A Simple Approach to the Multiple Source Identification of Information Diffusion

Xiaojie Li, Xin Yu, Chubing Guo, Yuxin Wang, Jianshe Wu

A Simple Approach to the Multiple Source Identification of Information Diffusion

Xiaojie Li[1], Xin Yu[1], Chubing Guo[1], Yuxin Wang[1], Jianshe Wu[1]

[1]Xidian University, China

[2]The 20th Research Institute of China Electronics Technology Group Corporation - Xidian University Joint Laboratory of Artificial Intelligence, China

Abstract. This paper studies the problem of identifying multiple information sources in networks. Assuming that the information diffusion follows a Susceptible-Infected (SI) model which allowing all nodes in the network are in the susceptible state or infected state. The number of information sources is known, we propose a simple method to identify multiple diffusion sources. After the diffusion started, any node can be infected very quickly, for an arbitrary node, it is infected by its closest source in terms of spreading time. Therefore, to identify multiple diffusion sources, we partition the information diffusion network and minimize the sum of all partition propagation times. Then in each partition we can find a node which has the minimum spreading time as diffusion source. Furthermore, we also give a new method to estimate the spreading time which can improve the proposed multiple sources identification algorithm. Simulation results show that the proposed method has distinct advantages in identifying multiple sources in various real-world networks.

Keywords: Complex networks, information diffusion, multiple sources identification, SI model.

1 INTRODUCTION

Interconnection in networks brings us many conveniences, but it also makes us vulnerable to various network risks. For instance, disease or epidemic spread in the human society[1],[2], rumors spread incredibly fast in online social networks such as WeChat, Facebook, and Twitter[3], etc. To contain these network risks, we need to identify the source of the propagation accurately[4]. By accurately identifying the sources, we can predict its further spread and develop timely mitigation strategies.

In the past few years, some methods are designed to work on tree-like networks whose propagation follows the classical Susceptible-Infected (SI) model [5]-[9]. Some other methods to identify diffusion sources in tree-like networks but with different epidemic models, such as the Susceptible-Infected- Recovery (SIR) model and the Susceptible-Infected-Susceptible (SIS) model [10]-[12]. These methods mainly aim at the identification of a single diffusion source. It is

shown that even in tree networks the source identification problem is NP-complete[5]. Recently, some heuristic methods are designed to relax the constraints from tree-like topologies to general networks [13]-[16].

Usually there is not a single source in the initial stage of the diffusion, such as infection disease can start from multiple locations. However, only a few of existing methods are designed to identifying multiple diffusion sources, such as the multi-rumor-center method[12], dynamic age method[17] and *K*-Center method[18]. In this paper, we adopt the SI model and propose a simple method named KST method to identify multiple diffusion sources in general networks. In general, a node is infected by its closest source in terms of spreading time, so we can identify diffusion sources by minimize the sum of all partition propagation times. At first we partition the diffusion network into several parts, then we find a node for each partition which has the minimum spreading time as the diffusion sources. We evaluate the KST method in the North American Power Grid and the Yeast protein-protein interaction network. Experiments show that our method outperforms the other competing methods.

2 RELATED WORKS AND MOTIVATIONS

We briefly introduce several state-of-the-art methods about multiple diffusion sources identification.

2.1 Related Methods

Dynamic Age. Fioriti *et al.* propose a method to identify multiple diffusion sources under the SI model[17]. The method only requires the topology of an undirected network and the infected network. However, due to the high complexity $O(N^3)$, the dynamic age method is not suitable in large-scale networks.

Multiple Rumor Center. Shah and Zaman introduce a rumor centrality method for single source identification in tree-like networks[5],[7]. The rumor centrality is defined as the number of distinct propagation paths originating from the source. The node with the maximum rumor centrality is called rumor center. For regular trees, the rumor center is considered as the propagation origin. Based on the definition of rumor centrality for a single node, Luo *et al.* extend the rumor centrality to a set of nodes [8], which is defined as the number of distinct propagation paths originating from the node set.

The computational complexity of this method is $O(n^k)$, where n is the number of infected nodes and k is the number of sources. Similar as the dynamic age, the method does not consider the propagation probabilities.

K-Center. Jiang *et al.* propose the *K*-Center method on the SI model to identify multiple diffusion sources in general networks [18]. They adopt a measure named as effective distance proposed in[13] to transform propagation probability to distance. The effective distance is defined as

$$e(i,j) = 1 - \log(\eta_{ij}) \tag{1}$$

where η_{ij} is the propagation probability from v_i to v_j. The concept of effective distance reflects the idea that a small propagation probability is equivalent to a large distance between them, and vice versa.

Based on the altered network, they derive an objective function for the multiple source identification:

$$\min f = \sum_{i=1}^{k} \sum_{v_j \in C_i} d(v_j, s_i) \tag{2}$$

where v_j is the infection node associated with source s_i, and $d(v_j, s_i)$ is the shortest path distance in terms of effective distance between v_j and s_i. By using the effective distance, the problem is simplified and *K*-Center has higher accuracy.

2.2 Motivations

Motivation 1. Existing source identification methods only take the probability of propagation into consideration and not consider the propagation time. In this paper, we take the propagation time into consideration and define an objection function for the multiple source identification problems by minimizing the sum of all partition propagation times.

Motivation 2. The multiple rumor center methods introduced above has computational complexity $O(n^k)$, and the complexity of Dynamic Age method is also $O(n^3)$. In most cases, identifying propagation sources quickly is of great significance in the real world.

3 Preliminaries and Problem Formulation

3.1 Susceptible-Infected (SI) Model

In the SI model, the nodes in a network have two possible states: susceptible (S) and infected (I). All the nodes are in the susceptible state initially except the diffusion sources. Infected nodes are those nodes that possess the infection and will remain infected throughout. Susceptible nodes are uninfected nodes, but may receive the infection from their infected neighbors and become infected.

Suppose the diffusion start at time $T = 0$, and we use $P_S(i,t)$ and $P_I(i,t)$ to denote the probability of node v_i being susceptible and infected at time $T = t$ respectively. At time $T = 0$, all infected nodes, namely the diffusion sources have the initial state of $P_S(i,t) = 0$ and $P_I(i,t) = 1$. Similarly all susceptible nodes have the initial state of $P_S(i,0) = 1$ and $P_I(i,t) = 0$. Then using following formulas, we can obtain the probability of each node in various states at an arbitrary time.

$$P_S(i,t) = [1 - Inf(i,t)] \cdot P_S(i,t-1) \tag{3}$$

$$P_I(i,t) = Inf(i,t) \cdot P_S(i,t-1) + P_I(i,t-1) \tag{4}$$

where $Inf(i,t)$ denotes the probability of node v_i to be infected by its neighbors at time $T = t$, which can be computed by

$$Inf(i,t) = 1 - \prod_{j \in N_i} [1 - \eta_{ji} \cdot P_I(j,t-1)] \tag{5}$$

where the η_{ji} denotes the propagation probability from node v_j to its neighboring node v_i, and N_i denotes the set of neighbors of node v_i.

3.2 Problem Formulation

Suppose there are $k(k \geq 1)$ sources: $S^* = \{s_1, s_2, \dots s_k\}$, and these sources start diffusion simultaneously at time $T = 0$. After the diffusion sustains for several time ticks, there are N nodes infected. These nodes form a connected networks $G(V,E)$, which we call the infection network.

Given an infection network G and the propagation probability between any two connected nodes, the problem is to identify a set of diffusion sources S^*.

4 KST METHOD

Suppose the diffusion source number k is known, we propose KST method to identify multiple diffusion sources in general networks. We firstly analyze the problem formulation for the KST method. Then we introduce the KST method for detail including two variables that need to be calculated in advance and the specific iterative process of KST method.

4.1 Analysis

According to the previous introduction, the diffusion sources start diffusion simultaneously, and for each source s_i, it has its infection region $C_i(\subseteq V)$. In other words, the infected network is composed of k regions, $C^* = \{C_1, C_2, \dots C_k\}$, supposing $C^* = \bigcup_{i=1}^{k} C_i$ and $C_i \cap C_j = \emptyset$ for $i \neq j$. In each region, there is only one source and the infection of other nodes in the region can be traced back to the corresponding source.

To identify the multiple diffusion sources S^*, we suppose the time of infection process is very short, therefore for an arbitrary infected node v_j in V we consider it is infected by the closest source in terms of spreading time. According to previous analysis, we can divide the infection network G into k partitions, so that each infected node belonging to the partition has the shortest spreading time to the corresponding sources. Then we get a partition of G and we think each partition is similar to the real region of infection network. The source node should be the node from which the diffusion takes the shortest spreading time to cover all the partition.

From the above analysis, we define an objective function as follow, which aims to find a partition of the infection network minimizing the sum of all partition propagation times as much as possible.

$$\min_{C^*} f = \sum_{i=1}^{k} t_i \tag{6}$$

where t_i is the spreading time of partition C_i, which is computed by

$$t_i = \max\{h(s_i, v_j) | v_j \in C_i\} \tag{7}$$

where node v_j belongs to partition C_i associated with source s_i and $h(s_i, v_j)$ is the minimum number of hops between s_i and v_j. The formula suggests that we can find a partition which can minimize the sum of spreading time of each partition, then for each partition the node with the minimum spreading time is the diffusion source.

4.2 KST Method

We first introduce two elementary knowledge including hop-based spreading time and propagation probability. These two variables need to be obtained in advance. Secondly, from a set of initial nodes in the infection network, an iterative approach is proposed to locate the diffusion sources.

4.2.1 Hops-Based Spreading Time

In SI model, it at least takes one time tick for the diffusion from one node to its neighbor, which means the spreading time can be estimated by the minimum number

of hops[18].

Given an infection network $G(V,E)$, where $V=\{v_1,v_2,\dots,v_n\}$ and $E=\{e_{ij}\},i,j\in\{1,\dots,n\}$, are the sets of nodes and edges, respectively. Suppose the network has one information source s, for an arbitrary node v_i, let $h(s,v_i)$ denote the minimum number of hops between s and v_i, which can be simply considered as the spreading time between s and v_i. Then, the spreading time is

$$t=\max\{h(s,v_i)|v_i\in V\} \tag{8}$$

The hops number of shortest path between node v_i and node v_j is regard as the spreading time between node v_i and node v_j.

4.2.2 Propagation Probability

For the infection network $G(V,E)$, there exists a propagation probability η_{ij} between any pair of neighbor nodes v_i and v_j. For any pair of unconnected nodes, the propagation probability is computed by the shortest path based method. That is to say, we assume that the propagation of information between any two nodes in the network is along the shortest path, and the propagation probability of the shortest path is the propagation probability between these nodes. In general, the shortest path has the highest propagation probability compared to other paths.

In order to calculate the propagation probability between any two nodes in the network, we first make the following conversion on each edge in the network:

$$q(i,j)=-\log(\eta_{ij}) \tag{9}$$

where η_{ij} is the probability of propagation on the edge. Using $q(i,j)$ as the weight of the edge, we can find the shortest path between by using a shortest path algorithm, e.g. the Dijkstra algorithm used in this paper. Then convert $q(i,j)$ to the probability of propagation between v_i and v_j by Eq.(10):

$$p(i,j)=e^{-q(i,j)} \tag{10}$$

4.2.3 Initialize Sources Set

First of all, we need to select k nodes as initial sources. Here, we choose them with the maximum distance because these nodes most likely are associated with different sources. We say an infected node v is associated with source s if the infection of v is traced back to source s. To select the initial sources, we first select a pair of infected nodes with the maximum distance, then we select other $k-2$ nodes greedily. The detail is shown in Algorithm 1.

4.2.4 Network Partition with Multiple Sources

Given an infection network G_n and a set of sources $S^*=\{s_1,s_2,\dots s_k\}$, we should partition G_n into k partitions. According to our previous analysis, each node $v_j\in G_n$ should be classified into partition C_i associated with source s_i, such that

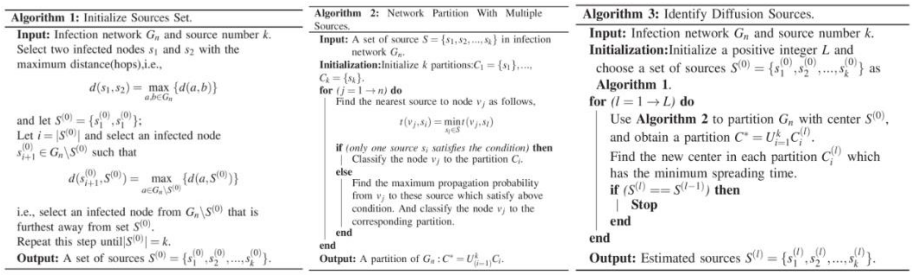
$$t(v_j,s_i)=\min_{s_l\in S} t(v_j,s_l) \tag{11}$$

The detail is presented in Algorithm 2. For each node v_j, we find the minimum spreading time from v_j to these sources. When there is only one source, denoted s_i, which satisfies above condition, we classify the node v_j into partition C_i associated with source s_i. If there are equal or greater than two sources which satisfy the condition, we further compare the propagation probability from v_j to these sources and classify the v_j to the source which has the maximum propagation probability

between itself and v_j.

4.2.5 Identifying Diffusion Sources

The complete process of KST is presented in this subsection. We first use Algorithm 1 to select an initial set of sources, then partition the infection network into k partitions by Algorithm 2. The partition algorithm can find a local optimal solution for (6). In order to further minimize the objective function, our method selects a new source for each partition which has the minimum spreading time in each partition. We call the method as K Shortest spreading Time nodes (KST). The detailed process of the KST is shown in Algorithm 3. The main computation is the calculation of the shortest path between node pairs. In the simulations, the Dijkstra algorithm is used to calculate the shortest path, whose computational complexity is $O(n^2)$. Therefore, the computational complexity of the KST is $O(n^2)$.



Algorithm 1: Initialize Sources Set.

Input: Infection network G_n and source number k.
Select two infected nodes s_1 and s_2 with the maximum distance(hops),i.e.,

$$d(s_1,s_2)=\max_{a,b\in G_n}\{d(a,b)\}$$

and let $S^{(0)}=\{s_1^{(0)},s_1^{(0)}\}$;
Let $i=|S^{(0)}|$ and select an infected node $s_{i+1}^{(0)}\in G_n\backslash S^{(0)}$ such that

$$d(s_{i+1}^{(0)},S^{(0)})=\max_{a\in G_n\backslash S^{(0)}}\{d(a,S^{(0)})\}$$

i.e., select an infected node from $G_n\backslash S^{(0)}$ that is furthest away from set $S^{(0)}$.
Repeat this step until$|S^{(0)}|=k$.
Output: A set of sources $S^{(0)}=\{s_1^{(0)},s_2^{(0)},\ldots,s_k^{(0)}\}$.

Algorithm 2: Network Partition With Multiple Sources.

Input: A set of source $S=\{s_1,s_2,\ldots,s_k\}$ in infection network G_n.
Initialization: Initialize k partitions:$C_1=\{s_1\},\ldots,C_k=\{s_k\}$.
for $(j=1\rightarrow n)$ **do**
  Find the nearest source to node v_j as follows,

$$t(v_j,s_l)=\min_{s_l\in S}t(v_j,s_l)$$

  if *(only one source s_i satisfies the condition)* **then**
    Classify the node v_j to the partition C_i.
  else
    Find the maximum propagation probability from v_j to these source which satisfy above condition. And classify the node v_j to the corresponding partition.
  end
end
Output: A partition of $G_n: C^*=U_{i=1}^{k}C_i$.

Algorithm 3: Identify Diffusion Sources.

Input: Infection network G_n and source number k.
Initialization: Initialize a positive integer L and choose a set of sources $S^{(0)}=\{s_1^{(0)},s_2^{(0)},\ldots,s_k^{(0)}\}$ as **Algorithm 1**.
for $(l=1\rightarrow L)$ **do**
  Use **Algorithm 2** to partition G_n with center $S^{(0)}$, and obtain a partition $C^*=U_{i=1}^{k}C_i^{(l)}$.
  Find the new center in each partition $C_i^{(l)}$ which has the minimum spreading time.
  if $(S^{(l)}==S^{(l-1)})$ **then**
    Stop
  end
end
Output: Estimated sources $S^{(l)}=\{s_1^{(l)},s_2^{(l)},\ldots,s_k^{(l)}\}$.

5 KST-IMPROVED METHOD

In KST, the shortest path is used to estimate the spreading time between two nodes, without considering the influences of propagation probability. In this subsection, propagation probability is used to estimate the spreading time between two nodes, which is defined as follows:

$$et(i,j)=-\log(p_{ij}) \tag{12}$$

where p_{ij} is the propagation probability between v_i and v_j, refer to subsection 4.2 for the calculation method. We call the new spreading time as the effective spreading time.

Then we partition the infection network G_n into k partitions according to the effective spreading time. For $v_j\in G_n$, it should be classified into partition C_i associated with source s_i, such that

$$et(v_j,s_i)=\min_{s_l\in S}et(v_j,s_l) \tag{13}$$

We denote the improved KST method as KST-Improved. Experiments show that the effective propagation time can improve accuracy of multiple source identification.

6 EVALUATION

6.1 Experiments Settings

The North American Power Grid network and the Yeast protein-protein interaction network are used. Table 1 gives the basic statics of the networks. Previous works[19]

show that the accuracy of the SI model cannot be affected by the distribution of propagation probability on each edge, therefore we set the propagation probability on each edge, η_{ij}, uniformly distributed in (0, 1). We set the number of infection source k to be 2 to 5 and perform 100 simulation runs for each k. In each simulation, we randomly select k nodes from network as the initial source and simulate the propagation process using the SI model. Each simulation terminates when the number of infected node is greater than a number such as 100 or the spreading time is equal a number such as 5.

6.2 Accuracy of Identifying Sources

We match the estimated sources with the real sources so that the sum of the error distances is minimized [错误!未找到引用源。,8]. The average error distance is given by

$$\Delta = \frac{1}{|S^*|}\sum_{i=1}^{|S^*|} h(s_i, \hat{s}_i) \tag{21}$$

where $s_i \in S^* = \{s_1, s_2, \dots s_k\}$, and $\hat{s}_i$ is the estimated sources corresponding to s_i.

The simulation result is given in Table 2, which show than KST and KST-Improved have the smaller average error distances. KST-Improved is better than KST method, which proves that the effective spreading time is more precise than hop-based spreading time.

Table 1. Statistics of data collected in the experiments

Dataset	Power Grid	Yeast
Number of nodes	4941	2361
Number of edges	13188	13554
Average degree	2.67	5.74
Maximum degree	19	64

Table 2. Average error distance of three methods

Experiment settings		Average error distance			
Network	*k*	KST	KST -Improved	K-Center	Dynamic Age
Power Grid	2	1.39	1.31	1.78	2.593
	3	1.67	1.687	2.152	3.434
	4	1.775	1.6	2.373	3.957
	5	1.893	1.741	2.725	4.467
Yeast	2	0.925	0.91	1.721	3.057
	3	0.98	0.936	2.222	3.823
	4	1.15	1.099	2.383	3.7
	5	1.226	1.158	3.04	4.133

7 CONCLUSION

We provided a simple method for general network to detect multiple information sources, its computational complexity is $O(n^2)$, which is much less than other methods. We propose a new measure to estimate the spreading time between nodes from the propagation probability, which improves the accuracy of source identification.

References

1. Neumann, G, T. Noda, and Y. Kawaoka. "Emergence and pandemic potential of swine-origin H1N1 influenza virus. " *Nature* 459.7249(2009):931-9.
2. Hvistendahl, M, D. Normile, and J. Cohen. "Influenza. Despite large research effort, H7N9 continues to baffle. " Science 340.6131(2013):414.
3. B. Doerr, M. Fouz, and T. Friedrich, "Why rumors spread so quickly in social networks," Commun. ACM, vol. 55, no. 6, pp. 70–75, Jun. 2012.
4. J. Jiang, S. Wen, S. Yu, Y. Xiang, and W. Zhou, "Identifying propagation sources in networks: State-of-the-art and comparative studies," *IEEE Commun. Surv. Tuts.*
5. D. Shah and T. Zaman, "Rumors in a network: Who's the culprit?" *IEEE Transactions on Information Theory*, vol. 57, pp. 5163 – 5181, 2011.
6. N. Karamchandani and M. Franceschetti, "Rumor source detection under probabilistic sampling," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, 2013, pp. 2184–2188.
7. D. Shah and T. Zaman, "Rumor centrality: A universal source detector," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 1, pp. 199–210, Jun. 2012.
8. W. Luo, W. P. Tay, and M. Leng, "Identifying infection sources and regions in large networks," *Signal Processing, IEEE Transactions on*, vol. 61, no. 11, pp. 2850–2865, 2013.
9. C. Milling, C. Caramanis, S. Mannor, and S. Shakkottai, "On identifying the causative network of an epidemic," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE, 2012, pp. 909–914.
10. K. Zhu and L. Ying, "Information source detection in the sir model: A sample path based approach," in *Information Theory and Applications Workshop (ITA)*, 2013, pp. 1–9.
11. K. Zhu and L. Ying, "A robust information source estimator with sparse observations," *arXiv preprint arXiv:1309.4846*, 2013.
12. W. Luo and W. P. Tay, "Finding an infection source under the sis model," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 2930 – 2934.
13. D. Brockmann and D. Helbing, "The hidden geometry of complex, network-driven contagion phenomena," *Science*, vol. 342, no. 6164, pp. 1337–1342, 2013.
14. F. Altarelli, A. Braunstein, L. DallAsta, A. Lage-Castellanos, and R. Zecchina, "Bayesian inference of epidemics on networks via belief propagation," *Physical review letters*, vol. 112, no. 11, p. 118701, 2014.
15. Z. Chen, K. Zhu, and L. Ying, "Detecting multiple information sources in networks under the sir model," in Information Sciences and Systems (CISS), 2014 48th Annual Conference on. IEEE, 2014, pp. 1–4.
16. W. Luo and W. P. Tay, "Identifying multiple infection sources in a network," in Signals, Systems and Computers (ASILOMAR), 2012 Conference Record of the Forty Sixth Asilomar Conference on. IEEE, 2012, pp. 1483–1489.
17. V. Fioriti, M. Chinnici, and J. Palomo, "Predicting the sources of an outbreak with a spectral technique," *Appl. Math. Sci.*, vol. 8, no. 135, pp. 6775–6782, 2014.
18. J. Jiang, S. Wen, S. Yu, Y. Xiang and W. L. Zhou, "K-Center: An Approach on the Multi-Source Identification of Information Diffusion" IEEE Trans. Information Forensics and Security, vol. 10, no. 12, pp. 2616-2626, Dec. 2015.
19. S. Wen, W. Zhou, J. Zhang, Y. Xiang, W. Zhou, and W. Jia, "Modeling propagation dynamics of social network worms," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 8, pp. 1633–1643, Aug. 2013.