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
Accepted author manuscript

Published in
Chaos

Citation (APA)

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.
Temporal information gathering process for node ranking in time-varying networks

Cunquan Qu\textsuperscript{a}, Xiuxiu Zhan\textsuperscript{b}, Guanghui Wang\textsuperscript{a}, Jianliang Wu\textsuperscript{a}, Zike Zhang\textsuperscript{c,d}

\textsuperscript{a}Shandong University, School of Mathematics, Jinan, 250110, P.R.China
\textsuperscript{b}Delft University of Technology, Intelligent Systems, Delft, 2600GA, the Netherlands
\textsuperscript{c}Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou, 311121, P.R.China
\textsuperscript{d}Institute of Automation, Shanghai Jiaotong University, Shanghai, 200030, P.R.China

Abstract

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 vital node identification in static networks, there is a lack of research in temporal systems. And how the temporal information influence node ranking is still unknown. In this work, we propose a temporal information gathering (TIG) process for temporal networks. TIG-process, as a node’s importance metric, can be used to do the node ranking. As a framework, TIG-process can be applied to explore the impact of temporal information on nodes’ significance. The key point of the TIG-process is that node’s importance relies on the importance of its neighborhood. There are four parameters, including the temporal information gathering depth $n$, temporal distance matrix $D$, initial information $c$ and weighting function $f$. We observe that TIG-process can degenerate to classic metrics both from static and temporal networks by proper combination of these four parameters. In addition, fastest arrival distance based TIG-process ($fad$-tig) performs much better in quantifying nodes’ efficiency and nodes’ spreading influence than the one based on temporal shortest distance. For the $fad$-tig process, we can find an optimal gathering depth $n$ which makes the TIG-process perform the best when $n$ is

\*Corresponding author

Email address: ghwang@sdu.edu.cn (Guanghui Wang)
1. Introduction

Vital nodes identification has attracted increasing attention lately due to its great significance as well as valuable applications [1, 2, 3]. As a matter of fact, a small amount of influential nodes can affect mechanisms like cascading, spreading and synchronizing in complex systems[4]. In the view of application, finding vital nodes can help to optimize product information diffusion in viral marketing [5], to control the spread of rumors [6], to prevent a catastrophic outage in power grids or the Internet[7], etc.

Many metrics have been defined in finding important nodes in static networks and most of them show good performance [8, 9, 10]. However, in the real world, there are many systems in which the topology of their corresponding networks varies over time. We refer to such networks as temporal networks or time-varying networks[11, 12, 13, 14, 15]. The study of identifying vital nodes in temporal networks can be more challenge than that of static networks, as we need to consider the time dimension of the nodes’ influence. There are some pioneering researches concentrated on ranking nodes in temporal networks [16, 17, 18]. 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 [16, 17]. 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 in [16] belong to this class of methods. Even though these methods may gain some improvement in finding vital nodes compared to static metrics in temporal networks, cutting the temporal networks into slides and taking the average value of all the slides may loss much temporal information. Therefore, it is necessary to define more reasonable node ranking metrics, which can describe the evolution of the nodes’ influence.

Node ranking metrics using local information of the nodes (e.g., degree and H-index) have shown good performance in identifying important nodes [8, 19]. On the other hand, 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 [20], closeness [21] and k-core centrality [22] are designed to capture the global information.

In this paper, we propose temporal information gathering based process on the context that each node is attributed an initial information, since for example, when a person first joined a new group, it has its own attribute. After communicating with other members, her/his importance is changing and can be estimated by her/his colleagues(neighborhood). For simplify, we denote the neighborhood of node \( v_i \) as \( N \leq l(i) \), which indicates the nodes with a temporal distance less than or equal to \( l \). Throughout the paper, we use \( TIG \)-process to denote the Temporal Information Gathering process and \( tig \)-score represents node importance obtained from \( TIG \)-process. \( TIG \)-process is controlled by four parameters, 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 the following sections, 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 temporal shortest distance. And for the former one, we can get an optimal gathering depth \( n \), regardless of the initial information[18], including static degree, static closeness, static strength, static betweenness, eigenvector centrality and PageRank centrality[16]. As the depth \( n \) increases, the performance will be degraded. In addition, many basic metrics can be derived from \( TIG \)-process by proper combinations of the four parameters mentioned above.

The rest of the paper is organized as follows. In Section 2, we give the definition of the \( TIG \)-process. We describe the benchmark metrics and two evaluation methods in Section 3. The datasets used in this paper are given in Section 4 and the results are shown in Section 5. We discuss and conclude in Section 6.

2. Temporal Information Gathering Process

In this section, we give a detailed illustration of the \( TIG \)-process.

2.1. 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 $(u, v, t_0, \lambda)$, where $u, v \in V$, $t_0$ is the start time of the event, $\lambda$ is the lasting time, and $t_0 + \lambda$ is the ending time. In this paper, we assume $\lambda = 0$, which means we only consider the 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 $t$. Additionally, the unweighted integrated static network of $G^T$ is expressed as $G = (V, E)$, where $E$ is the static edge set at the final time $T$. The adjacent matrix of $G$ is denoted as $A$ and the distance matrix is $M$. The entry $M(i, j)$ indicates the distance between the two corresponding nodes $v_i$ and $v_j$.

A temporal path in the temporal network $G^T$ is a sequence of nodes $P = \langle v_1, v_2, \cdots, v_k, v_{k+1} \rangle$, where $(v_i, v_{i+1}, t_i) \in E^T$ is the $i$-th event on $P$ for $1 \leq i \leq k$. Then the start time of $P$ is $t_{\text{start}}(P) = t_1$ and the end time of $P$ is $t_{\text{end}}(P) = t_k$. We define the temporal length of $P$ as $l(P) = t_{\text{end}}(P) - t_{\text{start}}(P) + 1$. Given a time period $[t_\alpha, t_\omega]$, let $P(x, y, [t_\alpha, t_\omega]) = \{ P : P \text{ is a temporal path from } x \text{ to } y \text{ such that } t_{\text{start}}(P) > t_\alpha \text{ and } t_{\text{end}}(P) < t_\omega \}$. 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 [23]. In this paper, we introduce two distance definitions for temporal networks, i.e., the fastest arrival distance and temporal shortest distance.

Fastest arrival path: The fastest arrival path between node $x$ and $y$ is the path that goes from $x$ to $y$ taking the minimum elapsed time counted from $t = 1$. That is to say, $P \in P(x, y, [t_\alpha, t_\omega])$ is a fastest arrival path if $t_{\text{end}}(P) = \min\{t_{\text{end}}(P') : P' \in P(x, y, [t_\alpha, t_\omega])\}$. And the fastest arrival distance between node $x$ and node $y$ is measured by the length of the fastest arrival path between them, denoted as $\phi(x, y)$. An example of the fastest arrival path is shown in Fig. 1(B) from the toy temporal network given in Fig. 1(A). The fastest arrival distance between node $v_1$ and node $v_4$ is 3.

Temporal shortest path: The temporal shortest path from $x$ to $y$ is path for which the overall traversal time needed is shortest. Therefore, $P \in P(x, y, [t_\alpha, t_\omega])$ is a temporal shortest path if $l(P) = \min\{l(P') : P' \in P(x, y, [t_\alpha, t_\omega])\}$. The temporal shortest distance between node $x$ and node $y$ is the length of the temporal shortest path between them, denoted as $\theta(x, y)$. Figure 1(C) shows the temporal shortest path between $v_1$ and $v_4$, and $\theta(v_1, v_4) = 2$.
Figure 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 node \(v_1\) and \(v_4\). (C) The shortest temporal path between node \(v_1\) and \(v_4\).

**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 \(\Phi\) and the temporal shortest distance matrix \(\Theta\).

**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.
2.2. The TIG-process

Recall that the temporal information gathering process is denoted by TIG-process for simplification. The ranking score of node \( v_i \) obtained from TIG-process is defined as \( tig \)-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 in the distance matrix, i.e., \( g^{(1)} = D_1g^{(0)} \). Similarly, the \( n \)th-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., \( N_{\leq n}(i) \). Thus, the \( n \)th-order TIG-process is written as

\[
g^{(n)} = \sum_{j=0}^{n} f_j \cdot D_j \cdot g^{(0)},
\]  

where \( f_j \) is the weighting function for \( j \)th-order neighbors. The \( n \)th-order \( tig \)-score is denoted by \( g^{(n)}_i \). We use \( g^{(n)}_i \) to indicate the ranking score of node \( v_i \). Obviously, a larger value of \( g^{(n)}_i \) implies node \( v_i \) is more important in the network.

From Eq. 2, we know that the TIG-process can be denoted as a quadruple \( (n, f, D, c) \) and these four parameters are independent of each other. For example, the parameter \( n \) controls the information gathering depth, which varies from 1 to \( T \). The weighting function \( f \) which weights the distance effect on the nodes importance can also take different formations. 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 state 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 parameters. We show in Figure 2 and Table 1 the relationship between TIG-process and some classic metrics which will be described in the next section.

3. Methods

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

3.1. Benchmark metrics

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

\[
SD(i) = \sum_j A(i,j).
\]

(3)

Static strength centrality (SS) of node \( v_i \) counts the number of occurrence of each node appeared in the temporal network.

\[
SS(i) = \sum_{t=1}^{T} \sum_j A_t(i,j).
\]

(4)

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

\[
SB(i) = \sum_{h \neq i \neq j} \frac{\sigma_{hj}(i)}{\sigma_{hj}}.
\]

(5)
where $\sigma_{hj}$ 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

$$SC(i) = \frac{N - 1}{\sum_{v_j \in V \setminus v_i} d_{ij}}, \quad (6)$$

where $d_{ij}$ 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)**[24] 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

$$FAC(i) = \frac{N - 1}{\sum_{v_j \in V \setminus v_i} \phi(i, j)}, \quad (7)$$

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

$$STC(i) = \frac{N - 1}{\sum_{v_j \in V \setminus v_i} \theta(i, j)}, \quad (8)$$

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

Table 1: The detailed combination of the parameters in TIG-process in order to get the classic metrics.
3.2. Network efficiency

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

\[
E(G) = \frac{1}{|V|(|V| - 1)} \sum_{v_i \neq v_j \in G} \frac{1}{M(v_i, v_j)},
\]

(9)

where \(M\) 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 \(M\) with some temporal distance matrices. We use the fastest arrival distance matrix \(\Phi\) or the temporal shortest distance matrix \(\Theta\) instead of \(M\) in Eq.(9) to define the efficiency \(E_{fad}\) or \(E_{std}\) 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 \(NE(V') = E(G) - E(G\setminus V')\). For each node \(v_i\) in a network, we define the node efficiency as \(NE(i) = E(G) - E(G\setminus v_i)\). 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 node efficiency and TIG-score with different initial information. The Kendall correlation coefficient \(\tau[26, 27]\) 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

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

(10)

If \(\tau\) takes the value of +1, then the agreement of the two rankings is perfect. If \(\tau\) is −1, then one list is the reverse of the other. If \(\tau\) is close to zero,
then the two rankings are independent. Therefore, the higher $\tau$ indicates the better the node ranking metric used to predict important nodes in terms of network efficiency.

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

3.3. Spreading influence

Another performance evaluation method for node ranking is based on the spreading process [28, 29, 30, 31]. 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 $\beta$, and each infected node can recover from disease with probability $\mu$. 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. Firstly, we simulate the SIR spreading by following the time order of the interactions. And 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 $\mathcal{R}(v_i) = \{(t^j_{v_i}, R^j_{v_i}) \mid t^j_{v_i} \text{ is the occurrence time of node } v_i\}$, 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.
4. Datasets

Eight real-world networks are studied in this paper, 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 2.

- High school 2011(2012) dynamic contact network[32]. The dataset records the contacts between students in a high school in Marseilles, France.

- Primary school temporal network[33]. The dataset contains the temporal network of contacts between the children and teachers in a primary school.

- Hospital ward dynamic contact network[34]. The dataset contains the temporal network of contacts between patients, patients and healthcare workers (HCWs) and among HCWs in a hospital ward in Lyon, France.

- Contact network in a workplace[35]. The dataset contains the temporal network of contacts between individuals in an office building.

- Email-Eu-core temporal network[36]. The network is generated using email data from a large European research institution.

- Manufacturing emails[37]. This network is the internal email communication network between employees of a mid-size manufacturing company.

- CollegeMsg temporal network[38]. This network is comprised of private messages sent on an online social network at the University of California, Irvine.

5. Results

For the experiments in this paper, we take the weighting function \( f \) as 1, which means for each node \( v_i \), we treat all the nodes in \( N_{\leq n}(i) \) equally.
The fastest arrival distance matrix ($\Phi$) and temporal shortest distance matrix ($\Theta$) are considered as the temporal distance matrix $D$ respectively. And we call these two kinds of TIG-process as FAD-based $tig$-process and STD-based $tig$-process, denoted as $fad$-$tig$ and $std$-$tig$, for 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).

5.1. Quantifying node efficiency

Recall that in Section 3, we introduced the definition of 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 $fad$-$tig$, $NE_{std}$ and $std$-$tig$ are denoted as $\tau_{ff}$, $\tau_{fs}$, $\tau_{sf}$ and $\tau_{ss}$ respectively.

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

The case of using $fad$-$tig$ to estimate $NE_{std}$ is similar with the one of using $fad$-$tig$ to estimate $NE_{fad}$. But, in Fig. 4, we can see the $\tau_{fs}$ is increasing as $n$ increases in general. Dissimilar with Fig. 4, the $\tau_{ff}$ decrease or keep

| Network          | $N$ | $T$ | $|E|$  | $C_{fad}$ | $C_{std}$ |
|------------------|-----|-----|-------|-----------|-----------|
| High School2011  | 126 | 42  | 28,561| 0.5798    | 0.3405    |
| High School2012  | 180 | 87  | 45,047| 0.6196    | 0.3664    |
| Primary School   | 242 | 20  | 125,773| 0.5288    | 0.1188    |
| Workplace        | 92  | 108 | 9,827 | 0.6191    | 0.4102    |
| Hospital Contact | 75  | 90  | 32,424| 0.8411    | 0.7956    |
| Eu core          | 771 | 68  | 38,328| 1.2913    | 0.6522    |
| Manu factory     | 167 | 268 | 82,927| 0.9081    | 0.6629    |
| OC communication | 1898| 188 | 61,726| 2.4579    | 1.0645    |

Table 2: Basic features of the real-world networks. The number of nodes ($N$), the original length of the observation time window ($T$ in number of steps), the total number of contacts ($|E|$), $C_{fad}$ 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.

The fastest arrival distance matrix ($\Phi$) and temporal shortest distance matrix ($\Theta$) are considered as the temporal distance matrix $D$ respectively. And we call these two kinds of $TIG$-process as FAD-based $tig$-process and STD-based $tig$-process, denoted as $fad$-$tig$ and $std$-$tig$, for 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).
steady when \( n \) is large enough. In other words, the performance of \( \text{fad-tig} \) will be degraded if \( n \) is too large regarding \( \text{FAD-based node efficiency} \).

Now we will check the performance of \( \text{std-tig} \). From Fig. 5 and Fig. 6, we can see that except the first fewer steps, the \( \tau_{sf} \) and \( \tau_{ff} \) decrease to a steady state quickly. Moreover, the optimal value of \( \tau_{sf} \) is smaller than \( \tau_{ff} \). The phenomena might be because of the following two reasons.

Firstly, from Fig. 16, we know that most of the temporal distances are relatively small, and the coefficient of variation \( C_{std} \) (see in Table 2) is small. When doing the \( \text{TIG-process} \), the majority of nodes will be taken into account in the first few steps. This explanation can be further confirmed by Fig. 6, the \( \tau_{ss} \) shows a better performance in Eu-core and Oc commu networks, and the \( C_{std} \) of these two datasets are relatively higher than the others. Another reason might because of the difference of 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 \( \Theta \) is quite similar with the adjacent matrix \( A \) of the static abstraction of the temporal networks, which means less temporal information contained in \( \Theta \) compared with \( \Phi \).

This kind of phenomena is also observed in the following sections. Now we will explore how the nodes with top-ranked \( \text{tig-scores} \) affect the network.
efficiency.

5.2. Quantifying network efficiency

In this section, we will see the evolution of 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 FAD-based network efficiency and STD-based network efficiency respectively.

Fig. 7 shows the changing of FAD-based network efficiency 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 $n$. Obviously, $NE_{fad}$ gets the best performance and $E_{fad}$ decreases most slowly when the nodes are randomly removed. What’s more, for most tig-scores with different basic metrics as the initial information, the $NE_{std}$ performs even worse. Simultaneously, Fig. 8 shows the decreasing trend of STD-based network efficiency is similar for fad-tig and std-tig. Both fad-tig and std-tig work well, which further confirms our observation in Section 5.1. The FAD matrix which of much more temporal information performs better in predicting important nodes in terms of network efficiency.
Figure 5: The evolution of Kendall ranking correlation coefficient $\tau_{sf}$ between STD-based TIG-score and FAD-based network efficiency with information gathering depth $n$ for networks.

5.3. Quantifying nodes’ spreading influence.

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

As is mentioned in Section 3, we have three different ways to measure the spreading influence for each node. The three ways are defined in terms of different situations. We can not say which way is better or which one is the best to measure the spreading influence for temporal networks. In the following, we can see the tig-score has a similar relationship with the three measurements of spreading influence.

We still use the Kendall ranking correlation coefficient to evaluate the performance. The Kendall ranking correlation coefficient between fad-tig score and $R_{\text{norm}}$ is denoted as $\tau_{f\text{Norm}}$. Similarly, $\tau_{f\text{Max}}$ and $\tau_{f\text{Mean}}$ indicate the Kendall coefficient between fad-tig score and $R_{\text{max}}$ and $R_{\text{mean}}$ respectively. For the std-tig process, the notations are defined in the same way. $\tau_{s\text{Norm}}$, $\tau_{s\text{Max}}$, and $\tau_{s\text{Mean}}$ denote the Kendall coefficient between std-tig score and $R_{\text{norm}}$, $R_{\text{max}}$, and $R_{\text{mean}}$, respectively.

Furthermore, the tig-score is not highly related with the spreading influence as that with node efficiency, which means the TIG-process can predict the important more effectively regarding the network efficiency. But the
Figure 6: The evolution of Kendall ranking correlation coefficient $\tau_{ks}$ between STD-based TIG-score and STD-based network efficiency with information gathering depth $n$ for networks.

Figure 7: The evolution of FAD-based network efficiency $\xi_{fad}$ as the nodes’ removal.
Figure 8: The evolution of STD-based network efficiency $E_{std}$ as the nodes’ removal overall trend is similar.

In addition, from Fig. 9 to Fig. 11, we find that the fad-tig score with 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. That is might 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.

Finally, as we can see in Fig. 12 to Fig. 14, the std-tig process performs worse than fad-tig, regardless of the way to measure the spreading influence. As we discussed in Section 5.1, the STD matrix contains less temporal information than 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 high-order neighbors. Therefore, when $n$ is small, we are gathering information from neighboring nodes that is 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
Figure 9: $\tau_{f,\text{Norm}}$

Figure 10: $\tau_{f,\text{Max}}$
when $n$ is very large implies that the neighbors that are far away from the current node have small influence on its importance ranking.

6. Discussions

Even though many works have been done for 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 paper, we take the idea that node importance relies on the importance of its neighborhood, which has been verified by researchers [8]. 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 distance from the target node, $D$ is a distance matrix and $c$ is the initial score. We show that by different combinations of these four parameters, the TIG-process can degenerate to classic node ranking metrics, such as static degree, static closeness, temporal degree, temporal closeness (Figure 2).

We verify the performance of the TIG-process by using the performance evaluation methods, that is network efficiency based one and SIR spreading based one, on real-world temporal networks. We observe that the fastest
Figure 12: $\tau_{sNorm}$

Figure 13: $\tau_{sMax}$
Figure 14: $\tau_{\text{Mean}}$

Figure 15: Distribution of FAD.
arrival distance based TIG process performs much better than the one based on temporal shortest distance. In addition, there is an optimal gathering depth $n$ which makes FAD based TIG-process perform best.

Actually, the main contribution of this paper is not to proposed an exact metric to do the node ranking problem. 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 nodes’ significance.

Through this framework, we observe that FAD-based tig-process is more functional in predicting the significant nodes compared with STD-based one. Firstly, the FAD matrix captures more temporal information, which means it fits to temporal networks better. On the other hand, from the definitions of these two kinds of distance, the former one can be calculated from any time of the observation time window. The later one is more like a temporal metric but based on the final state of networks. There is no doubt that there might exist some 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 physical or spatial distance and the distance in temporal networks as temporal distance, then which one is more significant in measuring nodes’ influence? In addition, in Fig. 2 and Table 1, an Iterative
TIG process was introduced, which means at each step in the TIG-process, we gather the current tig-score instead of the initial information. This metric will be discussed in future works. What’s more, for the datasets used in this paper, we cannot get the true initial information. With the rapid increasing of the amount of data, our proposed TIG-process can be further explored.

Acknowledgements

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


