Multi-source localization on complex networks with limited observers

epl A LETTERS JOURNAL EXPLORING THE FRONTIERS OF PHYSICS

January 2016

EPL, **113** (2016) 18006
doi: 10.1209/0295-5075/113/18006

www.epljournal.org

Multi-source localization on complex networks with limited observers

Ling Fu, Zhesi Shen, Wen-Xu Wang[a], Ying Fan[b] and Zengru Di

School of Systems Science, Beijing Normal University - Beijing, 100875, PRC

received 12 October 2015; accepted in final form 11 January 2016
published online 3 February 2016

PACS 89.20.Hh – World Wide Web, Internet
PACS 89.75.Fb – Structures and organization in complex systems

Abstract – Source localization is a significant task in the contagion process. In this paper, we study the problem of locating multiple sources in complex networks with limited observations. We propose a backward diffusion-based source localization method and apply it on several networks, finding that multiple sources can be located with high accuracy even when the fraction of observers is small and the time delay along the links are not known exactly. By comparing different observer placement strategies, we find that choosing small-degree nodes as observers is better than the other strategies.

Introduction. – Many diffusion dynamics taking place on complex networks are initiated from a small number of nodes, *e.g.*, an epidemic spreading among contact relationships, a rumour propagating through social networks, and a computer virus spreading through the Internet. Locating the sources of diffusion efficiently and accurately has very important applications in practice. To date, several approaches have been proposed to identify the source. Shah and Zaman [1] were pioneers in systematically studying the problem of estimating infection sources in a susceptible-infected (SI) model. Later, several methods [2–7] were successively proposed to locate the source using SI, SIR and SIS models. Recently, some scholars extended their methods to the multi-source localization problem in an SI model situation [3,8]. Although the listed works answer some key questions about source localization, the assumption that the complete snapshot of the states of all nodes is available is expensive to obtain and is impossible for large networks. To overcome this limitation, many researchers [9–13] proposed several source detection methods based on partial observations in which only a limited fraction of nodes can be observed. However, most of these works with partial observations focus on the single-source detection problem, and little attention has been paid to multiple source localization problems. When dealing with multiple sources, the computational complexity of previous approaches grows exponentially with an increasing number of sources.

In this paper, we study the multiple source localization problem with partial observations. We propose a backward diffusion-based multi-source localization approach and test its performance on several networks. We also study the effect of observer placement strategy on localization accuracy.

The following paper is organized as follows. We first introduce the propagation model and propose the multi-source localization method. Then, we show the performance of our method with respect to fractions of observers, network structures and time delays along the links. Finally, we investigate the performance of different observer placement strategies.

Model and method. –

Propagation model. We first introduce the network diffusion model. The underlying network is defined as a finite, undirected graph $G = \{V, E\}$, which consists of a set of nodes V and a set of edges E. The topology of the network is assumed to be known and static during the diffusion process. The diffusion source set, $S = \{s_m\}_{m=1}^{M}$, is the set of the nodes that initiates the diffusion.

In previous works, numerous information/virus spread models are used to model the diffusion process, *e.g.*, SI, SIS, and SIR. Here, we use a simple diffusion model associated with a time delay to model the propagation that is also used in refs. [9,11,14]. The details are described as follows and are shown in fig. 1.

1) In the network, at time t, all nodes are in two states: informed (the node has the information) or uninformed

[a] E-mail: wenxuwang@bnu.edu.cn
[b] E-mail: yfan@bnu.edu.cn

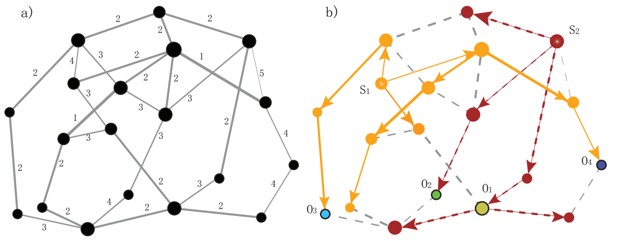

Fig. 1: (Color online) An example of the diffusion process. (a) is a weighted undirected graph G. The numbers on the edges are the corresponding time delays. The size of each node is in proportion to its degree. (b) is the diffusion process initiated from sources $S = \{s_1, s_2\}$. The spreading paths from the two sources s_1, s_2 are highlighted with orange arrows and brown arrows, respectively. The arrival time at observers $O = \{O_1, O_2, O_3, O_4\}$ is recorded.

(the node does not have the information). Only the source nodes are in the informed state at the initial time t^0.

2) The informed nodes forward the information to their uninformed neighbours. For clarity, let us assume that node i receives the information for the first time and becomes informed at time t_i. Then, i will transfer the information to all its neighbours, denoted as $H(i)$, so that each uninformed neighbour node $j \in H(i)$ receives the information at time $t_i + \theta_{L_{ij}}$ where $\theta_{L_{ij}}$ is the transmission time delay associated with edge L_{ij}. The time delay for different edges follows a known joint distribution, *e.g.*, Gaussian or uniform. Thus, for any node $i \in V$, the time t_i that i is informed is

$$t_i = t^0 + \min\{\Delta_{s_1,i}, \Delta_{s_2,i}, \cdots, \Delta_{s_M,i}\}, \tag{1}$$

where $\Delta_{s_m,i}$ is the shortest transmission time between s_m and i.

3) The process continues until all of the nodes in V have been informed.

Let $O = \{o_k\}_{k=1}^{K}$ denote the set of K observers, whose informed times are known. Differing from the settings in ref. [9] (they require the time and from whom they obtain the information), each observer can only measure the time it becomes informed. Specifically, the observers' informed time is $\{t_{o_1}, t_{o_2}, \ldots, t_{o_K}\}$.

Multisource localization. Once we obtain the final infected graph and enough observers, we can locate the sources with high accuracy.

For node i and source node s_m, if $\Delta_{s_m,i}$ is the minimum of $\{\Delta_{s_1,i}, \Delta_{s_2,i}, \ldots \Delta_{s_M,i}\}$, we say that node i is diffused by source s_m, that is $i \in \Pi_{s_m}$, where Π_{s_m} denotes the diffusion set of source s_m. Thus, the original vertex set V can be decomposed into M diffusion set and one source node set, *i.e.*, $V = \Pi_{s_1} + \Pi_{s_2} + \cdots + \Pi_{s_M} + S$.

Then, we can obtain the relationship between the informed time and the diffusion set. For any node $i \in \Pi_{s_m}$, according to eq. (1), the informed time is

$$t_i = t^0 + \min\{\Delta_{s_1,i}, \Delta_{s_2,i}, \ldots, \Delta_{s_M,i}\} = t^0 + \Delta_{s_m,i}. \tag{2}$$

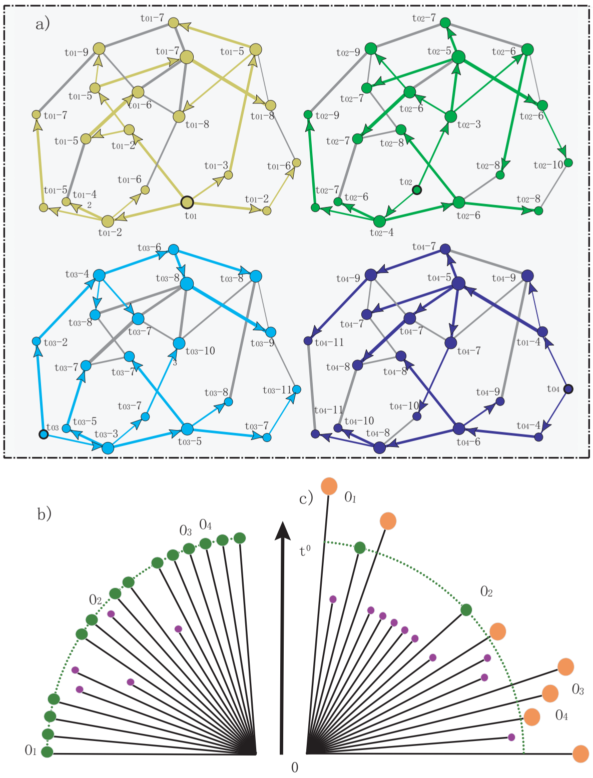

Fig. 2: (Color online) Illustration of the backward diffusion-based multi-source localization method. (a) Backward diffusions from each observer; the number marked at each node is $\Gamma(i, o_k) = t_{o_k} - \Delta_{i,o_k}$. (b) For a source node s_m, all values of $\Gamma(s_m, i)$, $\forall i \in V$ are arranged in a circle with radius proportional to $\Gamma(s_m, i)$. All the points are bounded in a circle whose radius is t^0, that is $\Gamma_{s_m}^{\max} = t^0$. (c) For a normal node $j \notin S$, some values of $\Gamma(j, i)$ will exceed the boundary, resulting in $\Gamma_j^{\max} > t^0$. The sources can be then identified as the nodes whose $\Gamma_s^{\max}$ are minimal.

Thus, for any pair of informed node and source node, *e.g.*, i and s, respectively, the estimated informed time of s from the viewpoint of node i is

$$\Gamma(s, i) \equiv t_i - \Delta_{s,i} \le t^0. \tag{3}$$

When $i \in \Pi_s$,

$$\begin{aligned} t_i &= t^0 + \min\{\Delta_{s_1,i}, \Delta_{s_2,i}, \ldots, \Delta_{s_M,i}\} = t^0 + \Delta_{s,i}, \\ \Gamma(s, i) &= t_i - \Delta_{s,i} = t^0; \end{aligned}$$

when $i \notin \Pi_s$,

$$\begin{aligned} t_i &= t^0 + \min\{\Delta_{s_1,i}, \Delta_{s_2,i}, \ldots, \Delta_{s_M,i}\} < t^0 + \Delta_{s,i}, \\ \Gamma(s, i) &= t_i - \Delta_{s,i} < t^0, \end{aligned}$$

then, for s, the maximum of $\Gamma(s, i)$ for all $i \in V$, denoted as $\Gamma_s^{\max}$, is t^0. However, for a node $j \notin S$, there must exist some $i \in V$, such that $\Gamma(j, i) = t_i - \Delta_{j,i} > t^0$, for example, the node j itself,

$$\Gamma(j, j) = t_j - \Delta_{j,j} = t_j > t^0,$$

so $\Gamma_j^{\max} > t^0$.

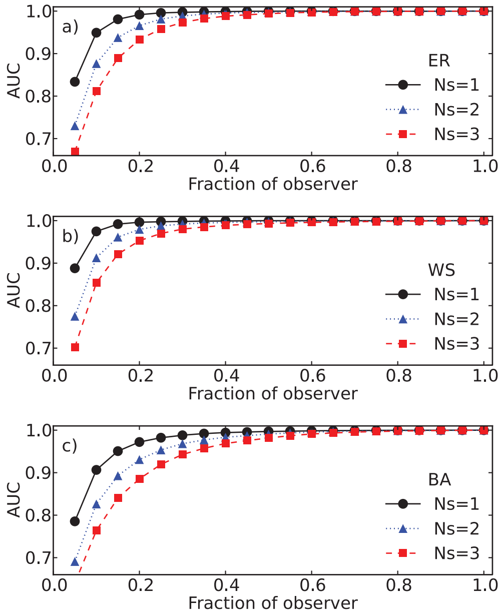

Fig. 3: (Color online) Performance of multiple source localization on model networks. (a) ER random network. (b) WS small-world network; the rewiring probability is 0.1. (c) BA scale-free network. The network size used here is $N = 100$, and the average degree is $\langle k \rangle = 4$. The time delay along the links follows a Gaussian distribution $N(1, 0.25^2)$ and only the mean delay of all links is used to identify the sources.

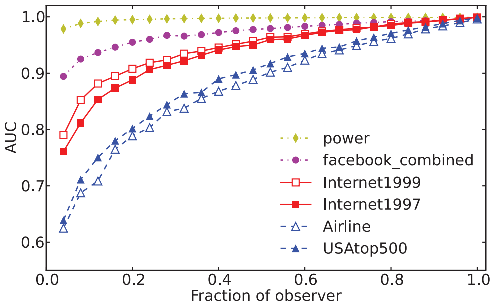

Fig. 4: (Color online) Performance of multiple source localization on real networks. The number of sources is 3 and the time delay associated with each link follows $N(1, 0.25^2)$. Detailed information from all the networks is shown in table 1.

Table 1: Description of real networks analysed in this paper. N and L denote the total numbers of nodes and links, respectively.

Type	Name	N	L
Internet	Internet1997 [15]	3015	5156
	Internet1999 [15]	5357	10328
Transportation	USAtop500 [16]	500	2890
	Airline [17]	332	2126
Power Grid	Power [18]	4941	6594
Social network	Facebook [19]	4039	88234

Figure 2(a) shows the backward diffusion initiated from each observer, and the estimated $\Gamma(*, o_k)$ is marked at each node. As for a source node s_m, by arranging all the values of $\Gamma(s_m, i)$ $\forall i \in V$ in a circle, shown in fig. 2(b), all the points are bounded in the boundary representing t^0; while for a node, j, that is not a source, shown in fig. 2(c), some values of $\Gamma(j, i)$ will exceed the boundary, resulting in $\Gamma_j^{\max} > t^0$. According to the above analysis, we can see that the values of the sources' $\Gamma_s^{\max}$ are minimal; then, we can identify the sources based on this rule. Specifically, for each node in the network, we can calculate the value of $\Gamma^{\max}$ against all the observers and select the nodes with the smallest value of $\Gamma^{\max}$ as the source nodes,

$$\begin{aligned} \hat{S}(O) &= \mathrm{argmin}_{i \in V}\, \Gamma_i^{\max} \\ &= \mathrm{argmin}_{i \in V} \{\mathrm{argmax}_{o \in O}\, \Gamma(i, o)\}. \end{aligned} \tag{4}$$

The computational complexity of our method is $O(KN \log N)$ and is independent of the number of sources.

Results. – To quantify the validity and efficiency of our multiple source localization approach in terms of the fraction of observers, we study the success rate of locating source nodes for homogeneous and heterogeneous network structures, including the Erdös-Rényi (ER) random network, the Watts-Strogatz (WS) small-world network, the Barabási-Albert (BA) scale-free network and some real networks. Here the area under the receiver operating characteristic curve (AUC) is used to quantify the localization performance of our approach. We first rank the nodes based on their values of $\Gamma_i^{\max}$ in ascending order. The true positive rate (TPR) and false positive rate (FPR) that are used to calculate AUC are defined as follows:

$$TPR(l) = \frac{TP(l)}{M} \tag{5}$$

where $TP(l)$ is the number of true positives in the top l predictions in the candidate list, and M is the number of sources,

$$FPR(l) = \frac{FP(l)}{N - M}, \tag{6}$$

where $FP(l)$ is the number of false positives in the top l predictions in the candidate list. The higher the value of AUC is, the better localization performance will be.

Figure 3 shows the localization accuracy against a different fraction of observers for a different number of sources. The sources and observers are selected randomly. As we can see, for all the network structure and number of sources, only approximately 20% of observers are needed, rendering relatively high localization accuracy ($AUC \geq 0.9$). Additionally, the increase of sources enhances the challenge of successfully locating the sources. Compared with the results of the ER network and the WS network, the sources of diffusion in the BA network are more challenging to locate.

We also use several types of real networks to test the performance of our method, including the power

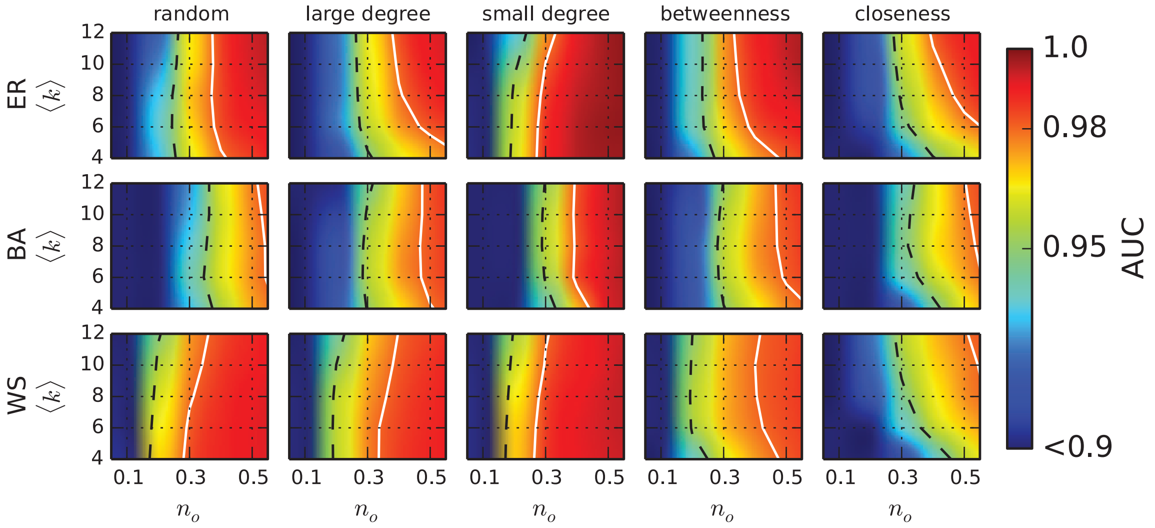

Fig. 5: (Color online) Performance of different observer placement strategies for ER, BA and WS networks. The white solid lines and black dashed lines are the boundaries with an $AUC = 0.98$ and $AUC = 0.95$, respectively. The network size used here is $N = 100$, and the average degree is $\langle k \rangle = 4$. The rewiring probability of the WS network is 0.1.

Table 2: Minimum fraction of observers. The minimum fraction $n_o^{\min}$ of randomly selected observers that assures $AUC_{loc} = 0.9$ of locating the sources of spreading dynamics in ER, WS and BA networks. The time delay of the links follow a Gaussian distribution $N(1, 0.25^2)$, $N(1, 0.5^2)$ with different standard variance and uniform distribution in the range $(0.5, 1.5)$. The mean delay of all links is used to identify the sources. The number of sources is $M = 2$, the network size is $N = 100$ and the average node degree is $\langle k \rangle = 4$. The results are obtained by averaging over 1000 independent realizations.

	Network type		
Delay distribution	ER	WS	BA
$N(1.0, 0.25^2)$	0.15	0.11	0.22
$N(1.0, 0.50^2)$	0.30	0.19	0.44
$N(1.0, 1.00^2)$	0.93	0.90	0.94
$U(0.75, 1.25)$	0.13	0.09	0.16
$U(0.50, 1.50)$	0.17	0.12	0.25
$U(0.25, 1.75)$	0.39	0.17	0.51

grid network, social communication network, internet and transportation network, shown in fig. 4. The performance of localization on these types of networks differs significantly. The power grid network is the easiest one because of its regular structure; the localization accuracy achieves $AUC = 0.9$ with only approximately 2% of observers. However, for the air transportation networks, up to approximately 50% of observers are needed for achieving an $AUC = 0.90$ in our simulation.

Infection time delay. The time delay of links will affect the source localization performance. Table 2 displays the minimum fraction $n_o^{\min}$ of observers for achieving $AUC_{loc} = 0.9$ for locating the sources in homogenous and inhomogeneous networks with a Gaussian distribution and a uniform distribution of time delay along the links. The results imply that our algorithm is effective when the variation of the time delays on the links is limited.

Performance of the observer placement strategy. In this section we focus on the performance of different observer placement strategies with different network topological properties. Here, we adopt several node centrality-based observer placement strategies, including large degree, small degree, large betweenness, and large closeness.

The localization accuracy of the basic random strategy and the above four observer placement strategies are investigated in ER, WS and BA networks, as shown in fig. 5. In this figure, the black dashed lines are the boundary of $AUC < 0.95$ and $AUC > 0.95$, and the white solid lines separate the regions of $AUC < 0.98$ and $AUC > 0.98$. As we can see, generally, the five observer placement strategies show similar performance on each network topology; each strategy shows a similar performance pattern across the three types of networks. Compared with all the other strategies, the *small-degree* strategy achieves a relatively higher accuracy.

Discussions. – In this work, we investigated a multiple source localization problem and proposed a source localization method based on backward diffusion. The computational complexity of our method is comparatively low and is independent of the number of sources. Simulations on different networks show that we can obtain highly accurate estimations in identifying the sources. Regardless of this, the method could still be improved. First, simulations on networks with weight following a Gaussian distribution or a uniform distribution are satisfying, while other distributions remain to be studied. Second, we still do not know how to select the minimum number of observers in an arbitrary network. In this paper, we compare the performance of several observer placement strategies; the results imply that none of these strategies show a dominant performance. The recent developed observability [20] of complex networks is a promising method for efficiently placing the observers. We may overcome such obstacles by using the recently developed compressed sensing-based network reconstruction method [21,22]. Third, incorporating side information remains undone. In some practical conditions, we can obtain the direction of diffusion among neighbours and modify the diffusion graph for further investigations.

* * *

All authors contributed equally to this paper. This work was supported by NSFC under Grant Nos. 61174150, 61573065, 61573064 and 61074116, the Fundamental Research Funds for the Central Universities and the Beijing Nova Programme.

REFERENCES

[1] Shah D. and Zaman T., *Finding sources of computer viruses in networks: Theory and experiment,* in *Proceedings of ACM Sigmetrics*, Vol. **15** (ACM) 2010, pp. 5249–5262.

[2] Shah D. and Zaman T., *IEEE Trans. Inf. Theory*, **57** (2011) 5163.

[3] Luo W., Tay W. P. and Leng M., *IEEE Trans. Signal Process.*, **61** (2013) 2850.

[4] Zhu K. and Ying L., *Information source detection in the sir model: A sample path based approach,* in *Proceedings of Information Theory and Applications Workshop (ITA) 2013* (IEEE) 2013, pp. 1–9.

[5] Altarelli F., Braunstein A., DallAsta L., Lage-Castellanos A. and Zecchina R., *Phys. Rev. Lett.*, **112** (2014) 118701.

[6] Lokhov A. Y., Mézard M., Ohta H. and Zdeborová L., *Phys. Rev. E*, **90** (2014) 012801.

[7] Luo W. and Tay W. P., *Finding an infection source under the sis model,* in *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) 2013, pp. 2930–2934.

[8] Zang W., Zhang P., Zhou C. and Guo L., *Proc. Comput. Sci.*, **29** (2014) 443.

[9] Pinto P. C., Thiran P. and Vetterli M., *Phys. Rev. Lett.*, **109** (2012) 068702.

[10] Brockmann D. and Helbing D., *Science*, **342** (2013) 1337.

[11] Zhu K., Chen Z. and Ying L., arXiv preprint, arXiv:1412.4141 (2014).

[12] Zhu K. and Ying L., *Comput. Soc. Netw.*, **1** (2014) 1.

[13] Seo E., Mohapatra P. and Abdelzaher T., *Identifying rumors and their sources in social networks,* in *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR III, Proceedings of SPIE*, Vol. **8389** (International Society for Optics and Photonics) 2012, p. 83891I.

[14] Shen Z., Cao S., Fan Y., Di Z., Wang W.-X. and Stanley H. E., arXiv preprint, arXiv:1501.06133 (2015).

[15] `http://pil.phys.uniroma1.it/∼gcalda/cosinsite/extra/data/internet/nlanr.html`.

[16] Colizza V., Pastor-Satorras R. and Vespignani A., *Nat. Phys.*, **3** (2007) 276.

[17] Batagelj V. and Mrvar A., *Pajek datasets* (2006).

[18] Watts D. J. and Strogatz S. H., *Nature*, **393** (1998) 440.

[19] Leskovec J. and Mcauley J. J., *Learning to discover social circles in ego networks,* in *Proceedings of Advances in Neural Information Processing Systems* (NIPS) 2012, pp. 539–547.

[20] Liu Y.-Y., Slotine J.-J. and Barabási A.-L., *Proc. Natl. Acad. Sci. U.S.A.*, **110** (2013) 2460.

[21] Shen Z., Wang W.-X., Fan Y., Di Z. and Lai Y.-C., *Nat. Commun.*, **5** (2014).

[22] Tibshirani R., *J. R. Stat. Soc., Ser. B*, **58** (1996) 267.