Temporal information gathering process for node ranking in time-varying networks

Qu, Cunquan; Zhan, Xiuxiu; Wang, Guanghui; Wu, Jianliang; Zhang, Zi-ke

DOI
10.1063/1.5086059

Publication date
2019

Document Version
Final published version

Published in
Chaos

Temporal information gathering process for node ranking in time-varying networks

Cite as: Chaos **29**, 033116 (2019); https://doi.org/10.1063/1.5086059
Submitted: 17 December 2018 . Accepted: 19 February 2019 . Published Online: 08 March 2019

Cunquan Qu, Xiuxiu Zhan, Guanghui Wang, Jianliang Wu, and Zi-ke Zhang

Temporal information gathering process for node ranking in time-varying networks

Cite as: Chaos **29**, 033116 (2019); doi: 10.1063/1.5086059
Submitted: 17 December 2018 · Accepted: 19 February 2019 ·
Published Online: 8 March 2019

Cunquan Qu,[1] Xiuxiu Zhan,[2,a)] Guanghui Wang,[1,b)] Jianliang Wu,[1] and Zi-ke Zhang[3,4]

AFFILIATIONS

[1]School of Mathematics, Shandong University, Jinan 250110, People's Republic of China
[2]Intelligent Systems, Delft University of Technology, Delft 2600GA, The Netherlands
[3]Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou 311121, People's Republic of China
[4]Engineering Research Center of Mobile Health Management System, Ministry of Education, Hangzhou 311121, People's Republic of China

[a)]**Electronic mail:** x.zhan@tudelft.nl
[b)]**Electronic mail:** ghwang@sdu.edu.cn

ABSTRACT

Many systems are dynamic and time-varying in the real world. Discovering the vital nodes in temporal networks is more challenging than that in static networks. In this study, we proposed a temporal information gathering (TIG) process for temporal networks. The *TIG*-process, as a node's importance metric, can be used to do the node ranking. As a framework, the *TIG*-process can be applied to explore the impact of temporal information on the significance of the nodes. The key point of the *TIG*-process is that nodes' importance relies on the importance of its neighborhood. There are four variables: temporal information gathering depth n, temporal distance matrix D, initial information c, and weighting function f. We observed that the *TIG*-process can degenerate to classic metrics by a proper combination of these four variables. Furthermore, the fastest arrival distance based *TIG*-process (*fad-tig*) is performed optimally in quantifying nodes' efficiency and nodes' spreading influence. Moreover, for the *fad-tig* process, we can find an optimal gathering depth n that makes the *TIG*-process perform optimally when n is small.

Vital node identification is crucial for understanding the topology of network structures as well as controlling the spreading process in complex systems. Even though many node ranking metrics have been designed for static networks, there is a lack of research in temporal systems. Also, how the temporal information influences node ranking is still unknown. In this study, we proposed a temporal information gathering (TIG) process. On the one hand, the *TIG*-process can be used to design a node ranking measurement. On the other hand, as a framework, the *TIG*-process can be applied to explore the impact of temporal information on nodes' importance. Many basic metrics can be derived from the *TIG*-process. Furthermore, we found that there exists an optimal gathering depth that makes the *TIG*-process perform optimally. Also, the fastest arrival distance based *TIG*-process works better than the other kinds of distance, which capture less temporal information.

I. INTRODUCTION

Vital node identification has attracted increasing attention lately due to its great significance as well as valuable applications.[1–4] As a matter of fact, a small number of influential nodes can affect mechanisms like cascading, spreading, and synchronizing in complex systems.[5] In the view of the application, finding vital nodes can help one to promote products in viral marketing,[6] to control the spread of rumors,[7] to prevent a catastrophic outage in power grids or the Internet,[8] etc.

Researchers have defined a series of important node identification metrics in static networks recently, such as neighborhood-based centrality metrics, path-based centrality metrics, etc.[9–11] One of the most representative neighborhood-based centrality metrics is degree centrality, which is efficient. However, the degree centrality considers only the direct contacts of each node.[12] Path-based centrality metrics show high accuracy as they consider the global information of the

network but usually with high computational complexity, such as Katz centrality,[13] which is difficult to be used in large-scale networks.[5] Despite the achievement in defining the node's important metrics for static networks, there is still a large gap in identifying important nodes in temporal networks. Most complex systems in the real world are changing over time and their corresponding networks are called temporal networks or time-varying networks.[14–19]

The study of identifying vital nodes in temporal networks can be more challenging than that in static networks, as the network is always changing with time. In temporal networks, a node has different roles in the different time step, which means that the importance of nodes also varies with time. There are some pioneering research studies concentrated on ranking nodes in temporal networks.[20–22] For example, some researchers first cut the temporal networks into a series of static snapshots and then estimate a node's topological importance using the average value of its centrality over all static snapshots.[20,21] The node ranking metrics obtained by this way are the generalization of the static ones; for instance, the temporal degree, temporal closeness, and temporal betweenness[20] belong to this class of methods. Even though these methods may gain some improvement in finding vital nodes compared in static metrics in temporal networks, cutting the temporal networks into slides and taking the average value of all the slides may result in a loss of some temporal information, such as the order of slices will be ignored in this process. Therefore, it is necessary to define node ranking metrics that can describe the evolution of the nodes' influence or capture more temporal information.

Node ranking metrics using local information of the nodes (e.g., degree and H-index) have shown good performance in identifying important nodes.[9,23] Meanwhile, some researchers claim that the global structure or the position of the nodes in the network should be considered in node ranking methodologies. Therefore, metrics like betweenness,[24] closeness,[25] and k-core centrality[26] are designed to capture global information.

In this study, we propose a temporal information gathering based process in the context that each node is attributed a piece of initial information, since, for example, when a person first joined a new group, she/he has its own attribute. After communicating with other members, her/his importance is changing and can be estimated by her/his colleagues (neighborhood). To simplify, we denote the neighborhood of node v_i as $\mathcal{N}_{\le l}(i)$, which indicates the nodes with a temporal distance less than or equal to l. Throughout the paper, we use the *TIG*-process to denote the **T**emporal **I**nformation **G**athering process and *tig*-score represents node importance obtained from the *TIG*-process. The *TIG*-process is controlled by four variables, i.e., (n, f, D, c), where n illustrates the temporal gathering depth, f is the weighting function, D is the temporal distance matrix, and c describes the initial information. In Sec. V, we take some basic centrality metrics as initial information to conduct the experiments.

We find that the fastest arrival distance based *TIG*-process performs much better than the one based on the temporal shortest distance. Also, for the former, we can get an optimal gathering depth n, regardless of the initial information,[22] including static degree, static closeness, static strength, static betweenness, Eigenvector centrality, and PageRank centrality.[20] As the depth n increases, the performance will be degraded. In addition, many basic metrics can be derived from the *TIG*-process by proper combinations of the four parameters mentioned above.

The rest of the paper is organized as follows. In Sec. II, we give the definition of the *TIG*-process. We describe the benchmark metrics and two evaluation methods in Sec. III. The datasets used in this paper are given in Sec. IV, and the results are shown in Sec. V. We discuss and conclude in Sec. VI.

II. TEMPORAL INFORMATION GATHERING PROCESS

In this section, we give a detailed illustration of the *TIG*-process.

A. Basic notations and definitions

Firstly, we give some basic notations and definitions used in this paper.

Let $G^T = (V, E^T)$ be a temporal network observed on $[1, T]$, where V is the node set, E^T is the event set, and $[1, T]$ is the observation time window. An event $e^T \in E^T$ is defined by a quadruple (u, v, t_0, λ), where $u, v \in V$, t_0 is the start time of the event, λ is the lasting time, and $t_0 + \lambda$ is the ending time. In this paper, we assume $\lambda = 0$, which means we only consider instant events. At each time $t \in [1, T]$, the adjacent matrix is denoted as A_t, where $A_t(i, j) = 1$ if there is a contact between node v_i and v_j at time t. In addition, the unweighted integrated static network of G^T is expressed as $G = (V, E)$, where E is the static edge set. The adjacent matrix of G is denoted as A and the distance matrix is $\mathcal{M}$. The entry $\mathcal{M}(i, j)$ indicates the distance between the two corresponding nodes v_i and v_j.

- **Temporal path**: A temporal path in the temporal network G^T is a sequence of nodes $P = \langle v_1, v_2, \ldots, v_k, v_{k+1} \rangle$, where $(v_i, v_{i+1}, t_i) \in E^T$ is the i-th event on P for $1 \le i \le k$. Then, the start time of P is $t_{start}(P) = t_1$ and the end time of P is $t_{end}(P) = t_k$. We define the temporal **length** of P as $l(P) = t_{end}(P) - t_{start}(P) + 1$. Given a time period $[t_\alpha, t_\omega]$, let $\mathbf{P}(u, v, [t_\alpha, t_\omega]) = \{P : \text{P is a temporal path from } u \text{ to } v \text{ such that } t_{start}(P) > t_\alpha \text{ and } t_{end}(P) < t_\omega\}$.

  In static networks, the distance between two nodes is defined by the length of the shortest path between them. However, in temporal networks, we have many ways to define the distance between nodes with regard to the physical distance as well as the duration time.[27]

  In this paper, we introduce two distance definitions for temporal networks, i.e., the fastest arrival distance and the temporal shortest distance.
- **Fastest arrival path**: The fastest arrival path between node u and v is the path that goes from u to v taking the minimum elapsed time counted from $t = 1$. In other words, the fastest arrival path is the first arrival path from the starting node u to the destination node v. That is to say, $P \in \mathbf{P}(u, v, [1, T])$ is the fastest arrival path if $t_{end}(P) = \min\{t_{end}(P') : P' \in \mathbf{P}(u, v, [1, T])\}$.

  Also, the fastest arrival distance between node u and node v is measured by the length of the fastest arrival path between them, denoted as $\phi(u, v)$.

  An example of the fastest arrival path is shown in Fig. 1(a) from the toy temporal network given in Fig. 1(b). The fastest arrival distance between node v_1 and node v_4 is 3.

- **Temporal shortest path**: The temporal shortest path from u to v is a path for which the overall traversal time needed is shortest. Therefore, $P \in \mathbf{P}(u, v, [t_\alpha, t_\omega])$ is a temporal shortest path if $l(P) = \min\{l(P') : P' \in \mathbf{P}(u, v, [t_\alpha, t_\omega])\}$. The temporal shortest distance between node u and node v is the length of the temporal shortest path between them, denoted as $\theta(u, v)$. Figure 1(c) shows the temporal shortest path between v_1 and v_4, and $\theta(v_1, v_4) = 2$.
- **Temporal distance matrix**: The temporal distance matrix of G^T is given by $D^{|V|\times|V|}$, where $D = \{D(i,j) = d(v_i, v_j), v_i, v_j \in V\}$. According to the temporal distance defined above, we have two distance matrices, i.e., the fastest arrival distance matrix Φ and the temporal shortest distance matrix Θ.
- **Distance index matrix**: We define a distance index matrix $D_s^{|V|\times|V|}$ as a 0-1 matrix, where

$$D_s(i,j) = \begin{cases} 1, & d(v_i, v_j) = s, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Obviously, $D = \sum_{s=0}^{+\infty}(s \cdot D_s)$. It should be noted that due to the time dependency of the temporal paths, the distance matrix D and the index matrix D_s are both asymmetric.

- **Coefficient of variation**: The coefficient of variation is used to measure the extent of variability in relation to the mean value of a dataset, which is also known as the relative standard deviation. The coefficient of variation is defined as the ratio of the standard deviation to the mean: $C = (\textit{standard deviation})/(\textit{mean value})$.
- **Kendall correlation coefficient**:[28] The Kendall correlation coefficient τ is defined as follows. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be the observations of two joint random variables X and Y. Then, Kendall ranking correlation coefficient $\tau \in [-1, 1]$ is defined as

$$\tau = \frac{1}{n(n-1)} \sum_{i \neq j} \text{sgn}(x_i - x_j)\text{sgn}(y_i - y_j). \tag{2}$$

If τ takes the value of $+1$, then the agreement of the two rankings is perfect. If τ is -1, then one list is the reverse of the other. If τ is close to zero, then the two rankings are independent.

B. The TIG-process

Recall that the temporal information gathering process is denoted by the *TIG*-process for simplification. The ranking score of node v_i obtained from the *TIG*-process is defined as a *tig*-score, denoted as g_i. Assume that each node v_i has an initial score c_i, which is also viewed as the 0-order *tig*-score g_i^0. Therefore, $g^{(0)} = (g_1^{(0)}, g_2^{(0)}, \ldots, g_{|V|}^{(0)}) = (c_1, c_2, \ldots, c_{|V|})$. The *TIG*-process is conducted based on these initial scores. Therefore, the 1st-order *TIG*-process for each node is calculated by gathering the information from its neighbors, i.e., $g^{(1)} = D_1 g^{(0)}$. Similarly, the nth-order *TIG*-process for node v_i is gathering the information of its neighborhood with a distance equal to or less than n from v_i, i.e., $\mathcal{N}_{\le n}(i)$. Thus, the nth-order *TIG*-process is written as

$$g^{(n)} = \sum_{j=0}^{n} f(j) \cdot D_j \cdot g^{(0)}, \tag{3}$$

where f is a function of j, which weighs the significance of jth-order neighbors and D_j is the distance index matrix, that is, $D_j(u, v) = 1$ if $d(u, v) = j$. The nth-order *tig*-score is denoted by $g^{(n)}$. We use $g_i^{(n)}$ to indicate the ranking score of node v_i. Obviously, a larger value of $g_i^{(n)}$ implies node v_i is more important in the network.

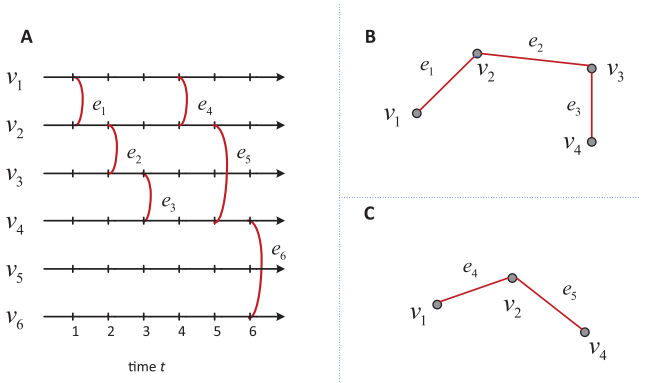

FIG. 1. (a) A schematic representation of a temporal network with nodes $\{v_1, v_2, \ldots, v_6\}$ and events $\{e_1, e_2, \ldots, e_6\}$. There are two paths between node v_1 and v_4. (b) The fastest arrival path between nodes v_1 and v_4. (c) The temporal shortest path between nodes v_1 and v_4.

From Eq. (3), we know that the *TIG*-process can be denoted as a quadruple (n, f, D, c) and these four variables are independent of each other. The variable n controls the information gathering depth, which varies from 1 to T. The weighting function f is a function of j, which weighs the distance effect on the nodes' importance, and it can take different formations, such as $f_j = 1/j$, $f_j = 1$, and so forth. The distance matrix, as we mentioned above, can be defined differently, such as the fastest arrival distance and temporal shortest distance matrix, and so forth. For the initial information c, in the real world, it can be estimated according to the actual situation. However, in the experiments of this paper, we treat some basic metrics as the initial information, such as random values, the degree, the closeness, etc. Many existing metrics can be derived by different combinations of these four variables. We show in Fig. 2 and Table I the relationship between the *TIG*-process and some classic metrics, which will be described in Sec. III.

III. METHODS

Aiming at illustrating the performance of the *TIG*-process, we start by introducing the benchmark metrics used in this study. Also, two performance evaluation metrics, i.e., the network efficiency and the *SIR* spreading influence, will be given at last.

A. Benchmark metrics

- *Static degree centrality (SD)* of node v_i is defined as the degree in the unweighted integrated network G, i.e.,

$$SD(i) = \sum_j A(i,j). \tag{4}$$

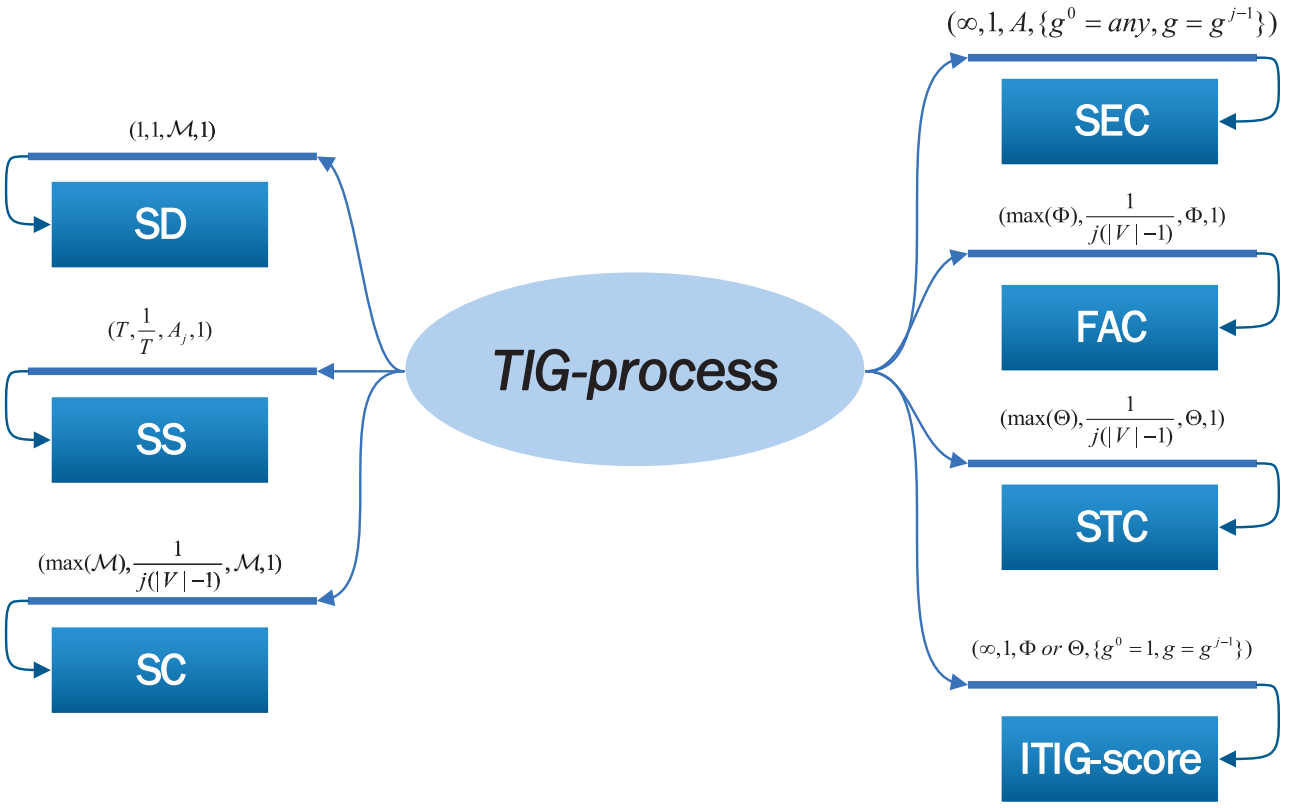

FIG. 2. The relationship between the *TIG*-process and some classic node ranking metrics. The combination of the parameters used in the *TIG*-process is given in Table I.

- *Static strength centrality (SS)* of node v_i counts the number of occurrences of each node that appeared in the temporal network,

$$SS(i) = \sum_{t=1}^{T} \sum_{j} A_t(i,j). \tag{5}$$

TABLE I. The detailed combination of the four parameters in *TIG*-process in order to get the classic metrics.

Benchmark metric	n	f	D	c
SD	1	1	$\mathcal{M}$	$\mathbb{1}$
SS	T	1	A_j	$\mathbb{1}$
SC	$\max(\mathcal{M})$	$\frac{1}{j(\|V\|-1)}$	$\mathcal{M}$	$\mathbb{1}$
SEC	∞	1	A	$g=g^{j-1}$ $c=any$
FAC	$\max(\Phi)$	$\frac{1}{j(\|V\|-1)}$	Φ	$\mathbb{1}$
STC	$\max(\Theta)$	$\frac{1}{j(\|V\|-1)}$	Θ	$\mathbb{1}$
Iterative TIG process	∞	1	Φ or Θ	$g=g^{j-1}$ $c=\mathbb{1}$

- *Static betweenness (SB)* of node v_i is the proportion of shortest paths passing through it, defined as

$$SB(i) = \sum_{h\neq i\neq j} \frac{\sigma_{hj}(i)}{\sigma_{hj}}, \tag{6}$$

where σ_{hj} is the total number of shortest paths from v_h to v_j and $\sigma_{hj}(i)$ is the number of paths passing through v_i in static networks.
- *Static closeness (SC)* of node v_i is given by the reciprocal of the sum of its distances from all the other nodes, namely,

$$SC(i) = \frac{|V|-1}{\sum_{v_j\in V\setminus v_i} \mathcal{M}(i,j)}, \tag{7}$$

where $\mathcal{M}(i,j)$ is the distance between nodes v_i and v_j in G and $V\setminus v_i$ indicates the node set except v_i.
- *Temporal closeness (TC)*[29] at time t of node v_i is the sum of inverse temporal distances to all other nodes in $V\setminus v_i$ in $[t,T]$. Thus, in this paper, the fastest arrival closeness (*FAC*) of node v_i is defined as

$$FAC(i) = \frac{|V|-1}{\sum_{v_j\in V\setminus v_i} \Phi(i,j)}, \tag{8}$$

where $\Phi(i,j)$ is the fastest arrival distance between v_i and v_j in the time interval $[1,T]$. Similarly, the temporal shortest closeness (STC) is defined by

$$STC(i) = \frac{|V|-1}{\sum_{v_j \in V \setminus v_i} \Theta(i,j)}, \tag{9}$$

where $\Theta(i,j)$ indicates the temporal shortest distance between node v_i and v_j.

- *Static eigenvector centrality (SEC).*[30] Given the adjacent matrix A of static network G, $SEC(v_i)$ is equal to the v_i-th component of the eigenvector corresponding to the greatest eigenvalue.
- *Static PageRank centrality (SPR)*[31] is an algorithm used by Google Search to rank web pages. A page gets a higher SPR if there are more links from other pages where the number of links on those other pages and the SPR of those other pages are also important.

B. Network efficiency

The network efficiency[32] is defined based on the assumption that the information in a network passes only through shortest paths. Therefore, we use it to measure how well nodes exchange information. The efficiency $\mathcal{E}(G)$ of the static network G is defined as

$$\mathcal{E}(G) = \frac{1}{|V|(|V|-1)} \sum_{v_i \neq v_j \in G} \frac{1}{\mathcal{M}(v_i,v_j)}, \tag{10}$$

where $\mathcal{M}$ is the distance matrix in static networks. In addition, removing a node or a set of nodes may decrease the efficiency of the network, as it can make the network disconnected. Therefore, the reduction of the efficiency after nodes' removal is used to measure the importance of the nodes in static networks.

When it comes to the temporal network, the efficiency can be defined similarly by replacing $\mathcal{M}$ with some temporal distance matrices. We use the fastest arrival distance matrix Φ or the temporal shortest distance matrix Θ instead of $\mathcal{M}$ in Eq. (10) to define the efficiency $\mathcal{E}_{fad}$ or $\mathcal{E}_{std}$, respectively. Consequently, the node(s) efficiency, denoted as NE, i.e., the importance of the node(s) V' in terms of the network efficiency, is given by $NE(V') = \mathcal{E}(G) - \mathcal{E}(G \setminus V')$. For each node v_i in a network, we define the node efficiency as $NE(i) = \mathcal{E}(G) - \mathcal{E}(G \setminus v_i)$. Similarly, NE_{fad} and NE_{std} indicate the FAD and STD based node efficiency, respectively.

We use the node efficiency as a performance evaluation method to test whether the TIG-process can well predict the node ranking in temporal networks. The evaluation is measured by computing the Kendall ranking correlation coefficient between the node efficiency and the TIG-score with different initial information.

Therefore, the higher τ indicates the better node ranking metric that is used to predict important nodes in terms of the network efficiency.

What is more, since the removal of nodes can reduce the network efficiency, we further explore the changing of the network efficiency as the removing of the top-ranked nodes. Obviously, the better the metric performs, the faster the network efficiency reduces.

C. Spreading influence

Another performance evaluation method for node ranking is based on the spreading process.[33–36] In this paper, we use the SIR spreading model to evaluate the spreading influence of each node in temporal networks. There are three states in an SIR spreading process, i.e., susceptible (S), infected (I), and recovered (R). The infected nodes can infect their susceptible neighbors with the infection probability β, and each infected node can recover from the disease with probability μ. In static networks, the spreading influence of node v_i is usually defined as the spread range R_i, calculated by the number of infected nodes and recovered nodes at the steady states of the SIR process.

However, it is quite different for temporal networks, since each node occurs many times and the occurrence time for each node is different as well. Thus, we do the SIR spreading simulation as follows. We simulate the SIR spreading by following the time order of the interactions. Also, for each node, we do realizations starting from each of its occurrence time, respectively. Finally, for each node at one occurrence time, the result is based on the average of 1000 independent realizations. Therefore, for example, for node v_i, the results can be recorded as $\mathcal{R}(v_i) = \{(t^j_{v_i}, R^j_{v_i}) \mid t^j_{v_i}$ *is the occurrence time of node* $v_i\}$, where $R^j_{v_i}$ represents the spreading range of node v_i, which occurs at time $t^j_{v_i}$. Here, we introduce three different definitions on nodes' spreading influence. The maximal spreading influence of v_i is defined as the largest spreading range over all the occurrence time, denoted as $R_{max}(i)$. The mean spreading influence is calculated by the mean value of the spreading range over all the occurrence time, written as $R_{mean}(i)$. The normalized spreading influence is denoted as $R_{norm}(i)$, which is given by the mean value of $\frac{R^j_{v_i}}{T-t^j_{v_i}}$ over all the occurrence time.

Similarly, we apply the Kendall ranking correlation coefficient between the tig-score and the three kinds of spreading influence mentioned above to measure the ranking performance regarding the spreading influence.

IV. DATASETS

Eight real-world networks are studied in this study, including five face-to-face contact networks and three email communication networks, which are given as follows. For the face-to-face contact networks, the time bin is one day. For the email communication networks, the time window is one week. The basic structural statistics are listed in Table II.

- High school 2011 (2012) dynamic contact network.[37] The dataset records the contacts between students in a high school in Marseilles, France.
- Primary school temporal network.[38] The dataset contains the temporal network of contacts between the children and teachers in a primary school.
- Hospital ward dynamic contact network.[39] The dataset contains the temporal network of contacts between patients, patients and health-care workers (HCWs), and among HCWs in a hospital ward in Lyon, France.
- Contact network in a workplace.[40] The dataset contains the temporal network of contacts between individuals in an office building.
- Email-Eu-core temporal network.[41] The network is generated using email data from a large European research institution.

TABLE II. Basic features of the real-world networks. The number of nodes ($|V|$), the length of the observation time window (T), the total number of contacts ($|E|$), and C_{fad} denotes the coefficient of variation of the average fastest arrival distance from each node to the others. C_{std} indicates the coefficient of variation of the temporal shortest distance from each node to the others.

Network	$\|V\|$	T	$\|E\|$	C_{fad}	C_{std}
High school 2011	126	42	28 561	0.579 8	0.340 5
High school 2012	180	87	45 047	0.619 6	0.366 4
Primary school	242	20	125 773	0.528 8	0.118 8
Workplace	92	108	9827	0.619 1	0.410 2
Hospital contact	75	90	32 424	0.841 1	0.795 6
Eu core	771	68	38 328	1.291 3	0.652 2
Manu factory	167	268	82 927	0.908 1	0.662 9
OC communication	1898	188	61 726	2.457 9	1.064 5

- Manufacturing emails.[42] This network is the internal email communication network between employees of a mid-size manufacturing company.
- CollegeMsg temporal network.[43] This network is comprised of private messages sent on an online social network at the University of California, Irvine.

V. RESULTS

For the experiments in this study, we take the weighting function f as 1, which means for each node v_i, we treat all the nodes in $\mathcal{N}_{\leq n}(i)$ equally. The fastest arrival distance matrix (Φ) and temporal shortest distance matrix (Θ) are considered as the temporal distance matrix D, respectively. Also, we call these two kinds of *TIG*-process as *FAD*-based *tig*-process and *STD*-based *tig*-process, denoted as *fad-tig* and *std-tig*, to simplify. For the initial information, some basic node ranking metrics are taken into account, including static degree (SD), static betweenness (SB), static closeness (SC), static strength (SS), Eigenvector centrality (SEC), and Pagerank centrality (SPR).

A. Quantifying node efficiency

Recall that in Sec. III, we introduced the definition of the node efficiency. Here, we denote the *FAD*-based and *STD*-based node efficiency as NE_{fad} and NE_{std}, respectively. Similarly, the Kendall ranking correlation coefficients between NE_{fad} and *fad-tig*, NE_{fad} and *std-tig*, NE_{std} and *fad-tig*, and NE_{std} and *std-tig* are denoted as τ_{ff}, τ_{fs}, τ_{sf}, and τ_{ss}, respectively.

Figure 3 shows the changing of τ_{ff} as the gathering depth n increases. The τ_{ff} can get a maximal value when n is small, especially for the three email communication networks. Furthermore, since the concept of the node efficiency is based on the shortest paths, the *tig*-score with the initial information of static closeness centrality gets the best performance.

The case of using *fad-tig* to estimate NE_{std} is similar to the one using *fad-tig* to estimate NE_{fad}. However, in Fig. 4, we can see that the τ_{fs} is increasing as n increases in general. Dissimilar with Fig. 4, the τ_{ff} decreases or keeps steady when n is large enough. In other words,

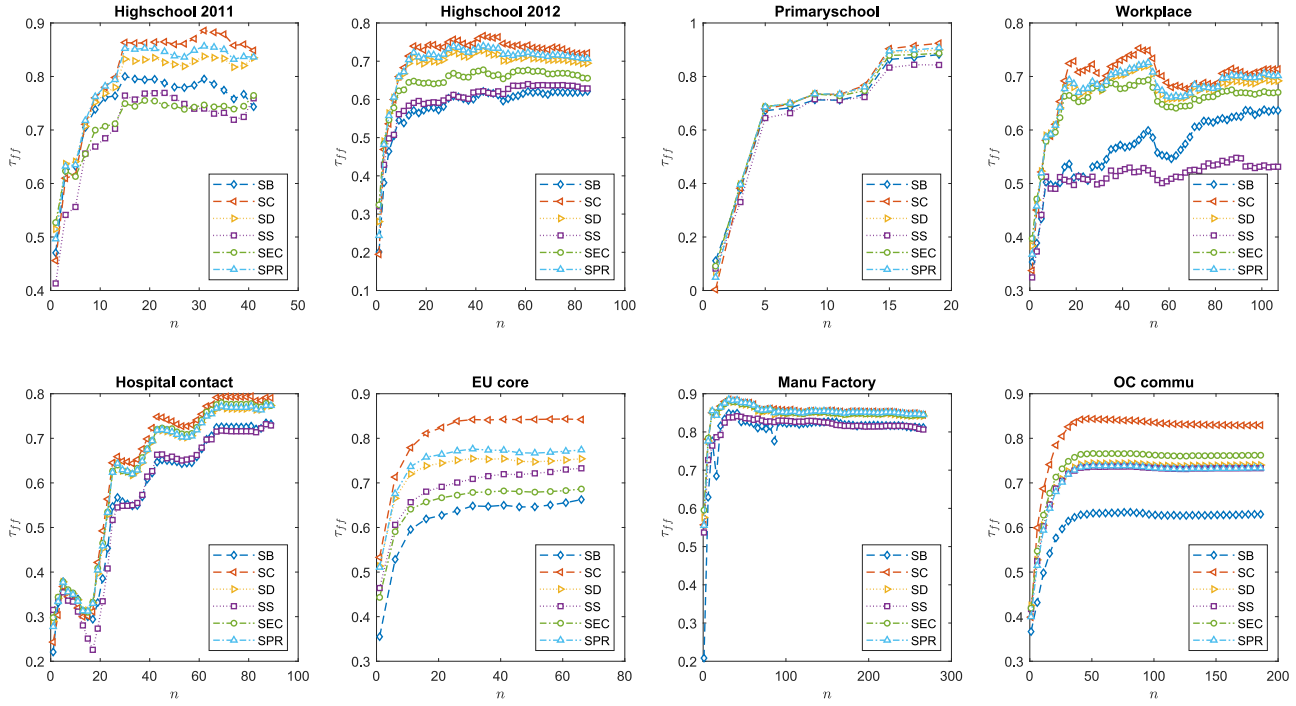

FIG. 3. The evolution of the Kendall ranking correlation coefficient τ_{ff} between *fad-tig*-score and NE_{fad} with the information gathering depth n increasing.

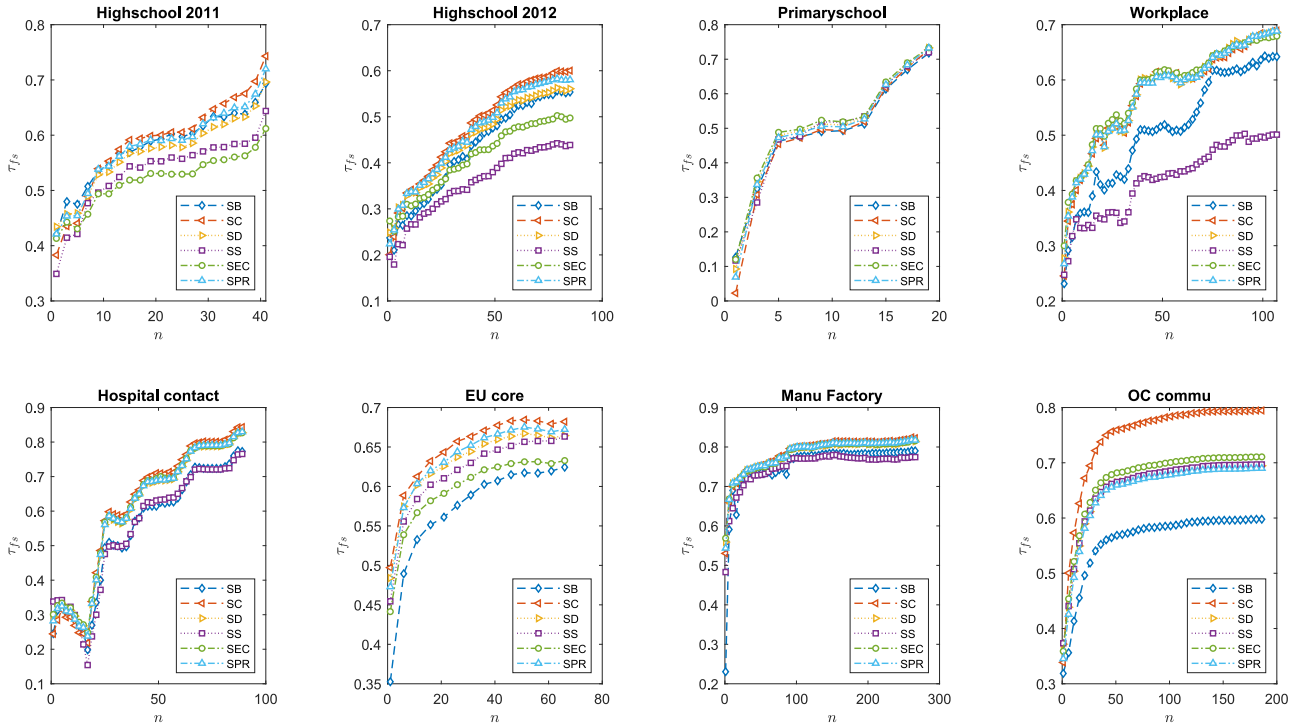

FIG. 4. The evolution of the Kendall ranking correlation coefficient τ_{fs} between *fad-tig*-score and NE_{std} with the information gathering depth n increasing.

the performance of *fad-tig* will be degraded if n is too large regarding NE_{fad}.

Now, we will check the performance of *std-tig*. From Figs. 5 and 6, we can see that the τ_{sf} and τ_{ss} decrease to a steady state quickly. Moreover, the optimal value of τ_{sf} is smaller than τ_{ff} and τ_{ss} is smaller than τ_{fs}. The phenomena might be due to the following two reasons.

In Appendix A, we plot the histogram of *FAD* and *STD*. Firstly, from Fig. 12, we know that most of the temporal distances are relatively small, and the coefficient of variation C_{std} (see in Table II) is small. When doing the *TIG*-process, the majority of nodes will be taken into account in the first few steps. This explanation can be confirmed by Fig. 6, the τ_{ss} shows a better performance in Eu-core and Oc commu networks, and the C_{std} of these two datasets is relatively higher than the others.

Another reason might be because of the difference in the amount of temporal information contained in the two types of temporal distance. Since the face-to-face contact networks are much denser than email communication networks, the Θ is quite similar to the adjacent matrix A of the static abstraction of the temporal networks, which means less temporal information contained in Θ compared with Φ.

B. Quantifying network efficiency

In this section, we will see the evolution of the network efficiency as the removing of top-ranked nodes. It is well known that the problem of influential maximization is *NP*-hard. Here, we treat NE_{fad} and NE_{std} as the best metrics in terms of $\mathcal{E}_{fad}$ and $\mathcal{E}_{std}$, respectively.

Figure 7 shows the changing of $\mathcal{E}_{fad}$ as the removal of top-ranked nodes. For each network, we remove at most 50% nodes. For each basic metric as the initial information, we choose the optimal gathering depth n, which is listed in Table III. Obviously, NE_{fad} gets the best performance and $\mathcal{E}_{fad}$ decreases most slowly when the nodes are randomly removed. What is more, for most *tig-scores* with different basic metrics as the initial information, the NE_{std} performs even worse. Simultaneously, Fig. 8 shows that the decreasing trend of $\mathcal{E}_{std}$ is similar for NE_{std} and NE_{fad}. Both NE_{std} and NE_{fad} work well, which further confirms our observation in Sec. V A. The *FAD* matrix, which is of much more temporal information, performs better in predicting important nodes in terms of the network efficiency.

C. Quantifying nodes' spreading influence

In this section, we will check the validation of our proposed process to quantify the *SIR* spreading influence. For *SIR* simulation, we set the infection rate β as 0.1 and recovery rate μ as 0.01.

As is mentioned in Sec. III, we have three different ways to measure the spreading influence for each node. The three ways are defined in terms of different situations. We cannot say which one is better to measure the spreading influence for temporal networks. In Table IV, we list the Pearson correlation coefficients of these three spreading measurements. We still use the Kendall ranking

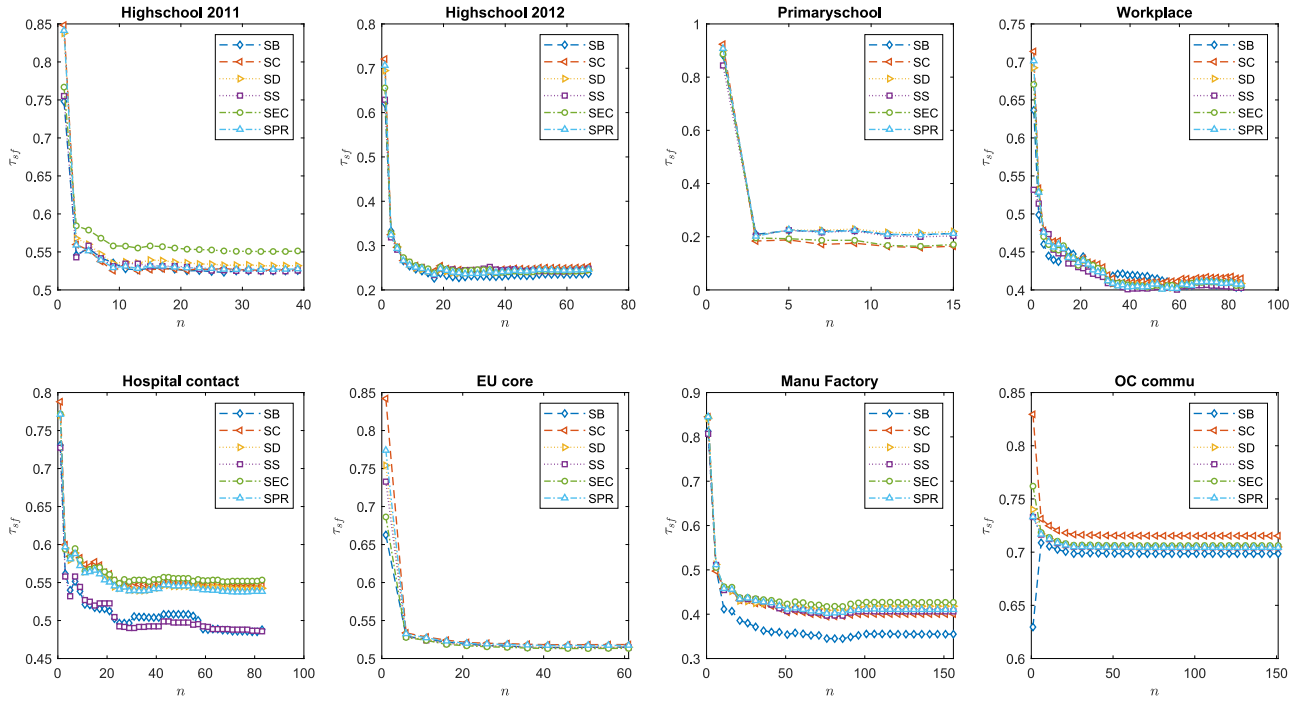

FIG. 5. The evolution of the Kendall ranking correlation coefficient τ_{sf} between *std-tig*-score and NE_{fad} with information gathering depth n increasing.

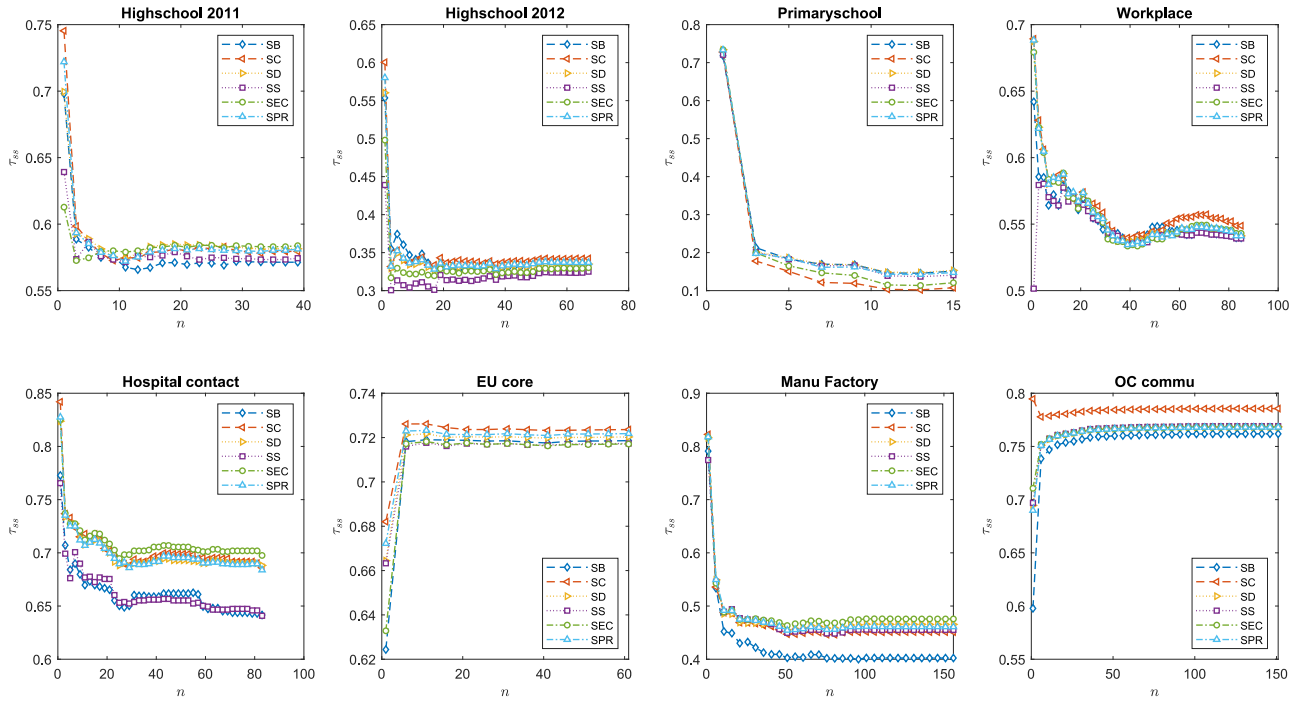

FIG. 6. The evolution of the Kendall ranking correlation coefficient τ_{ss} between *std-tig*-score and NE_{std} with information gathering depth n increasing.

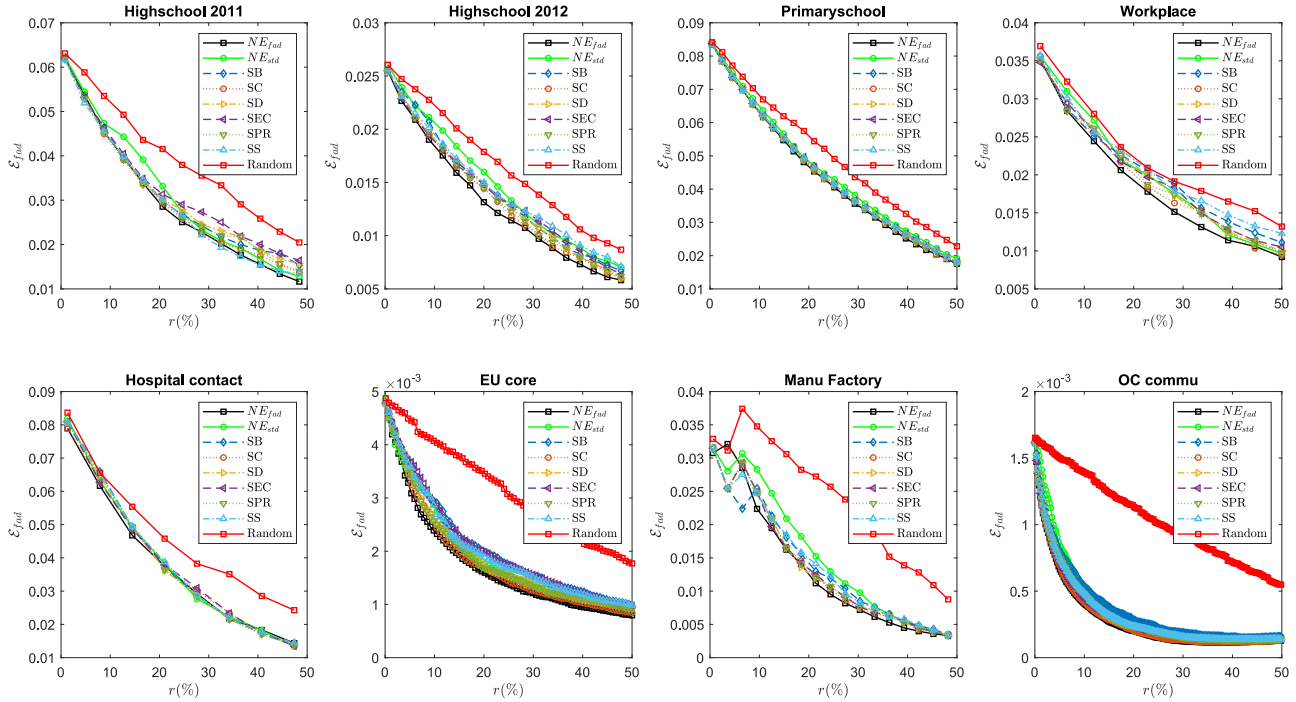

FIG. 7. The evolution of $\mathcal{E}_{fad}$ as the top-ranked nodes' removal.

correlation coefficient to evaluate the performance. The *tig*-score performs similarly in evaluating the three types of spreading influence. Thus, we show the result for R_{norm} here, and the others are listed in Appendix C.

The Kendall ranking correlation coefficient between *fad-tig* score and R_{norm} is denoted as τ_{fNorm}. For the *std-tig* process, the notations are defined in the same way. τ_{sNorm} denotes the Kendall coefficient between *std-tig* score and R_{norm}.

Note that the *tig*-score is not highly related to the spreading influence as that with the node efficiency, which means the *TIG*-process can predict the importance more effectively regarding the network efficiency. However, the overall trend is similar.

From Fig. 9, we find that the *fad-tig* score with the initial information of static strength performs the best compared with the other kinds of initial information and the one with static eigenvector centrality as initial information takes the second place. This might be because the static strength centrality is equivalent to temporal average degree centrality. In other words, the static strength centrality captures more temporal information than the others. At the same time, the *SIR* process is simulated step by step, time by time, which captures the most amount of temporal information as well.

TABLE III. The optimal gathering depth *n* for each *TIG*-process with different basic metric as initial information.

Network	SB	SC	SD	SEC	SPR	SS
High school 2011	15	32	34	42	34	23
High school 2012	44	43	43	43	43	61
Primary school	20	19	20	20	20	18
Workplace	92	48	48	51	51	90
Hospital contact	88	80	88	69	88	87
Eu core	68	54	32	68	32	65
Manu factory	34	31	31	31	31	39
OC communication	83	44	52	52	52	78

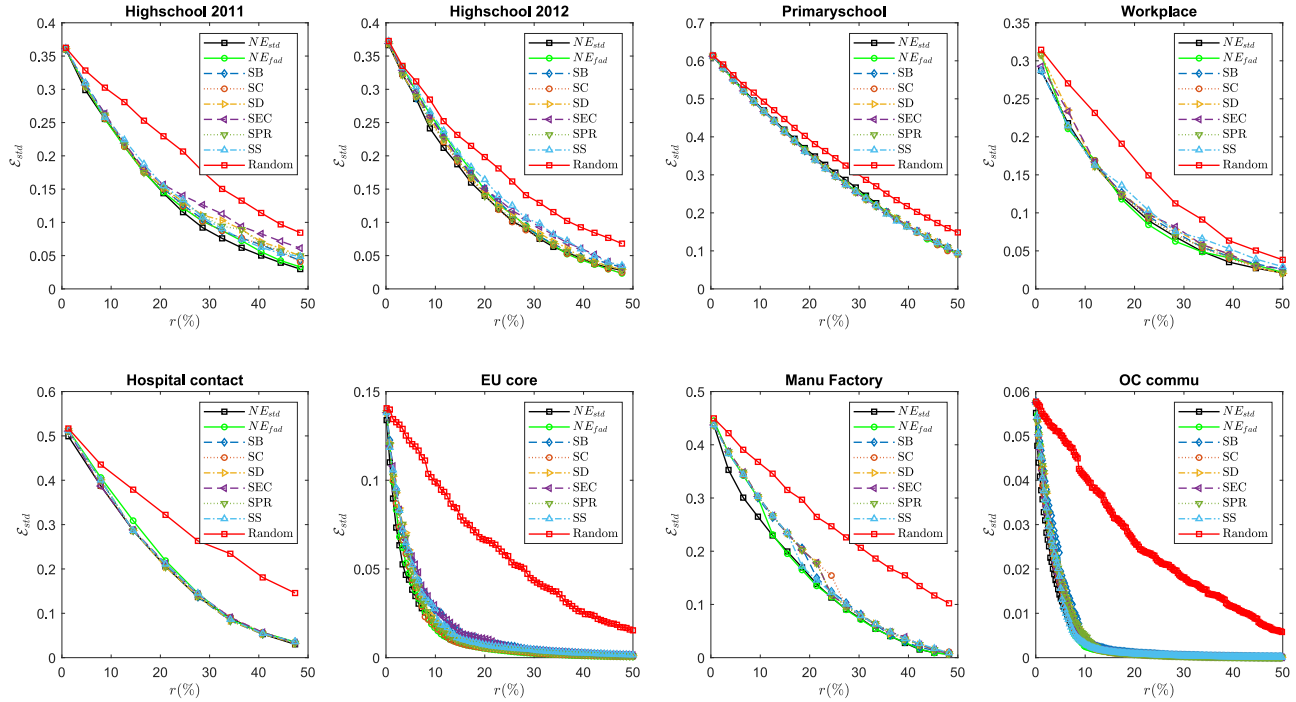

FIG. 8. The evolution of $\mathcal{E}_{std}$ as the top-ranked nodes' removal.

Finally, as we can see in Figs. 10–16, the *std-tig* process performs worse than *fad-tig*, regardless of the way to measure the spreading influence. As we discussed in Sec. V A, the *STD* matrix contains less temporal information than the *FAD* matrix. The *FAD* is defined by considering both the time proximity and path length between nodes. The assumption of the information gathering process is based on the fact that the importance of the nodes is related to their temporal neighbors, not only immediate neighbors but also higher-order neighbors. Therefore, when n is small, we are gathering information from neighboring nodes that are close to the current node both in time and the number of hop count. When n is large, neighboring nodes that are far away are also included. Therefore, the decrease of the performance when n is large implies that the neighbors that are far away from the current node have a small influence on its importance ranking.

VI. DISCUSSIONS

Even though many works have been done for the node ranking problem in static networks, there is still a lack of deep study for that in temporal networks. The evolution of the topology makes it impossible to use the static node ranking metrics in temporal networks.

In this study, we take the idea that node importance relies on the importance of its neighborhood, which has been verified by researchers.[9] We proposed a temporal information gathering (*TIG*) process to identify vital nodes in temporal networks. In the *TIG*-process, there are four parameters (n, f, D, c), in which n represents the information gathering depth, f is the weighting function that controls the influence of neighbors with different distances from the target node, D is a distance matrix, and c is the initial score. We show that by different combinations of these four variables, the *TIG*-process can degenerate to classic node ranking metrics, such as static degree, static closeness, temporal degree, and temporal closeness (Fig. 2).

We verify the performance of the *TIG*-process by using the performance evaluation methods, that is, the network efficiency based one and the *SIR* spreading based one, on real-world temporal networks. We observe that the fastest arrival distance based *TIG* process performs much better than the one based on the temporal shortest distance. In addition, there is an optimal gathering depth n, which makes the *FAD* based *TIG*-process perform optimally.

Actually, the main contribution of this paper is not to propose an exact metric to do the node ranking. In other words, as a node's importance metric, *TIG*-process can be used to rank the nodes for temporal networks. At the same time, as a framework, it can be used to explore the impact of temporal information on the significance of the nodes.

Through this framework, we observe that the *FAD*-based *tig*-process is more functional in predicting the significant nodes

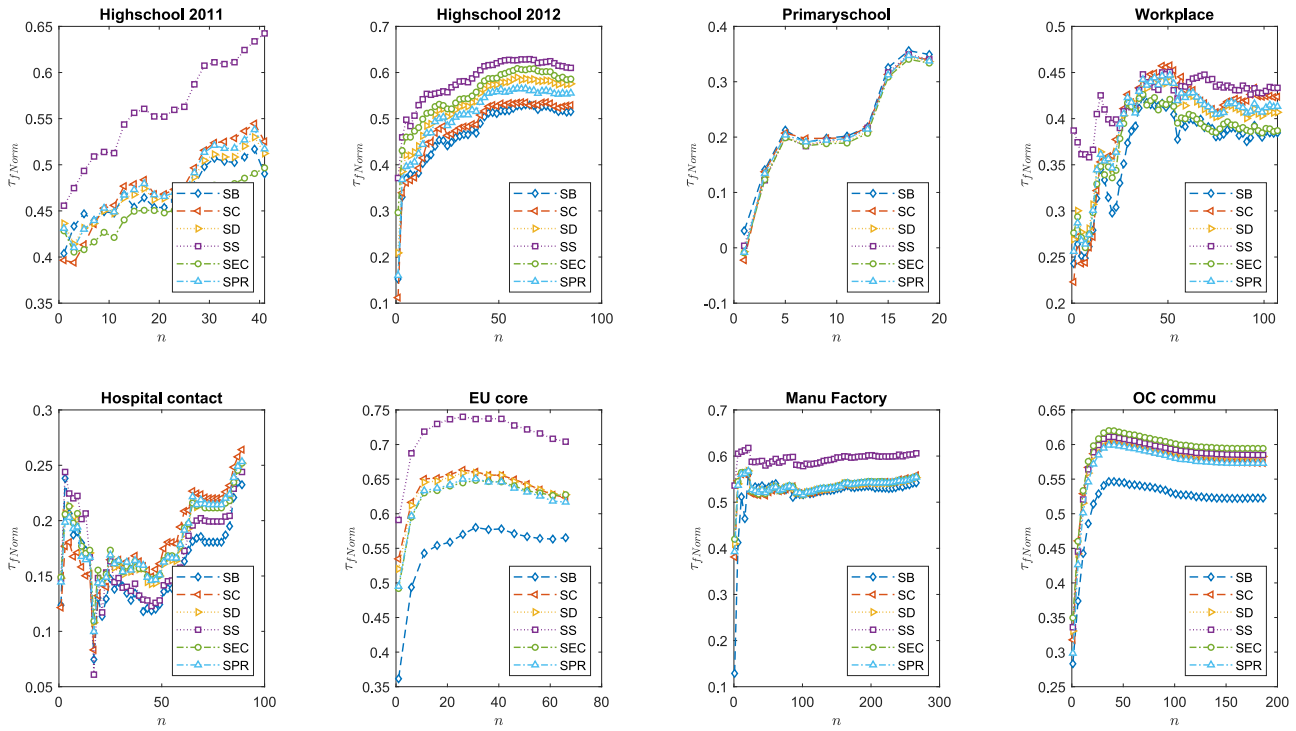

FIG. 9. The evolution of τ_{fNorm} between *fad-tig*-score and R_{norm} as the gathering depth n increases.

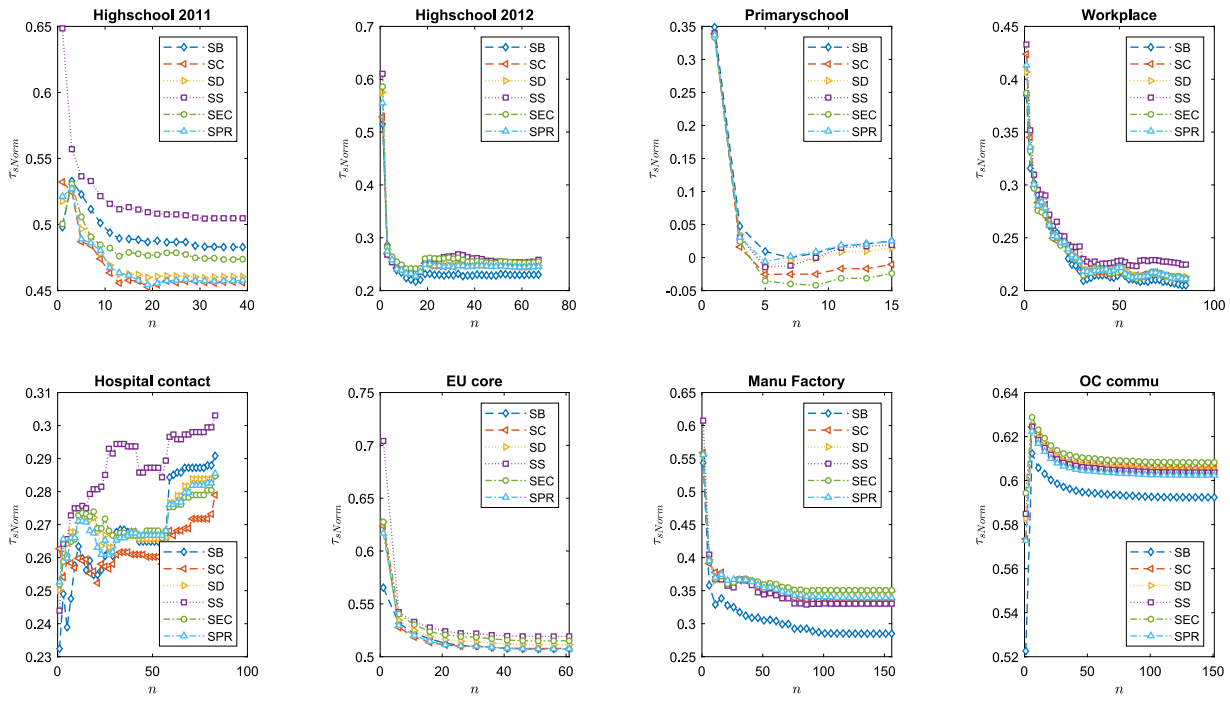

FIG. 10. The evolution of τ_{sNorm} between *std-tig*-score and R_{norm} as the gathering depth n increases.

compared with the *STD*-based one. Firstly, the *FAD* matrix captures more temporal information, which means it fits temporal networks better. Furthermore, from the definitions of these two kinds of distances, the former can be calculated from any time of the observation time window and the latter is more like a temporal metric but based on the final state of networks. There is no doubt that there might exist more suitable distance matrices that can be used in the *TIG*-process.

This work opens new challenging questions like, if we consider the distance in static networks as a physical or spatial distance and the distance in temporal networks as a temporal distance, then which one is more significant in measuring nodes' influence? In addition, in Fig. 2 and Table I, an *Iterative TIG process* was introduced, which means we gather the updated *tig-score* instead of the initial information at each step in the *TIG*-process. This metric will be discussed in future works. Moreover, for the datasets used in this work, we cannot get the true initial information. With the rapid increase in the amount of data, our proposed *TIG*-process can be further explored.

ACKNOWLEDGMENTS

The authors would like to thank the National Natural Science Foundation of China (NNSFC) (Grant Nos. 11601430, 11631014, and 11871311) for support. Z.K.Z. was supported by the National Science Foundation of China (NSFC) (Grant No. 61673151) and ZJNSF (Grant No. LR18A050001).

APPENDIX A: THE DISTRIBUTION OF FAD AND STD

In this section, we give the histograms of the fastest arrival distance and the temporal shortest arrival distance.

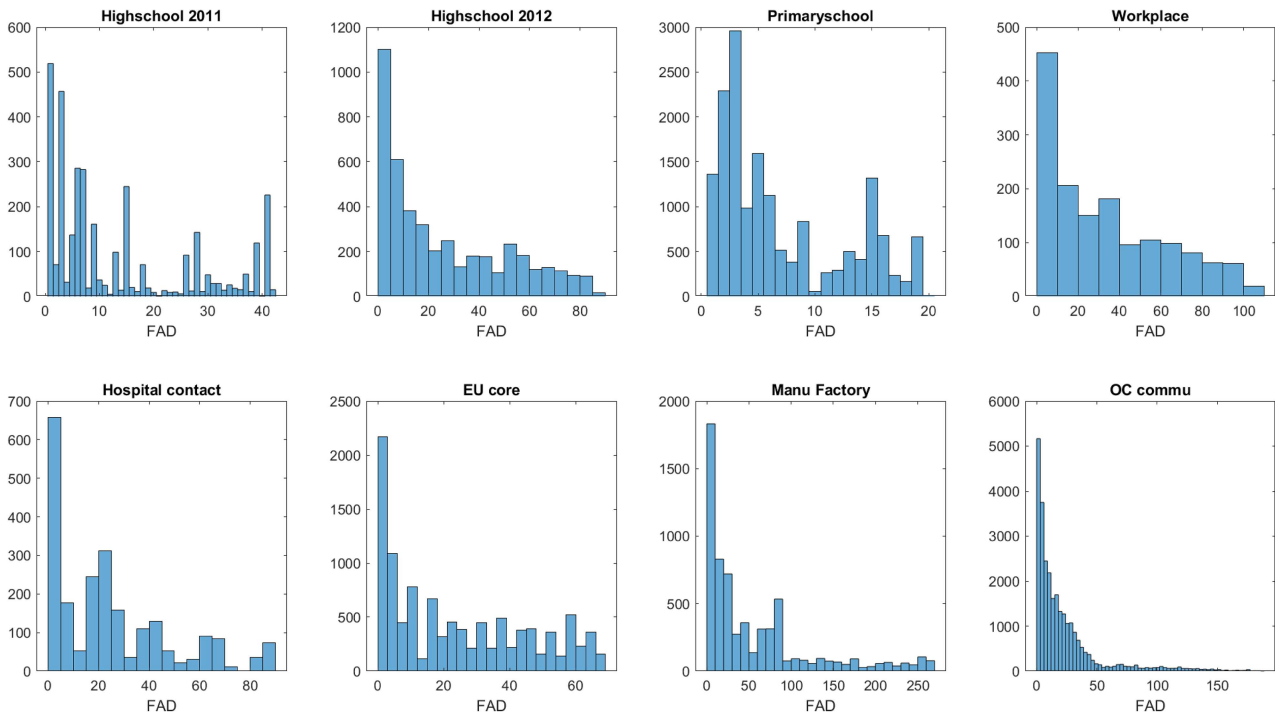

FIG. 11. Distribution of *FAD*.

APPENDIX B: THE RELATIONSHIP BETWEEN DIFFERENT SPREADING MEASUREMENTS

Note that for some datasets, the three measurements are highly correlated. However, for some datasets, they are quite different from each other.

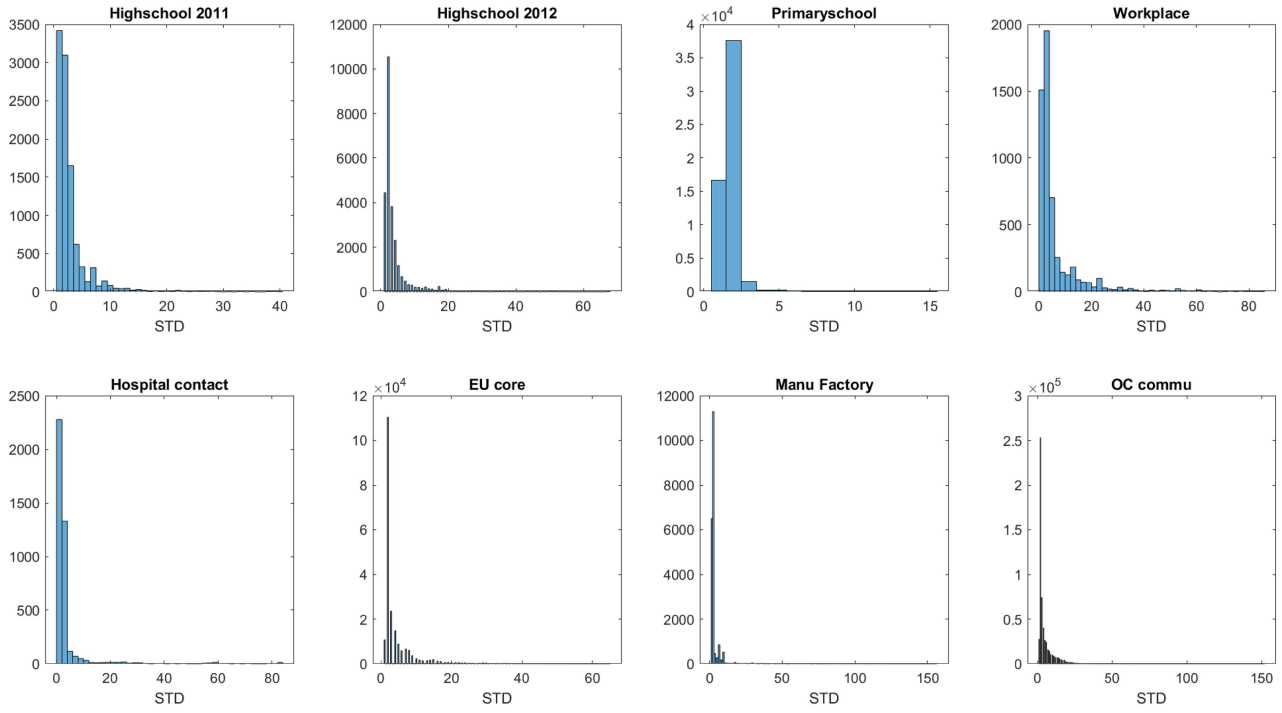

FIG. 12. Distribution of *STD*.

TABLE IV. The Pearson correlation coefficient between different spreading measurements.

Network	R_{mean} *vs* R_{max}	R_{max} *vs* R_{norm}	R_{norm} *vs* R_{mean}
High school 2011	0.853 2	0.698 8	0.732 2
High school 2012	0.692 8	0.624 2	0.775 0
Primary school	0.536 9	0.177 2	0.612 4
Workplace	0.800 9	0.770 1	0.845 0
Hospital contact	0.866 0	−0.471 3	−0.525 3
Eu core	0.954 3	0.933 9	0.990 7
Manu factory	0.955 4	0.806 9	0.864 2
OC communication	0.910 5	0.921 3	0.994 9

APPENDIX C: THE RESULTS FOR R_{max} AND R_{mean}

τ_{fMax} and τ_{fMean} indicate the Kendall coefficient between *fad-tig* score and R_{max} and R_{mean}, respectively. τ_{sMax} and τ_{sMean} denote the Kendall coefficient between *std-tig* score and R_{max} and R_{mean}, respectively.

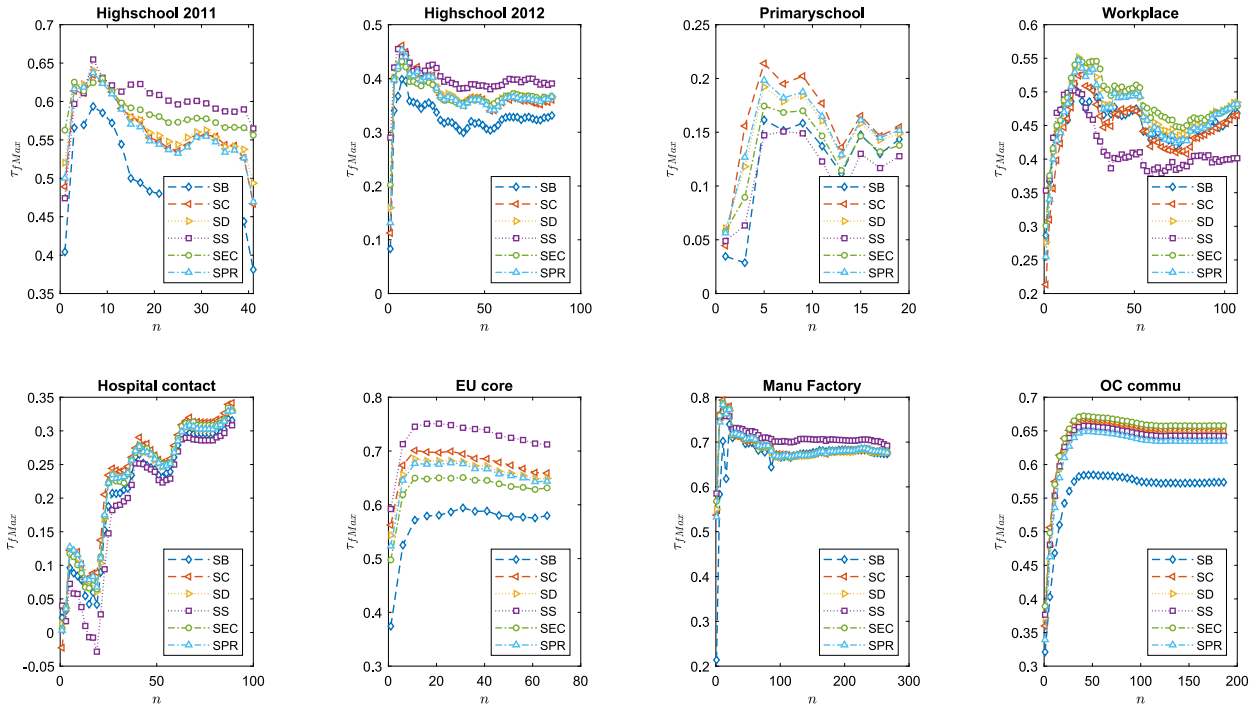

FIG. 13. The evolution of τ_{fMax} between *fad-tig*-score and R_{max} as the gathering depth n increases.

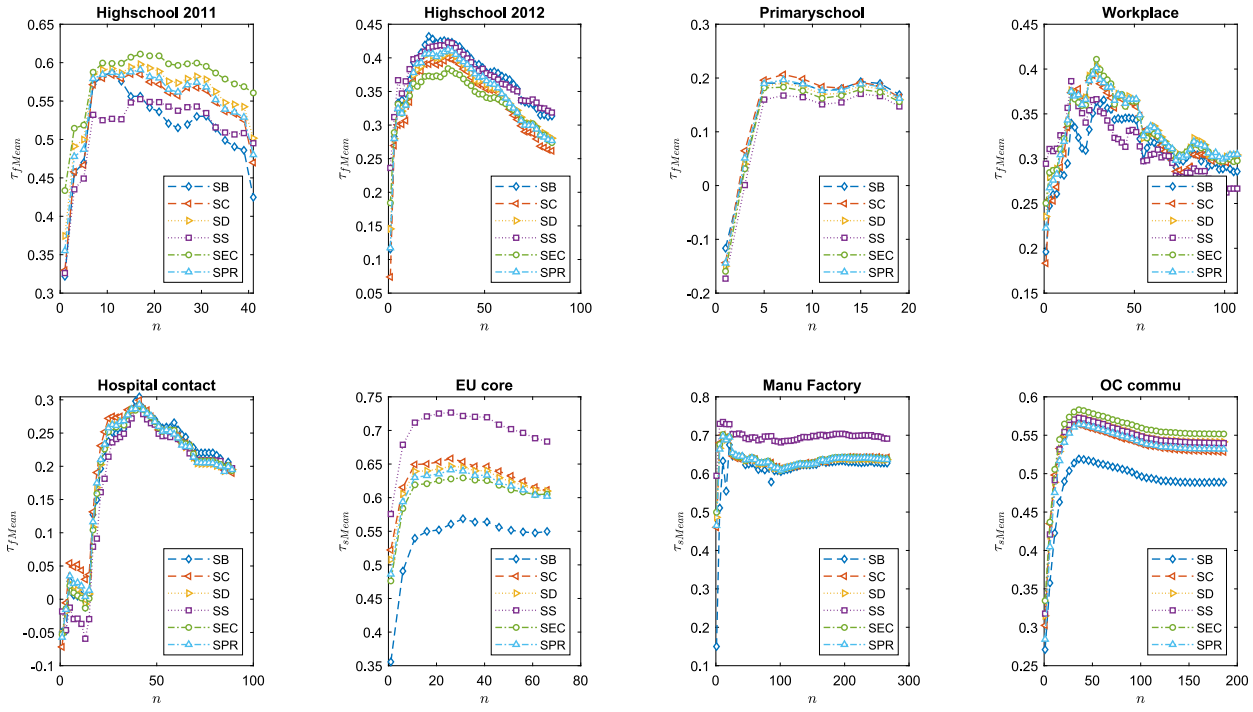

FIG. 14. The evolution of τ_{fMean} between *fad-tig*-score and R_{mean} as the gathering depth n increases.

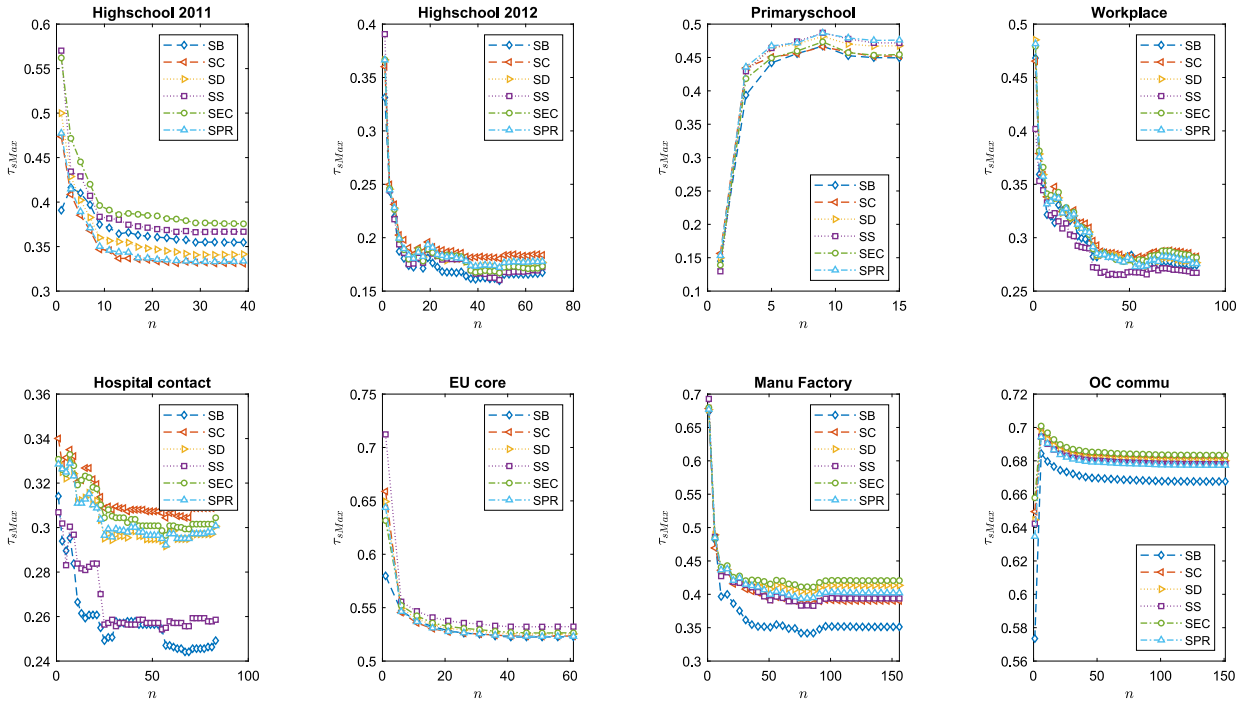

FIG. 15. The evolution of τ_{sMax} between *std-tig*-score and R_{max} as the gathering depth n increases.

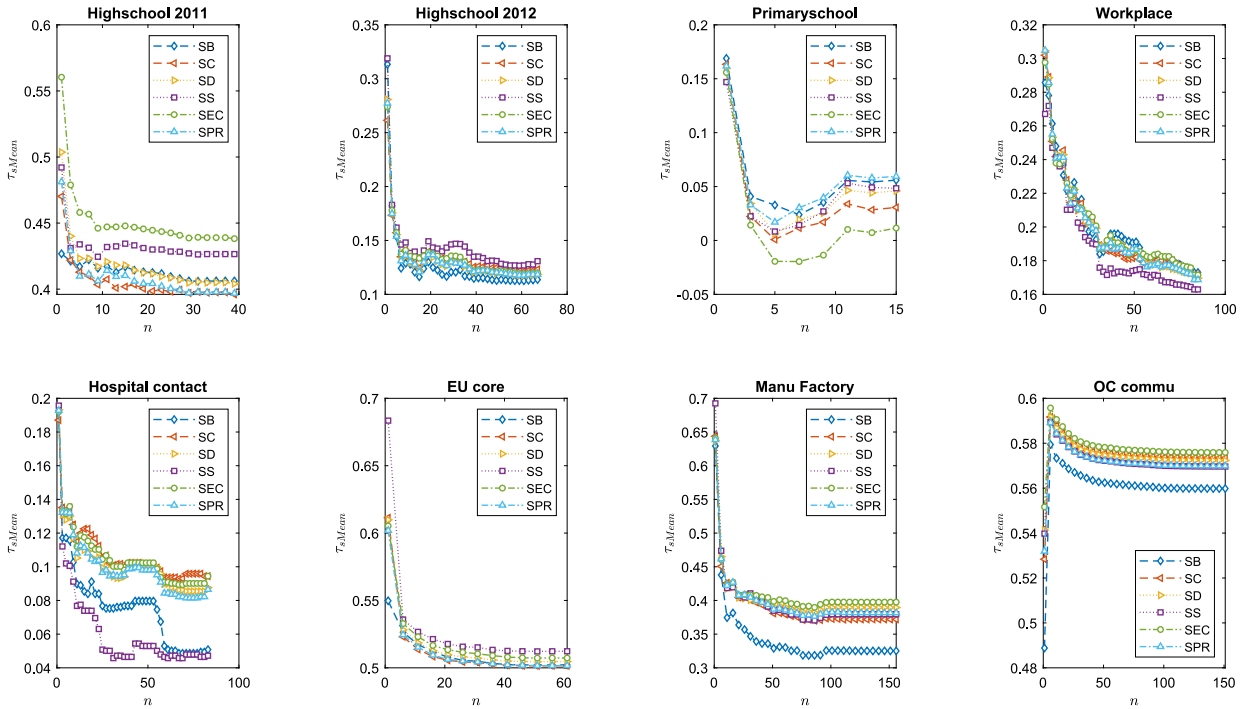

FIG. 16. The evolution of τ_{sMean} between *std-tig*-score and R_{mean} as the gathering depth n increases.

REFERENCES

[1] F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," Nature **524**, 65 (2015).
[2] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2009), pp. 807–816.
[3] X. Zhang, J. Zhu, Q. Wang, and H. Zhao, "Identifying influential nodes in complex networks with community structure," Knowl. Based Syst. **42**, 74–84 (2013).
[4] X.-X. Zhan, C. Liu, G. Zhou, Z.-K. Zhang, G.-Q. Sun, J. J. Zhu, and Z. Jin, "Coupling dynamics of epidemic spreading and information diffusion on complex networks," Appl. Math. Comput. **332**, 437–448 (2018).
[5] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, "Vital nodes identification in complex networks," Phys. Rep. **650**, 1–63 (2016).
[6] Z. Zhu, "Discovering the influential users oriented to viral marketing based on online social networks," Phys. A Stat. Mech. Appl. **392**, 3459–3469 (2013).
[7] N. Zhou, X.-X. Zhan, Q. Ma, S. Lin, J. Zhang, and Z.-K. Zhang, "Identifying spreading sources and influential nodes of hot events on social networks," in *International Workshop on Complex Networks and Their Applications* (Springer, 2017), pp. 946–954.
[8] A. E. Motter, "Cascade control and defense in complex networks," Phys. Rev. Lett. **93**, 098701 (2004).
[9] S. Xu, P. Wang, and J. Lü, "Iterative neighbour-information gathering for ranking nodes in complex networks," Sci. Rep. **7**, 41321 (2017).
[10] J. G. Restrepo, E. Ott, and B. R. Hunt, "Characterizing the dynamical importance of network nodes and links," Phys. Rev. Lett. **97**, 094102 (2006).
[11] C. M. Taniguchi, B. Emanuelli, and C. R. Kahn, "Critical nodes in signalling pathways: Insights into insulin action," Nat. Rev. Mol. Cell Biol. **7**, 85 (2006).
[12] A. Landherr, B. Friedl, and J. Heidemann, "A critical review of centrality measures in social networks," Bus. Inf. Syst. Eng. **2**, 371–385 (2010).
[13] L. Katz, "A new status index derived from sociometric analysis," Psychometrika **18**, 39–43 (1953).
[14] P. Holme, and J. Saramäki, "Temporal networks," Phys. Rep. **519**, 97–125 (2012).
[15] F. Kuhn and R. Oshman, "Dynamic networks: Models and algorithms," ACM SIGACT News **42**, 82–96 (2011).
[16] D. G. Rand, S. Arbesman, and N. A. Christakis, "Dynamic social networks promote cooperation in experiments with humans," Proc. Natl. Acad. Sci. U.S.A. **108**, 19193–19198 (2011).
[17] M. G. Zimmermann, V. M. Eguíluz, and M. San Miguel, "Coevolution of dynamical states and interactions in dynamic networks," Phys. Rev. E **69**, 065102 (2004).
[18] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," Proc. Natl. Acad. Sci. U.S.A. **106**, 21544–21549 (2009).
[19] X.-X. Zhan, A. Hanjalic, and H. Wang, "Information diffusion backbones in temporal networks," preprint arXiv:1804.09483 (2018).
[20] H. Kim and R. Anderson, "Temporal node centrality in complex networks," Phys. Rev. E **85**, 026107 (2012).
[21] R. K. Pan and J. Saramäki, "Path lengths, correlations, and centrality in temporal networks," Phys. Rev. E **84**, 016105 (2011).
[22] T. Takaguchi, N. Sato, K. Yano, and N. Masuda, "Importance of individual events in temporal networks," New J. Phys. **14**, 093003 (2012).
[23] L. Lü, T. Zhou, Q.-M. Zhang, and H. E. Stanley, "The h-index of a network node and its relation to degree and coreness," Nat. Commun. **7**, 10168 (2016).
[24] L. C. Freeman, "A set of measures of centrality based on betweenness," Sociometry **40**, 35–41 (1977).
[25] G. Sabidussi, "The centrality index of a graph," Psychometrika **31**, 581–603 (1966).
[26] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, "A model of internet topology using k-shell decomposition," Proc. Natl. Acad. Sci. U.S.A. **104**, 11150–11154 (2007).
[27] H. Wu, J. Cheng, S. Huang, Y. Ke, Y. Lu, and Y. Xu, "Path problems in temporal graphs," Proc. VLDB Endowment **7**, 721–732 (2014).
[28] M. G. Kendall, "A new measure of rank correlation," Biometrika **30**, 81–93 (1938).
[29] A. E. Sizemore and D. S. Bassett, "Dynamic graph metrics: Tutorial, toolbox, and tale," NeuroImage **180**, 417–427 (2017).
[30] M. E. Newman, "Mathematics of networks," *The New Palgrave Dictionary of Economics* (Springer, 2016), pp. 1–8.
[31] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," Comput. Netw. ISDN Syst. **30**, 107–117 (1998).
[32] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," Phys. Rev. Lett. **87**, 198701 (2001).
[33] Z.-K. Zhang, C. Liu, X.-X. Zhan, X. Lu, C.-X. Zhang, and Y.-C. Zhang, "Dynamics of information diffusion and its applications on complex networks," Phys. Rep. **651**, 1–34 (2016).
[34] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili, "Theory of rumour spreading in complex social networks," Phys. A Stat. Mech. Appl. **374**, 457–470 (2007).
[35] A. Saumell-Mendiola, M. Á. Serrano, and M. Boguná, "Epidemic spreading on interconnected networks," Phys. Rev. E **86**, 026106 (2012).
[36] R. Parshani, S. Carmi, and S. Havlin, "Epidemic threshold for the susceptible-infectious-susceptible model on random networks," Phys. Rev. Lett. **104**, 258701 (2010).
[37] J. Fournet and A. Barrat, "Contact patterns among high school students," PLoS ONE **9**, e107878 (2014).
[38] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina *et al.*, "High-resolution measurements of face-to-face contact patterns in a primary school," PLoS ONE **6**, e23176 (2011).
[39] P. Vanhems, A. Barrat, C. Cattuto, J.-F. Pinton, N. Khanafer, C. Régis, B.-a. Kim, B. Comte, and N. Voirin, "Estimating potential infection transmission routes in hospital wards using wearable proximity sensors," PLoS ONE **8**, e73970 (2013).
[40] M. Génois, C. L. Vestergaard, J. Fournet, A. Panisson, I. Bonmarin, and A. Barrat, "Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers," Netw. Sci. **3**, 326–347 (2015).
[41] A. Paranjape, A. R. Benson, and J. Leskovec, "Motifs in temporal networks," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (ACM, 2017), pp. 601–610.
[42] R. Michalski, S. Palus, and P. Kazienko, "Matching organizational structure and social network extracted from email communication," in *International Conference on Business Information Systems* (Springer, 2011), pp. 197–206.
[43] P. Panzarasa, T. Opsahl, and K. M. Carley, "Patterns and dynamics of users' behavior and interaction: Network analysis of an online community," J. Am. Soc. Inf. Sci. Technol. **60**, 911–932 (2009).