal orders of the contacts in determining how many other nodes a node is able to reach directly? Fig. 3 shows the degree distributions in G_W and $G_B(1)$ respectively for each temporal network dataset with its longest observation window as given in Table S1 in APPENDIX A when these two backbones differ the most. We find that the degree distributions in these two backbones respectively indeed share a similar shape, which again support the strong linear correlation between the degrees of a node in these two backbones. However, not all networks G_W have a power-law degree distribution. The strong degree correlation between G_W and $G_B(1)$ exists even when G_W has a relatively homogeneous degree distribution. This observation motivates us to explore whether a node pair with

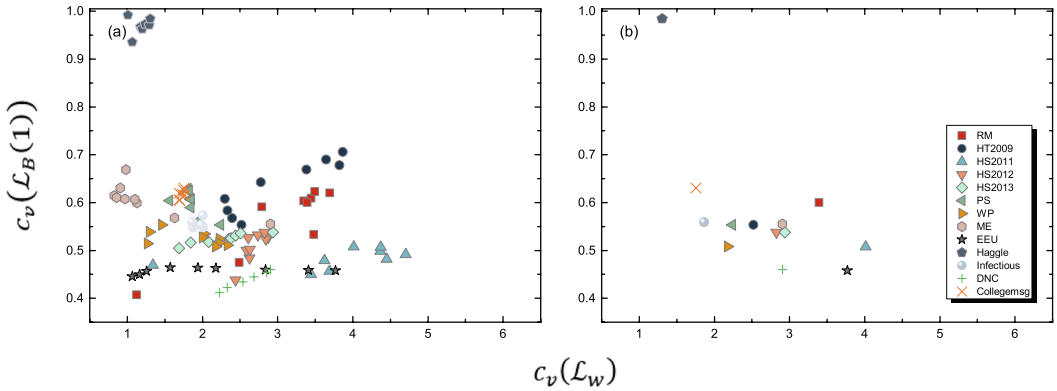

Figure 7. The relationship between the coefficient of variation c_v of the weight distribution in G_W and $G_B(1)$ for (**a**) all the networks with observation windows given in Table S1 in APPENDIX A; (**b**) all the networks with longest observation windows.

a high degree product in G_W thus also in $G_B(1)$ tends to be connected in $G_B(1)$ in Section Relationship between Local Features and the Diffusion Backbone $G_B(1)$.

The degree of a node j in $G_B(1)$ tells maximally how many nodes it could propagate the information directly to given that each node is possibly the source of the information, but not necessarily how frequently this node contributes or engages in an information diffusion process when $\beta = 1$. The latter is reflected from the node strength of a node in $G_B(1)$: $\sum_{k=1}^{N} w_{jk}^B(\beta = 1)$.

Link Weight Variance in Different Backbones. The standard deviation of link weights in a backbone indicates how much the links differ in their probability of appearing in a diffusion process. We compare the standard deviation of a link weight normalized by its mean $c_v = \frac{\sqrt{\mathrm{Var}[W^B]}}{\mathrm{E}[W^B]}$ (which is called the coefficient of variation) in $G_B(1)$ and $G_B(0)$. Figure 7 shows that the link weights in $G_B(0)$ or equivalently G_W are more heterogeneous than those in $G_B(1)$ for almost all the networks we considered. The relatively homogeneous link weights in $G_B(1)$ implies that predicting which node pairs tend to have a high weight in $G_B(1)$ can be challenging.

Identifying the Diffusion Backbone $G_B(1)$. In this section, we investigate how to identify the (high weight) links in the backbone $G_B(1)$ based on local and temporal connection features of each node pair. The key objective to understand how a node pair's local and temporal connection features are related to its role in the global diffusion backbone $G_B(1)$. Our investigation may also allow us to approximate the backbone, whose computational complexity is high ($\mathcal{O}(N^3T)$) base on local temporal features whose computational complexity is low.

We propose to consider systematically a set of local temporal features for node pairs and examine whether node pairs having a higher value of each feature/metric tend to be connected in the backbone $G_B(1)$. Some of these features are derived from the integrated network G_W whereas the feature *Time-scaled Weight* that we will propose encodes also the time stamps of the contacts between a node pair. These node pair features or metrics include:

- *Time-scaled Weight* of a node pair (j, k) is defined as

$$\phi_{jk}(\alpha) = \sum_{m=1}^{n} \left(\frac{1}{t_{jk}^{(m)}} \right)^{\alpha} \tag{3}$$

where n is the total number of contacts between j and k over the given observation window and $t_{jk}^{(i)}$ is the time stamp when the *i-th* contact occurs and α is the scaling parameter to control the contribution of temporal information. For the node pairs that have no contact, we assume their temporal weights to be zero. This metric is motivated by the intuition that when each node is set as the seed of the diffusion process at time $t = 0$, the contacts that happen earlier have a higher probability to be used for the actual information diffusion, thus appear in $G_B(1)$. When $\alpha = 0$, $\phi_{jk}(0) = w_{jk}^B(\beta = 0)$ degenerates to the weight of the node pair in G_W. Larger α implies the node pairs with early contacts have a higher time-scaled weight.

- *Degree Product* of a node pair (j, k) refers to $d_j(\beta = 0) \cdot d_k(\beta = 0)$, the product of the degrees of j and k in the integrated network G_W. If two nodes are not connected in G_W, their degree product is zero. The motivation for this measure is as follows. Given the degree of each node in $G_B(1)$ and if the links are randomly placed, the probability that a node pair (j, k) is connected in $G_B(1)$ is proportional to $d_j(\beta = 1) \cdot d_k(\beta = 1)$. We have observed in Section Relationship between Diffusion Backbones that the degrees of a node in G_W and $G_B(1)$ are strongly and positively correlated. Moreover, only node pairs connected in G_W are possible to appear or be connected in $G_B(1)$. If the connections in $G_B(1)$ are as random as in the configuration model[40], node pairs with a high degree product $d_j(\beta = 0) \cdot d_k(\beta = 0)$ tend to appear in $G_B(1)$.

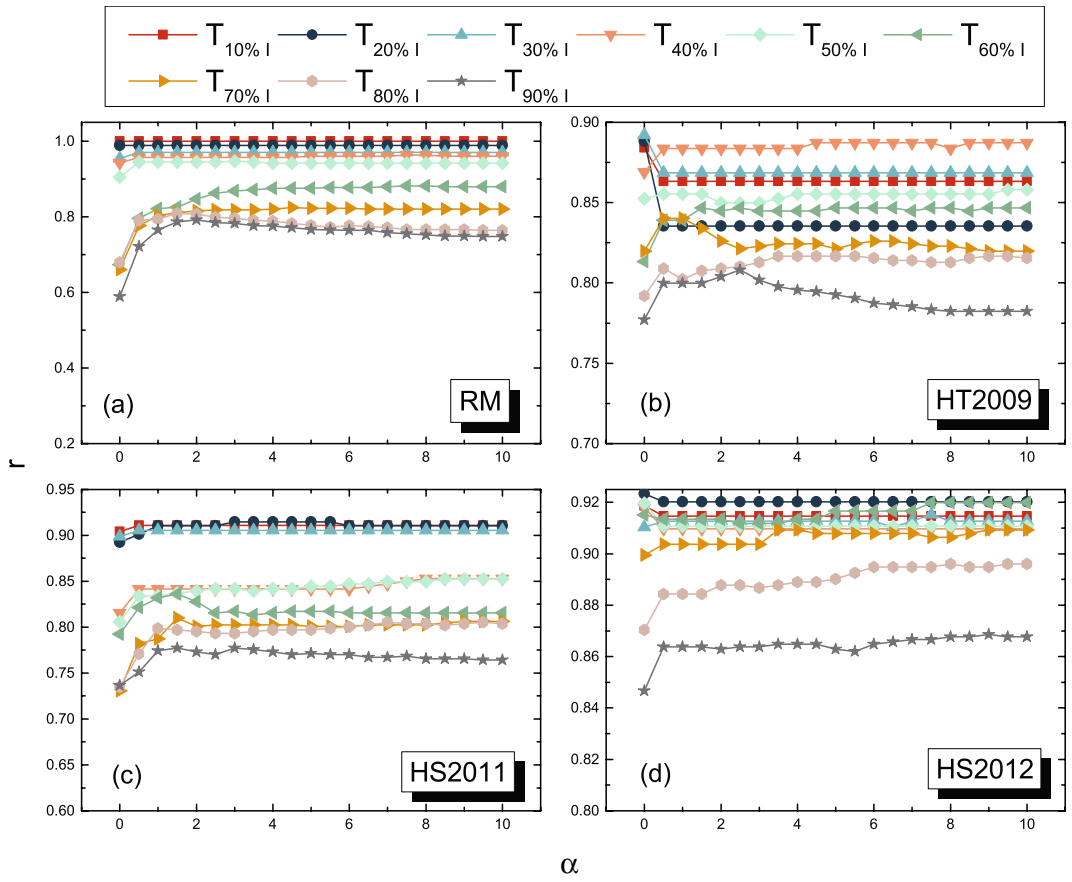

Figure 8. The quality of identifying links in $G_B(1)$ by using the time-scaled weight $\phi_{jk}(\alpha)$ as a function of α in temporal networks derived from datasets (**a**) *RM*, (**b**) *HT*2008, (**c**) *HS*2011 and (**d**)*HS*2012.

- *Strength Product* of a node pair (j, k) refers to $s_j(\beta=0)\cdot s_k(\beta=0)$, the product of the node strengths of j and k in the integrated network G_W, where the node strength $s_j(\beta = 0) = \sum_{i\in\mathcal{N}} A(j, i)$ of a node in G_W equals the total weight of all the links incident to this node[41,42]. If two nodes are not connected in G_W, their strength product is zero. This measure is an extension of the degree product to weighted networks.
- *Betweenness* of a link in G_W counts the number of shortest paths between all node pairs that traverse the link. The distance of each link, based on which the shortest path is computed, is considered to be $\frac{1}{w^B_{jk}(\beta=0)}$, inversely proportional to its link weight in G_W, since a node pair with more contacts tend to propagate information faster[43,44]. Node pairs that are not connected in G_W have a betweenness 0. Betweenness is not local, but considered here as a benchmark feature that has been widely studied.

We explore further whether these node pair features could well identify the connection of node pairs in $G_B(1)$. According to the definition of the aforementioned centrality metrics, a higher value of a metric may suggest the connection of the corresponding node pair in $G_B(1)$. According to each metric, we rank the node pairs and the $|\mathcal{L}_B(1)|$ node pairs with the highest values are identified as the links in $G_B(1)$. The identification quality of a metric, e.g., the time-scaled weight $\phi_{jk}(\alpha)$, is quantified as the overlap $r(\phi_{jk}(\alpha),1)$ between the identified link set and the link set $\mathcal{L}_B(1)$ in $G_B(1)$, as defined by Eq. (2).

Before we compare all the metrics in their identification powers, we examine first how the scaling parameter α in the time-scaled weight $\phi_{jk}(\alpha)$ influences its identification quality. Figure 8 and Figure S3 in the APPENDIX D show that the quality differs mostly when $0 \le \alpha \le 2$ and remains relatively stable when $\alpha \ge 2$ in all the temporal networks. Hence, we will confine ourselves to the range $0 \le \alpha \le 2$.

The quality r by using each metric versus the ratio $\frac{|\mathcal{L}_B(1)|}{|\mathcal{L}_W|}$ of the number of links in $G_B(1)$ to that in G_W are plotted in Fig. 9 for all the empirical temporal networks, with different lengths of the observation time windows. The diagonal curve $r = \frac{|\mathcal{L}_B(1)|}{|\mathcal{L}_W|}$ corresponds to the quality of the random identification, where $|\mathcal{L}_B(1)|$ links are randomly selected from the links in G_W as the identification for the links in $G_B(1)$. Degree product, strength product and betweenness perform, in general, worse than or similarly to the random identification. Even if the connections in $G_B(1)$ were random given the degree of each node in $G_B(1)$, the quality r of identifying links in $G_B(1)$ by using the degree product is close to that of the random identification, if the distribution of the degree product is relatively homogeneous or if the $\frac{|\mathcal{L}_B(1)|}{|\mathcal{L}_W|}$ is large. The degree distribution in $G_B(1)$ is indeed relatively homogeneous

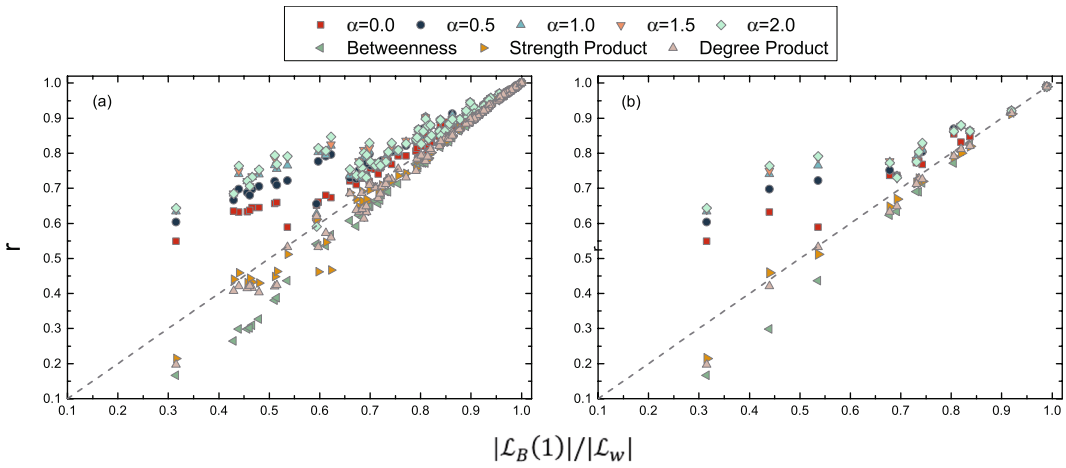

Figure 9. The quality of identifying links in $G_B(1)$ by using each metric for (**a**) all the networks with observation windows given in Table S1 in APPENDIX A; (**b**) all the networks with longest observation windows. The time-scaled weight with different α values are considered.

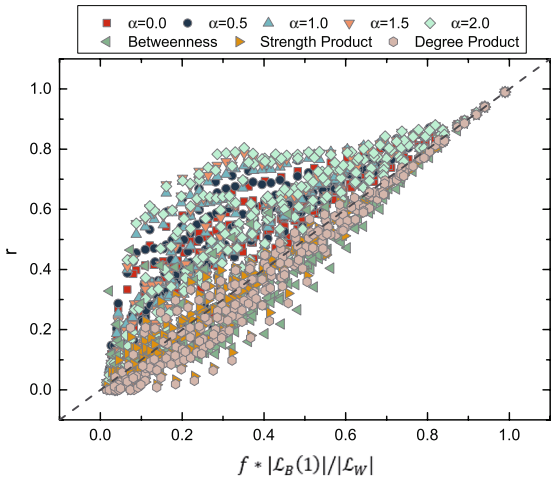

Figure 10. The quality r of identifying top weighted links in $G_B(1)$ by using each metric for all the networks with longest observation windows in each dataset. The time-scaled weight with different α values are considered.

and $\frac{|\mathcal{L}_B(1)|}{|\mathcal{L}_W|}$ is large in most empirical networks. This explains why the degree product performs similarly to the random identification.

The link weight in G_W, equivalently, $\phi_{jk}(\alpha = 0)$, outperforms the random identification, whereas the time-scaled weight $\phi_{jk}(\alpha)$ with a larger α performs better. Node pairs with many contacts that occur early in time tend to contribute to the actual information propagation, i.e., be connected in $G_B(1)$. This observation suggests that the temporal information is essential in determining the role of nodes in a spreading process.

We investigate also whether these metrics can identify the links with the highest weights in $G_B(1)$. The quality r, as defined earlier, of identifying the top f fraction of links with the highest weight in $G_B(1)$ is plotted in Fig. 10. We choose the top $f_*|\mathcal{L}_B(1)|$ node pairs according to each metric as the identification of the top $f_*|\mathcal{L}_B(1)|$ links in $G_B(1)$ with the highest weights. We consider the networks with the longest observation window from each dataset. The diagonal curve $r = f_* \frac{|\mathcal{L}_B(1)|}{|\mathcal{L}_W|}$ corresponds to the quality of the random identification. Similar to the identification of all the links in $G_B(1)$, the time-scaled weight $\phi_{jk}(\alpha)$ with a large α performs the best in identifying highly weighted links in $G_B(1)$, emphasizing again the important role of the temporal information of contacts.

Discussion

Much effort has been devoted to understand how temporal network features influence the prevalence of a diffusion process. In this work, we addressed the further question: node pairs with what kind of local and temporal connection features tend to appear in a diffusion trajectory or path, thus contribute to the actual information diffusion? We consider the Susceptible-Infected spreading process with an infection probability β per contact on a

temporal network as the starting point. We illustrate how to construct the information diffusion backbone $G_B(\beta)$ where the weight of each link tells the probability that a node pair appears in a diffusion process starting from a random node. We unravel how these backbones corresponding to different infection probabilities relate to each other with respect to their topology (overlap in links), the heterogeneity of the link weight, and the correlation in node degree. These relations point out the importance of two extreme backbones: $G_B(1)$ and the integrated network $G_B(0) = G_W$, between which $G_B(\beta)$ varies. We find that the temporal node pair feature that we proposed could better identify the links in $G_B(1)$ as well as the high weight links than the features derived from the integrated network. This universal finding across all the empirical networks highlights that temporal information is crucial in determining a node pair's role in a diffusion process. A node pair with many early contacts tends to appear in a diffusion process. We have also used rank correlation like Kendall and Spearman to evaluate the quality of time-scaled weight in identifying the precise weight ranking of all the links in $G_B(1)$. However, we found that the time-scaled weight when $\alpha = 0$ performs the best, which means the temporal node pair feature is not ideal to identify the exact importance of the links in the backbone $G_B(1)$. Therefore, how to predict the ranking of the link weights in the backbone remains as an interesting future question.

This work reminds us the studies a decade ago about the information transportation via the shortest path on a static network. How frequently a link appears in a shortest path thus contributes to the transportation of information is reflected by the weight of the link in the backbone or overlay, the union of shortest paths between all node pairs[45]. This weight equals the betweenness, which has a high computational complexity, thus motivated the exploration of how a node pair's local connection features are related to its betweenness.

The study of information diffusion paths on a temporal network is more complex due to the extra dimension of time. Our finding that early contacts with a quadratic decay in weight over time indicates the appearance of a node pair in a diffusion path, suggests the possibility to identify the appearance of a node pair in a diffusion path in a long period based on its early contacts within a short period, an interesting follow-up question. This work opens new challenging questions like which nodes tend to be reached early and more likely by the information, how such heterogeneous features at node or link level are related to local temporal connection features. In addition, other spreading models like social contagions and coevolution spreading models can be further considered beyond the *SI* spreading model studied here[1,2,46–49]. Our findings may inspire the exploration of optimization problems such as which node pairs or contacts should be stimulated (e.g. added) in order to maximize the prevalence of an information diffusion process. Stimulating early contacts seems essential but adding them between which node pairs and when is non-trivial.

References

1. Watts, D. J. A simple model of global cascades on random networks. *Proc. Natl Acad. Sci. USA* **99**, 5766–5771 (2002).
2. Granovetter, M. Threshold models of collective behavior. *Am. J. Sociol* **83**, 1420–1443 (1978).
3. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925 (2015).
4. Liu, C., Zhan, X.-X., Zhang, Z.-K., Sun, G.-Q. & Hui, P. M. How events determine spreading patterns: information transmission via internal and external influences on social networks. *New J. Phys.* **17**, 113045 (2015).
5. Zhang, Z.-K. *et al*. Dynamics of information diffusion and its applications on complex networks. *Phys. Rep* **651**, 1–34 (2016).
6. Wang, H. *et al*. Effect of the interconnected network structure on the epidemic threshold. *Phys. Rev. E* **88**, 022801 (2013).
7. Qu, B. & Wang, H. Sis epidemic spreading with heterogeneous infection rates. *IEEE Trans. Netw. Sci. Eng.* **4**, 177–186 (2017).
8. Barabási, A.-L. *Network science* (Cambridge university press, 2016).
9. Holme, P. & Saramäki, J. Temporal networks. *Phys. Rep* **519**, 97–125 (2012).
10. Holme, P. Modern temporal network theory: a colloquium. *Eur. Phys. J. B* **88**, 234 (2015).
11. Scholtes, I. *et al*. Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks. *Nat. Commun.* **5**, 5024 (2014).
12. Valdano, E., Ferreri, L., Poletto, C. & Colizza, V. Analytical computation of the epidemic threshold on temporal networks. *Phys. Rev. X* **5**, 021005 (2015).
13. Zhang, Y.-Q., Li, X. &Vasilakos, A. V. Spectral analysis of epidemic thresholds of temporal networks. *IEEE Trans. Cybern* (2017).
14. Karsai, M. *et al*. Small but slow world: How network topology and burstiness slow down spreading. *Phys. Rev. E* **83**, 025102 (2011).
15. Lambiotte, R., Tabourier, L. & Delvenne, J.-C. Burstiness and spreading on temporal networks. *Eur. Phys. J. B* **86**, 320 (2013).
16. Moinet, A., Starnini, M. & Pastor-Satorras, R. Burstiness and aging in social temporal networks. *Phys. Rev. Lett.* **114**, 108701 (2015).
17. Hethcote, H. W. The mathematics of infectious diseases. *SIAM review* **42**, 599–653 (2000).
18. Rocha, L. E. & Blondel, V. D. Bursts of vertex activation and epidemics in evolving networks. *PLOS Comput. Biol.* **9**, e1002974 (2013).
19. 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).
20. Starnini, M., Machens, A., Cattuto, C., Barrat, A. & Pastor-Satorras, R. Immunization strategies for epidemic processes in time-varying contact networks. *J. Theor. Biol.* **337**, 89–100 (2013).
21. Yang, Z. & Zhou, T. Epidemic spreading in weighted networks: an edge-based mean-field solution. *Phys. Rev. E* **85**, 056106 (2012).
22. Chu, X., Guan, J., Zhang, Z. & Zhou, S. Epidemic spreading in weighted scale-free networks with community structure. *J. Stat. Mech. Theory Exp.* **2009**, P07043 (2009).
23. Pfitzner, R., Scholtes, I., Garas, A., Tessone, C. J. & Schweitzer, F. Betweenness preference: Quantifying correlations in the topological dynamics of temporal networks. *Phys. Rev. Lett.* **110**, 198701 (2013).
24. Li, X. & Li, X. Reconstruction of stochastic temporal networks through diffusive arrival times. *Nat. Commun.* **8**, 15729 (2017).
25. Reality mining network dataset–KONECT, http://konect.uni-koblenz.de/networks/mit.
26. Eagle, N. & (Sandy) Pentland, A. Reality Mining: Sensing complex social systems. *Pers. Ubiquitous Comput* **10**, 255–268 (2006).
27. Hypertext 2009 network dataset–KONECT, http://konect.uni-koblenz.de/networks/sociopatterns-hypertext.
28. Isella, L. *et al*. What's in a crowd? analysis of face-to-face behavioral networks. *J. Theor. Biol.* **271**, 166–180 (2011).
29. Fournet, J. & Barrat, A. Contact patterns among high school students. *PloS One* **9**, e107878 (2014).
30. Mastrandrea, R., Fournet, J. & Barrat, A. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PloS One* **10**, e0136497 (2015).
31. Stehlé, J. *et al*. High-resolution measurements of face-to-face contact patterns in a primary school. *PloS One* **6**, e23176 (2011).
32. 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).
33. Haggle network dataset–KONECT, http://konect.uni-koblenz.de/networks/contact.

34. Chaintreau, A. *et al*. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Trans. Mob. Comput* **6**, 606–620 (2007).
35. Manufacturing emails network dataset–KONECT, http://konect.uni-koblenz.de/networks/radoslaw_email.
36. Michalski, R., Palus, S. & Kazienko, P. Matching organizational structure and social network extracted from email communication. In *Lecture Notes in Business Information Processing*, vol. 87, 197–206 (Springer Berlin Heidelberg, 2011).
37. Leskovec, J., Kleinberg, J. & Faloutsos, C. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discovery Data* **1**, 2 (2007).
38. Dnc emails network dataset–KONECT, http://konect.uni-koblenz.de/networks/dnc-temporalGraph.
39. Panzarasa, P., Opsahl, T. & Carley, K. M. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *J. Assoc. Inf. Sci. Technol* **60**, 911–932 (2009).
40. Newman, M. E., Strogatz, S. H. & Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118 (2001).
41. Wang, H. *et al*. Effect of tumor resection on the characteristics of functional brain networks. *Phys. Rev. E* **82**, 021924 (2010).
42. Grady, D., Thiemann, C. & Brockmann, D. Robust classification of salient links in complex networks. *Nat. Commun.* **3**, 864 (2012).
43. Newman, M. E. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E* **64**, 016132 (2001).
44. Wang, H., Hernandez, J. M. & Van Mieghem, P. Betweenness centrality in a weighted network. *Phys. Rev. E* **77**, 046105 (2008).
45. Van Mieghem, P. & Wang, H. The observable part of a network. *IEEE/ACM Trans. Netw.* **17**, 93–105 (2009).
46. Chen, X., Wang, W., Cai, S., Stanley, H. E. & Braunstein, L. A. Optimal resource diffusion for suppressing disease spreading in multiplex networks. *Journal of Statistical Mechanics: Theory and Experiment* **2018**, 053501 (2018).
47. Zhan, X.-X. *et al*. Coupling dynamics of epidemic spreading and information diffusion on complex networks. *Applied Mathematics and Computation* **332**, 437–448 (2018).
48. Wang, W., Cai, M. & Zheng, M. Social contagions on correlated multiplex networks. *Physica A: Statistical Mechanics and its Applications* **499**, 121–128 (2018).
49. Wang, W., Liu, Q., Liang, J., Hu, Y. & Zhou, T. Coevolution spreading in complex networks. *CoRR* abs/1901.02125 (2019).

Acknowledgements

This work has been partially supported by the China Scholarship Council (CSC).

Author Contributions

X.Z., A.H. and H.W. planed the study; X.Z. and H.W. performed the experiments, analyzed the data and prepared the figures. All authors wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-43029-5.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.