Knowledge-Based Systems 195 (2020) 105674

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Locating the propagation source in complex networks with a direction-induced search based Gaussian estimator☆

Fan Yang [a], Shuhong Yang [a], Yong Peng [a], Yabing Yao [b], Zhiwen Wang [a], Houjun Li [a], Jingxian Liu [a], Ruisheng Zhang [c], Chungui Li [a,*]

[a] School of Computer Science and Communication Engineering, Guangxi University of Science and Technology, Liuzhou 545006, China
[b] School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China
[c] School of Information Science and Engineering, Lanzhou University, Lanzhou 730030, China

ARTICLE INFO

Article history:
Received 19 July 2019
Received in revised form 27 December 2019
Accepted 18 February 2020
Available online 22 February 2020

Keywords:
Complex networks
Propagation source locating
Gaussian estimator (GE)
Direction-induced search (DIS)
Direction-induced search based Gaussian estimator (DISGE)

ABSTRACT

Locating the propagation source is crucial for developing strategies to control the spreading process taking on complex networks. Gaussian estimator (GE) is one of the most effective methods for propagation source locating in networks with limited observers. However, on general graphs, due to GE makes an approximation that the actual diffusion tree in spreading process is assumed to be a breadth-first search (BFS) tree, and thus ignores the effect of direction information recorded in observers. Therefore, the accuracy of GE is affected. In this paper, by utilizing the direction information in observers, we define, for the first time, a novel direction-induced search (DIS). Further, a direction-induced search based Gaussian estimator (DISGE) is proposed by combining DIS and the original GE. Experimental results on a series of synthetic and real networks show that the DISGE is feasible and effective in locating the propagation source with limited observers.

1. Introduction

In the modern world, the ubiquity of the spreading phenomena occurring on various networks has incurred huge losses to human society. Some typical examples include computer virus propagation [1], disease spreading [2] and rumours diffusion [3]. Obviously, it is of great theoretical and practical significance to develop effective strategies to control the spreading process on networks. As one of the significant measures, propagation source locating has attracted widespread attentions, many excellent methods are proposed in recent years [4]. For example, the Gaussian estimator [5], the effective distance [6] and so on. Meanwhile, how to further improve the accuracy of source locating methods is still a challenging task.

Gaussian estimator (GE) [5] is an effective strategy for propagation source locating in networks with limited observers. GE is optimal on general trees because making use of both the direction and timing information recorded in observers. However, on general graphs, due to assuming that the actual diffusion tree is a breadth-first search (BFS) tree, the direction information is ignored. Thus, its accuracy is affected. In this paper, following the GE, we focus on improving GE on general graphs by utilizing the direction information recorded in observers. We define, for the first time, a novel direction-induced search (DIS) that takes advantage of the direction information, and further propose a direction-induced search based Gaussian estimator (DISGE) with computational complexity $O(MN)$. The feasibility and effectiveness of DISGE are verified on a series of synthetic and real networks.

This paper is organized as follows. The related works is reviewed in Section 2. In Section 3, we introduce the original Gaussian estimator. The DISGE is proposed in Section 4. In Section 5, we verify the performance of DISGE. We conclude this work in Section 6.

2. Related work

The systematic study of propagation source locating was pioneered by Shah et al. [7,8], they designed a novel topological quantity and proposed a source estimator termed as Rumour Centrality. Zhu et al. [9] developed a sample path based method named Jordan Center. Luo et al. [10] investigated the performance of Jordan Center on different spreading model. Lokhov et al. [11] presented a source inference algorithm based on Dynamic Message Passing equations. Altarelli et al. [12] identified the origin

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2020.105674.

* Corresponding author.
E-mail addresses: fanyang2014@lzu.edu.cn (F. Yang), lcggxust@163.com (C. Li).
URL: http://www.gxust.edu.cn (F. Yang).

https://doi.org/10.1016/j.knosys.2020.105674
0950-7051/

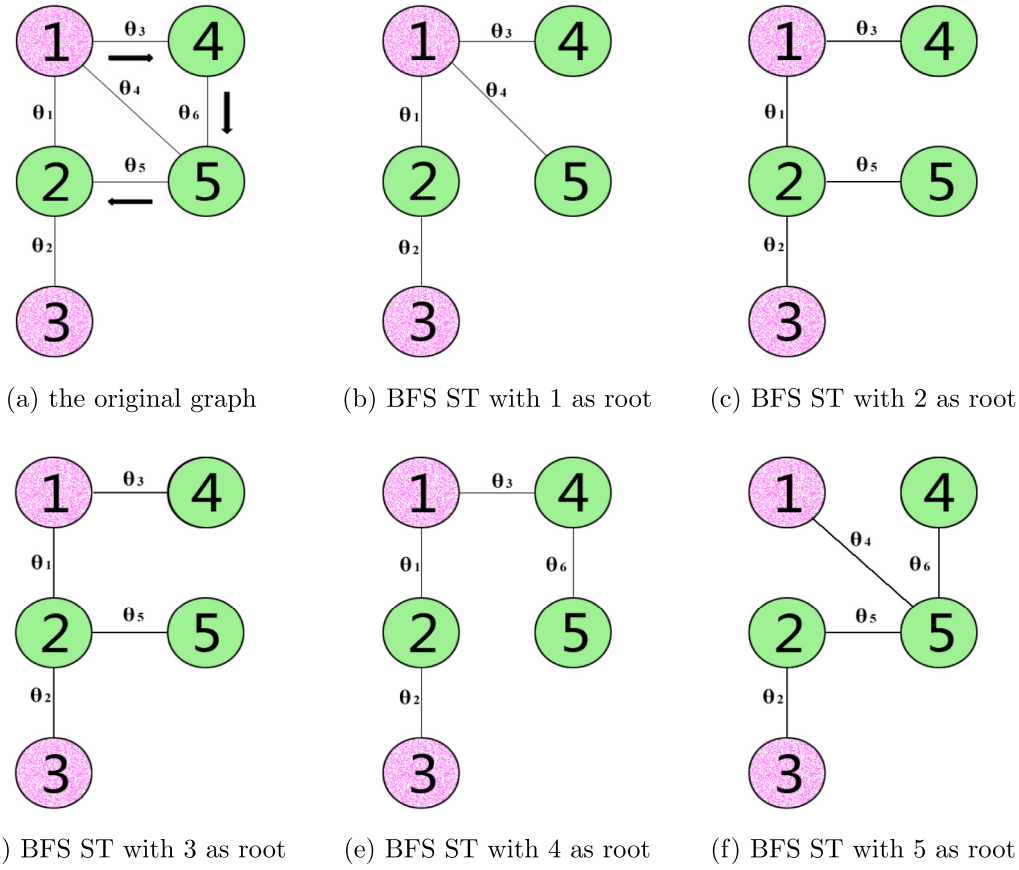


Fig. 1. (a) shows a schematic diagram, nodes 2, 4 and 5 are observers, all of them record the Direction and Timing information. The arrows in (a) show the Direction information. (b)–(f) show the different BFS spanning trees with nodes 1–5 as root, respectively. BFS ST denotes BFS spanning tree.

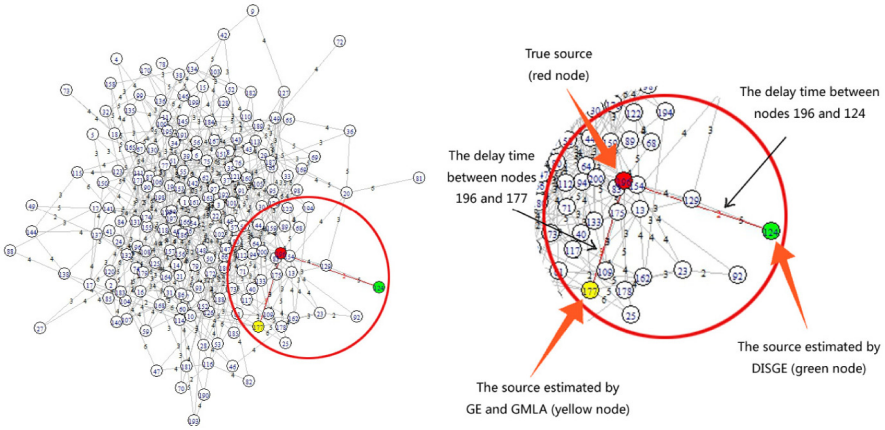
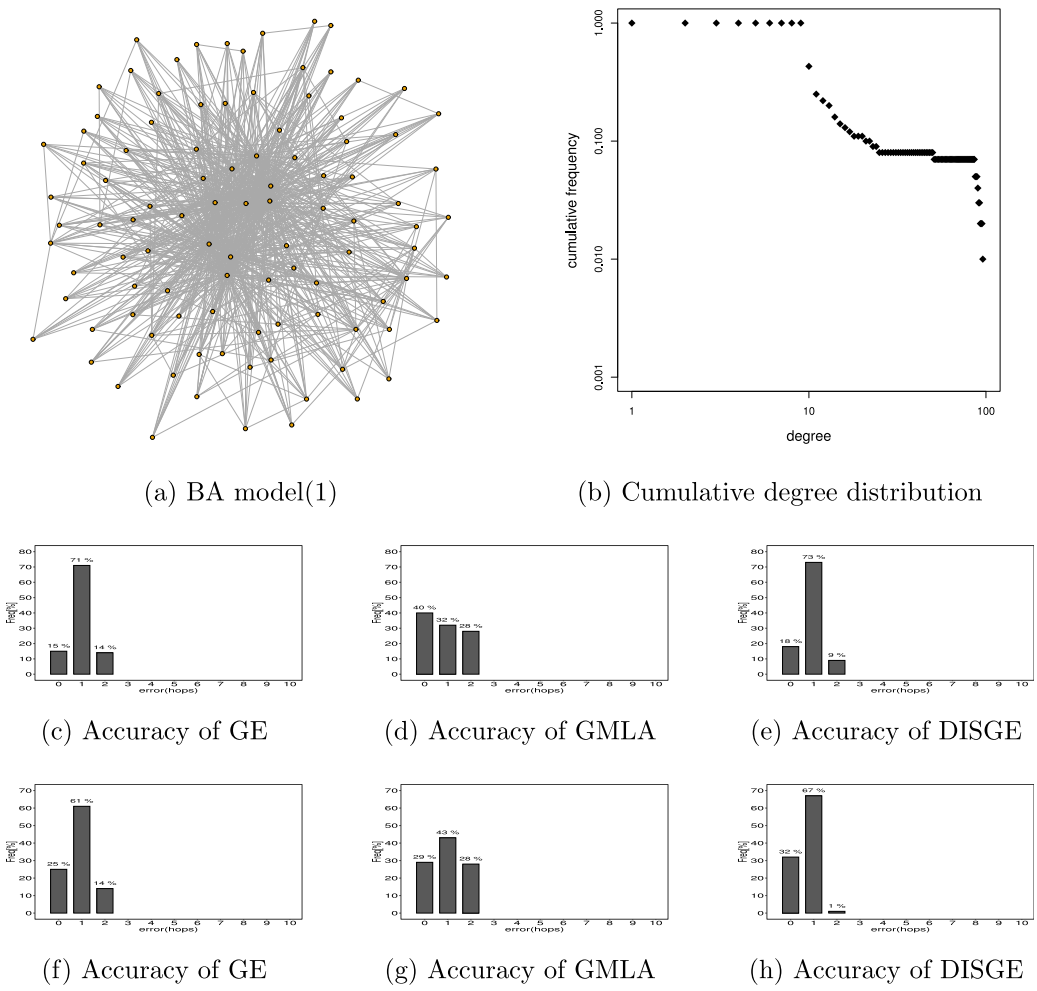


Fig. 2. The propagation source locating results of GE and GMLA (node 177, marked with yellow) and DISGE (node 124, marked with green). The true source is node 196(marked with red). The error hops between nodes 177 and 196 is 1, which is consistent with the error hops between nodes 124 and 196. However, the time delay between nodes 177 and 196 is 3 (marked with red), which is different from the time delay between nodes 124 and 196 (the value is 2, marked with red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of an epidemic outbreak based on Belief Propagation equations. Antulov et al. [13] presented a Soft-Margin estimator based on Monte-Carlo method. Yang et al. [14] defined a rationality observation value and proposed a Propagation Centrality. Cai et al. [15] proposed an estimation framework by which the location of source, spreading ratio and starting time can be estimated simultaneously. Refs. [7–14] mainly focused on locating a single source in networks. For multiple sources, Luo et al. [16] extended the Rumour Centrality [7,8]. Prakash et al. [17] proposed a method based on minimum description length. Although the above works can locate the propagation source effectively, they are mainly used for unweighted networks.

In practice, the edges in a network is usually associated with various weight, such as traffic, propagation delay and so on. Brockmann et al. [6] modelled the Global Mobility Network as weighted graph, and located the propagation source based on a novel effective distance. Based on the effective distance, Jiang et al. [18] proposed a K-centre method to identify multiple diffusion sources, Manitz et al. [19] applied the effective distance in public transportation networks. The effective distance based



(a) BA model(1) (b) Cumulative degree distribution

(c) Accuracy of GE (d) Accuracy of GMLA (e) Accuracy of DISGE

(f) Accuracy of GE (g) Accuracy of GMLA (h) Accuracy of DISGE

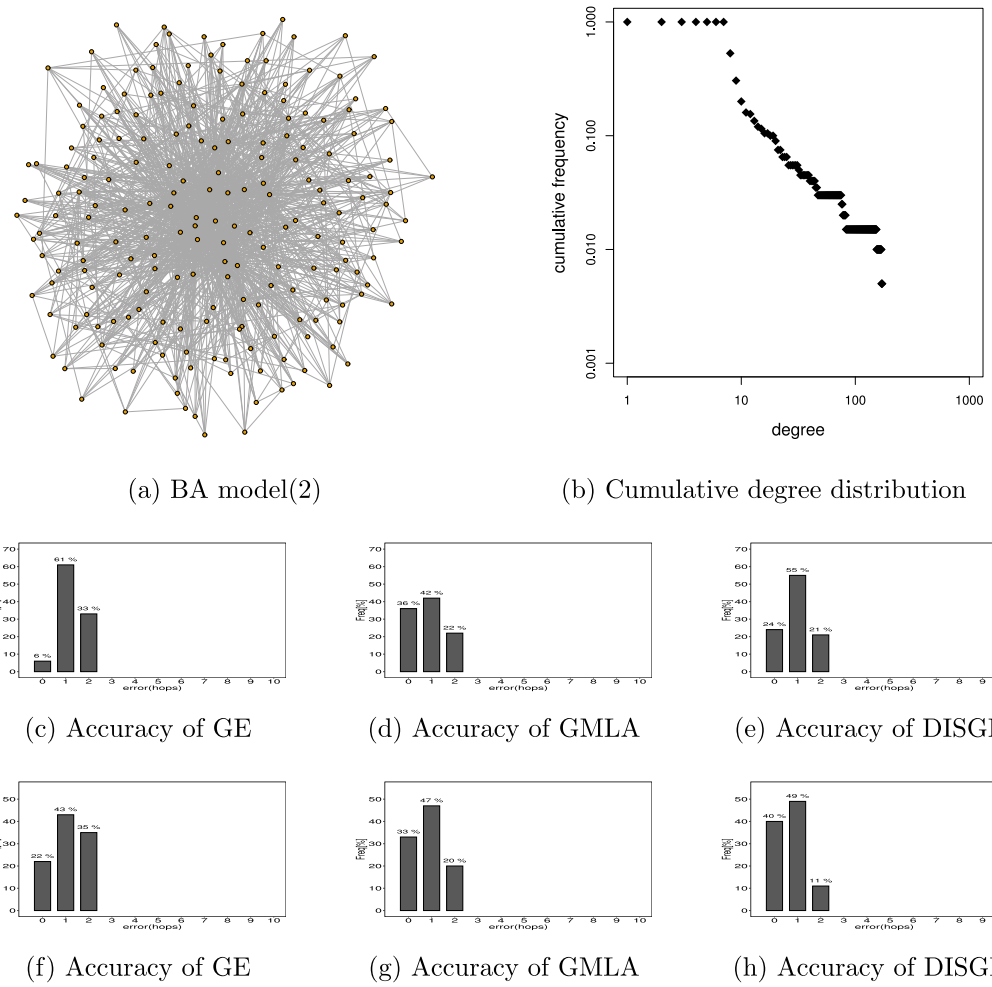
Fig. 3. The error hops between the estimated source and the true source obtained by GE, GMLA and DISGE. Each subplot is obtained by 100 runs. (c)–(e) show the error hops of $\beta = 0.25$. (f)–(h) show the error hops of $\beta = 0.50$.

methods must obtain the state of each node in a network. However, it is often the case that only limited nodes state can be observed in reality [20]. By assuming that the propagation delays associated with edges follows Gaussian distribution, Pinto et al. [5] proposed a Gaussian estimator for the networks with limited observers, which is optimal on general trees and sub-optimal on general graphs. Following the Gaussian estimator, Paluch et al. [21] ignored the observers with low quality information and thus reduced the complexity of Gaussian estimator. However, its average accuracy may be affected. Gajewski et al. [22] considered the existence of multiple shortest paths between nodes to improve the Gaussian estimator. While, the calculation of multiple shortest paths will increase the computational time. Besides, with limited observers, Shen et al. [23] developed a time-reversal backward spreading algorithm to locate multiple propagation sources. Li et al. [24] provided a probabilistic method to locate the source via parameter estimation and maximum likelihood estimation. Wang et al. [25] proposed an online regression model for real-time diffusion source identification. However, these methods [5,21–25] did not consider the effect of direction information reflected by observers. In fact, the direction information is beneficial to accurately locate the propagation source [26]. Another interesting work is presented by Ji et al. [27], in the process of source locating, they utilized Gromov matrices to construct reasonable spanning trees.

In this paper, by utilizing the direction information recorded in observers, we define, for the first time, a novel direction-induced search (DIS). Then, we combine DIS with the original Gaussian estimator to propose a direction-induced search based Gaussian estimator (DISGE). Different from the Ref. [27], we use DIS to generate reasonable spanning trees.

3. Gaussian estimator

Network model and spreading process. In the Ref. [5], the underlying network on which the spreading process occurs is modelled as a finite and undirected graph $\mathcal{G} = \{V, E, \boldsymbol{\theta}\}$, where V and E represent the nodes set and edges set, respectively, $\theta_{vu} \in \boldsymbol{\theta}$ denotes the random propagation delay associated with an edge $vu \in E$. The random variables $\{\theta_{vu}\}$ for different edges vu have a known arbitrary joint distribution. The spreading process is modelled as follows. Each node $v \in \mathcal{G}$ is only in one of the two states: (*i*) informed, if it has received the information from any one neighbour, or (*ii*) ignorant, if it has not been informed so far. The spreading process is initiated by a single source s^*, all nodes are ignorant except for s^* is informed. Let $\mathcal{N}(v)$ denote the neighbour(s) of node v. Suppose v is in the ignorant state at time t_v, and receives the information for the first time from one neighbour s, thus becoming informed. Then, v will attempt to retransmit the information to all its other neighbours, so that each neighbour $u \in \mathcal{N}(v) \setminus s$ receives the information at time $t_v + \theta_{vu}$. The spreading process is terminated when there is no retransmission occurs.



(a) BA model(2) (b) Cumulative degree distribution

(c) Accuracy of GE (d) Accuracy of GMLA (e) Accuracy of DISGE

(f) Accuracy of GE (g) Accuracy of GMLA (h) Accuracy of DISGE

Fig. 4. The error hops between the estimated source and the true source obtained by GE, GMLA and DISGE. Each subplot is obtained by 100 runs. (c)–(e) show the error hops of $\beta = 0.25$. (f)–(h) show the error hops of $\beta = 0.50$.

Gaussian estimator for general graphs. Let $\mathcal{O} \triangleq \{o_k\}_{k=1}^{K} \subset V$ denote the K observers. The location of $\mathcal{O}$ on $\mathcal{G}$ is known. Each observer measures *from which neighbour* and *at what time* it received the information. Generally, only a subset $V_\mathrm{a} \subseteq V$ will be informed, and only a subset $K_\mathrm{a} \leq K$ observers are active. Correspondingly, there is a subgraph $\mathcal{G}_\mathrm{a} \subseteq \mathcal{G}$ and $\mathcal{O}_\mathrm{a} \subseteq \mathcal{O}$, where $\mathcal{G}_\mathrm{a} = \{V_\mathrm{a}, E_\mathrm{a}, \boldsymbol{\theta}_\mathrm{a}\}$, $\boldsymbol{\theta}_\mathrm{a} \triangleq \{\theta_1, \ldots, \theta_{E_a}\}$, $\mathcal{O}_\mathrm{a} \triangleq \{o_k\}_{k=1}^{K_\mathrm{a}} \subset V_\mathrm{a}$. All the nodes state in $\mathcal{G}$ are unknown except for the K observers, and what can be used for source locating is nothing but only the K_a active observers and the information recorded by them. The goal is to locate the source s^* from the measurements taken at the active observers set $\mathcal{O}_\mathrm{a}$ on $\mathcal{G}$. In fact, when the information is spread on $\mathcal{G}$, there is a tree corresponding to the first time each node gets informed, which spans all nodes in $\mathcal{G}$. However, the number of spanning trees of $\mathcal{G}$ can be exponentially large. Thus, an approximation is made in Ref. [5] by assuming that the actual propagation tree is a breadth-first search (BFS) tree. Then, the Gauss estimator for general graphs can be written as,

$$\hat{s} = \underset{s \in \mathcal{G}}{\operatorname{argmax}}\ \mathcal{S}\left(s, \mathbf{d}, \mathcal{T}_{\mathrm{BFS},s}\right) \tag{1}$$

where $\mathcal{S}\left(s, \mathbf{d}, \mathcal{T}_{\mathrm{BFS},s}\right) = \boldsymbol{\mu}_s^{\mathrm{T}} \boldsymbol{\Lambda}_s^{-1} \left(\mathbf{d} - \frac{1}{2}\boldsymbol{\mu}_s\right)$, $\mathbf{d}$ is the observed delay vector, $\mathbf{d} \triangleq \left[d_1, \ldots, d_{K_a-1}\right]^{\mathrm{T}}$, where,

$$d_k \triangleq t_{k+1} - t_1 = \sum_{i \in \mathcal{P}\left(s^*, o_{k+1}\right)} \theta_i - \sum_{i \in \mathcal{P}\left(s^*, o_1\right)} \theta_i \tag{2}$$

where $t_k = t^* + \sum_{i \in \mathcal{P}(s^*, o_k)} \theta_i$, $\mathcal{P}(v, u)$ denote the set of edges (*path*) connecting node v and u. θ_i denotes the corresponding propagation delay. The parameter $\boldsymbol{\mu}_s$ is the deterministic delay vector, $\boldsymbol{\mu}_s \triangleq \left[d_1, \ldots, d_{K_a-1}\right]^{\mathrm{T}}$, where,

$$[\boldsymbol{\mu}_s]_k = \mu \cdot \left(|\mathcal{P}(s, o_{k+1})| - |\mathcal{P}(s, o_1)|\right) \tag{3}$$

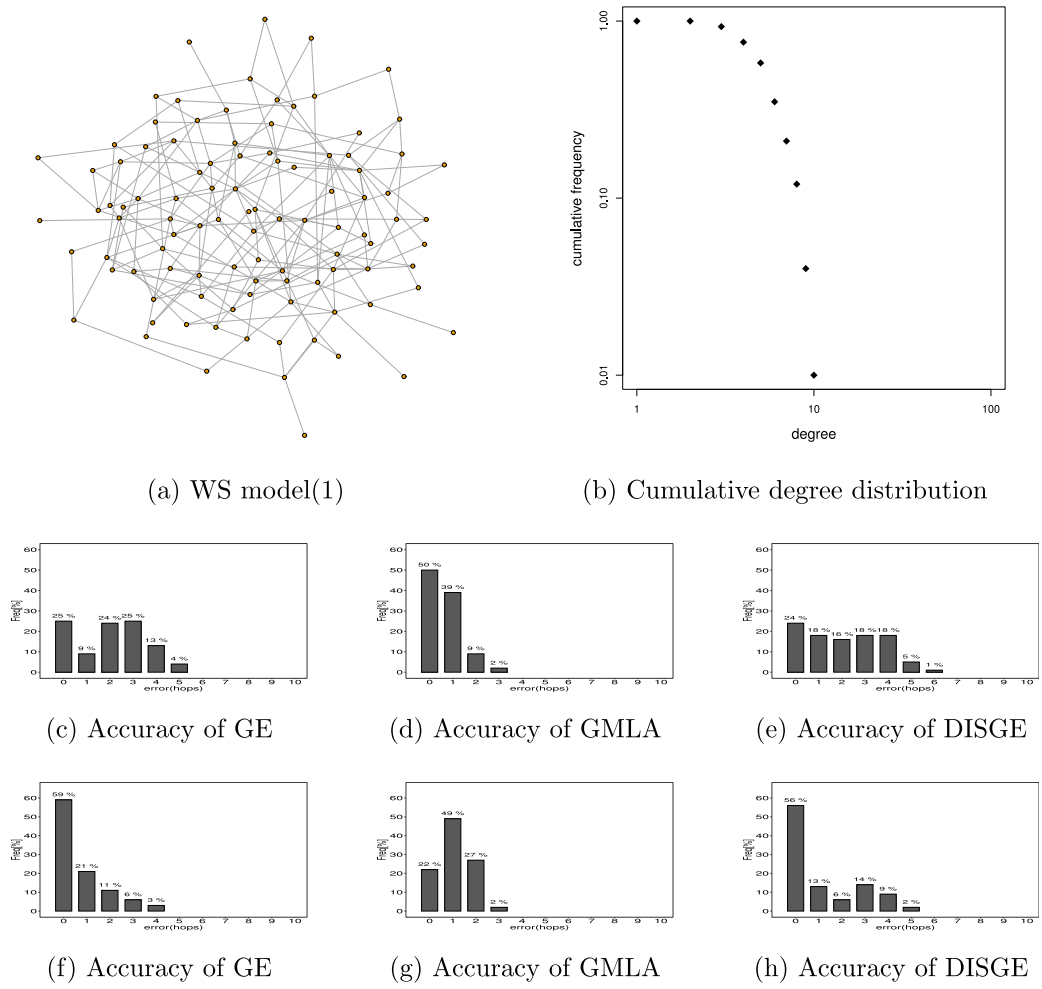
the parameter $\boldsymbol{\Lambda}_s$ is the delay covariance matrix,

$$[\boldsymbol{\Lambda}]_{k,i} = \sigma^2 \cdot \begin{cases} |\mathcal{P}(o_1, o_{k+1})|, & k = i \\ |\mathcal{P}(s, o_{k+1})| \cap |\mathcal{P}(s, o_{i+1})|, & k \neq i \end{cases} \tag{4}$$

for $k, i = 1, \ldots, K_a - 1$, and $|\mathcal{P}(v, u)|$ denoting the number of edges (*length*) of the path connecting nodes v and u. $\boldsymbol{\mu}_s$ and $\boldsymbol{\Lambda}_s^{-1}$ are computed with respect to the BFS spanning tree $\mathcal{T}_{\mathrm{BFS},s}$ rooted at s. The procedure for source locating of GE on general graphs is summarized in Algorithm 1.

4. Direction-induced search based Gaussian estimator

From the Refs. [5,28], we know that the observations made by the K_a active observers provide two types of information: *(i)the Direction in which information arrives to the active observers and, (ii)the Timing at which the information arrives to the active observers.* For general trees, Gaussian estimator takes advantage of the Direction and Timing information to estimate the propagation source, and get an optimal result. For general graphs, the authors assumed that the actual diffusion tree is a BFS tree, which makes



(a) WS model(1) (b) Cumulative degree distribution

(c) Accuracy of GE (d) Accuracy of GMLA (e) Accuracy of DISGE

(f) Accuracy of GE (g) Accuracy of GMLA (h) Accuracy of DISGE

Fig. 5. The error hops between the estimated source and the true source obtained by GE, GMLA and DISGE. Each subplot is obtained by 100 runs. (c)–(e) show the error hops of $\beta = 0.25$. (f)–(h) show the error hops of $\beta = 0.50$.

Algorithm 1: Gaussian estimator for general graphs [28]

```
01  select one arrival time as reference, and label it t_1
02  compute the delay vector d relative to t_1
03  for each node s ∈ G do
04    compute the spanning tree T_{bfs,s} rooted at s
05    compute the μ_s and Λ_s with respect to tree T_{bfs,s}
06    compute the source likelihood in formula (1) for node s
07  end for
08  pick ŝ according to the maximization in formula (1)
```

utilizing the correct Direction information impossible. A typical example is shown in Fig. 1. In Fig. 1(a), nodes 2, 4 and 5 are set as observers. All of them contain Direction and Timing information. However, in Fig. 1(b)–(f), with different node as root, there is no any BFS spanning tree could reflect the correct Direction information recorded in nodes 2, 4 and 5 simultaneously. Obviously, on general graphs, the correct Direction information recorded in observers is not utilized to generate spanning tree, and thus is helpless in locating the propagation source, only the Timing information is utilized. The above reason partly explains why the accuracy of GE may be affected on general graphs.

In this paper, *we focus on improving the original Gaussian estimator on general graphs by sufficiently utilizing the Direction information recorded in observers*. We also model the underlying network on which spreading process occurs as a finite and undirected graph $\mathcal{G} = \{V, E, \boldsymbol{\theta}\}$, and adopt the same spreading process in Section 3.

Suppose all nodes in $\mathcal{G}$ are set as observers (i.e. $\mathcal{O} = V$) and each node has been informed (i.e. $\mathcal{O}_{\mathrm{a}} = V$). When the spreading process starting with a single source s^* is terminated, a corresponding $\mathcal{G}_{\mathrm{a}} = \{V_{\mathrm{a}}, E_{\mathrm{a}}, \boldsymbol{\theta}_{\mathrm{a}}\}$ can be obtained. Then we have Proposition 1,

Proposition 1. *Given an arbitrary $\mathcal{G}$, the corresponding $\mathcal{G}_{\mathrm{a}}$, and an arbitrary observer set $\mathcal{O} \triangleq \{o_k\}_{k=1}^{K}$. Suppose $\mathcal{O} = V$ and $\mathcal{O}_{\mathrm{a}} = V$, with the Direction information recorded in each o_k, the spanning tree corresponding to each node $v \in V_{\mathrm{a}}$ first time got informed can be uniquely determined.*

The proof of Proposition 1 can be found in Appendix A.1.

Proposition 1 indicates that the true spreading tree can be reconstructed by utilizing the Direction information recorded in observers. Correspondingly, with Proposition 1, we propose a novel direction-induced search (DIS) which can be used for reconstructing the true spanning tree. DIS is summarized in Algorithm 2.

Analysis of Algorithm 2. The set $E_{\mathrm{a}}(\mathcal{T})$ in line 4 is to record the true propagation directions, where, the propagation directions are reflected by directed edges. Lines 5 to 15 implement the process of traversing $\mathcal{G}_{\mathrm{a}}$ as well as record the true propagation directions. In this process, the meaning of lines 7 to 14 is that, for

Algorithm 2: direction-induced search (DIS)

```
01  for each node s ∈ V_a do
02    let Q be a queue
03    Q.enqueue(s) and set s as visited
04    initialize a set E_a(T) = NULL
05    while Q is not empty do
06      v = Q.dequeue
07      for each neighbour u of v
08        if u has not been visited then
09          if the Direction recorded in u is v then
10            Q.enqueue(u) and set u as visited
11            E_a(T) = E_a(T) ∪ e_{v→u}
12          end if
13        end if
14      end for
15    end while
16    remove the edges not in E_a(T) from G_a and the
      remaining part of G_a is denoted by T
17    if (|E_a(T)| == |V_a|−1) then
18      mark T as DIS spanning tree T_{dis,s} and mark
        the current node s as the source s*
19      break
20    end if
21  end for
```

*where, $\mathcal{T}_{\text{dis},s}$ denotes a DIS spanning tree rooted at s. $e_{v\to u}$ in line 11 represents the true propagation direction from v to u.

Algorithm 3: direction-induced search Gaussian estimator

```
01  select one arrival time as reference, and label it t_1
02  compute the delay vector d relative to t_1
03  for each node s ∈ V do
04    let Q be a queue
05    Q.enqueue(s) and set s as visited
06    initialize a set E(T) = NULL
07    while Q is not empty do
08      v = Q.dequeue
09      for each neighbour u of v
10        if u has not been visited then
11          if u ∈ O_a then
12            if the Direction recorded in u is v then
13              Q.enqueue(u) and set u as visited
14              E(T) = E(T) ∪ e_{v→u}
15            end if
16          else
17            Q.enqueue(u) and set u as visited
18            E(T) = E(T) ∪ e_{vu}
19          end if
20        end if
21      end for
22    end while
23    remove the edges not in E(T) from G and the
      remaining part of G is denoted by T
24    if (|E(T)| == |V|−1) then
25      mark T as DIS spanning tree T_{dis,s}
26      compute the μ_s and Λ_s with respect to T_{dis,s}
27      compute the source likelihood in formula (5) for node s
28    end if
29  end for
30  pick ŝ according to the maximization in formula (5)
```

*where, $\mathcal{T}_{\text{dis},s}$ denotes a spanning tree rooted at s. The $e_{v\to u}$ in step 14 represents the true propagation direction from v to u, while, the e_{vu} in step 18 represents that the propagation direction is assumed from v to u.

each unvisited node u, there is only one true direction from which the information is passed to u. In lines 9 to 12, the corresponding directed edge containing the true Direction information will be recorded in $E_{\text{a}}(\mathcal{T})$. This process requires $O(M)$ computations, where $M \triangleq |E_{\text{a}}|$. Line 16 requires $O(M+N)$ computations, which can be reduced to $O(M)$, where, $N \triangleq |V_{\text{a}}|$. In lines 17 to 20, the $\mathcal{T}$ got in line 16 is marked as DIS spanning tree and the current node s is marked as the true propagation source s^* if and only if $(|E_{\text{a}}(\mathcal{T})| == |V_{\text{a}}| - 1)$. From Proposition 1, we know that DIS spanning tree is unique. Therefore, Algorithm 2 will be terminated after the DIS spanning tree being constructed. Finally, to reconstruct the true spreading tree, in the worst case, each node $s \in V_{\text{a}}$ may be used for constructing spanning tree. Thus, the complexity of Algorithm 2 is $O(MN)$.

With the unique spanning tree determined by DIS, we have the following proposition,

Proposition 2. *The unique true spreading tree reconstructed by DIS uniquely determines the true propagation source s^* in $\mathcal{G}_{\text{a}}$.*

The proof of Proposition 2 can be found in Appendix A.2.

Ideally, if $\mathcal{O} = V$ and $\mathcal{O}_{\text{a}} = V$, the Direction information recorded in each node can be obtained, then the true propagation source can be precisely located. However, it may be a huge cost to set so many nodes as observers in practice, and it is not necessary that all nodes are informed. Generally, for the active observer set $\mathcal{O}_{\text{a}} \triangleq \{o_i\}_{k=1}^{K_{\text{a}}}$, there is $K_{\text{a}} \ll |V|$. Thus, it is very difficult to precisely reconstruct the spreading tree corresponding to the first time each node got informed.

Although the accuracy of GE may be affected on general graphs, it is a good method in approximately estimating the propagation source with limited observers. Besides, since all nodes state are unknown except for the observers, thus, each node may be the possible propagation source candidate. Then, we modify the original DIS in Algorithm 2 and combine with GE to propose a direction-induced search based Gaussian estimator (DISGE), which can be written as follows,

$$\hat{s} = \underset{s\in\mathcal{G}}{\operatorname{argmax}}\ \mathcal{S}\left(s, \mathbf{d}, \mathcal{T}_{\text{dis},s}\right) \tag{5}$$

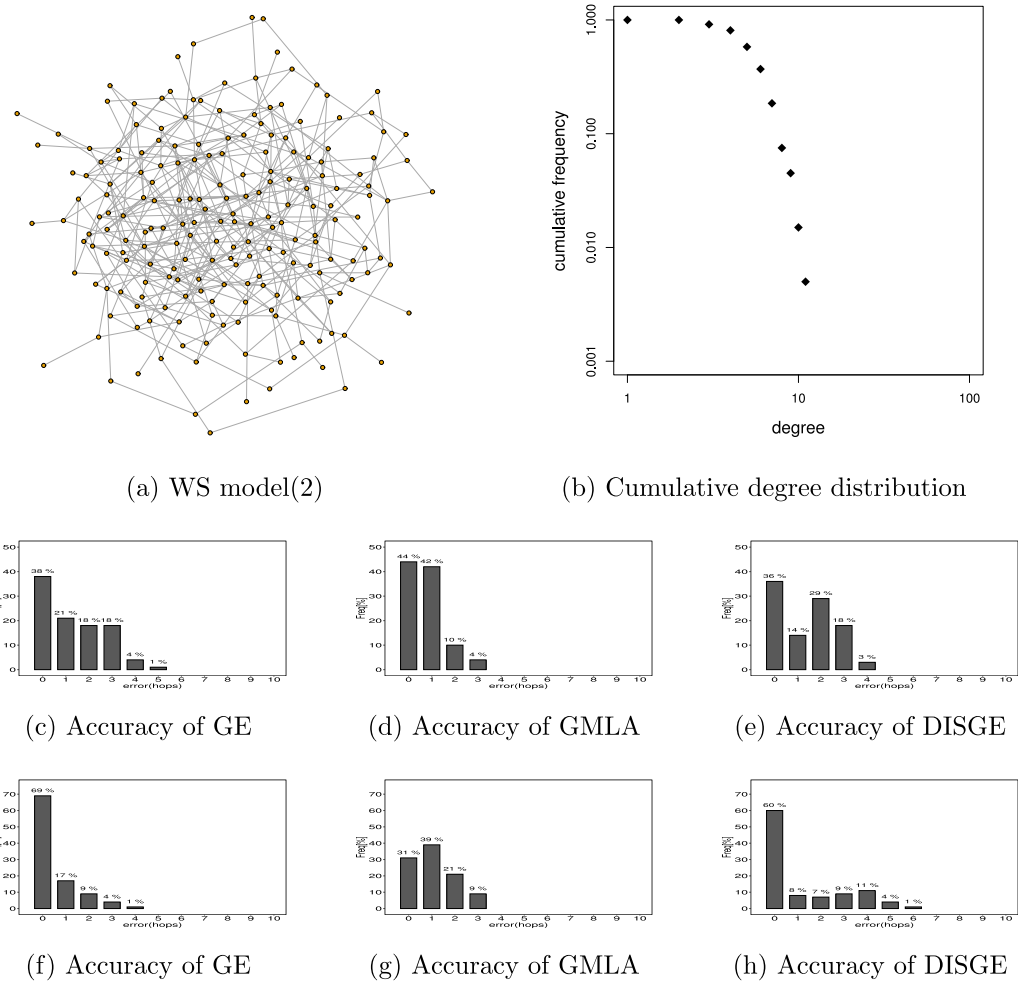
where $\mathcal{S}\left(s, \mathbf{d}, \mathcal{T}_{\text{dis},s}\right) = \boldsymbol{\mu}_s^{\mathrm{T}}\boldsymbol{\Lambda}_s^{-1}\left(\mathbf{d} - \frac{1}{2}\boldsymbol{\mu}_s\right)$, with parameters $\boldsymbol{\mu}_s$ and $\boldsymbol{\Lambda}_s^{-1}$ computed with respect to the DIS tree $\mathcal{T}_{\text{dis},s}$ rooted at s. $\mathbf{d}$ is the observed delay vector, $\boldsymbol{\mu}_s$ is deterministic delay vector, $\boldsymbol{\Lambda}_s$ is the delay covariance matrix. Correspondingly, we present a DISGE algorithm to locate the propagation source on general graphs, which is summarized in Algorithm 3.

Analysis of Algorithm 3. The set $E(\mathcal{T})$ in line 6 is to record the true propagation directions, where, the propagation direction is reflected by directed edges. Lines 7 to 22 is to implement the process of traversing $\mathcal{G}$ as well as record the true propagation directions, which requires $O(M)$ computations. Here, from line 11 to 19, for each unvisited node u, it has only one of the two different identity, observer or not. For the former, node u will be set as visited if and only if it is informed by current v (lines 11 to 15). For the latter, u will be set as visited by assuming that it is informed by current v (lines 16 to 19). Line 23 requires $O(M+N)$ computations, which can be reduced to $O(M)$. From line 24 to 28, $\mathcal{T}$ will be marked as $\mathcal{T}_{\text{dis},s}$ if and only if $(|E(\mathcal{T})| == |V| - 1)$. Meanwhile, the parameters of $\mathcal{T}_{\text{dis},s}$ and source likelihood of node s will be calculated. Obviously, the details of constructing the DIS spanning tree in Algorithm 3 is different from Algorithm 2. Since only part of nodes in $\mathcal{G}$ become active observers, only the Direction information recorded in these observers are utilized to locate the source. Thus, we relax the condition for constructing DIS spanning tree. Finally, taking the loop in line 3 into account, the time complexity of Algorithm 3 is $O(MN)$.

5. Experimental results

To verify the performance of DISGE, we compare it with the original GE [5] and its improved method GMLA [21]. Their time complexity are listed in Table 1.

Experimental environment. Hardware: Dell R740 with 2 Intel® Xeon® gold 6254 CPU and 256 GB RAM. Software: Cygwin 3.0.7 + Eclipse Cpp 2019 + igraph C 0.7.1 (used for computation). R 64x 3.3.3+igraph R 1.2.1 (used for generating synthetic networks and plotting figures).



(a) WS model(2) (b) Cumulative degree distribution

(c) Accuracy of GE (d) Accuracy of GMLA (e) Accuracy of DISGE

(f) Accuracy of GE (g) Accuracy of GMLA (h) Accuracy of DISGE

Fig. 6. The error hops between the estimated source and the true source obtained by GE, GMLA and DISGE. Each subplot is obtained by 100 runs. (c)–(e) show the error hops of $\beta = 0.25$. (f)–(h) show the error hops of $\beta = 0.50$.

Table 1
Time complexity of GE, GMLA and DISGE.

	GE	GMLA	DISGE
Complexity	$O(N^3)$	$O(N^2 log\ (N))$	$O(MN)$

M and N denote the size of nodes and edges of $\mathcal{G}$, respectively.

Table 2
The parameters for generating synthetic networks.

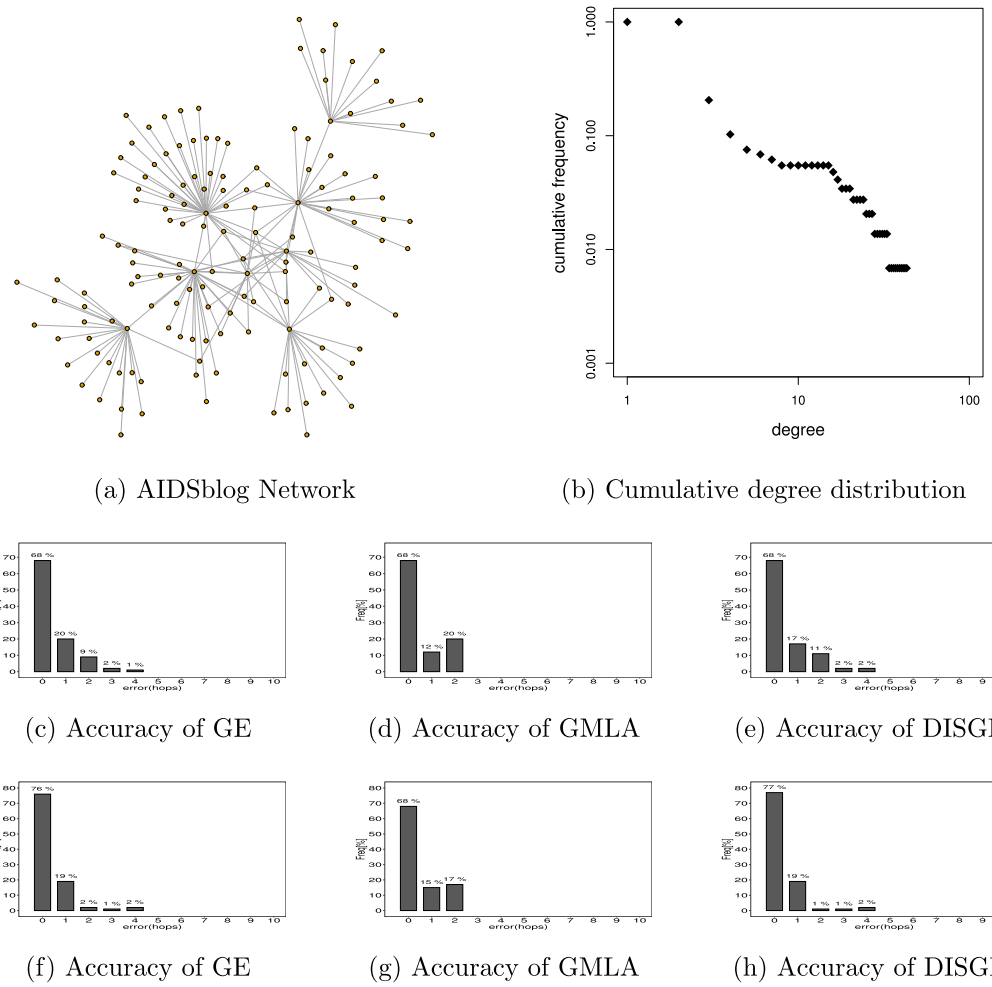
Networks	Parameters
BA model(1)	barabasi.game(100, power = 3, m = 8, directed = FALSE)
BA model(2)	barabasi.game(200, power = 2, m = 6, directed = FALSE)
WS model(1)	watts.strogatz.game(1, 100, 2, 1.0)
WS model(2)	watts.strogatz.game(1, 200, 2, 0.7)

Software: R 64x 3.3.3, igraph R 1.2.1.

Datasets. We evaluate the accuracy of the three methods on a series of synthetic and real networks. Synthetic networks include the small-world (WS) model [29] and scale-free (BA) model [30]. The parameters for generating synthetic networks are listed in Tables 2 and B.8. Real networks are selected from different fields, which can be obtained from Network Data Repository [31] and Koblenz Network Collection http://konect.uni-koblenz.de/ for free, including AIDSblog [32] http://math.bu.edu/people/kolaczyk/datasets/AIDSBlog.zip, PDZBase http://konect.uni-koblenz.de/networks/maayan-pdzbase, USAirlines http://networkrepository.com/inf-USAir97.php, NetScience http://networkrepository.com/ca-netscience.php, Celegans http://konect.uni-koblenz.de/networks/arenas-meta and Euroroad http://konect.uni-koblenz.de/networks/subelj_euroroad. The topology properties of these networks are shown in Table 3.

Parameter settings. For an arbitrary $\mathcal{G} = \{V, E, \boldsymbol{\theta}\}$, the propagation delays $\boldsymbol{\theta}$ associated with E are independent identically distributed (IID) Gaussian $\mathcal{N}\left(\mu, \sigma^2\right)$. Here, $\mu/\sigma = 4$, which is consistent with Ref. [5]. We adopt the same spreading process as Ref. [5] (which has been described in Section 3), and the propagation ratio is $\beta = \sigma/\mu = 0.25$. In addition, we also consider the case of $\beta = 0.5$. It is no doubt that, similar with the original GE, the accuracy of GMLA and DISGE will be affected by the observer density. To fairly compare the performance of the three methods, the size of parameter K0 in GMLA is equals to the size of observers in GE and DISGE. Besides, on different networks, with different propagation ratio β, the number of generated observers will be different. Thus, for different β, we set different observer density on different networks, the corresponding observer density are listed in Table 4. All observers are randomly selected.

Evaluation methodology. Generally, the distance between the estimated propagation source and the true source is the common metric to measure the accuracy of a propagation source locating method, which is termed as error hops. However, in a network, with propagation delay as the weight of edges, the source estimated by different methods may be consistent in their error hops



(a) AIDSblog Network (b) Cumulative degree distribution

(c) Accuracy of GE (d) Accuracy of GMLA (e) Accuracy of DISGE

(f) Accuracy of GE (g) Accuracy of GMLA (h) Accuracy of DISGE

Fig. 7. The error hops between the estimated source and the true source obtained by GE, GMLA and DISGE. Each subplot is obtained by 100 runs. (c)–(e) show the error hops of $\beta = 0.25$. (f)–(h) show the error hops of $\beta = 0.50$.

Table 3
The topology properties of networks.

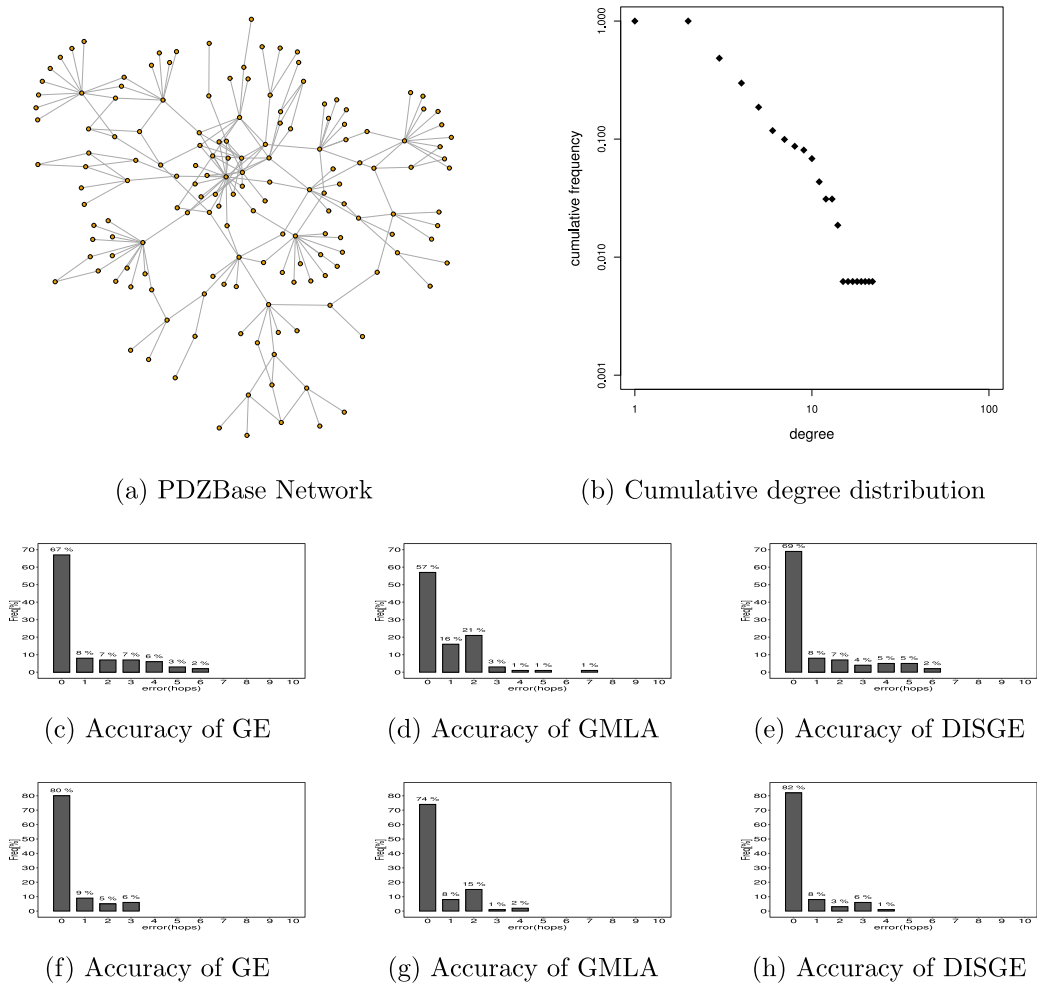
Networks	N	E	R	$\langle k\rangle$	$\langle k^2\rangle$	C	A	H	β_c	APL
BA model(1)	100	764	7.64	15.28	672.28	0.228	−0.660	2.88	0.023	1.85
BA model(2)	200	1179	5.90	11.79	593.70	0.111	−0.414	4.27	0.020	1.95
WS model(1)	100	200	2.00	4.00	19.44	0.035	−0.075	1.22	0.206	3.43
WS model(2)	200	400	2.00	4.00	19.22	0.012	−0.002	1.20	0.208	3.98
AIDSblog	146	180	1.23	2.47	36.44	0.023	−0.725	5.99	0.068	3.42
PDZBase	161	209	1.30	2.60	15.25	0.003	−0.466	2.63	0.170	5.33
USAirlines	332	2126	6.40	12.81	568.16	0.396	−0.208	3.46	0.023	2.74
NetScience	379	914	2.41	4.82	38.69	0.431	−0.082	1.66	0.125	6.04
Celegans	453	2025	4.47	8.94	358.49	0.124	−0.226	4.49	0.025	2.66
Euroroad	1039	1305	1.26	2.51	7.75	0.035	0.090	1.23	0.324	18.40

N and E denote the size of nodes set and edges set in a network. $R = \frac{E}{N}$. $\langle k\rangle$ and $\langle k^2\rangle$ denote the average degree and the 2-order average degree. C denotes the clustering coefficient. A denotes the assortative coefficient [33]. H denotes the degree heterogeneity [34–36], $H = \frac{\langle k^2\rangle}{\langle k\rangle^2}$. β_c denotes the theoretical epidemic threshold [37], $\beta_c = \frac{\langle k\rangle}{\langle k^2\rangle}$. APL denotes the average path length (the number of edges).

but different in their propagation delays. A typical example is shown in Fig. 2. We can see that GE, GMLA and DISGE are consistent in their error hops. However, GE and GMLA are consistent in their time delays, while both of them are different from DISGE. In this paper, we will adopt the error hops as well as the time delays to evaluate the performance of the three methods.

Table 5 shows the average error hops of GE, GMLA and DIGE on the ten networks. Table 6 shows the average time delays on the ten networks. Table 7 shows the average execution time ratio on the ten networks.

Figs. 3 and 4 show the experimental results of GE, GMLA and DISGE on the BA model(1) and BA model(2), respectively. In Fig. 3(c)–(e), when $\beta = 0.25$, the precise locating (0 error hops) rates of GE, GMLA and DISGE are 15 percent, 40 percent and 18 percent, respectively. From Tables 5 and 6, we can see that the average error hops of GE, GMLA and DISGE are 0.99, 0.88 and 0.91, respectively, and the average time delays are 4.23, 3.36 and 3.51, respectively. In Fig. 3(f)–(h), when $\beta = 0.50$, the precise locating (0 error hops) rates are 25 percent, 29 percent and 32 percent, respectively. From Table 5, we can see that the average error



(a) PDZBase Network (b) Cumulative degree distribution

(c) Accuracy of GE (d) Accuracy of GMLA (e) Accuracy of DISGE

(f) Accuracy of GE (g) Accuracy of GMLA (h) Accuracy of DISGE

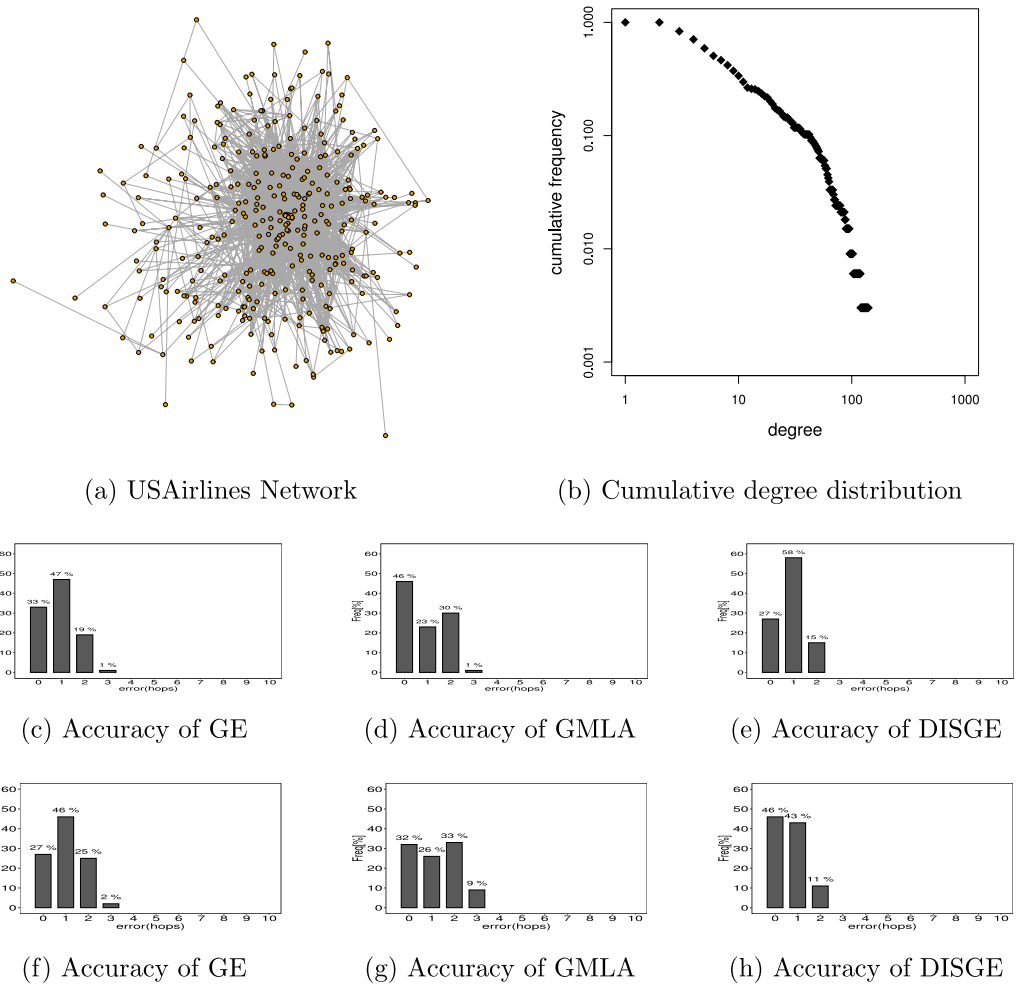
Fig. 8. The error hops between the estimated source and the true source obtained by GE, GMLA and DISGE. Each subplot is obtained by 100 runs. (c)–(e) show the error hops of $\beta = 0.25$. (f)–(h) show the error hops of $\beta = 0.50$.

Table 4
The observer density on ten networks.

Network	$\beta = 0.25$	$\beta = 0.50$
BA model(1)	0.1	0.3
BA model(2)	0.1	0.3
WS model(1)	0.1	0.3
WS model(2)	0.1	0.3
AIDSblog	0.1	0.3
PDZBase	0.05	0.2
USAirlines	0.1	0.2
NetScience	0.02	0.1
Celegans	0.1	0.2
Euroroad	0.01	0.02

hops of GE, GMLA and DISGE are 0.89, 0.99 and 0.69, respectively. From Table 6, we can see that the average time delays are 3.43, 3.42 and 2.81, respectively. In Fig. 4(c)–(e), when $\beta = 0.25$, the precise locating (0 error hops) rates of GE, GMLA and DISGE are 6 percent, 36 percent and 24 percent. From Tables 5 and 6, we can see that the average error hops of GE, GMLA and DISGE are 1.27, 0.86 and 0.97, respectively, and the average time delays are 5.09, 3.31 and 3.91, respectively. In Fig. 4(f)–(h), when $\beta = 0.50$, the precise locating (0 error hops) rates are 22 percent, 33 percent and 40 percent, respectively. From Table 5, we can see that the average error hops of GE, GMLA and DISGE are 1.13, 0.87 and 0.71, respectively. From Table 6, we can see that the average time delays are 4.51, 3.17 and 2.75, respectively. Overall, on the BA models generated with different parameters, when $\beta = 0.25$, the accuracy of DISGE is superior to GE but inferior to GMLA. However, when $\beta = 0.50$, DISGE is superior to both GE and GMLA.

Figs. 5 and 6 show the experimental results of GE, GMLA and DISGE on the WS model(1) and WS model(2), respectively. In Fig. 5(c)–(e), when $\beta = 0.25$, the precise locating (0 error hops) rates of GE, GMLA and DISGE are 25 percent, 50 percent and 24 percent. From Tables 5 and 6, we can see that the average error hops of GE, GMLA and DISGE are 2.04, 0.63 and 2.07, respectively, and the average time delays are 8.18, 2.31 and 8.30, respectively. In Fig. 5(f)–(h), when $\beta = 0.50$, the precise locating (0 error hops) rates are 59 percent, 22 percent and 56 percent, respectively. From Table 5, we can see that the average error hops of GE, GMLA and DISGE are 0.73, 1.09 and 1.13, respectively. From Table 6, we can see that the average time delays are 2.74, 4.14 and 4.45, respectively. In Fig. 6(c)–(e), when $\beta = 0.25$, the precise locating (0 error hops) rates of GE, GMLA and DISGE are 38 percent, 44 percent and 36 percent. From Tables 5 and 6, we can see that the average error hops of GE, GMLA and DISGE are 1.32, 0.74 and 1.38, respectively, and the average time delays are 5.08, 2.77 and 5.27, respectively. In Fig. 6(f)–(h), when $\beta = 0.50$, the precise locating (0 error hops) rates are 69 percent, 31 percent and 60 percent, respectively. From Table 5, we can see that the average error hops of GE, GMLA and DISGE are 0.51, 1.08 and 1.19, respectively. From Table 6, we can see that the average time delays are 2.14, 4.20 and 4.79, respectively. Overall, on the WS models generated



(a) USAirlines Network (b) Cumulative degree distribution

(c) Accuracy of GE (d) Accuracy of GMLA (e) Accuracy of DISGE

(f) Accuracy of GE (g) Accuracy of GMLA (h) Accuracy of DISGE

Fig. 9. The error hops between the estimated source and the true source obtained by GE, GMLA and DISGE. Each subplot is obtained by 100 runs. (c)–(e) show the error hops of $\beta = 0.25$. (f)–(h) show the error hops of $\beta = 0.50$.

Table 5
The average error hops of GE, GMLA and DISGE on ten networks.

Network	GE	GMLA	DISGE	GE	GMLA	DISGE
β	$\beta = 0.25$	$\beta = 0.25$	$\beta = 0.25$	$\beta = 0.50$	$\beta = 0.50$	$\beta = 0.50$
BA model(1)	0.99	0.88	0.91	0.89	0.99	0.69
BA model(2)	1.27	0.86	0.97	1.13	0.87	0.71
WS model(1)	2.04	0.63	2.07	0.73	1.09	1.13
WS model(2)	1.32	0.74	1.38	0.51	1.08	1.19
AIDSblog	0.48	0.52	0.53	0.34	0.49	0.32
PDZBase	0.94	0.83	0.91	0.37	0.49	0.36
USAirlines	0.88	0.86	0.88	1.02	1.19	0.65
NetScience	1.27	0.65	1.24	0.67	0.60	0.66
Celegans	1.48	1.13	1.44	1.33	1.18	0.77
Euroroad	1.30	0.91	1.27	0.98	1.22	1.29

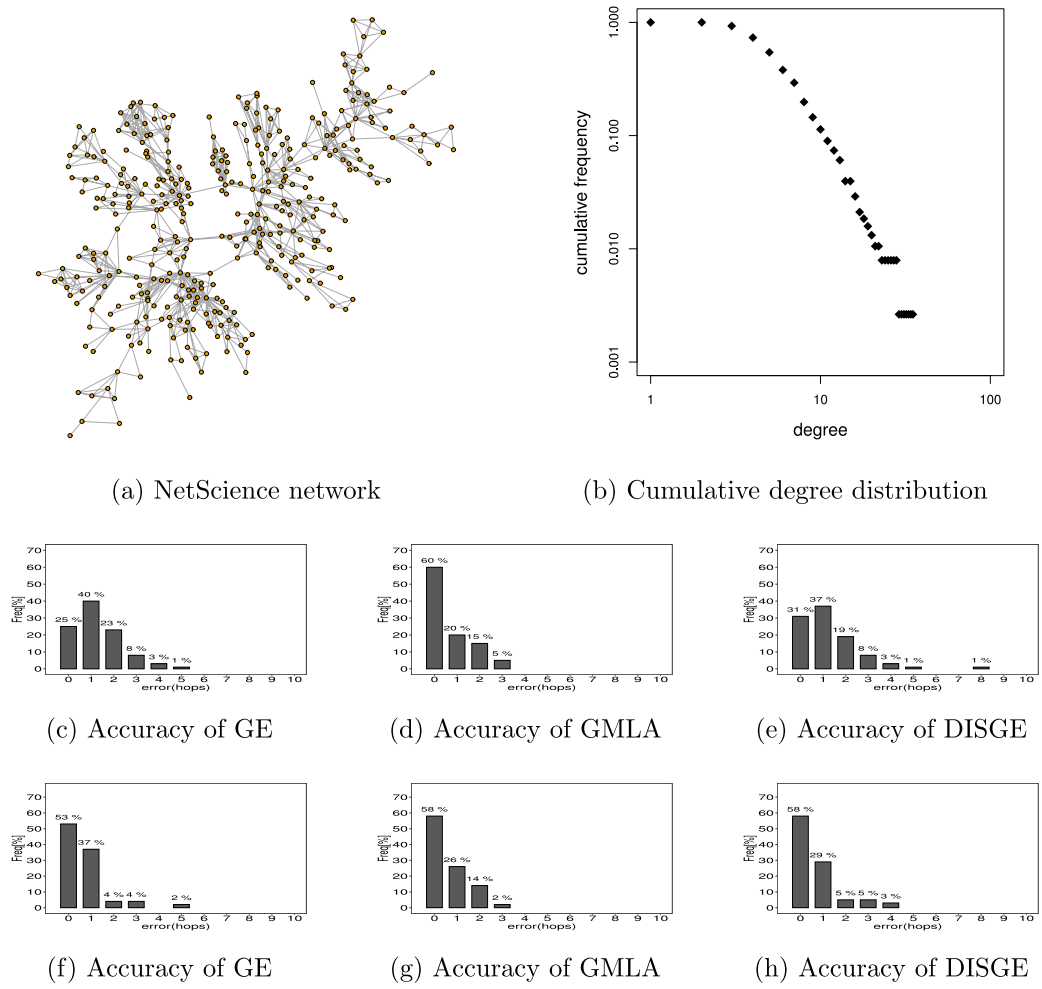
β denotes the propagation ratio in the spreading process.

with different parameters, when $\beta = 0.25$, DISGE is similar with GE but inferior to GMLA. When $\beta = 0.50$, GE exposes the best performance in the three methods.

From the results shown in Figs. 3–6 and Tables 5–6, we can see that the accuracy of DISGE on the BA models is obviously superior to the one on the WS models. Combining with Table 3, we know that the heterogeneity (H value) of BA models is obviously greater than the one of WS models. Further, by the experiments on a sequence of synthetic networks (which can be found in Appendix B), we find that, in the networks with $H > 2.5$, when $\beta = 0.50$, DISGE outperforms GE and GMLA. Next, we will further verify this conclusion on real networks.

5.1. Real networks

Fig. 7 shows the experimental results of GE, GMLA and DISGE on AIDSblog network. In Fig. 7(c)–(e), when $\beta = 0.25$, the precise locating (0 error hops) rates of GE, GMLA and DISGE are 68 percent, 68 percent and 68 percent, respectively. From Tables 5 and 6, we can see that the average error hops of GE, GMLA and DISGE are 0.48, 0.52 and 0.53, respectively, the average time delays are 1.87, 2.15 and 2.07, respectively. In Fig. 7(f)–(h), when $\beta = 0.50$, the precise locating (0 error hops) rates are 76 percent, 68 percent and 77 percent, respectively. From Table 5, we can see that the average error hops are 0.34, 0.49 and 0.32, respectively.



(a) NetScience network (b) Cumulative degree distribution

(c) Accuracy of GE (d) Accuracy of GMLA (e) Accuracy of DISGE

(f) Accuracy of GE (g) Accuracy of GMLA (h) Accuracy of DISGE

Fig. 10. The error hops between the estimated source and the true source obtained by GE, GMLA and DISGE. Each subplot is obtained by 100 runs. (c)–(e) show the error hops of $\beta = 0.25$. (f)–(h) show the error hops of $\beta = 0.50$.

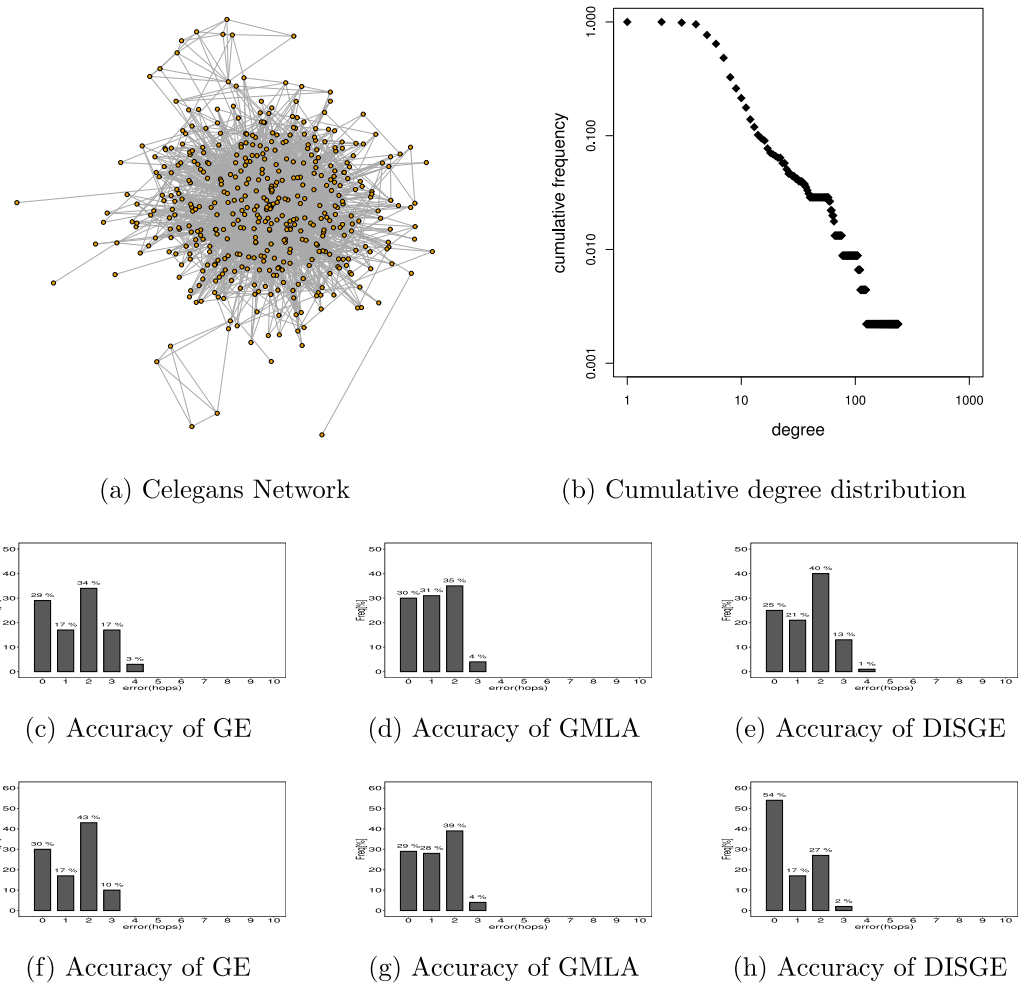
Table 6
The average time delays of GE, GMLA and DISGE on ten networks.

Network	GE	GMLA	DISGE	GE	GMLA	DISGE
β	$\beta = 0.25$	$\beta = 0.25$	$\beta = 0.25$	$\beta = 0.50$	$\beta = 0.50$	$\beta = 0.50$
BA model(1)	4.23	3.36	3.51	3.43	3.42	2.81
BA model(2)	5.09	3.31	3.91	4.51	3.17	2.75
WS model(1)	8.18	2.31	8.30	2.74	4.14	4.45
WS model(2)	5.08	2.77	5.27	2.14	4.20	4.79
AIDSblog	1.87	2.15	2.07	1.36	1.92	1.25
PDZBase	3.74	3.44	3.71	1.48	1.86	1.43
USAirlines	3.42	3.05	3.39	3.82	4.35	2.50
NetScience	5.15	2.65	5.09	2.71	2.24	2.73
Celegans	5.82	4.26	5.60	5.09	4.41	3.04
Euroroad	5.32	3.57	4.94	3.98	4.69	5.27

β denotes the propagation ratio in the spreading process.

From Table 6, we can see that the average time delays are 1.36, 1.92 and 1.25, respectively. When $\beta = 0.25$, the accuracy of DISGE is similar with GMLA but inferior to GE. When $\beta = 0.50$, DISGE is superior to both GE and GMLA. From Table 3, we know the H value of AIDSblog network is 5.99. Thus, the results is similar with the cases on synthetic networks with $H > 2.5$.

Fig. 8 shows the experimental results of GE, GMLA and DISGE on PDZBase network. In Fig. 8(c)–(e), when $\beta = 0.25$, the precise locating (0 error hops) rates of GE, GMLA and DISGE are 67 percent, 57 percent and 69 percent, respectively. From Tables 5 and 6, we can see that the average error hops of GE, GMLA and DISGE are 0.94, 0.83 and 0.91, respectively, the average time delays are 3.74, 3.44 and 3.71, respectively. In Fig. 8(e)–(h), when $\beta = 0.50$, the precise locating (0 error hops) rates are 80 percent, 74 percent and 82 percent, respectively. From Table 5, we can see that the average error hops of GE, GMLA and DISGE are 0.37, 0.49 and 0.36, respectively. From Table 6, we can see that the average time delays are 1.48, 1.86 and 1.43, respectively. When $\beta = 0.25$, the performance of DISGE is little better than GE but inferior to GMLA. When $\beta = 0.50$, DISGE is superior to GE and GMLA. From Table 3, we know the H value of PDZBase network is 2.63. The results is similar with the cases on synthetic networks with $H > 2.5$.



(a) Celegans Network (b) Cumulative degree distribution

(c) Accuracy of GE (d) Accuracy of GMLA (e) Accuracy of DISGE

(f) Accuracy of GE (g) Accuracy of GMLA (h) Accuracy of DISGE

Fig. 11. The error hops between the estimated source and the true source obtained by GE, GMLA and DISGE. Each subplot is obtained by 100 runs. (c)–(e) show the error hops of $\beta = 0.25$. (f)–(h) show the error hops of $\beta = 0.50$.

Table 7
The average execution time ratio on ten networks.

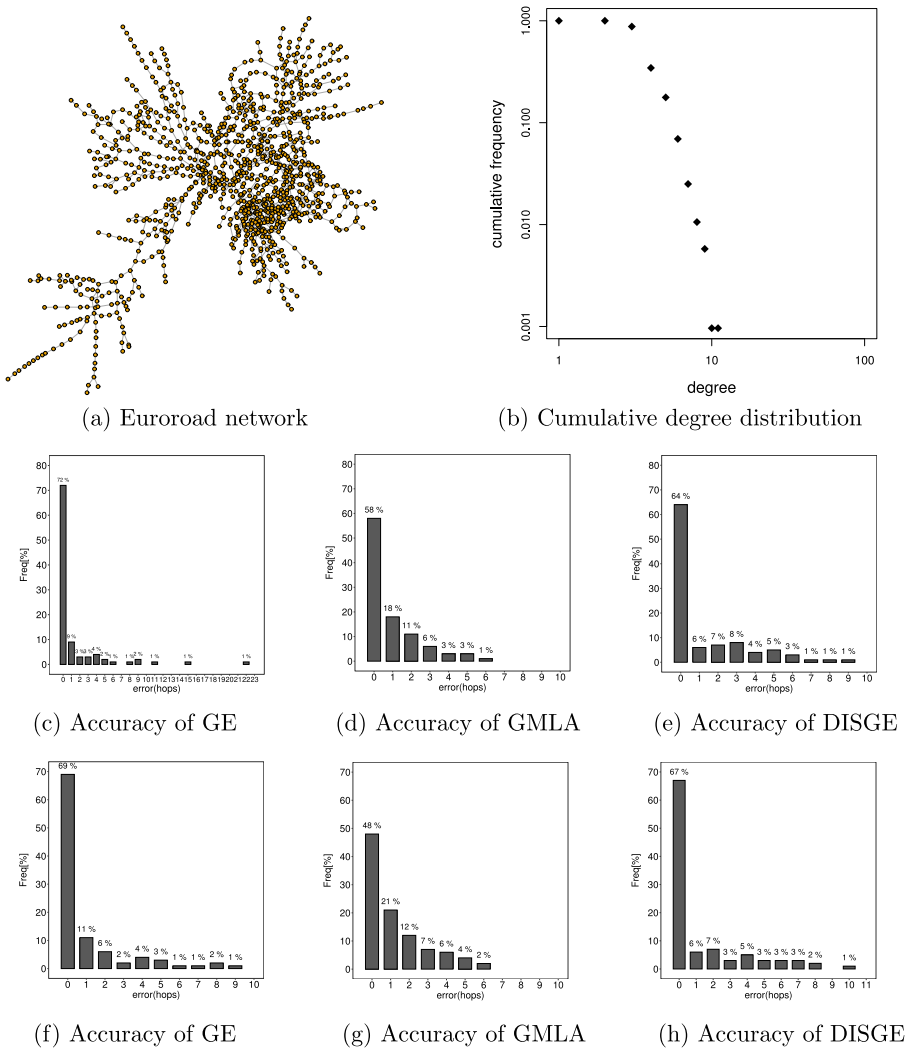
Network	$\frac{r(\text{GE})}{r(\text{DISGE})}$	$\frac{r(\text{GMLA})}{r(\text{DISGE})}$	$\frac{r(\text{GE})}{r(\text{DISGE})}$	$\frac{r(\text{GMLA})}{r(\text{DISGE})}$
β	$\beta = 0.25$	$\beta = 0.25$	$\beta = 0.50$	$\beta = 0.50$
BA model(1)	1.70	0.91	1.70	1.08
BA model(2)	1.54	0.56	1.81	0.81
WS model(1)	2.48	0.22	3.00	0.24
WS model(2)	2.91	0.14	5.35	0.22
AIDSblog	9.01	1.80	19.94	3.39
PDZBase	3.60	0.29	17.49	1.24
USAirlines	3.79	0.71	1.96	0.37
NetScience	3.57	0.15	6.77	0.19
Celegans	2.09	0.32	1.68	0.29
Euroroad	6.79	0.05	5.15	0.04

$r(\cdot)$ represents the execution time of a method.

Fig. 9 shows the experimental results of GE, GMLA and DISGE on USAirlines network. In Fig. 9(c)–(e), when $\beta = 0.25$, the precise locating (0 error hops) rates of GE, GMLA and DISGE are 33 percent, 46 percent and 27 percent, respectively. From Tables 5 and 6, we can see that the average error hops of GE, GMLA and DISGE are 0.88, 0.86 and 0.88, respectively, the average time delays are 3.42, 3.05 and 3.39, respectively. In Fig. 9(f)–(h), when $\beta = 0.50$, the precise locating (0 error hops) rates are 27 percent, 32 percent and 46 percent, respectively. From Table 5, we can see that the average error hops of GE, GMLA and DISGE are 1.02, 1.19 and 0.65, respectively. From Table 6, we can see that the average time delays are 3.82, 4.35 and 2.50, respectively. When $\beta = 0.25$, the three methods expose a similar performance. When $\beta = 0.50$, DISGE obviously outperforms GE and GMLA. From Table 3, we know the H value of USAirlines network is 3.46. Thus, this result is similar with the cases on synthetic networks with $H > 2.5$.

Fig. 10 shows the experimental results of GE, GMLA and DISGE on NetScience network. In Fig. 10(c)–(e), when $\beta = 0.25$, the precise locating (0 error hops) rates are 25 percent, 60 percent and 31 percent, respectively. From Table 5, we can see that the average error hops of GE, GMLA and DISGE are 1.27, 0.65 and 1.24, respectively. From Table 6, we can see the average time delays are 5.15, 2.65 and 5.09, respectively. In Fig. 10(e)–(f), when $\beta = 0.50$, the precise locating (0 error hops) rates are 53 percent, 58 percent and 58 percent. From Table 5, we can see that the average error hops of GE, GMLA and DISGE are 0.67, 0.60 and 0.66, respectively. From Table 6, we can see the average time delays are 2.71, 2.24 and 2.73, respectively. The accuracy of DISGE is similar with GE and inferior to GMLA. From Table 3, we know the H value of NetScience network is 1.66. Thus, it is similar with the cases on synthetic networks with $H < 2.5$.

Fig. 11 shows the experimental results of GE, GMLA and DISGE on Celegans network. In Fig. 11(c)–(e), when $\beta = 0.25$, the precise locating (0 error hops) rates of GE, GMLA and DISGE are 29 percent, 30 percent and 25 percent, respectively. From Tables 5 and 6, we can see that the average error hops of GE, GMLA and DISGE are 1.48, 1.13 and 1.44, respectively, the average time delays are 5.82, 4.26 and 5.60, respectively. In Fig. 11(f)–(h),



(a) Euroroad network (b) Cumulative degree distribution

(c) Accuracy of GE (d) Accuracy of GMLA (e) Accuracy of DISGE

(f) Accuracy of GE (g) Accuracy of GMLA (h) Accuracy of DISGE

Fig. 12. The error hops between the estimated source and the true source obtained by GE, GMLA and DISGE. Each subplot is obtained by 100 runs. (c)–(e) show the error hops of $\beta = 0.25$. (f)–(h) show the error hops of $\beta = 0.50$.

when $\beta = 0.50$, the precise locating (0 error hops) rates are 30 percent, 29 percent and 54 percent, respectively. From Table 5, we can see that the average error hops of GE, GMLA and DISGE are 1.33, 1.18 and 0.77, respectively. From Table 6, we can see that the average time delays are 5.09, 4.41 and 3.04, respectively. When $\beta = 0.25$, the accuracy of DISGE is a little better than GE but inferior to GMLA. When $\beta = 0.50$, DISGE obviously outperforms both GE and GMLA. From Table 3, we know the H value of Celegans network is 4.49. This result is similar with the cases on synthetic networks with $H > 2.5$.

Fig. 12 shows the experimental results of GE, GMLA and DISGE on Euroroad network. In Fig. 12(c)–(e), when $\beta = 0.25$, the precise locating (0 error hops) rates are 72 percent, 58 percent and 64 percent, respectively. From Tables 5 and 6, we can see that the average error hops of GE, GMLA and DISGE are 1.30, 0.91 and 1.27, respectively, the average time delays are 5.32, 3.57 and 4.94, respectively. In Fig. 12(f)–(h), when $\beta = 0.50$, the precise locating (0 error hops) rates are 69 percent, 48 percent and 67 percent, respectively. From Tables 5 and 6, we can see that the average error hops of GE, GMLA and DISGE are 0.98, 1.22 and 1.29, respectively, the average time delays are 3.98, 4.69 and 5.27, respectively. When $\beta = 0.25$, the accuracy of DISGE is similar with GE but inferior to GMLA. When $\beta = 0.50$, GE exposes the best performance. From Table 3, we know the H value of Euroroad network is 1.23. The result is similar with the cases on synthetic networks with $H < 2.5$.

Combining the results on real networks and the topology properties of networks listed in Table 3, we can see that the conclusion got in synthetic networks is verified, i.e. in the cases with $\beta = 0.50$, DISGE shows an advantage in dealing with the networks with $H > 2.5$.

Table B.8
The parameters for generating synthetic networks.

Networks	Parameters
BA model(1)	barabasi.game(100, power = 2.2, m = 8, directed = FALSE)
BA model(2)	barabasi.game(100, power = 2.4, m = 8, directed = FALSE)
BA model(3)	barabasi.game(100, power = 2.6, m = 8, directed = FALSE)
BA model(4)	barabasi.game(100, power = 2.8, m = 8, directed = FALSE)
BA model(5)	barabasi.game(100, power = 3.0, m = 8, directed = FALSE)

H denotes the degree heterogeneity [34–36], $H = \frac{\langle k^2 \rangle}{\langle k \rangle^2}$.
Software: R 64x 3.3.3, igraph R 1.2.1.

6. Conclusion

Locating the propagation source is one of the most important measures to control the spreading process on complex networks.

Table B.9
The average error hops and average time delays of GE, GMLA and DISGE on five BA models.

	H	Average error hops			Average time delays		
		GE	GMLA	DISGE	GE	GMLA	DISGE
		$\beta = 0.50$	$\beta = 0.50$	$\beta = 0.50$	$\beta = 0.50$	$\beta = 0.50$	$\beta = 0.50$
BA model(1)	2.23	0.90	0.77	0.78	3.59	2.79	3.24
BA model(2)	2.39	0.91	0.74	0.88	3.62	2.72	3.66
BA model(3)	2.53	0.86	0.91	0.85	3.59	3.02	3.67
BA model(4)	2.70	0.86	0.95	0.83	3.56	3.46	3.31
BA model(5)	2.88	0.89	0.99	0.69	3.43	3.42	2.81

β denotes the propagation ratio in the spreading process.
Each value is obtained by 100 runs.

In this paper, we define a novel direction-induced search (DIS) by which the Direction information recorded in observers can be fully utilized. Further, by combining DIS and the well-known Gaussian estimator (GE), a direction-induced search based Gaussian estimator (DISGE) is proposed. Experimental results reveal that, in the cases with a larger propagation ratio β, DISGE shows an advantage in dealing with the networks with $H > 2.5$. In fact, compared with a smaller β, a larger one will inform more nodes through more different paths in the spreading process, which makes distinguishing the true propagation source more difficult. However, DISGE exposes a good performance in locating the propagation source in the difficult environment, which indicates its feasibility and effectiveness. Besides, the DIS we defined is very helpful to recover the true diffusion tree in spreading process. Therefore, it can be introduced into other propagation source locating methods that set observers to record the Direction information in the spreading process. In the next step, we will generalize the DIS in more extensive scenarios.

CRediT authorship contribution statement

Fan Yang: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing. **Shuhong Yang:** Validation, Formal analysis, Writing - review & editing. **Yong Peng:** Software, Validation, Writing - review & editing. **Yabing Yao:** Software, Validation, Formal analysis. **Zhiwen Wang:** Writing - review & editing, Visualization. **Houjun Li:** Software, Writing - review & editing. **Jingxian Liu:** Writing - review & editing, Validation. **Ruisheng Zhang:** Writing - review & editing. **Chungui Li:** Conceptualization, Methodology.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61841202, 61962007), the Science and Technology Planning Project of Guangxi (Grant No. AD19245101), the Natural Science Foundation of Guangxi (Grant No. 2018GXNSFAA050020, 2018GXNSFDA294001), the 2019 Guangxi Education Department Program (Grant. No. 2019KY0372, 2018KY0322), the Doctoral Foundation of Guangxi University of Science and Technology (Grant No. 19Z06).

Appendix A. Proof

A.1. Proof of Proposition 1

Proof. Because $\mathcal{O} = V$ and $\mathcal{O}_a = V$, we have $\mathcal{O}_a = \mathcal{O}$, $K_a = |V_a| = |V|$ and $\mathcal{G}_a = \mathcal{G}$. The Direction information recorded in each active observer in $\mathcal{O}_a$ can be obtained, which indicates that each node has known where the information first time came from. Essentially, for an arbitrary node $u \in V_a$, the Direction information recorded in u is one of its neighbours (denoted by v) from which the information is passed to u for the first time, which can be represented by a directed edge $e_{v\to u}$. Here, if there are k neighbours having the same propagation delay to u, u can be informed by only one of the k neighbours for the first time. Thus, for each pair of nodes u and v, the true propagation direction passing the information from v to u can be uniquely determined by a corresponding directed edge $e_{v\to u}$. For all pairs of nodes, by retaining all the directed edges that uniquely reflect the true propagation direction while removing other edges, a spanning tree will be determined. Here, this spanning tree is denoted by *ST1*.

Further, we use the method of proof by contradiction. Suppose, *ST1* is not unique. Then there exists at least one another spanning tree, denoted by *ST2*, on which there exists at least one directed edge that is different from *ST1*. Without loss of generality, suppose there is only one different directed edge between *ST1* and *ST2*, they are denoted by $e_{v\to u}$ and $e_{w\to u}$, respectively. Therefore, node u is informed for the first time by two different directions v and w simultaneously. However, it is easy to know that for each node u in $\mathcal{G}_a$, the direction by which node u is first time informed is unique. By this contradiction, with the recorded Direction information, the spanning tree corresponding to each node first time got informed can be uniquely determined. Therefore, Proposition 1 is proved. □

A.2. Proof of Proposition 2

Proof. By Proposition 1, we know there exists an unique DIS spanning tree corresponding to each node in $\mathcal{G}_a$ first time got informed in the spreading process. From the line 18 in Algorithm 2, we know the true propagation source s^* will be determined when the true spreading tree (DIS spanning tree) is reconstructed. Further, the true propagation source s^* can be precisely found out. Therefore, Proposition 2 is proved. □

Appendix B. Determining the threshold of heterogeneity (H value)

To determine a point of H value by which DISGE can expose a good performance, the corresponding experiments are performed on a series of synthetic networks. The parameters used for generating synthetic networks are listed in Table B.8, the results are listed in Table B.9. From the results, we get a conclusion that, for the networks with $H > 2.5$, when the propagation ratio $\beta = 0.50$, the accuracy of DISGE outperforms GE and GMLA.

References

[1] Y. Wang, S. Wen, Y. Xiang, W. Zhou, Modeling the propagation of worms in networks: A survey, IEEE Commun. Surv. Tutor. 16 (2) (2014) 942–960, http://dx.doi.org/10.1109/SURV.2013.100913.00195.
[2] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Rev. Modern Phys. 87 (2015) 925–979, http://dx.doi.org/10.1103/RevModPhys.87.925.

[3] B. Doerr, M. Fouz, T. Friedrich, Why rumors spread so quickly in social networks, Commun. ACM 55 (6) (2012) 70–75, http://dx.doi.org/10.1145/2184319.2184338.
[4] J. Jiang, W. Sheng, Y. Shui, X. Yang, W. Zhou, Identifying propagation sources in networks: State-of-the-art and comparative studies, IEEE Commun. Surv. Tutor. 19 (1) (2017) 465–481, http://dx.doi.org/10.1109/COMST.2016.2615098.
[5] P.C. Pinto, T. Patrick, V. Martin, Locating the source of diffusion in large-scale networks, Phys. Rev. Lett. 109 (6) (2012) 068702, http://dx.doi.org/10.1103/PhysRevLett.109.068702.
[6] D. Brockmann, D. Helbing, The hidden geometry of complex, network-driven contagion phenomena, Science 342 (6164) (2013) 1337–1342, http://dx.doi.org/10.1126/science.1245200.
[7] D. Shah, T. Zaman, Detecting sources of computer viruses in networks: Theory and experiment, SIGMETRICS Perform. Eval. Rev. 38 (1) (2010) 203–214, http://dx.doi.org/10.1145/1811099.1811063.
[8] D. Shah, T. Zaman, Rumors in a network: Who's the culprit? IEEE Trans. Inform. Theory 57 (8) (2011) 5163–5181, http://dx.doi.org/10.1109/TIT.2011.2158885.
[9] K. Zhu, L. Ying, Information source detection in the sir model: A sample path based approach, in: Information Theory and Applications Workshop (ITA2013), 2013, pp. 1–9, http://dx.doi.org/10.1109/ITA.2013.6502991.
[10] W. Luo, W.P. Tay, M. Leng, How to identify an infection source with limited observations, IEEE J. Sel. Top. Sign. Proces. 8 (4) (2014) 586–597, http://dx.doi.org/10.1109/JSTSP.2014.2315533.
[11] A.Y. Lokhov, M. Mézard, H. Ohta, L. Zdeborová, Inferring the origin of an epidemic with a dynamic message-passing algorithm, Phys. Rev. E 90 (1) (2014) 012801, http://dx.doi.org/10.1103/PhysRevE.90.012801.
[12] F. Altarelli, A. Braunstein, L. Dall'Asta, A. Lage-Castellanos, R. Zecchina, Bayesian inference of epidemics on networks via belief propagation, Phys. Rev. Lett. 112 (11) (2014) 118701, http://dx.doi.org/10.1103/PhysRevLett.112.118701.
[13] N. Antulov-Fantulin, A. Lančić, T. Šmuc, H. Štefančić, M. Šikić, Identification of patient zero in static and temporal networks: robustness and limitations, Phys. Rev. Lett. 114 (24) (2015) 248701, http://dx.doi.org/10.1103/PhysRevLett.114.248701.
[14] F. Yang, R. Zhang, Y. Yao, Y. Yuan, Locating the propagation source on complex networks with propagation centrality algorithm, Knowl.-Based Syst. 100 (2016) 112–123, http://dx.doi.org/10.1016/j.knosys.2016.02.013.
[15] K. Cai, X. Hong, J.C.S. Lui, Information spreading forensics via sequential dependent snapshots, IEEE/ACM Trans. Netw. 26 (1) (2018) 478–491, http://dx.doi.org/10.1109/TNET.2018.2791412.
[16] W. Luo, W.P. Tay, M. Leng, Identifying infection sources and regions in large networks, IEEE Trans. Signal Process. 61 (11) (2013) 2850–2865, http://dx.doi.org/10.1109/TSP.2013.2256902.
[17] B.A. Prakash, J. Vreeken, C. Faloutsos, Efficiently spotting the starting points of an epidemic in a large graph, Knowl. Inf. Syst. 38 (1) (2014) 35–59, http://dx.doi.org/10.1007/s10115-013-0671-5.
[18] J. Jiang, S. Wen, S. Yu, Y. Xiang, W. Zhou, K-center: An approach on the multi-source identification of information diffusion, IEEE Trans. Inf. Forensics Secur. 10 (12) (2015) 2616–2626, http://dx.doi.org/10.1109/TIFS.2015.2469256.
[19] J. Manitz, J. Harbering, M. Schmidt, T. Kneib, A. Schöbel, Source estimation for propagation processes on complex networks with an application to delays in public transportation systems, J. R. Stat. Soc. Ser. C. Appl. Stat. 66 (3) (2017) 521–536, http://dx.doi.org/10.1111/rssc.12176.
[20] J.G. Caputo, A. Hamdi, A. Knippel, Inverse source problem in a forced network, Inverse Problems 35 (5) (2019) 055006, http://dx.doi.org/10.1088/1361-6420/aafcc6.
[21] R. Paluch, X. Lu, K. Suchecki, B.K. Szymaski, J.A. Holyst, Fast and accurate detection of spread source in large complex networks, Sci. Rep. 8 (1) (2018) 2508, http://dx.doi.org/10.1038/s41598-018-20546-3.
[22] L. Gajewski, K. Suchecki, J. Holyst, Multiple propagation paths enhance locating the source of diffusion in complex networks, Phys. A 519 (C) (2019) 34–41, http://dx.doi.org/10.1016/j.physa.2018.12.0.
[23] Z. Shen, S. Cao, W.-X. Wang, Z. Di, H.E. Stanley, Locating the source of diffusion in complex networks by time-reversal backward spreading, Phys. Rev. E 93 (3) (2016) 032301, http://dx.doi.org/10.1103/physreve.93.032301.
[24] X. Li, X. Wang, C. Zhao, X. Zhang, D. Yi, Locating the source of diffusion in complex networks via Gaussian-based localization and deduction, Appl. Sci. 9 (18) (2019) 3758, http://dx.doi.org/10.3390/app9183758.
[25] H. Wang, J. Wu, S. Pan, P. Zhang, L. Chen, Towards large-scale social networks with online diffusion provenance detection, Comput. Netw. 114 (2016) 154–166, http://dx.doi.org/10.1016/j.comnet.2016.08.025.
[26] Z.-L. Hu, Z. Shen, S. Cao, B. Podobnik, H. Yang, W.-X. Wang, Y.-C. Lai, Locating multiple diffusion sources in time varying networks from sparse observations, Sci. Rep. 8 (1) (2018) 2685, http://dx.doi.org/10.1038/s41598-018-20033-9.
[27] F. Ji, W. Tang, W.P. Tay, On the properties of gromov matrices and their applications in network inference, IEEE Trans. Signal Process. 67 (10) (2019) 2624–2638, http://dx.doi.org/10.1109/TSP.2019.2908133.
[28] P.C. Pinto, T. Patrick, V. Martin, Supplemental material of locating the source of diffusion in large-scale networks, Phys. Rev. Lett. 109 (6) (2012) http://dx.doi.org/10.1103/PhysRevLett.109.068702.
[29] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440, http://dx.doi.org/10.1038/30918.
[30] A.L. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512, http://dx.doi.org/10.1126/science.286.5439.509.
[31] R.A. Rossi, N.K. Ahmed, The network data repository with interactive graph analytics and visualization, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, pp. 4292–4293, URL http://networkrepository.com.
[32] S. Gopal, The evolving social geography of blogs, in: Societies and Cities in the Age of Instant Access, Vol. 88, 2007, pp. 275–293, http://dx.doi.org/10.1007/1-4020-5427-0_18.
[33] M.E.J. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (20) (2002) 208701, http://dx.doi.org/10.1103/PhysRevLett.89.208701.
[34] A. Zeng, C.J. Zhang, Ranking spreaders by decomposing complex networks, Phys. Lett. A 377 (14) (2013) 1031–1035, http://dx.doi.org/10.1016/j.physleta.2013.02.039.
[35] F. Yang, X. Li, Y. Xu, X. Liu, J. Wang, Y. Zhang, R. Zhang, Y. Yao, Ranking the spreading influence of nodes in complex networks: An extended weighted degree centrality based on a remaining minimum degree decomposition, Phys. Lett. A 382 (34) (2018) 2361–2371, http://dx.doi.org/10.1016/j.physleta.2018.05.032.
[36] F. Yang, R. Zhang, Z. Yang, R. Hu, M. Li, Y. Yuan, K. Li, Identifying the most influential spreaders in complex networks by an extended local K-Shell sum, Internat. J. Modern Phys. C 28 (01) (2017) 1750014, http://dx.doi.org/10.1142/S0129183117500140.
[37] C. Claudio, P.S. Romualdo, Thresholds for epidemic spreading in networks, Phys. Rev. Lett. 105 (21) (2010) 218701, http://dx.doi.org/10.1103/PhysRevLett.105.218701.