

Identify information sources with different start times in complex networks based on sparse observers

Yuan-Zhang Deng(邓元璋), Zhao-Long Hu(胡兆龙), Feilong Lin(林飞龙), Chang-Bing Tang(唐长兵), Hui Wang(王晖), and Yi-Zhen Huang(黄宜真)

Citation: Chin. Phys. B, 2024, 33 (**11**): 118901. DOI: 10.1088/1674-1056/ad7af4
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Identifying influential spreaders in complex networks based on density entropy and community structure

Zhan Su(苏湛), Lei Chen(陈磊), Jun Ai(艾均), Yu-Yu Zheng(郑雨语), and Na Bie(别娜)
Chin. Phys. B, 2024, 33 (**5**): 058901. DOI: 10.1088/1674-1056/ad20d6

A multilayer network diffusion-based model for reviewer recommendation

Yiwei Huang(黄羿炜), Shuqi Xu(徐舒琪), Shimin Cai(蔡世民), and Linyuan Lü(吕琳媛)
Chin. Phys. B, 2024, 33 (**3**): 038901. DOI: 10.1088/1674-1056/ad181d

Source localization in signed networks with effective distance

Zhi-Wei Ma(马志伟), Lei Sun(孙蕾), Zhi-Guo Ding(丁智国), Yi-Zhen Huang(黄宜真), and Zhao-Long Hu(胡兆龙)
Chin. Phys. B, 2024, 33 (**2**): 028902. DOI: 10.1088/1674-1056/ad1482

Identifying multiple influential spreaders in complex networks based on spectral graph theory

Dong-Xu Cui(崔东旭), Jia-Lin He(何嘉林), Zi-Fei Xiao(肖子飞), and Wei-Ping Ren(任卫平)
Chin. Phys. B, 2023, 32 (**9**): 098904. DOI: 10.1088/1674-1056/acac0b

Self-similarity of complex networks under centrality-based node removal strategy

Dan Chen(陈单), Defu Cai(蔡德福), and Housheng Su(苏厚胜)
Chin. Phys. B, 2023, 32 (**9**): 098903. DOI: 10.1088/1674-1056/acd3e2

Identify information sources with different start times in complex networks based on sparse observers

Yuan-Zhang Deng(邓元璋)[1], Zhao-Long Hu(胡兆龙)[1,†], Feilong Lin(林飞龙)[1], Chang-Bing Tang(唐长兵)[2], Hui Wang(王晖)[1], and Yi-Zhen Huang(黄宜真)[3]

[1] *School of Computer Science and Technology, Zhejiang Normal University, Jinhua* 321004, *China*
[2] *School of Physics and Electronic Information Engineering, Zhejiang Normal University, Jinhua* 321004, *China*
[3] *School of Information Engineering, Jinhua Polytechnic, Jinhua* 321016, *China*

(Received 14 May 2024; revised manuscript received 13 August 2024; accepted manuscript online 14 September 2024)

The dissemination of information across various locations is an ubiquitous occurrence, however, prevalent methodologies for multi-source identification frequently overlook the fact that sources may initiate dissemination at distinct initial moments. Although there are many research results of multi-source identification, the challenge of locating sources with varying initiation times using a limited subset of observational nodes remains unresolved. In this study, we provide the backward spread tree theorem and source centrality theorem, and develop a backward spread centrality algorithm to identify all the information sources that trigger the spread at different start times. The proposed algorithm does not require prior knowledge of the number of sources, however, it can estimate both the initial spread moment and the spread duration. The core concept of this algorithm involves inferring suspected sources through source centrality theorem and locating the source from the suspected sources with linear programming. Extensive experiments from synthetic and real network simulation corroborate the superiority of our method in terms of both efficacy and efficiency. Furthermore, we find that our method maintains robustness irrespective of the number of sources and the average degree of network. Compared with classical and state-of-the art source identification methods, our method generally improves the AUROC value by 0.1 to 0.2.

Keywords: complex networks, information spread, source identification, backward spread centricity

PACS: 89.75.–k, 87.23.Ge, 89.75.Fb **DOI:** 10.1088/1674-1056/ad7af4

1. Introduction

The rapid development of online social networks such as Facebook, Twitter, and Weibo has greatly facilitated information sharing and dissemination. With the advancement of information technology, various network platforms intertwine, further intensifying the breadth and depth of information dissemination. However, some malicious entities exploit this feature of the Internet to spread false information[1] and disseminate computer viruses,[2] which severely affects national security and the well-being of the public. Confronted with these threats, the identification of the sources of such detrimental activities has emerged as a critical and urgent imperative in the effort to formulate effective strategies to mitigate these risks and safeguard social stability.

Complex network theory, which can reveal information spread patterns from both micro and macro perspectives, has been introduced for information source localization in recent years. Due to the high similarity between information spread and infectious disease spread, scholars often employ spread dynamic models such as SI (susceptible–infectious), SIS (susceptible–infectious–susceptible), SIR (susceptible–infectious–recovered) or some derivative models to characterize information spread.[3] For example, the influence maximization,[4–9] the information spread and evolution.[10,11] Here we focus on a reverse problem: identify the location of the initial burst of information, which is also called as information source localization.

To precisely and swiftly identify the source, various source localization methods have been proposed under different assumptions. Depending on how information is acquired, these methods generally fall into two categories: snapshot-based source localization methods and observer-based or sensor-based source localization methods. The former approach involves inferring source(s) by analyzing the infection subgraph given the snapshot at a specific moment.[12–21] The latter method estimates source(s) in situations of limited observers (i.e., observed nodes).[22–36]

Source localization methods that rely on snapshots typically necessitate the status of nearly all nodes within a network. This requirement renders such methods impractical for application in extensive networks like social media platforms. In contrast, source localization techniques that operate with a limited number of observational nodes are not only more efficient but also more feasible for practical use. Among these, the backward spread source localization algorithm has been validated as an efficacious approach.[22,25,29,32,33] A more significant concern is that almost all the existing studies assume that information spread at the same moments, a presumption that does not align with real-world scenarios.[37,38] Although

[†]Corresponding author. E-mail: huzhaolong@zjnu.edu.cn
© 2024 Chinese Physical Society and IOP Publishing Ltd

Ji *et al.*, based on snapshot, studied the multi-source identification with different initial spread moments, both the state of all nodes in the network and the number of sources need to be known.[39] Nevertheless, the number of sources and the initial spread times are often unavailable, it remains a challenge to develop a method for identifying sources with different start times using limited observers.

To address these issues, we propose a new method for multi-source localization with different start times using limited observers. Our main contributions are as follows: (i) We proposed and proved the backward spread tree theorem and source centrality theorem, and further provided the backward spreading centrality (BSC) algorithm for source localization. (ii) Our method does not require a known number of sources and triggering time. (iii) Our method can estimate the number of sources, the initial spread moment and spread duration.

The structure of the remaining parts of this paper is organized as follows: In Section 2, we overview the related works. In Section 3, we introduce preliminaries and describe the objective to be studied. Section 4 proposes the BSC algorithm to address multi-source identification with different initial spread times. Subsequently, we verify and analyze the performance of our method for source localization by synthetic networks and real networks in Section 5. Finally, we summary the paper in Section 6.

2. Related works

The foundational research on source localization began with the snapshot-based methodology. In 2010, Shah and Zaman proposed that the source was the one with the most spread paths among all nodes and introduced the rumor centrality theory for single-source localization.[12] Whereafter, the Jordan centrality index was proposed, and it was found that the accuracy of source localization of Jordan centrality was often higher than rumor centrality.[14,40] Although multi-source identification is challenging, several innovative strategies have emerged. For instance, considering the uninfected neighbors of infected nodes, Prakash *et al.* introduced the minimum description length and proposed the NetSleuth algorithm to locate sources.[41] Jiang *et al.* proposed the K-center method to identify multiple sources based on equivalent distance and capacity-limited Voronoi diagrams.[42] Ali *et al.* considered both infected and uninfected neighbors of nodes, combined K-means clustering to divide spread regions, and subsequently located the sources.[43] Ji *et al.* divided the infection area and identified the overlapping areas by combinatorial optimization and joint estimation, and located the source in each infected area by the Jordan centrality.[39]

The above research related with source localization primarily relies on network topology and source centrality theory, in which sources are generally surrounded by other infected nodes. From a physical perspective, Lokhov *et al.* estimated the probability of state of each node at each moment using the DMP (dynamic message–passing) algorithm, and combined MLE (maximum likelihood estimation) to identify the source. This method has strong robustness, as it can achieve high performance of source localization even when the number of observers accounts for 70% of the network.[44] Since the DMP algorithm fails to address multi-source localization, Altarelli *et al.* extended the DMP algorithm and proposed the BP (belief propagation) algorithm to study multi-source localization.[45] This algorithm is highly innovative, as it can not only estimate spread model parameters but also can infer sources based on incomplete observation data. Considering that the spread dynamics model may be difficult to obtain, scholars developed the LPSI (label propagation-based source identification) algorithm.[46] Based on the LPSI algorithm, Peng *et al.* proposed the NLSI (node-level-based source identification) algorithm to locate the epidemic with SEIR (susceptible–exposed–infectious–recovered) model.[47] Assuming that the information spread path can be obtained, Zhu *et al.* proposed a path-based label back-propagation method to locate the source, namely, the observation node can obtain the neighbor from which the information is transmitted.[20] Subsequently, they also extended this method to locate the source in multi-layer networks.[48] With the same assumption, Yang *et al.* presented a method to locate the source with observers-based similarity.[33] The truth, however, is that sometimes we do not know who passed the information to us, and the spread of diseases through air is a typical example.

The aforementioned snapshot-based source localization methods necessitate full network observation or observing most of nodes in the network, which is impractical for large-scale networks. Consequently, various observer-based or sensor-based source localization theories and methods have been developed. The primary strategies for deploying observers include random selection,[17,22,25,29,49] based on centrality index ranking,[14,40] high coverage rate,[25,50] observability theory,[24,51] a greedy algorithm rooted in network structure differences,[52] a dynamic and static combination,[53] high variance observers.[54,55] The random selection is always the predominant approach among the strategies. Based on randomly deployed observation nodes, Pinto *et al.* used a breadth-first algorithm to approximate the information spread process, and proposed a backward spread method and combined it with the MLE to identify the source.[22] This research pioneered the backward spread source localization methods. According to the structural characteristics of signed networks, Ma *et al.* extended this method to source localization on signed networks by considering the effective distance, and found that the ratio of positive and negative edges affects the results of source localization.[56] Shen *et al.* constructed time-reversal

paths, and located the source with the variance of node infection times deduced from all observers.[49] Paluch *et al.* ranked the moments when observation nodes received information, and selected some early information-receiving nodes to locate the source with backward spread method, which refined the algorithm proposed in Ref. [22] and enhanced the efficiency.[29] Hu *et al.* analyzed differences in spread delays and provided a greedy optimization algorithm to identify the minimum set of observers.[52] These methods assumed that information travels along the shortest path, when locating the source with backward spread. Gajewski *et al.* pointed out that when there are multiple longer shortest paths, the information spread may be faster than that of shorter path. They also found that making full use of multiple shortest paths can improve the accuracy of source localization.[57]

Researchers further devised numerous source localization methods in response to the prevalent multi-source diffusion phenomenon. Literatures[25,58] expanded the backward spread source localization algorithm [30] and presented the MMM and IP algorithms to identify multiple information sources. Based on Gromov matrix, Tang *et al.* put forth a maximum likelihood estimation algorithm for identifying single source. They further employed tree partitioning methods to separate spread regions and located the source in each region applying the single-source localization algorithm, also called as the SCCE algorithm.[30] Recently, Wang *et al.* proposed a method for multi-source localization based on the positive correlation between infected moments of nodes and the shortest distance between nodes and sources.[32] Although the aforementioned observer-based backward spread source localization methods do not require full network observation, some of them do require predicting whether each node is a source.[22,25,32] If we can reduce the number of nodes that are unlikely to be sources, the efficiency of source identification can be improved. In addition, many studies have revealed that information may burst from multiple locations at various start times.[37,38] Nevertheless, those source localization studies overlooked the situation where information triggers the spread at different times. Despite the fact that Ji *et al.* conducted on multiple sources emitting information at different times, it required the knowledge of the state of all nodes and the number of sources.[39] However, one common situation is that the number of sources and the initial spread times are often unavailable, how to develop a method to identify all the sources with different start times using limited observers faces great challenges.

3. Model and methods

3.1. Information spread model

We model the information spread in a finite undirected graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ consisted of $|\boldsymbol{V}|$ nodes and $|\boldsymbol{E}|$ edges, where $\boldsymbol{V}$ represents the set of nodes and $\boldsymbol{E}$ represents the set of edges. To further analyze the dynamic network structures prevalent in real-world scenarios, we adopt a static network structure for this study. If there is contact or communication between a pair of nodes u and v, then there exists one edge (u,v) between the two nodes. We assume that the source node set $\boldsymbol{S} = \{s_1, s_2, \ldots, s_{|\boldsymbol{S}|}\}$ are randomly selected from the node set $\boldsymbol{V}$, where $|\boldsymbol{S}|$ is the number of sources.

During the information spread process, sources $s_1, s_2, \ldots, s_{|\boldsymbol{S}|}$ initiate their spread at unknown starting moments $t_{s_1}, t_{s_2}, \cdots, t_{s_{|\boldsymbol{S}|}}$ and disseminate information to their neighbors based on corresponding spread delays. We only adopt the moment when the node first received the information, which is also called the infection moment. When one node receives information, it forwards the information to all its neighbor based on corresponding spread delays. In the end, all nodes will receive the information. Define t_v as the infection moment of node v, $\Gamma(v)$ as the set of neighbors of node v, and $\theta(u,v)$ as the spread delay of the edge (u,v). Because we lack empirical data to support the exact information propagation delays across each edge, therefore, like Ref. [22], we also assume that the random spread delays associated with the edges in $\boldsymbol{G}$ are independent and follow the identical Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with $\mu = 1$. Although this assumption may not reflect reality, it provides a paradigm for investigation. The information spread process is similar to those in Refs. [22,25,29,49,52], except that sources can spread information at different times.

Define $\Delta t_{s_i,v}$ as the shortest spread delays for node v receiving information from source s_i. Since the information travels along the path with the shortest spread delay, therefore the infection moment of node v can be expressed as

$$t_v = \min_{s_i \in \boldsymbol{S}} t_{s_i} + \Delta t_{s_i,v}. \tag{1}$$

Definition 1 (time-reversal path) Based on the assumption that information always spreads along the shortest path in the network, we construct a time-reversal path to predict the source using the concept of backward spread. The time-reversal path $P(o,v)$ represents the shortest path in reverse order for an observer o to infer node v as the information source.

For the network system, the shortest path delay matrix $\boldsymbol{D} \in \mathbb{R}^{|\boldsymbol{O}| \times (|\boldsymbol{V}| - |\boldsymbol{O}|)}$ describes the shortest path delays between each pair of nodes. The element $D_{o,v}$ in the matrix represents the time-reversal path delays from observer o to non-observer v. $|\boldsymbol{O}|$ is the number of observers. For example, in Fig. 1(a), assuming that the spread delay of each edge is 1, and node v_{A} is the source and v_{L} is the observer. The time-reversal path $P(v_{\mathrm{L}}, v_{\mathrm{A}})$ is $v_{\mathrm{L}} \to v_{\mathrm{O}} \to v_{\mathrm{A}}$ or $v_{\mathrm{L}} \to v_{\mathrm{J}} \to v_{\mathrm{A}}$. The shortest path delay $D_{v_{\mathrm{L}}, v_{\mathrm{A}}}$ is 2, thus the initial spread time for v_{A} is $t_{v_{\mathrm{L}}} - 2$.

The initial spread moment $t^*(o,v)$ for node v to act as the source can be estimated by using the moment when observa-

tion node o received information and the time-reversal path $P(o,v)$, namely,

$$t^*(o,v) = t_o - D_{o,v}. \tag{2}$$

Since we cannot obtain the real spread delay, then we use the shortest path delay $D_{o,v}$ instead of the spread delay $d(o,v)$ when estimating $t^*(o,v)$.

Definition 2 (critical duration) In order to minimize the probability that each source can spread information without being covered by other sources, we define a critical duration l. For example, assumed the spread delay between source s_1 and source s_2 is 3, then the start moment for them should satisfy $\mathrm{abs}(t_{s_1} - t_{s_2}) < 3$. Here $\mathrm{abs}(x)$ is the function that takes the absolute value of x. Therefore, the shortest path delay $d(s_i,s_j)$ between any two sources s_i and s_j should satisfy

$$\min_{s_i,s_j \in \boldsymbol{S}} d(s_i,s_j) - \mathrm{abs}(t_{s_i} - t_{s_j}) > l. \tag{3}$$

For example, in Fig. 1, the shortest distance between v_A and v_B is 3. To better understand critical duration, we assume that the shortest distance is equal to the propagation delay. If the start times of v_A and v_B are 1 and 5 respectively, it results in the spread region of v_B being covered by v_A, then it makes no sense for v_B to be the source.

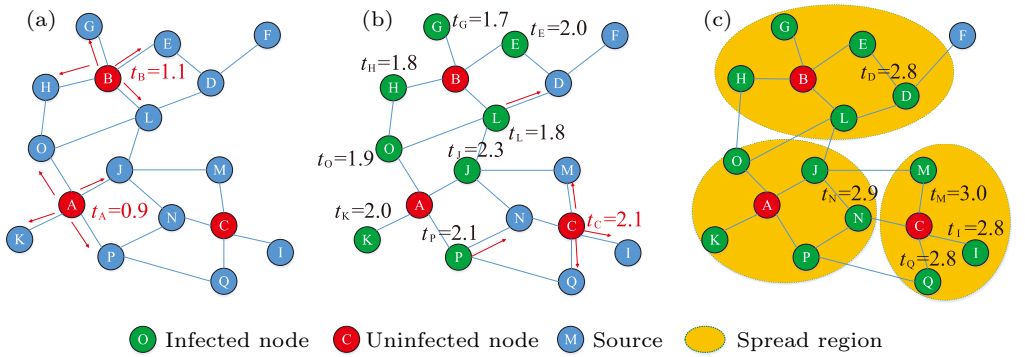


Fig. 1. A simple diagram of multi-source spread with different start times. (a) Source v_A and source v_B send information to their neighbours at $t = 0.9$ and $t = 1.1$, respectively. The red arrow represents the direction of information dissemination. (b) Source v_C sends information to their neighbours at $t = 2.1$. The two infected nodes v_L and v_P send information to its uninfected neighbors. (c) At $t = 3.0$, three spread regions are formed for sources v_A, v_B, and v_C. Here v_F is not included because it has not infected yet. The random spread delay between nodes conforms to a Gaussian distribution $N(1, 0.3^2)$.

3.2. Research objective

Given the network structure $\boldsymbol{G}$ and the infection moment $\boldsymbol{T_O} = \{t_{o_1}, t_{o_2}, \ldots, t_{o_{|\boldsymbol{O}|}}\}$ for $|\boldsymbol{O}|$ random observers, our goal is to propose one method to identify the information sources with limited observers under the following two conditions: (i) Both the number of sources and initial spread moments are unknown. (ii) The initial spread moments may differ.

3.3. Methods

Some observer-based backward spread methods for source localization require checking whether each node is the source,[22,25,32] which is computationally expensive. Here we firstly propose and prove the backward spread trees theorem and source centrality theorem. Based on the two theorems, the backward spread centrality (BSC) algorithm is presented which provides a sources search strategy to capture potential sources. Secondly, by combining the BSC algorithm and linear programming, we develop a method to locate the sources and further present a method for predicting the start time and spread duration.

3.3.1. Infer potential sources with the BSC algorithm

Definition 3 (backward spread tree) The backward spread tree $\mathcal{T}_v$ is a shortest path tree, rooted at node v and composed of a set of time-reversal paths $P_v = \{P(o_1,v), P(o_2,v), \ldots, P(o_c,v)\}$ within the spread region of v. Here $o_1, o_2, \ldots, o_c$ are the observers within the spread region of v.

A detailed example of a backward spread tree is shown in Fig. 2. There are two sources, v_J and v_L. The nodes surrounded by two yellow dashed circles are their spread regions, remarked by $\boldsymbol{C}_{v_\mathrm{J}}$ and $\boldsymbol{C}_{v_\mathrm{L}}$, respectively. The solid yellow line with arrows represents backward spread trees of the two sources.

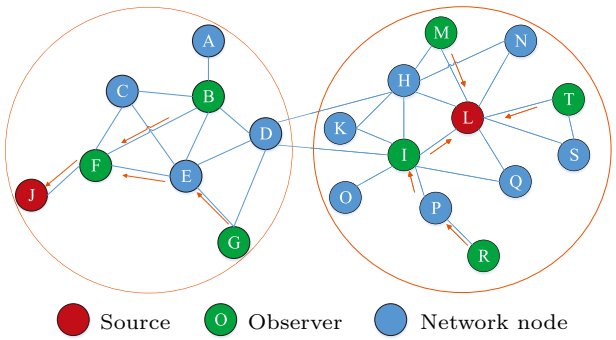


Fig. 2. A simple diagram of backward spread trees of the sources. The network includes two sources, v_J and v_L. Observers v_F, v_B, v_G are in the spread region of source v_J. Observers v_M, v_T, v_R, v_I are in the spread region of source v_L. The backward spread trees of source v_J and source v_L are $\mathcal{T}_{v_\mathrm{L}}$ and $\mathcal{T}_{v_\mathrm{J}}$, respectively, where $\mathcal{T}_{v_\mathrm{J}} = P(v_\mathrm{B}, v_\mathrm{J}) \cup P(v_\mathrm{G}, v_\mathrm{J})$ and $\mathcal{T}_{v_\mathrm{L}} = P(v_\mathrm{M}, v_\mathrm{L}) \cup P(v_\mathrm{T}, v_\mathrm{L}) \cup P(v_\mathrm{R}, v_\mathrm{L})$.

In order to locate sources with different triggering times,

a key step lies in inferring the moment when the node becomes infected. Let us define $t^*(v)$ as the estimated infection moment of node v. If the node $v \in \boldsymbol{O}$, the estimation of infection moment $t^*(v)$ is its infection moment t_v. If node $v \notin \boldsymbol{O}$, then $t^*(v)$ is the maximum estimated infection moment among the observers $o_1, o_2, \ldots, o_{|\boldsymbol{O}|}$ along the time-reversal paths $P(o_1, v), P(o_2, v), \ldots, P(o_{|\boldsymbol{O}|}, v)$ speculating on node v, expressed by

$$t^*(v) = \begin{cases} t_v, & v \in \boldsymbol{O}, \\ \max\limits_{o \in \boldsymbol{O}} t^*(o, v), & v \notin \boldsymbol{O}, \end{cases} \tag{4}$$

where $t^*(o, v)$ can be obtained by Eq. (2).

Theorem 1 (backward spread trees theorem) In the underlying network $\boldsymbol{G}$, $\boldsymbol{C}_i \subseteq \boldsymbol{G}$ and $\mathcal{T}_v \subseteq \boldsymbol{C}_i$, where $\boldsymbol{C}_i$ is the spread region of source s_i. If there exists an observer $o \in \mathcal{T}_v$, then the expectation of infection moment of node v is the infection moment, i.e., $E(t^*(v)) = t_v$.

Proof We consider two cases to prove Theorem 1.

If node v is an observer, according to Eq. (4), $v \in \boldsymbol{O}$ and $t^*(v) = t_v$. Therefore, $E(t^*(v)) = E(t_v) = t_v$.

Now we analyze the case that node v is not an observation node. Suppose there exists a node $u \in \mathcal{T}_v$ and observer $o^* \notin \mathcal{T}_v$ on the same time-reversal path $P(o^*, v)$. The time-reversal path $P(o^*, v)$ can be decomposed into two paths: $P^*(o^*, u)$ and $P^*(u, v)$ with spread delays $d^*(o^*, u)$ and $d^*(u, v)$, respectively. Since $o^* \notin \mathcal{T}_v$, the following relationship holds:

$$t_{o^*} < t_u + d^*(o^*, u), \tag{5}$$

which yields

$$t_{o^*} - d^*(o^*, u) - d^*(u, v) < t_u - d^*(u, v), \tag{6}$$

$$t_{o^*} - d(o^*, v) < t_u - d^*(u, v). \tag{7}$$

By using Eq. (2), we have

$$t^*(o^*, v) < t^*(u, v). \tag{8}$$

Because $o, u \in \mathcal{T}_v$, then $E(t^*(u, v)) = t^*(o, v)$, which leads to

$$E(t^*(o^*, v)) < E(t^*(u, v)) = E(t^*(o, v)). \tag{9}$$

According to Eq. (4), we have

$$E(t^*(v)) = \max_{o \in \boldsymbol{O}} E(t^*(o, v)) = t_v. \tag{10}$$

Because each source has its own spread region, then for the backward spread tree $\mathcal{T}_{s_i}$, the source s_i is earlier infected than all its neighbors. However, for any other node v, there must exists at least a neighbor that was infected earlier than v. In this paper, we extend this characteristic of the backward spread tree to general graphs. If node i has more neighbors whose infection moment is higher than node i, then node i is more likely to be the source. To quantify the probability of a node being a source, we define the backward spread centrality. The backward spread centrality $\mathcal{C}_v$ of node v is calculated by

$$\mathcal{C}_v = \frac{1}{|\Gamma(v)|} \sum_{u \in \Gamma(v)} \delta(t_v, t_u), \tag{11}$$

where the function $\delta(x, y)$ is used to compare x and y. If $x < y$, then $\delta(x, y) = 1$; otherwise, $\delta(x, y) = 0$.

Theorem 2 (sources centrality theorem) For any $w \in v \cup \Gamma(v)$, if there exists an observer $o \in \mathcal{T}_w$, then condition p is a necessary and insufficient condition for condition q.

Condition p: $\mathcal{C}_v > \mathcal{C}_u$ is satisfied for node v and any its neighbor u.

Condition q: Node v is the source.

Proof We prove Theorem 2 in two steps.

Firstly, we prove that p is a necessary condition for q. Node v is a source and is surrounded by many non-source nodes. If node $u \in \Gamma(v)$, then $u \notin \boldsymbol{S}$. For any node $w \in v \cup \Gamma(v)$ and $o \in \mathcal{T}_w$, according to Theorem 1, the infection moment of node w can be estimated with $t^*(w)$. Since node u and source v are adjacent to each other, then $t^*(v) < t^*(u)$ and $\mathcal{C}_u < 1$. Because $t^*(v) < t^*(u)$ for the source v and any neighbor u, then we have $\mathcal{C}_v = 1$. If node v is a source, then there exists a relationship $\mathcal{C}_v > \mathcal{C}_u$ for any neighboring node u. In summary, p is a necessary condition for q.

Second, we prove that p is not a sufficient condition for q. Similarly, we can estimate the infection moment using the observation moment according to Theorem 1. When source v and node i satisfy the conditions $\Gamma(v) = \Gamma(i)$ and $\mathcal{T}_v \setminus v = \mathcal{T}_i \setminus i$, there exists the relation $\mathcal{C}_v > \mathcal{C}_u$ and $\mathcal{C}_i > \mathcal{C}_u$. But the source v and node i cannot be distinguished by the backward spread method. Therefore, p is not a sufficient condition for q.

Theorem 2 implies that if $\mathcal{C}_v > \mathcal{C}_u$, node v may be not the true source. For example, in Fig. 2, v_{B} and v_{G} are observers. Assumed that a network consists of v_{B}, v_{G}, v_{D}, and v_{E}. Based on the two observers, it is inferred that the initial spread time of node v_{D} and node v_{E} is the same through backward spread, so it is impossible to distinguish which of node v_{D} and node v_{E} is the source.

Based on Theorems 1 and 2, we propose a sources search strategy. Firstly, we estimate the infection time for each node according to Theorem 1, and calculate the corresponding backward spread centrality for each node using Eq. (11). Secondly, according to Theorem 2, we add nodes with higher centrality than all their neighbors as potential sources to the potential source set $\boldsymbol{U}$.

3.4. Multi-source identification with the BSC algorithm and linear programming

To distinguish sources in the potential source set $\boldsymbol{U}$ from other nodes, we use the measurement information from observation nodes and the theorem of backward spread trees. The true source, acting as the root node, will account for more observation nodes than any non-source. Therefore, we select the minimum subset $\boldsymbol{S}^*$ from the potential source set $\boldsymbol{U}$ that covers the maximum number of observation nodes as the predicted source set. Thus the multi-source identification can be formulated as a linear programming problem, which yields

$$\begin{aligned} &\min ||\boldsymbol{x}||_1 \\ &\text{s.t. } \boldsymbol{A}\boldsymbol{x} > \boldsymbol{b}, \end{aligned} \tag{12}$$

where $||\boldsymbol{x}||_1$ represents the L1-norm of the vector $\boldsymbol{x}$. The relationship matrix $\boldsymbol{A} \in \mathbb{R}^{|O|\times|U|}$ with $\boldsymbol{A}_{o,u}$ reflects the credibility relationship between observer o and potential source u. The element b_o in constraint vector $\boldsymbol{b}$ ensures that each observer o is covered by a finite number of potential sources. The element x_u with 1 or 0 in vector $\boldsymbol{x}$ describes whether potential source u is a predicted source.

An observation node o has a credible relationship with potential source u, if u is the closest potential source to o. Then the elements $A_{o,u}$ and b_o can be expressed as

$$A_{o,u} = \gamma(o,u), \tag{13}$$

$$b_o = \min(1, \sum_{u\in\boldsymbol{U}} A_{o,u}), \tag{14}$$

where $\gamma(o,u)$ is 1 if u is the closest potential source to o, otherwise is 0.

With the above equations, we can obtain the vector $\boldsymbol{x}$ such that we can determine whether a potential source u is predicted as a source. If the element x_u corresponding to the potential source u is equal to 1, we add u to the source prediction set $\boldsymbol{S}^*$ to obtain the final set of predicted sources $\boldsymbol{S}^* = \{s_1^*, s_2^*, \ldots, s_{|\boldsymbol{S}^*|}^*\}$.

Algorithm 1: Backward spread centrality algorithm

Input: Network structure $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, observers set $\boldsymbol{O}$, infection time vector $\boldsymbol{T}_{\mathrm{O}}$, Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$.

Output: Sources prediction set $\boldsymbol{S}^*$, predicted start time $t_{\boldsymbol{S}^*}$, and predicted spread duration T^*.

Initialization $\boldsymbol{S}^* \leftarrow \{\}$, $\boldsymbol{C}^* \leftarrow \{\}$, $\boldsymbol{U} \leftarrow \{\}$;
Calculate the matrix $\boldsymbol{D}$, namely, the shortest path time delay from the observers to non-observers;
for $v \in \boldsymbol{V}$ **do**
    Compute backward spread centrality $\mathcal{C}_v$ of node v according to Eq. (11).
end
for $v \in \boldsymbol{V}$ **do**
    if $\mathcal{C}_v > \mathcal{C}_u, \forall u \in \Gamma(v)$ **then**
        $\boldsymbol{U} \leftarrow \{\boldsymbol{U}, v\}$
    end
end
Obtain $\boldsymbol{A}$ and $\boldsymbol{b}$ by Eqs. (13) and (14);
Compute $\boldsymbol{x}$ by Eq. (12);
for $x_u \in \boldsymbol{x}$ **do**
    if $x_u = 1$ **then**
        $\boldsymbol{S}^* \leftarrow \{\boldsymbol{S}^*, u\}$
    end
end
for $s_i^* \in \boldsymbol{S}^*$ **do**
    $C_i^* \leftarrow \{s_i^*\}$;
    Based on Eq. (4), infer the initial spread time of source $t^*(s_i^*)$;
end
for $v \in \boldsymbol{V} \setminus \boldsymbol{S}^*$ **do**
    Find the source s_i^* meets $\arg\min_{s_i^*\in\boldsymbol{S}^*} t^*(s_i^*) + t(v, s_i^*)$;
    $C_i^* \leftarrow \{C_i^*, v\}$;
end
$\boldsymbol{C}^* \leftarrow \{C_1^*, C_2^*, ..., C_{|\boldsymbol{S}^*|}^*\}$, obtain $t_{\boldsymbol{S}^*}$ and T^* according to Eqs. (15) and (16), respectively.

The start time of information spread and the spread duration are two important roles in information spread. Therefore, we further provide one method for predicting the start time $t_{\boldsymbol{S}}$ and spread duration T. Firstly, each source s_i^* in $\boldsymbol{S}^*$ is set as the center of network partition C_i^*. Secondly, based on the predicted source and Eq. (4), we can infer the initial spread time of each source. Thirdly, for each source and corresponding initial spread time, construct the spread region of each source by the breadth-first algorithm. With the above three steps, we can obtain the prediction set of network partition $\boldsymbol{C}^* = \{C_1^*, C_2^*, \ldots, C_{|\boldsymbol{S}^*|}^*\}$. Finally, based on the predicted network partition $\boldsymbol{C}^*$ and the predicted source set $\boldsymbol{S}^*$, the start time $t_{\boldsymbol{S}}$ and spread duration T can be calculated by

$$t_{\boldsymbol{S}^*} = \min_{s_i^* \in \boldsymbol{S}^*} t^*(s_i^*), \tag{15}$$

$$T^* = \max_{v \in \boldsymbol{C}^*, s_i^* \in \boldsymbol{S}^*} t^*(v) - t^*(s_i^*). \tag{16}$$

The BSC algorithm is presented in Algorithm 1. The time complexity of the BSC algorithm mainly depends on the shortest paths computation and linear programming. The time complexity of constructing the shortest paths time delay matrix $\boldsymbol{D}$ is $O(|\boldsymbol{O}||\boldsymbol{V}|\log(|\boldsymbol{V}|))$. The highest time complexity of 0–1 integer programming is $O(2^{|\boldsymbol{U}|})$ through the violent enumeration. After the above analysis, we conclude that the time complexity of the BSC algorithm is approximate to $O(|\boldsymbol{O}||\boldsymbol{V}|\log(|\boldsymbol{V}|) + 2^{|\boldsymbol{U}|})$.

4. Experimental results and analysis

4.1. Evaluation indicators for multi-source identification

To quantify the performance of the algorithm for source localization, we introduce the area under a receiver operating characteristic (AUROC), where a larger AUROC value indicates higher source localization performance.[59] The AUROC is a typical measure of performance used in dichotomous classification, and it has been widely used in network reconstruction,[60,61] link prediction,[62,63] source localization.[31,32] The true positive rate $\mathrm{TPR}(l)$ and false positive rate $\mathrm{FPR}(l)$ that are used to calculate AUROC are defined as follows:

$$\mathrm{TPR}(l) = \frac{\mathrm{TP}(l)}{|\boldsymbol{S}|}, \tag{17}$$

$$\mathrm{FPR}(l) = \frac{\mathrm{FP}(l)}{|\boldsymbol{V}| - |\boldsymbol{S}|}, \tag{18}$$

where $\mathrm{TP}(l)$ and $\mathrm{FP}(l)$ denote the number of true and false sources in the first l nodes in the candidate list.

4.2. Data description

Here we analyze the effectiveness of the proposed multi-source identification method with both synthetic and real networks. The synthetic networks used for testing include the Erdös–Rényi (ER) network[64] and the Barabási–Albert (BA) network.[65] For real networks, we consider the IEU (ia-email-univ) network,[66] SWLC (socfb-Wake73-largest-compoent) network,[67] the SWV (soc-wiki-vote) network,[68] and the WP (web-polblogs) network,[69] which were derived from real online social networks. The four real networks datasets are public and can be available at https://icon.colorado.edu/#!/networks, http://www-personal.umich.edu/m̃ejn/netdata/ and https://networkrepository.com/.

The SWLC network is a largest connected subgraph of the original network which is consist of 5372 nodes and 279191 edges. Table 1 presents some topological properties of the four real networks, including the number of nodes and edges, the average degree, the diameter of the network and clustering coefficient. With the assortativity coefficient formula in Ref. [70], we find that the ER network, IEU network, and SWLC network are homogeneous, while the other three networks are heterogeneous.

Table 1. Network topology properties of four real networks.

Datasets	$\lvert V\rvert$	$\lvert E\rvert$	$\langle k\rangle$	Diameter	Clustering coefficient
IEU	1131	5451	9.62	8	0.22
SWLC	5366	279186	104.06	9	0.27
SWV	889	2914	6.56	13	0.15
WP	643	2284	7.10	10	0.23

4.3. Experimental evaluation for multi-source identification

Based on snapshot, Ji *et al.* studied the problem of multi-source identification with different initial spread moments. However, all nodes needed to be observed and the number of sources needed to be known.[39] There is a lack of multi-source identification methods based on partial observation nodes with different initial spread times and unknown number of sources. There are some multi-source identification methods based on backward spread and partial observation nodes, such as Refs. [25,30,32,58] where reference[58] can only be used in the case of integer spread delay. More recently, scholars have proposed methods for multi-source identification,[20,21,33] which are based on snapshot. Therefore, here we adopt the three methods (i.e., the SCCE algorithm,[30] the MMM algorithm,[25] and the PS algorithm[32]) as the baseline method. For the PS algorithm, we set the threshold $h \in [2, 30]$, which is a broad range to ensure that the best results are included. Unless otherwise stated, the average degree $\langle k\rangle$ of ER and BA networks is 12 and network size $|\boldsymbol{V}|$ is set to 1000, the number of sources $|\boldsymbol{S}| = 3$, and the random spread delay along each edge follows a Gaussian distribution $\mathcal{N}(1, 0.3^2)$. The source start to spread information at the time interval $[t^0, t^0 + \mu]$ with the critical duration $l = \mu = 1$, where t^0 is the earliest possible triggering time.

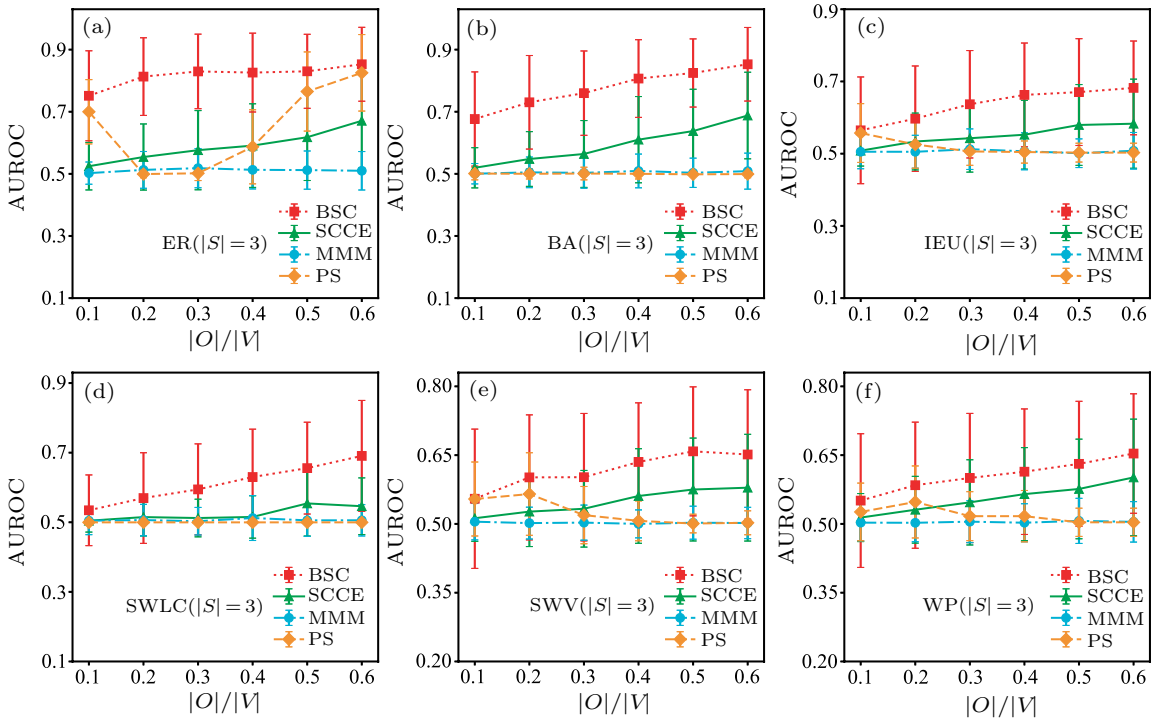


Fig. 3. The relationship between AUROC and the fraction of observers $|\boldsymbol{O}|/|\boldsymbol{V}|$ under the four algorithms in two synthetic networks [(a) and (b)] and four real networks [(c)–(f)]. The network size of ER network and BA network $|\boldsymbol{V}| = 1000$, and the network average degree $\langle k\rangle = 12$. The source sends information at time intervals $[t^0, t^0+\mu]$ with the critical duration $l = \mu = 1$. The number of diffusion sources $|\boldsymbol{S}| = 3$, and the random spread delay of the edge conforms to $\mathcal{N}(1, 0.3^2)$. The results are obtained with 200 times independent simulations, and the error bar represents the standard deviation.

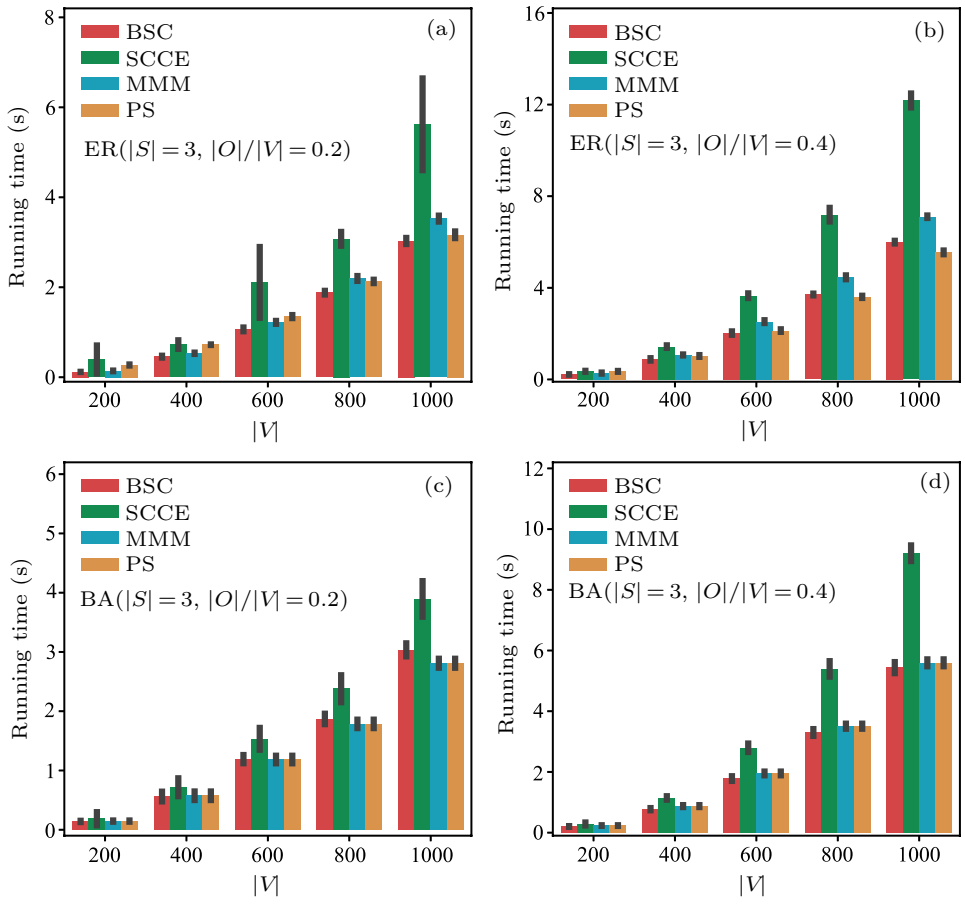


Fig. 4. The relationship between the running time and the network size $|\boldsymbol{V}|$ under the four algorithms with different $|\boldsymbol{O}|/|\boldsymbol{V}|$ in ER networks [(a) and (b)] and in BA networks [(c) and (d)]. The network average degree $\langle k\rangle = 12$. The source sends information at time intervals $[t^0, t^0+\mu]$ with the critical duration $l = \mu = 1$. The number of diffusion sources $|\boldsymbol{S}| = 3$, and the random spread delay of the edge conforms to $\mathcal{N}(1, 0.3^2)$. The results are obtained with 200 times independent simulations, and the error bar represents the standard deviation.

All the results are average over 200 independent simulations, running on a 3.6-GHz workstation with 16 GB of RAM, Intel(R) Core(TM) i7-7700 CPU.

(i) Multi-source identification with different methods

From Fig. 3, it can be found that the performance of the BSC algorithm and SCCE algorithm gradually improve with the increase of the observer ratio $|\boldsymbol{O}|/|\boldsymbol{V}|$. Compared with the other three algorithms, the BSC algorithm has a significant advantage in source localization. Surprisingly, the MMM algorithm often shows high AUROC of multi-source identification with the same initial spread times,[25] but its AUROC is almost close to a random guess once the initial spread times are different. The PS algorithm has a good performance when the observer ratio is high (e.g., 0.5 or 0.6) in ER networks, while it shows bad performance in other case.

(ii) Efficiency of multi-source identification with different methods

Upon examining the running time illustrated in Fig. 4, it becomes evident that the efficiency of SCCE is the worst, while the efficiency of the other three algorithms are almost the same. In fact, the running times of these algorithms are within a few seconds. The BSC algorithm is composed of two parts: backward spread centrality and linear programming. We then analyzed the consumption time of the two parts of the BSC algorithm. The running time of backward spread centrality and linear programming are marked by BSC_SP and BSC_LP, respectively. From Fig. 5, we can find that BSC_SP takes more time than BSC_LP.

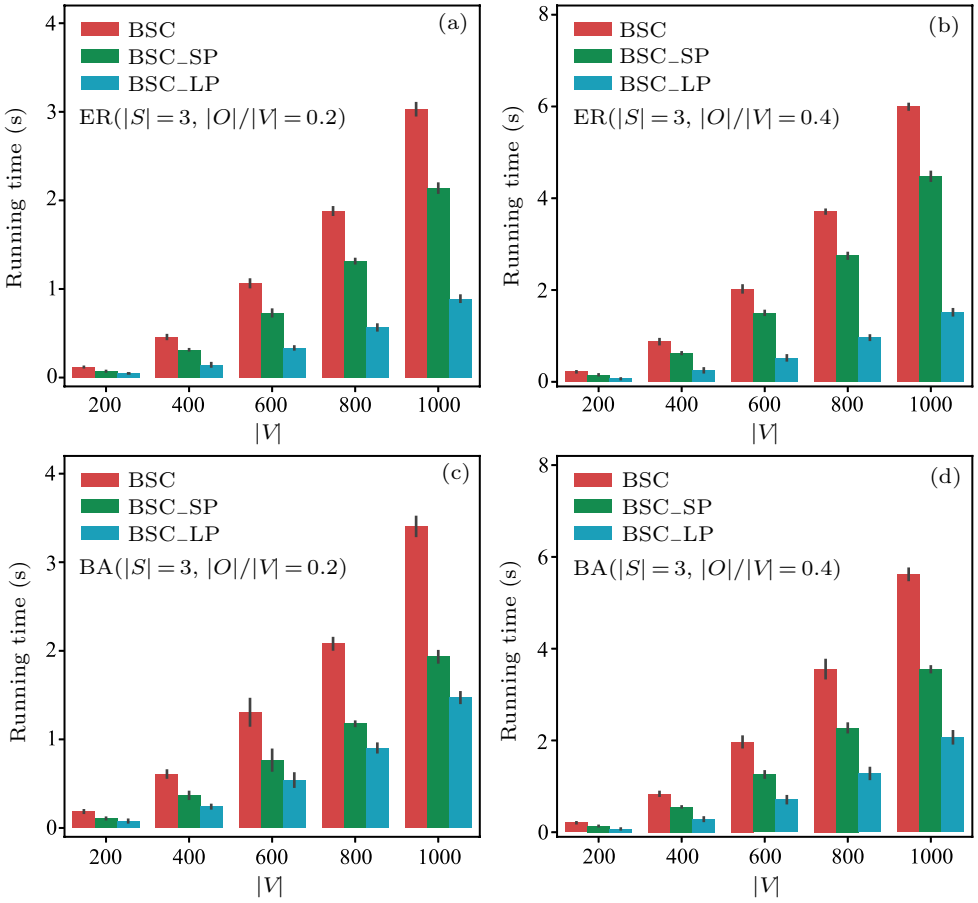


Fig. 5. The relationship between the running time and the network size $|\boldsymbol{V}|$ under the BSC algorithm with different $|\boldsymbol{O}|/|\boldsymbol{V}|$ in ER networks [(a) and (b)] and in BA networks [(c) and (d)]. The running times of BSC_SP and BSC_LP are that of backward spread centrality and linear programming, respectively. The network average degree $\langle k\rangle = 12$. The source sends information at time intervals $[t^0, t^0+\mu]$ with the critical duration $l = \mu = 1$. The number of diffusion sources $|\boldsymbol{S}| = 3$, and the random spread delay of the edge conforms to $\mathcal{N}(1, 0.3^2)$. The results are obtained with 200 times independent simulations, and the error bar represents the standard deviation.

(iii) Prediction of the start time and spread duration

Figure 6 presents the predicted start time error $\Delta t_{\boldsymbol{S}} = |t_{\boldsymbol{S}^*} - t_{\boldsymbol{S}}|$ and the predicted spread duration error $\Delta T = |T^* - T|$ under different observer ratios $|\boldsymbol{O}|/|\boldsymbol{V}|$. It is obvious that as the observation ratio increases, the performance of the BSC algorithm on predicting the start time and spread duration gradually improves. The predicted start time error is smaller than 0.6 for the SWLC network with a large scale, and the error is smaller than 0.3 for other networks. Because the mean of random spread delay of each edge is 1, and the predicted errors for the two factors are smaller than 1, which suggests the BSC algorithm works well on predicting the start time and spread duration of sources.

(iv) Effects of start time intervals on source localization

In this section, we furthermore analyze how the initial spread time span of the source affects source localization.

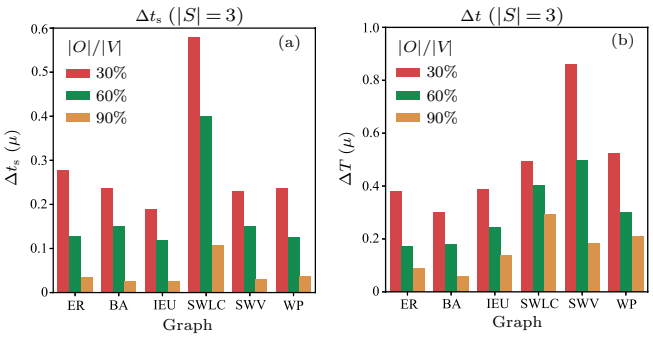


Fig. 6. The average error of the BSC algorithm on predicting start time (a) and spread duration (b) with different $|\boldsymbol{O}|/|\boldsymbol{V}|$ in two synthetic networks and four real networks. The network size of ER network and BA network $|\boldsymbol{V}| = 1000$, and the network average degree $\langle k\rangle = 12$. The source sends information at time intervals $[t^0, t^0+\mu]$ with the critical duration $l = \mu = 1$. The number of diffusion sources $|\boldsymbol{S}| = 3$, and the random spread delay of the edge conforms to $\mathcal{N}(1, 0.3^2)$. The results are obtained with 200 times independent simulations.

Figure 7 shows the AUROC of source localization when the sources trigger the spread at different time intervals in the ER network and the BA network. We can find that when the time intervals is small, say in the range $[t^0, t^0+\mu]$, the AUROC of source localization is high. The AUROC decreases when the range of time interval is larger, however, the AUROC does not significantly decrease as the span of the source's initial spread time increases, say the range $[t^0, t^0+2\mu]$ and $[t^0, t^0+3\mu]$, which indicates that the BSC algorithm can be applied to the multi-source localization with different start times.

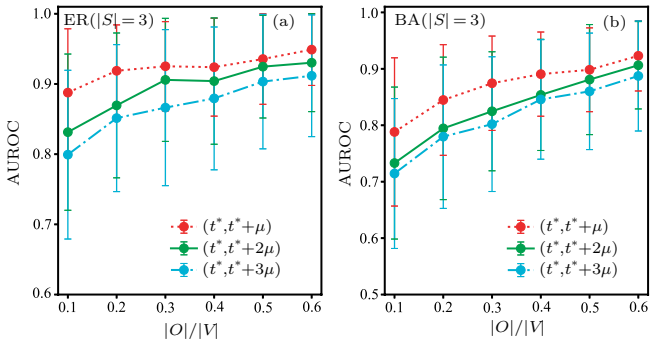


Fig. 7. The relationship between AUROC and the fraction of observers $|\boldsymbol{O}|/|\boldsymbol{V}|$ under the BSC algorithm with different start time intervals in ER networks (a) and in BA networks (b). The network size of ER network and BA network $|\boldsymbol{V}| = 1000$, and the network average degree $\langle k\rangle = 12$. The critical duration $l = \mu = 1$. The number of diffusion sources $|\boldsymbol{S}| = 3$, and the random spread delay of the edge conforms to $\mathcal{N}(1, 0.2^2)$. The results are obtained with 200 times independent simulations, and the error bar represents the standard deviation.

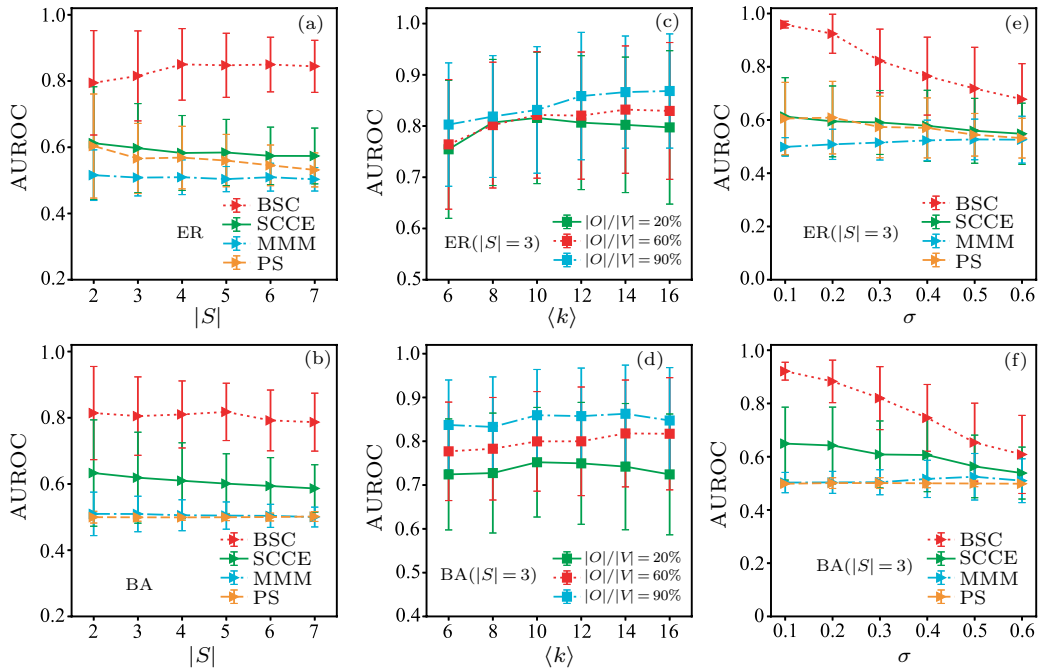


Fig. 8. The relationship between AUROC and the number of sources $|\boldsymbol{S}|$ under the four algorithms with $|\boldsymbol{O}|/|\boldsymbol{V}| = 0.4$ in ER networks (a) and in BA networks (b). The relationship between AUROC and the network average degree $\langle k\rangle$ under the BSC algorithms with different $|\boldsymbol{O}|/|\boldsymbol{V}|$ in ER networks (c) and in BA networks (d). The relationship between AUROC and the standard deviation σ under the four algorithms with $|\boldsymbol{O}|/|\boldsymbol{V}| = 0.4$ in ER networks (e) and in BA networks (f). The network size of ER network and BA network $|\boldsymbol{V}| = 1000$, and the network average degree $\langle k\rangle = 12$. The source sends information at time intervals $[t^0, t^0+\mu]$ with the critical duration $l = \mu = 1$. The number of diffusion sources $|\boldsymbol{S}| = 3$, and the random spread delay of the edge conforms to $\mathcal{N}(1, 0.3^2)$. The results are obtained with 200 times independent simulations, and the error bar represents the standard deviation.

(v) Effects of the number of sources, average degree and spread delay fluctuation on source localization

In this section, we investigate the robustness of the proposed BSC algorithm against the number of sources $|\boldsymbol{S}|$, network structure and spread delay volatility. As depicted from Figs. 8(a) and 8(b), the BSC algorithm maintains the highest performance. With the change of the number of sources, we can see that the AUROC of our method behaves smoothly, while the AUROC of SCCE algorithm shows a downward trend. Figures 8(c) and 8(d) show the performance of the BSC algorithm in ER and BA networks with different $\langle k\rangle$. The experimental results demonstrate that the AUROC of source localization of the BSC algorithm seems to be a gradual upward trend with the increase of network average degree $\langle k\rangle$

when $|\boldsymbol{O}|/|\boldsymbol{V}|$ is high, say 0.4 or 0.6. But in general, the network average degree has little influence on the AUROC, suggesting our method has strong robustness to the network density. Finally, we analyze the impact of the standard deviation σ of spread delay along each edge on source localization. Figures 8(e) and 8(f) illustrate that the AUROC of source localization decreases with the increase of the standard deviation σ, and the downward trend is approximately linear. Compared with these four methods, the AUROC value of the BSC algorithm is the highest.

Next, we examine how the number of sources and propagation delay noise affect the outcomes of source localization in real networks. Figures 9(a)–9(d) illustrates that in the SWV network, AUROC tends to decrease with an increase in the number of sources. Similar trends are observed in the other three real networks compared to the synthetic networks, where the number of sources minimally impacts results of source localization (see Figs. 8(a) and 8(b)). Figures 9(e)–9(h) reveal that, akin to synthetic networks findings, higher noise levels correlate with reduced accuracy of source localization.

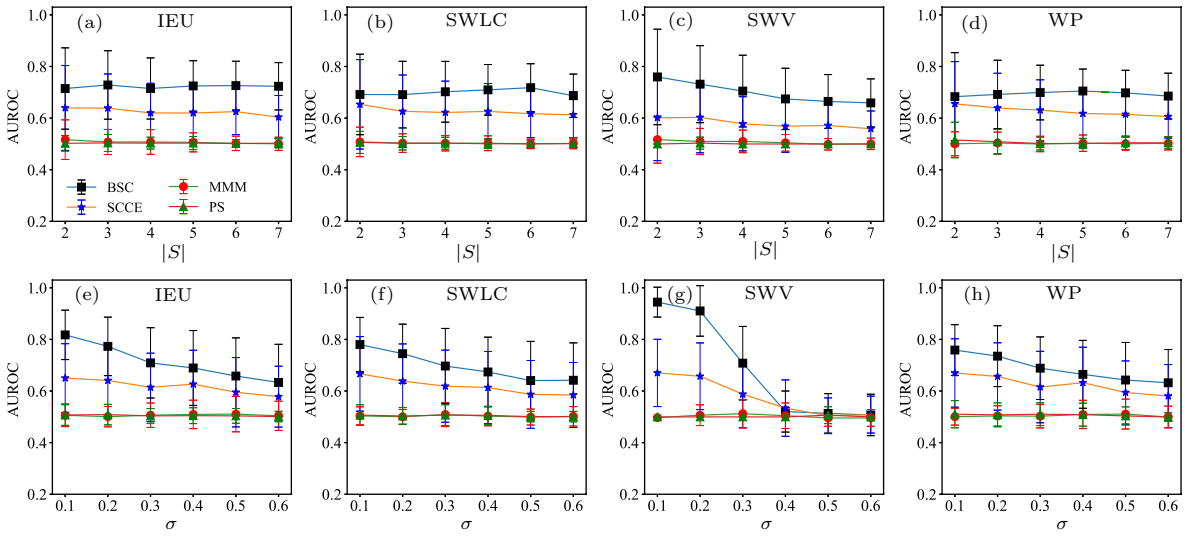


Fig. 9. (a)–(d) The relationship between AUROC and the number of sources $|\boldsymbol{S}|$ under the four algorithms with $|\boldsymbol{O}|/|\boldsymbol{V}| = 0.4$ in the real networks. Here the random spread delay of the edge conforms to $\mathcal{N}(1, 0.3^2)$. (e)–(h) The relationship between AUROC and the standard deviation σ under the four algorithms with $|\boldsymbol{O}|/|\boldsymbol{V}| = 0.4$ in the real networks. Here the number of diffusion sources $|\boldsymbol{S}| = 3$, and the random spread delay of the edge conforms to $\mathcal{N}(1, \sigma^2)$. The source sends information at time intervals $[t^0, t^0 + \mu]$ with the critical duration $l = \mu = 1$. The results are obtained with 200 times independent simulations, and the error bar represents the standard deviation.

Overall, from the simulations illustrated by Figs. 3–9, we can find that our proposed method for source identification is feasible, efficient, and robust. Besides, compared with scale-free network structure, random network structure is more conducive to source identification.

5. Conclusion and prospects

In this paper, in order to effectively control the malicious spread and avoid the harm of malicious information, we have proposed the BSC algorithm based on the backward spread tree theorem and source centrality theorem. The BSC algorithm utilizes the source centrality theorem to infer potential sources in the network, and identifies the sources from the potential sources using linear programming. Besides, The main conclusions can be summarized as follows.

(i) The BSC algorithm has higher AUROC than the other algorithms. An incredible result is that the AUROC of the MMM, which performs well when the initial spread time is the same, is actually close to random guess when the initial spread time is different. As the observer ratio $|\boldsymbol{O}|/|\boldsymbol{V}|$ increases, the BSC algorithm has a significant advantage over the PS and SCCE algorithms with higher performance. Compared with the SCCE and PS algorithms, the AUROC value of our algorithm is generally improved by 0.1 to 0.2.

(ii) The BSC algorithm is robust to network average degree and the number of sources. Despite the performance of source localization decreases with the increase of spread delay noise, the BSC algorithm always works best. Besides, the BSC algorithm works well on predicting the spread duration and the start time. An interesting finding is that, random network structure is more beneficial for source localization, compared with scale-free network structure.

Combined with the results of Table 1 and Fig. 3, it is difficult to see whether the clustering coefficient is related to source localization. In future study, we will adopt synthetic networks and make in-depth analysis by adjusting the clustering coefficient to investigate whether or how the clustering coefficient affects the results of source localization. In this paper, we identify the source only through the shortest path by using the backward spread method, without considering the influence of multiple shortest paths. If multiple shortest paths are also taken into account, the accuracy may be further improved.[57] We assumed that the network topology

remained unchanged during the information spread process. However, the spread of information in real-world is often accompanied by the change of network structure.[51,71–73] Additionally, we assumed that all nodes received the information, whereas in practical applications, identifying the source is necessary when some nodes receive the malicious information to prevent it from breaking out. As posited in Ref. [22], it is assumed that the propagation delay across each edge conforms to a Gaussian distribution with a mean of 1. However, real-world propagation delays may adhere to different distributions, suggesting that alternative methods could potentially offer more practical descriptions. In the future, we will try to study source localization for the real propagation, such as the spread of COVID-19. However, the real propagation often contains the time-varying characteristics and incomplete measurement information, which prompts us to optimize the algorithm to meet the requirements of practical applications. In fact, in online social networks, information usually has multimodal characteristics, such as text, pictures, videos, how to detect false information[74] and locate the source is also worth further research.[75,76]

Acknowledgements

Project supported by the National Natural Science Foundation of China (Grant Nos. 62103375, 62006106, 61877055, and 62171413), the Philosophy and Social Science Planning Project of Zhejinag Province, China (Grant No. 22NDJC009Z), the Education Ministry Humanities and Social Science Foundation of China (Grant No. 19YJCZH056), and the Natural Science Foundation of Zhejiang Province, China (Grant Nos. LY23F030003, LY22F030006, and LQ21F020005).

References

[1] Vosoughi S, Roy D and Sinan Aral S 2018 *Science* **359** 1146
[2] Zarin R, Khaliq H, Khan A, Khan D, Akgül D and Humphries U W 2022 *Results Phys.* **33** 105130
[3] Pastor-Satorras R, Castellano C, Van Mieghem P and Vespignani A 2015 *Rev. Mod. Phys.* **87** 925
[4] Morone F and Makse H A 2015 *Nature* **524** 65
[5] Su Z, Chen L, Ai J, Zheng Y Y and Bie N 2024 *Chin. Phys. B* **33** 058901
[6] Xie M, Zhan X X, Liu C and Zhang Z K 2023 *Inf. Process. Manag.* **60** 103161
[7] Hu Z L, Liu J G, Yang G Y and Ren Z M 2014 *Europhys. Lett.* **106** 18002
[8] Shao Y, Chen L, Chen Y, Liu W and Dai C 2023 *Inf. Sci.* **635** 375
[9] Yang P L, Zhao L J, Dong C, Xu G X and Zhou L X 2023 *Chin. Phys. B* **32** 058901
[10] Liu C and Zhang Z K 2014 *Commun. Nonlinear Sci.* **19** 896
[11] Gao L, Wang W, Pan L, Tang M and Zhang H 2016 *Sci. Rep.* **6** 1
[12] Shah D and Zaman T 2010 *ACM SIGMETRICS Perform. Eval. Rev.* **38** 203
[13] Xu X, Qian T, Xiao Z, Zhang N, Wu J and Zhou F 2024 *Expert Syst. Appl.* **238** 122028
[14] Comin C H and Fontoura Costa L 2011 *Phys. Rev. E* **84** 056105
[15] Shi C, Zhang Q and Chu T 2022 *Chin. Phys. B* **31** 070203
[16] Luo W, Tay W P and Leng M 2013 *IEEE Trans. Signal Process.* **61** 2850
[17] Louni A and Subbalakshmi K 2018 *IEEE Trans. Comput. Soc. Syst.* **5** 335
[18] Chen Z, Zhu K and Ying L 2016 *IEEE Trans. Netw. Sci. Eng.* **3** 17
[19] Liu Y, Li W, Yang C and Wang J 2022 *Sci. Rep.* **12** 5467
[20] Zhu P, Cheng L, Gao C, Wang Z and Li X 2022 *IEEE Trans. Netw. Sci. Eng.* **9** 1853
[21] Yuan S, Liu W, Zeng H and Wang C 2023 *Europhys. Lett.* **141** 60001
[22] Pedro C P, Patrick T and Martin V 2012 *Phys. Rev. Lett.* **109** 068702
[23] Zhang X, Zhang Y, Lv T and Yin Y 2016 *Physica A* **442** 100
[24] Hu Z L, Han X, Lai Y C and Wang W X 2017 *Roy. Soc. Open Sci.* **4** 170091
[25] Fu L, Shen Z, Wang W X, Fan Y and Di Z 2016 *Europhys. Lett.* **113** 18006
[26] Hu Z L, Shen Z, Han J, Peng H, Lu J F, Jia R, Zhu X B and Zhao D 2019 *Physica A* **527** 121262
[27] Wang H J, Zhang F F and Sun K J 2021 *Phys. Lett. A* **393** 127184
[28] Hu Z L, Han J, Peng H, Lu J F, Zhu X B, Jia R and Li M 2022 *IEEE Trans. Netw. Sci. Eng.* **9** 3515
[29] Robert P, Lu X, Krzysztof S, Szymański B K and Hołyst J A 2018 *Sci. Rep.* **8** 2508
[30] Tang W, Ji F and Tay W P 2018 *IEEE Trans. Inf. Foren. Sec.* **13** 3035
[31] Wang C, Man Z, Yuan S and Zhang G 2022 *Europhys. Lett.* **139** 21002
[32] Wang H J, Hu Z L, Tao L, Shao S and Wang S Z 2023 *Sci. Rep.* **13** 5692
[33] Yang F, Li C, Peng Y, Liu J, Yao Y, Wen J and Yang S 2023 *Soft Comput.* **27** 16059
[34] Wang H J, Hu Z L, Shao S Y, Zhang F F and Tao L 2024 *Phys. Rev. E* **109** 014311
[35] Hu Z L, Wang H J, Sun L, Tang C B and Li M 2024 *Expert Syst. Appl.* **256** 124946
[36] Ma Z W, Wang H J, Hu Z L, Zhu X B, Huang Y Z and Huang F L 2024 *Phys. Lett. A* **523** 129772
[37] Yin F, Shao X, Ji M and Wu J 2021 *J. Med. Int. Res.* **23** e25734
[38] Nsoesie E O, Cesare N, Müller M and Ozonoff A 2020 *J. Med. Int. Res.* **22** e24425
[39] Ji F, Tay W P and Varshney L R 2017 *IEEE Trans. Signal Process.* **65** 2517
[40] Zhu K and Ying L 2014 *IEEE/ACM Trans. Netw.* **24** 408
[41] Prakash B A, Vreeken J and Faloutsos C 2014 *Knowl. Inf. Syst.* **38** 35
[42] Jiang J, Wen S, Yu S, Xiang Y and Zhou W 2015 *IEEE Trans. Inf. Foren. Sec.* **10** 2616
[43] Ali S S, Anwar T, Rastogi A and Rizvi S A M 2019 *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, November 3–7, 2019, Beijing, China, p. 891
[44] Lokhov A Y, Mézard M, Ohta H and Zdeborová L 2014 *Phys. Rev. E* **90** 012801
[45] Altarelli F, Braunstein A, Alejandro L D, Lage-Castellanos A and Zecchina R 2014 *Phys. Rev. Lett.* **112** 118701
[46] Wang Z, Wang C, Pei J and Ye X 2017 *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, February 4–9, 2017, San Francisco, California, USA, p. 217
[47] Peng S L, Wang H J, Peng H, Zhu X B, Li X, Han J, Zhao D and Hu Z L 2023 *Chaos* **33** 083125
[48] Cheng L, Li X, Han Z, Luo T, Ma L and Zhu P 2022 *Chaos, Solitons and Fractals* **159** 112139
[49] Shen Z, Cao S, Wang W X, Di Z and Stanley H E 2016 *Phys. Rev. E* **93** 032301
[50] Paluch R, Gajewski L G, Hołyst J A and Szymanski B K 2020 *Future Gener. Comput. Syst.* **112** 1070
[51] Hu Z L, Shen Z, Cao S, Podobnik B, Yang B, Wang W X and Lai Y C 2018 *Sci. Rep.* **8** 2685
[52] Hu Z L, Wang L and Tang C B 2019 *Chaos* **29** 063117
[53] Spinelli B, Celis L E and Thiran P 2019 *IEEE Trans. Netw. Sci. Eng.* **6** 86
[54] Wang H J and Sun K J 2020 *Europhys. Lett.* **131** 48001
[55] Gajewski Ł, Paluch R, Suchecki K, Sulik A, Szymanski B K and Hołyst J A 2022 *Sci. Rep.* **12** 5079
[56] Ma Z W, Sun L, Ding Z G, Huang Y Z and Hu Z L 2024 *Chin. Phys. B* **33** 028902
[57] Gajewski Ł, Szymanski B K and Hołyst J A 2019 *Physica A* **519** 34
[58] Hu Z L, Shen Z, Tang C B, Xie B B and Lu J F 2018 *Phys. Lett. A* **382** 931

[59] Hanley J A and McNeil B J 1982 *Radiology* **143** 29
[60] Prasse B and Van Mieghem P 2020 *IEEE Trans. Netw. Sci. Eng.* **7** 2755
[61] Ma C, Chen H S, Li X, Lai Y C and Zhang H F 2020 *SIAM J. Appl. Dyn. Syst.* **19** 124
[62] Pech R, Hao D, Lee Y L, Yuan Y and Zhou T 2019 *Physica A* **528** 121319
[63] Ahmad I, Akhtar M U, Noor S and Shahnaz A 2020 *Sci. Rep.* **10** 364
[64] Erdős P and Rényi A 1960 *Publ. Math. Inst. Hungarian Acad. Sci.* **5** 17
[65] Barabási A and Albert R 1999 *Science* **286** 509
[66] Guimera R, Danon L, Diaz-Guilera A, Giralt F and Arenas A 2003 *Phys. Rev. E* **68** 065103
[67] Rossi R and Ahmed N 2015 *Proceedings of the AAAI conference on artificial intelligence*, January 25–30, 2015, Austin, Texas, USA, p. 4292
[68] Leskovec J, Huttenlocher D and Kleinberg J 2010 *Proceedings of the SIGCHI conference on human factors in computing systems*, April 10–15, 2010, Atlanta, Georgia, USA, p. 1361
[69] Adamic L A and Glance N 2005 *Proceedings of the 3rd international workshop on Link discovery*, August 21–25, 2005, New York, USA, p. 36
[70] Newman M E 2002 *Phys. Rev. Lett.* **89** 208701
[71] Holme P and Saramäki J 2012 *Phys. Rep.* **519** 97
[72] Jiang J, Wen S, Yu S, Xiang Y and Zhou W 2018 *IEEE Trans. Depend. Sec. Comput.* **15** 166
[73] Chai Y, Wang Y and Zhu L 2021 *IEEE Trans. Inf. Foren. Sec.* **16** 2621
[74] Zhang H, Qian S, Fang Q and Xu C 2020 *IEEE Trans. Multimedia* **23** 4441
[75] Ling C, Jiang J, Wang J and Liang Z 2022 *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, August 14–18, 2022, Washington, DC, USA, p. 1010
[76] Wan P, Wang X, Pang G, Wang L and Min G 2023 *Expert Syst. Appl.* **213** 119239