Identifying Sources of Random Walk-Based Epidemic Spreading in Networks

Bo Qin and Cunlai Pu(✉)

School of Computer Science and Engineering,
Nanjing University of Science and Technology, Nanjing 210094, China
pucunlai@njust.edu.cn

Abstract. Identifying the sources of epidemic spreading is of critical importance to epidemic control and network immunization. However, the task of source identification is very challenging, since in real situations the dynamics of the spreading process is usually not clear. In this paper, we formulate the multiple source epidemic spreading process as the multiple random walks, which is a theoretical model applicable to various spreading processes. Considering the different influence of distinct epidemic sources on the observed infection graph, we derive the maximum likelihood estimator of the multiple source identification problem. Simulation results on real-world networks and network models, such as the Price model and Erdös-Rényi (ER) model, demonstrate the efficiency of our estimator. Furthermore, we find that the efficiency of our estimator increases with the enhancement of network sparsity and heterogeneity.

Keywords: Source identification · Random walk · Maximum likelihood (ML) · Network heterogeneity

1 Introduction

Epidemic spreading is an universal process in nature and man-made systems, such as the spread of cyber viruses in communication networks, rumors in social networks and diseases in biological networks. Identifying the sources of these harmful spreading processes is of practical interest to researchers as well as system administrators for forensic purposes. In addition, early recognition of the epidemic sources helps to block the epidemic spreading promptly and eventually decrease the loss. Accurately identifying the sources of epidemic spreading is a very challenging task, since we usually have very limited information, such as the observed network structure and states of nodes. A more relaxed problem is to estimate the likelihood of nodes being the spreading sources based on the

Supported in part by National Natural Science Foundation of China (No. 61872187, No. 61871444, No. 61773215) and Major Special Project of Core Electronic Devices, Hign-end Generic Chips and Basic Software (No. 2015ZX01041101).

X. B. Zhai et al. (Eds.): MLICOM 2019, LNICST 294, pp. 375–386, 2019.
https://doi.org/10.1007/978-3-030-32388-2_32

maximum likelihood estimate (MLE) [1,2]. In recent years, many source identification methods for this problem have been proposed. A large portion of them is based on some sort of node centrality. For example, Shah et al. proposed the first algorithm of source identification according to the rumor centrality [3–5]. Later, Luo et al. extended the work of Shah to the case of multiple sources [6–8]. Zhu et al. presented a novel identification method based on the Jordan infection center [9,10]. By calculating the dynamical age of nodes in the infection graph, Fioriti et al. obtained that old nodes are more likely to be the infection sources than young nodes [11,12]. Moreover, Comin provided an identification algorithm based on the unbiased betweenness centrality [13], and then Zang extended this algorithm to the multiple source scenario [14].

In addition to the centrality-based methods, there are some other remarkable identification methods. For instance, Lokhov et al. gave the dynamic message-passing (DMP) algorithm to estimate the likelihood of every node [15]. Altarelli et al. proposed the belief propagation algorithm based on the factor graph to infer the origin of an epidemic [16]. Antulov-Fantulin et al. used the Monte-Carlo method to simulate the possible propagation process to conjecture the source [17]. Prakash et al. proposed the NetSleuth algorithm which transforms the source identification problem into an optimization problem that solves the minimum description length [18,19]. Most of these methods depend on some sort of approximation in the construction of maximum likelihood probability, and in most of the cases, they require either large computational complexity to find near-optimal solutions, or simplified heuristics to achieve suboptimal performance.

Random walk is a general process that can be used to describe many diffusion processes on networks. In particular, the spread of some viruses in reality can be modeled as random-walk diffusion, such as the spread of Bluetongue virus driven by the random movements of insect vectors and the cyber virus driven by the random transmission of information packets. Most recently, Abigail et al. proposed a single source identification algorithm by considering the random walk of virus [20]. Their calculation takes account all the possible spreading paths, while the previous work considers only the shortest paths or some high-probability paths. Their simulation results on the food supply network data demonstrated the efficiency of the algorithm.

Enlightened by Abigail's work, we investigate the case of multiple infection sources in networks. It is apparent that the source of an infected node can be anyone in the source set. Different source nodes have different possibilities to infected a node. Accordingly, in the calculation, we provide two ways to approximate this possibility. We further calculate the likelihood of a candidate set to be the source by considering all the possible infection paths from the set, and obtain our maximum likelihood estimator. The set of maximum likelihood is supposed to be the source.

To validate our estimator, we apply it to the complex network models, such as the ER and Price models, and some real-world networks. Source sets of two and three nodes are considered. The experimental results demonstrate the efficiency

of our estimator. Furthermore, we observe that the efficiency of our estimator increases with the enhancement of network sparsity and heterogeneity.

2 Problem Formulation and Our Method

We consider the problem of identifying multiple infection sources. Previous work usually uses the susceptible-infected (SI) model to describe the virus spreading process. Differently, we consider the diffusion-based spread of virus. In particular, we assume that the spread of virus is based on multiple random walks (simultaneous random walks starting from multiple sources). In this section, we first present our diffusion model, and then derive the maximum likelihood estimator of our model.

2.1 Diffusion Model

Let $G(V, E)$ be a directed graph, where V and E represent the set of all nodes and the set of all edges, respectively. Assume that the set of terminal nodes, which have zero out-degree, is V_T, and the set of non-terminal nodes is V_N. We randomly select m (> 1) non-terminal nodes to be the infection sources, which form the source set $s^* = \{s_1, s_2, \ldots, s_m\}, \forall s_i \in V_N$. At the beginning, every infection source propagates one copy of the virus to one of its neighbor nodes. The movements of different virus copies are assumed to be independent and identical random walks. A node is infected if it has ever been visited by the virus; otherwise, it is healthy, while susceptible to the virus. We do not consider the recovery of infected nodes. At each time step, every infected node spreads one copy of virus to one of its neighbor nodes. After a period of time, we observe that the set of infected terminal nodes is $\Theta = \{o_1, o_2, \ldots, o_K\}, \forall o_k \in V_T$. Our aim is to estimate the source set s^* based on the set of infected terminal nodes Θ.

Since the movement of virus copy is a random walk process, i.e., Markov process, each neighbor of an infected node will receive the virus copy with equal probability. For instance, if the out-degree of the infected node i is k_i, then its neighbor node j receives the virus copy with probability $p_{ij} = 1/k_i$. Considering all node pairs, we have the one-step probability transition matrix P, of which the element is p_{ij}. We further partition matrix P into a 4×4 block matrix, which is as follows:

$$P = \begin{bmatrix} P_N & P_T \\ 0 & I_T \end{bmatrix}, \tag{1}$$

where P_N is the $|V_N| \times |V_N|$ submatrix concerning the transition probabilities between non-terminal nodes, and P_T is the $|V_N| \times |V_T|$ submatrix, which consists of the transition probabilities from non-terminal nodes to terminal nodes. The submatrix I_T is an identity matrix of order $|V_T|$, since we assume that the self-transition probability of a terminal node is 1. Note that the transition probability from terminal node to non-terminal node is 0, which corresponds to the zero submatrix of P. According to the principle of Markov process, the n-step

transition probability is the nth power of the one-step transition matrix. For instance, the n-step transition probability from node i to j is $\{P^n\}_{ij}$.

2.2 Multiple Source Estimator: Maximum Likelihood (ML)

In the source identification problem, the available information is the network G and the set of infected terminal nodes Θ. We utilize the maximum likelihood estimate to infer the set of infection sources, which is expressed as

$$\hat{s} = \arg\max_{s \in \Omega} P(\Theta | s^* = s), \tag{2}$$

where Ω represents the set of all possible combinations of m non-terminal nodes, and s is a candidate source set. Equation (2) implies that our target set is the one which maximizes the condition probability of the set of infected terminal nodes given the candidate source set.

To facilitate the calculation of $P(\Theta | s^* = s)$, we make the following denotations:

- γ_{so_k}: A path starting from an arbitrary node in set s to an infected terminal node o_k, $o_k \in \Theta$.
- Γ_{so_k}: The set of all paths starting from nodes in set s to the infected terminal node o_k, $\Gamma_{so_k} = \{\gamma_{so_k}\}$.
- π_s: A specific permutation of K paths, which start from nodes of set s to the K infected terminal nodes (one-one correspondence), $\pi_s = (\gamma_{so_1}, \ldots, \gamma_{so_K})$. Actually, π_s is an element of the Cartesian product of all $\{\Gamma_{so_k}\}_{o_k \in \Theta}$, i.e., $\pi_s \in \Gamma_{so_1} \times \cdots \times \Gamma_{so_K}$.
- Π_s: The set of all possible path permutations $\{\pi_s\}$, i.e., $\Pi_s = \{\pi_s\} = \Gamma_{so_1} \times \cdots \times \Gamma_{so_K} = \{(\gamma_{so_1}, \ldots, \gamma_{so_K}) : \gamma_{so_k} \in \Gamma_{so_k}\}$.

Similar to [20], we consider all possible spreading paths starting from the nodes of the given candidate source set. Each path permutation π_s has some probability to be the actual one. Thus, we have

$$\begin{aligned} P(\Theta | s^* = s) &= \sum_{\pi_s \in \Pi_s} P(\pi_s | s) \\ &= P(\Pi_s | s) \\ &= P(\Gamma_{so_1} \times \cdots \times \Gamma_{so_K} | s). \end{aligned} \tag{3}$$

Since we assume that the virus copies perform random walks independently, the infection of terminal nodes is also independent. Thus, we have

$$\begin{aligned} P(\Gamma_{so_1} \times \cdots \times \Gamma_{so_K} | s) &= P(\prod_{o_k \in \Theta} \Gamma_{so_k} | s) \\ &= \prod_{o_k \in \Theta} P(\Gamma_{so_k} | s), \end{aligned} \tag{4}$$

where $P(\Gamma_{so_k}|s)$ is equal to the sum of probabilities of all possible infection paths starting from set s to node o_k, $\sum_{\gamma_{so_k} \in \Gamma_{so_k}} P(\gamma_{so_k}|s)$. Combining (3) and (4), we can get [20]:

$$P(\Theta|s^* = s) = \prod_{o_k \in \Theta} P(\Gamma_{so_k}|s), \tag{5}$$

In order to calculate $P(\Gamma_{so_k}|s)$, we need to enumerate all possible paths. Although we use Γ_{so_k} to indicate all paths of source set s infecting o_k, actually o_k is infected by a certain node or several nodes in the collection s and the probability that different node in s infects o_k is different. So we have the following definition:

$$P(\Gamma_{so_k}|s) = \sum_{i=1}^{|s^*|} P(\Gamma_{s_i o_k}|s_i) P(s_i|s). \tag{6}$$

Therefore, $P(\Gamma_{so_k}|s)$ can be expressed as

$$P(\Gamma_{so_k}|s) = \sum_{i=1}^{|s^*|} \sum_{n=0}^{\infty} \sum_{l \in V_N} p_{s_i} p_{s_i l}^{(n)} p_{l o_k}, \tag{7}$$

where p_{s_i} represents the probability that among the $|s^*|$ source nodes, node s_i infects terminal node o_k. $p_{s_i l}^{(n)}$ denotes the probability that source node s_i infects non-terminal node l in exactly n steps, and it equals the value of the $(s_i, l)th$ element of matrix $(P_N)^n$. $p_{l o_k}$ represents the probability that non-terminal node l infects terminal node o_k in exactly one step, and it is equivalent to the value of the $(s_i, l)th$ element of matrix P_T. The right side of Eq. (7) indicates that we consider all spreading paths from a node of the source set to a terminal node and the probability of the former to be the exact source node of the latter.

In Eq. (7), p_{s_i} is hard to obtain, and its value might be different for different nodes in set s. Here we propose two approximation methods of p_{s_i}. The first one is based on the transition probability matrix P. We assume that p_{s_i} is proportional to the sum of one-step transition probabilities to all the terminal nodes, which can be written as

$$p_{s_i} = \frac{\sum_{o_k \in \Theta} (p_{s_i o_k}^{(1)} + \varepsilon)}{\sum_{j=1}^{|s^*|} \sum_{o_k \in \Theta} (p_{s_j o_k}^{(1)} + \varepsilon)}, \tag{8}$$

where $p_{s_i o_k}^{(1)}$ represents the $(s_i, o_k)th$ element of matrix P. ε is a small positive number to ensure the denominator is non-zero. In non-sparse networks, this approximation method works well. However, in sparse network, many elements of the one-step transition matrix is zero, and in this case p_{s_i} will be equal for all the source nodes. This is contradictory to the assumption that each node in s might has different possibility to be the source.

To better quantify the possibility of a node to be the source, we propose another approximation method, which considers the mean first passage time

(FPT) [21] of a random walk. We assume that p_{s_i} is proportional to the reciprocal of the sum of FPT from s_i to all the terminal nodes, which is

$$p_{s_i}^{'} = \frac{1/\sum_{o_k \in \Theta} t_{s_i o_k}}{\sum_{j=1}^{|s^*|} 1/\sum_{o_k \in \Theta} t_{s_j o_k}}, \tag{9}$$

where $t_{s_i o_k}$ is the FPT from node s_i to node o_k. The smaller $t_{s_i o_k}$, the more likely that s_i is the source. The pseudocode of FPT calculation is shown in Algorithm 1.

Algorithm 1. FPT Calculation

Input: directed graph $G(V, E)$, iteration times n, diffusion time t, source node s_i, target node o_k.
for $i := 1$ to n **do**
  for $j := 1$ to t **do**
    for each non-terminal node **do**
      if the node is infected **then**
        Randomly select a neighbour node to infect.
      End If
      if node o_k is infected **then**
        Record the current time as the FPT of this iteration, and cease this iteration.
      End If
    End For
  End For
  if there is no path between s_i and o_k **then**
    $t_{s_i o_k} := t$.
  End If
End For
Calculate the mean of all iteration results to get $t_{s_i o_k}$.
Output: $t_{s_i o_k}$.

Furthermore, we can get the matrix form of Eq. (7)

$$A = (\sum_{i=1}^{|s^*|} P_{w_i}) \sum_{n=0}^{\infty} P_N^n P_T, \tag{10}$$

where P_{w_i} is a $|C_{|V_N|}^{|s^*|}| \times |V_N|$ matrix. It is apparent that the quantity of rows in this matrix represents the quantity of possible combinations of all non-terminal nodes. Let $\bar{s}$ be a certain source node combination and $\bar{s}_i$ be the ith node in this combination. The $\bar{s}_i th$ element of each line of matrix P_{w_i} is $p_{\bar{s}_i}$ and the remaining elements are zero.

According to the summation formula of geometric series, we obtain

$$A = (\sum_{i=1}^{|s^*|} P_{w_i})(I - P_N)^{-1} P_T, \tag{11}$$

where $I - P_Q$ has an inverse matrix [22]. It should be noted that $\{A\}_{ij}$ represents the sum of probabilities of all possible infection paths from the ith source combination s to terminal node o_j. In other words, $\{A\}_{ij}$ is equivalent to $P(\Gamma_{so_j}|s)$. Combining (5) and (11), we can get

$$P(\Theta|s^* = s) = \prod_{o_k \in \Theta} \{(\sum_{i=1}^{|s^*|} P_{w_i})(I - P_N)^{-1} P_T\}_{so_k}. \tag{12}$$

Based on Eq. (12), we can obtain the likelihood of each candidate node combination. Combining (2) and (12), we finally obtain

$$\hat{s} = \arg\max_{s \in \Omega} \prod_{o_k \in \Theta} \{(\sum_{i=1}^{|s^*|} P_{w_i})(I - P_N)^{-1} P_T\}_{so_k}. \tag{13}$$

It worths mentioning that our solution considers all possible infection paths. In addition, due to the fact that we need to enumerate all possible node combinations, the time complexity of our solution is $O(n^{|s^*|+1})$.

3 Results

In this section, we evaluate the efficiency of our estimator. In the experiments, we only consider cases of two and three sources. We use complex network models such as the ER and Price models and some real-world networks, including the GD96_d network and the power-494-bus network. We test the efficiency of our estimator on different networks. In addition, we investigate how network heterogeneity and density impact efficiency of our estimator.

To quantify the efficiency of our estimator, we provide a metric, i.e., minimum error distance, which is given as follows:

$$\Delta = \sum_{i=1}^{|s^*|} d(\hat{s}_i, s_i^*),$$

where $d(\hat{s}_i, s_i^*)$ represents the shortest distance between each source node $\hat{s}_i$ in the prediction result and node s_i^* which is the closest node to $\hat{s}_i$ in the real sources set. In the experiments, we mainly check the distribution of Δ to evaluate our estimator.

3.1 Synthetic Networks

We do experiments on synthetic networks. First, we use the ER model to generate random networks. In this model, every node pair has the same connection probability. The pseudocode of ER model is shown in Algorithm 2, in which p_{ER} controls the density of the network. The larger p_{ER}, the denser network. We set network size $N = 500$, and then change the network density to investigate how

Algorithm 2. ER Network Generation Algorithm

Input: total number of nodes N, connection probability p_{ER}.
Initialization: $N := 500$ and $p_{ER} \in [0, 1]$.
for each nodes pair $(i\ ,\ j)$ **do**
    Generate a random number $r \in [0, 1]$.
    if $r < p_{ER}$ **then**
        Add an edge between node pairs $(i\ ,\ j)$.
    End If
End For
Output: $G(N, p_{ER})$.

the minimum error distance distribution change. The number of source nodes is 2. The first approximation method (Eq. (8)) is employed in our estimator with $\varepsilon = 0.0001$. We perform 1000 independent runs. In Fig. 1(a), we see that when $p_{ER} = 0.001$, the probability of $\Delta = 0$ is more than 0.5, which means the identification accuracy is more than half. However, when p_{ER} increases, the identification accuracy decreases accordingly.

Algorithm 3. Price Network Generation Algorithm

Input: strongly connected graph with m_0 nodes, total number of nodes N, priority connection probability p_{pri}, the number of added edges for each new node m.
Initialization: add the end nodes of all directed edges in the initial network to array $Array$, $m_0 := 5$ $N := 500$, $m := 3$ and $p_{pri} \in [0, 1]$.
for $N - m_0$ remaining nodes **do**
    for i:=i to m **do**
        Generate a random number $r \in [0, 1]$.
        if $r < p_{pri}$ **then**
            Randomly select a node in $Array$ and make sure that newly selected m nodes are unique.
        else
            Randomly select a node that already exists in the network, and make sure that newly selected m nodes are unique.
        End If
    End For
    For each selected node, add a directed edge pointing to the newly added node, and add the selected m nodes to $Array$.
End For
Output: $G(N, p_{pri})$.

Then, we use the Price model to generate the scale-free networks with 500 nodes. The pseudocode of the Price model is given in Algorithm 3, in which the network heterogeneity is controlled by p_{pri}. The larger p_{pri}, the larger network heterogeneity. As shown in Algorithm 3, we slightly modify the Price model by reversing the direction of links, which is originally pointing from newly added

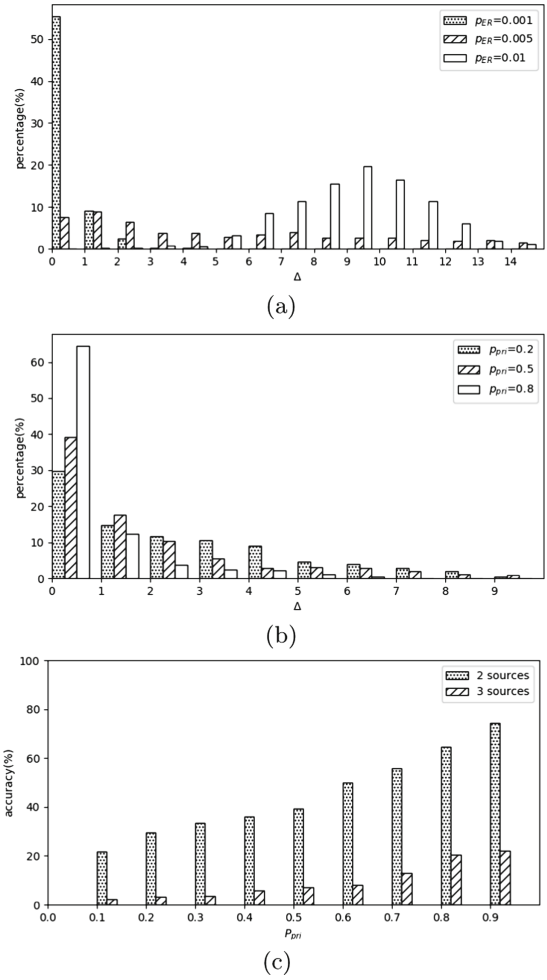

Fig. 1. (a) Distribution of minimum error distance for the ER network. (b) Distribution of minimum error distance for the Price network. (c) Accuracy of identification under different p_{pri} values in the Price network with 2 sources and 3 sources. p_{pri} reflects the heterogeneity of price network.

nodes to old nodes. This modification maintains the statistical property of the model, while lead to the emerge of terminal nodes, which are fit for our model setting, but do not exist in the original model. The number of source nodes is 2. The first approximation method (Eq. (8)) is employed in our estimator with $\varepsilon = 0.0001$. We perform 1000 independent runs and calculate the distribution of Δ for different network heterogeneity. In Fig. 1(b), we can see that $\Delta < 10$, and this indicates the source nodes identified based on our estimator are topologically very close to the real source nodes. Moreover, the larger p, the larger

identification accuracy. For instance, when $p_{pri} = 0.8$, Δ is more than 0.6, while when $p_{pri} = 0.5$, Δ decreases to less than 0.4. This means that heterogeneous networks facilitate source identification.

Next, we increase the number of source nodes to 3. As shown in Fig. 1(c), we obtain the same conclusion as in Fig. 1(b) that the identification accuracy increases with network heterogeneity. Also, we can infer that the identification accuracy decreases when the number of source nodes increases.

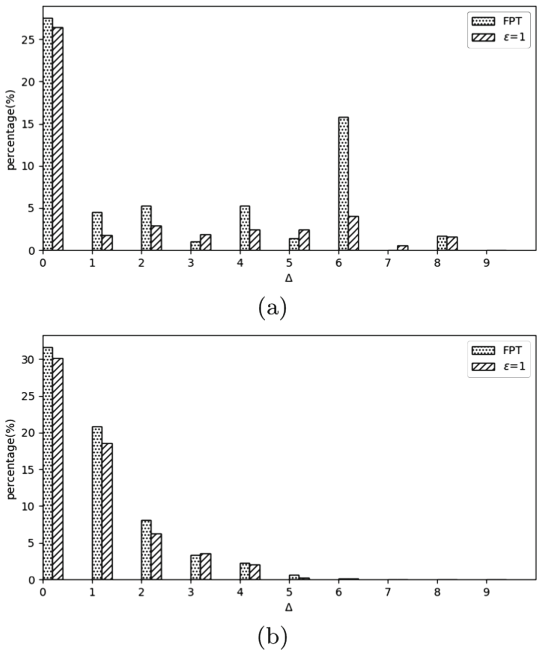

Fig. 2. (a) Distribution of minimum error distance for the GD96_d network. (b) Distribution of minimum error distance for the power-494-bus network.

3.2 Real-World Network

Finally, we do experiments on real-world networks including the GD96_d network [23] and the power-494-bus network [23]. The GD96_d network contains 180 nodes and 229 edges, and the power network has 494 nodes, 1381 edges. Based on these two networks, we compare the two approximation methods, Eqs. (8) and (9), to see which one is better when employing in our estimator. For the first approximation, we set $\varepsilon = 1$. For each network, 1000 independent runs are performed. The simulation result on GD96_d network is shown in Fig. 2(a). We can see that the identification accuracy of the FPT-based approximation is larger than the ε-based approximation. The distribution of minimum error distance of

the former is more to the left than the latter, which means in general the error distance of FPT-based approximation is smaller than the ε-based approximation.

The initial power network dose not have any terminal nodes. Thus, we set all the nodes with only one out-edge as terminal nodes, and randomly choose two non-terminal nodes as the real infection sources. As shown in Fig. 2(b), we can also see that the FPT-based approximation is better than the ε-based approximation for our estimator in terms of identification accuracy and minimum error distance. The advantage of FPT-based approximation is that it can differentiate the impacts of different source nodes better than the ε-based approximation.

4 Conclusion

In summary, we propose a method to identify multiple sources of random walk-based epidemic spreading process. Our method is based on the maximum likelyhood estimate. When deriving the estimator, we consider the different possibilities of different source nodes infecting a terminal node and all possible spreading paths from a source node to a terminal node. We propose two approximation methods to quantifying the possibility of a certain source node in set s infecting terminal nodes, which are ε-based approximation and FPT-based approximation. We validate our method on model networks and real-world networks by investigating the distribution of minimum error distance. Experimental results show that the performance of our method increases with network heterogeneity, while decreases with network density. Moreover, the FPT-based approximation is better than the ε-based approximation. Since we enumerate all possible node combinations in the calculation, the time complexity of our identification method is $O(n^{|s^*|+1})$. In the future, we will develop fast algorithms to solve the multiple-source identification problem.

References

1. Brightwell, G., Winkler, P.: Counting linear extensions. Order **8**(3), 225–242 (1991)
2. Valiant, L.G.: The complexity of enumeration and reliability problems. SIAM J. Comput. **8**(3), 410–421 (1979)
3. Shah, D., Zaman, T.: Detecting sources of computer viruses in networks: theory and experiment. In: ACM SIGMETRICS Performance Evaluation Review, vol. 38, pp. 203–214. ACM (2010)
4. Shah, D., Zaman, T.: Rumors in a network: who's the culprit? IEEE Trans. Inf. Theory **57**(8), 5163–5181 (2011)
5. Shah, D., Zaman, T.: Rumor centrality: a universal source detector. In: ACM SIGMETRICS Performance Evaluation Review, vol. 40, pp. 199–210. ACM (2012)
6. Luo, W., Tay, W.-P., Leng, M.: Identifying infection sources and regions in large networks. IEEE Trans. Signal Process. **61**(11), 2850–2865 (2013)
7. Luo , W., Tay, W.P.: Identifying multiple infection sources in a network. In: 2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), pp. 1483–1489. IEEE (2012)

8. Luo, W., Tay, W.P.: Identifying infection sources in large tree networks. In: 2012 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON), pp. 281–289. IEEE (2012)
9. Zhu, K., Ying, L.: Information source detection in the SIR model: a sample-path-based approach. IEEE/ACM Trans. Netw. (TON) **24**(1), 408–421 (2016)
10. Luo, W., Tay, W.P.: Estimating infection sources in a network with incomplete observations. In: 2013 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 301–304. IEEE (2013)
11. Restrepo, J.G., Ott, E., Hunt, B.R.: Characterizing the dynamical importance of network nodes and links. Phys. Rev. Lett. **97**(9), 094102 (2006)
12. Fioriti, V., Chinnici, M.: Predicting the sources of an outbreak with a spectral technique. arXiv preprint arXiv:1211.2333 (2012)
13. Comin, C.H., da Fontoura Costa, L.: Identifying the starting point of a spreading process in complex networks. Phys. Rev. E **84**(5), 056105 (2011)
14. Zang, W., Peng, Z., Zhou, C., Li, G.: Locating multiple sources in social networks under the SIR model: a divide-and-conquer approach. J. Comput. Sci. **10**, 278–287 (2015)
15. Lokhov, A.Y., Mézard, M., Ohta, H., Zdeborová, L.: Inferring the origin of an epidemic with a dynamic message-passing algorithm. Phys. Rev. E **90**(1), 012801 (2014)
16. Altarelli, F., Braunstein, A., et al.: Bayesian inference of epidemics on networks via belief propagation. Phys. Rev. Lett. **112**(11), 118701 (2014)
17. Antulov-Fantulin, N., Lančić, A., Šmuc, T., Štefančić, H., Šikić, M.: Identification of patient zero in static and temporal networks: robustness and limitations. Phys. Rev. Lett. **114**(24), 248701 (2015)
18. Prakash, B.A., Vreeken, J., Faloutsos, C.: Efficiently spotting the starting points of an epidemic in a large graph. Knowl. Inf. Syst. **38**(1), 35–59 (2014)
19. Prakash, B.A., Vreeken, J., Faloutsos, C.: Spotting culprits in epidemics: how many and which ones? In: 2012 IEEE 12th International Conference on Data Mining (ICDM), pp. 11–20. IEEE (2012)
20. Horn, A.L., Friedrich, H.: Locating the source of large-scale diffusion of foodborne contamination. arXiv preprint arXiv:1805.03137 (2018)
21. Wang, S.P., Pei, W.J.: First passage time of multiple Brownian particles on networks with applications. Phys. A **387**(18), 4699–4708 (2008)
22. Kemeny, J.G., Snell, J.L.: Finite Markov Chains: With a New Appendix "Generalization of a Fundamental Matrix". Springer, New York (1983)
23. Rossi, R.A., Ahmed, N.K.: The network data repository with interactive graph analytics and visualization. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)