Detecting Multiple Information Sources in Networks under the SIR Model

Zhen Chen, Kai Zhu and Lei Ying
School of Electrical, Computer and Energy Engineering
Arizona State University
Tempe, AZ, United States, 85287
Email: {zchen113, kzhu17, lei.ying.2}@asu.edu

***Abstract*—In this paper, we study the problem of detecting multiple information sources in networks under the Susceptible-Infected-Recovered (SIR) model. First, assuming the number of information sources is known, we develop a sample-path-based algorithm, named clustering and localization, for trees. For g-regular trees, the estimators produced by the proposed algorithm are within a constant distance from the real sources with a high probability. We further present a heuristic algorithm for general networks and an algorithm for estimating the number of sources when the number of real sources is unknown.**

I. Introduction

Recently, there have been a lot of interests in the problem of detecting information sources in networks. The solutions to this problem have important applications in practice, such as identifying the sources of infectious diseases and finding the sources of leaked confidential information. Shah and Zaman analytically studied this problem under the SI model and developed the rumor centrality estimator [1]–[3]. Detecting multiple information sources using the rumor centrality estimator has been investigated in [4], [5], and detecting a single information source with partial observations by using the rumor centrality estimator has been considered in [6]. In [7], the detection rate of the rumor centrality estimator when a priori distribution of the source node is given has been evaluated.

Besides the SI model, information source detection under the SIR model has also been studied. Zhu and Ying developed a sample-path-based estimator in [8] for detecting a single information source under the SIR model. They later proved that the sample path estimator remains to be an effective estimator even with sparse observations [9]. The effectiveness of the sample path estimator for the SI model with partial observations and for the SIS model have been investigated in [10] and [11], respectively.

Several other source detecting algorithms have also been proposed recently, including an eigenvalue-based estimator [12], a dynamic message-passing algorithm based on a mean-field-like approximation for the SIR model [13], a fast Monte Carlo algorithm [14], and an algorithm that utilizes sparsely placed observers [15].

This paper considers the problem of detecting multiple information sources under the SIR model. It is not uncommon to have multiple information sources. For example, confidential information can be leaked from different sources and an infectious disease can start from multiple locations. We study this multi-source detection problem under the SIR model. Motivated by the sample-path-based estimator proposed in [8], we first present a clustering and localization algorithm for tree networks, where the number of real sources is assumed to be known. We then prove that on g-regular trees, the distances between the estimators given by the algorithm and the real sources are upper bounded by a constant with a high probability. We further present a heuristic algorithm for general networks and an algorithm for estimating the number of sources when it is unknown.

II. Basic Model

In this section, we introduce the SIR model and the multi-source detection problem.

A. SIR Model

The network is defined to be an undirected graph $G(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges. Each node in graph G may represent a person, a computer or a mobile device. An edge represents a communication channel such that information can be transmitted from one node to another if there is an edge between these two nodes.

We consider a multi-source Susceptible-Infected-Recovered (SIR) model [16], [17] for information diffusion in networks. In the SIR model, every node has three states: susceptible (S), infected (I) and recovered (R) such that:

- a susceptible node may be infected by his/her infected neighbors,
- an infected node may recover, and
- a recovered node cannot be infected again.

We consider a time slotted system. At the beginning of each time slot, a susceptible node is infected by each of its infected neighbors with probability q, and each infected node recovers with probability p. Assuming the number of infected neighbors of a susceptible node is n, the probability the node becomes infected is $1-(1-q)^n$. Initially, at $t = 0$, all nodes are in the susceptible state except a set of source nodes, denoted by $\mathcal{S}$.

B. Problem Description

The objective of this paper is to locate the set of sources $\mathcal{S}$ given a snapshot of the network in which we can distinguish infected nodes from other nodes, but cannot distinguish susceptible nodes and recovered nodes. We say a node is healthy if the node is in either the susceptible state or recovered state.

Define X_v to be the state of node v in the given snapshot and $\mathbf{X} = \{X_v; v \in \mathcal{V}\}$, where

$$X_v = \begin{cases} 1, & \text{if } v \text{ is in the infected state;} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Denote by $\mathcal{V}_I$ the set of observed infected nodes in the snapshot. We further assume the number of sources ($S = |\mathcal{S}|$) is known.

III. Main Results

We summarize the main results in this section.

A. Multi-Source Detection on Tree Networks

In this section, we present a multi-source detection algorithm for tree networks, named clustering and localization (CL). For tree networks, we define the distance between two nodes a and b to the length of the path that connects node a and b, denoted by $d(a, b)$. We further define the distance between a node a and a set of nodes $\mathcal{N}$ to be the minimum distance from node a to any node in $\mathcal{N}$, i.e.,

$$d(a, \mathcal{N}) = \min_{b \in \mathcal{N}} d(a, b).$$

The algorithm is presented in Algorithm 1.

Algorithm 1 Clustering and Localization (CL)

1: Select two infected nodes e_1 and e_2 with the maximum distance, i.e.,
$$d(e_1, e_2) = \max_{a,b \in \mathcal{V}_I} d(a, b),$$
and let $\mathcal{B} = \{e_1, e_2\}$.
2: Let $i = |\mathcal{B}|$ and select an infected node $e_{i+1} \in \mathcal{V}_I \setminus \mathcal{B}$ such that
$$d(e_{i+1}, \mathcal{B}) = \max_{a \in \mathcal{V}_I \setminus \mathcal{B}} d(a, \mathcal{B}),$$
i.e., selecting an infected node from $\mathcal{V}_I \setminus \mathcal{B}$ that is furthest away from set $\mathcal{B}$. Repeat this step until $|\mathcal{B}| = \min\{S, |\mathcal{V}_I|\}$.
3: Without loss of generality, assume $|\mathcal{B}| = S$. Partition the set of infected nodes into S sets: $\mathcal{V}_I^{(s)}$ for $s = 1, \cdots, S$. An infected node a is assigned to set $\mathcal{V}_I^{(s)}$ if
$$d(a, e_s) = \min_{j=1,\cdots,S} d(a, e_j).$$
Ties are broken arbitrarily.
4: For each $\mathcal{V}_I^{(s)}$, compute the infection radius
$$r_s = \left\lfloor \max_{a,b \in \mathcal{V}_I^{(s)}} \frac{d(a, b)}{2} \right\rfloor.$$
Furthermore, compute the maximum infection radius
$$r_{\max} = \max_{s=1,\ldots,S} r_s.$$
5: Consider the tree $\mathcal{T}$ formed by the set of nodes in $\mathcal{B}$. For each $e_i \in \mathcal{B}$, find a node γ_i on tree $\mathcal{T}$ such that $d(e_i, \gamma_i) = r_{\max}$, and add node γ_i into $\tilde{\mathcal{S}}$.
6: $\tilde{\mathcal{S}}$ is the set of source estimators.

In the first step of Algorithm 1, we select a pair of infected nodes with the maximum distance because most likely these two nodes are associated with two different information sources, where we say an infected node a is "associated" with source s if node a is on the information spreading tree starting from node s. The second step of the algorithm is to select S infected nodes in a greedy fashion to maximize the pairwise distances of these S nodes. These S infected nodes are likely to be associated with different sources, and are likely to be the leaf nodes of the corresponding information spreading trees. The third step divides the set of infected nodes into S sets according to their distances to the selected S nodes. The purpose is to cluster the infected nodes according to their associated sources. The fourth step estimates the maximum distance $r_{\max}$ from a source to any observed infected node associated with the source, which can be used to approximate the depth of the information spreading trees. In the final step, a tree $\mathcal{T}$ with nodes in $\mathcal{B}$ as the leaf nodes is constructed; and for each $e_i \in \mathcal{B}$, we select a node $\gamma_i \in \mathcal{T}$ that is $r_{\max}$ hops away from e_i as the corresponding source estimator. Note that γ_i is close to the real source associated with infected node e_i when e_i is close to a leaf node on the corresponding information spreading tree and $r_{\max}$ is close to the depth of the tree.

The following theorem shows that the distance between a detected source and the closest real source can be bounded by a constant with a high probability under Algorithm 1, where the constant is independent of the size of the infected subnetwork. The detailed proof can be found in [18].

Theorem 1. *Consider a $(g+1)$-regular tree with infinite number of levels where $g > 2$. Assume that $gq > 1$ and the distance between any two sources is larger than C for some large enough constant C. Then given any $\epsilon > 0$, there exists a constant d_ϵ such that the distance between each estimator and its closest real source is upper bounded by d_ϵ with a probability at least $1 - \epsilon$, where d_ϵ is independent of the size of the infected subnetwork.* □

B. Heuristic for General Network Topologies

Locating multiple information sources in general networks is a much more complicated problem. Algorithm 1 is not directly applicable to a general network because after set $\mathcal{B}$, multiple trees can be constructed with nodes in $\mathcal{B}$ as leaf nodes. Therefore, we propose the following heuristic algorithm, which use the Jodan infection center defined by $\mathcal{V}_I^{(s)}$ as the estimator associated with e_s.

Algorithm 2 Clustering and Reverse Infection (CRI)

1: Step 1 to 3 of Algorithm 1.
2: For $\mathcal{V}_I^{(s)}$, use the reverse infection algorithm in [8] to find a Jordan infection center for $\mathcal{V}_I^{(s)}$, named γ_s, and then add γ_s to $\tilde{\mathcal{S}}$. A Jordan infection center γ_s for $\mathcal{V}_I^{(s)}$ is defined to be
$$\gamma_s \in \arg\min_{b \in \mathcal{V}} \left(\max_{a \in \mathcal{V}_I^{(s)}} d(b, a) \right).$$

The reverse infection algorithm was proposed in [8] to localize the source in the single-source SIR model. If the infected nodes in $\mathcal{V}_I^{(s)}$ are associated with the same source, we can expect the reserve infection algorithm restricted to $\mathcal{V}_I^{(s)}$ to output a good estimator.

C. Heuristic for Estimating S

Both Algorithms 1 and 2 require the knowledge of the number of real sources, which may be difficult to know in practice. We further propose the following heuristic for estimating the number of real sources when the number of real sources is unknown.

Algorithm 3 An Algorithm for Approximating S

1: Choose a large number $\bar{S}$. For each k such that $1 \le k \le \bar{S}$, use the steps 1-4 in Algorithm 1 to compute $w_k = r_{\max}$ by assuming the number of real sources is k.
2: Set

$$\tilde{S} = \arg \max_{k:1 \le k \le \bar{S}-2} w_k - w_{k+1} - (w_{k+2} - w_{k+1}),$$

and claim $\tilde{S}$ to be the number of real sources.

Assume the information spreading trees never die out. Consider the case where k is smaller than the number of real sources. Then after the clustering step of Algorithm 1, there exists a set $\mathcal{V}_I^{(s)}$ which contains infected nodes from at least two different real sources. In such a set, the maximum distance between two nodes will be significantly larger than the distance of two infected nodes that are associated with the same source, assuming the real sources are not close to each other. Then under Algorithm 2, we will observe a significant decrease in w_k when change the value of k from $S-1$ to S. Based on this observation, we use k that maximizes

$$w_k - w_{k+1} - (w_{k+1} - w_{k+2})$$

as the estimator of S.

IV. Simulations

In this section, we evaluate the performance of our algorithms using simulations.

A. Tree Networks

In this set of simulations, we assumed the number of real sources is known. We tested our Algorithm 1 on g-regular trees and binomial random trees. For each simulation, we randomly chose 4 nodes in the network as the sources, the probability of infection, q, was chosen uniformly at random from $(0, 0.3)$, and the probability of recovery, p, was chosen uniformly at random $(0, 0.2)$.

In Figure 1 and 2 , we plotted the the average distance between real sources and the estimators versus the degree of the trees. We can see that as the degree of the tree increases, the distance decreases. For regular trees, the average distance is smaller than 3 when the degree is 5 or larger; and for binomial random trees, the average distance is smaller than 4 when the average degree is 5 or larger.

In Figure 3, we plotted the detection rate of Algorithm 1, which is the fraction of estimators that are real sources. As the degree increases, we can see that the detection rates of both regular trees and binomial trees improves.

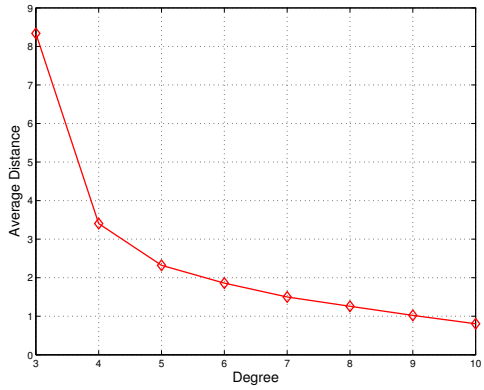

Figure 1. The Average Distance from Sources to the Estimators on Regular Trees

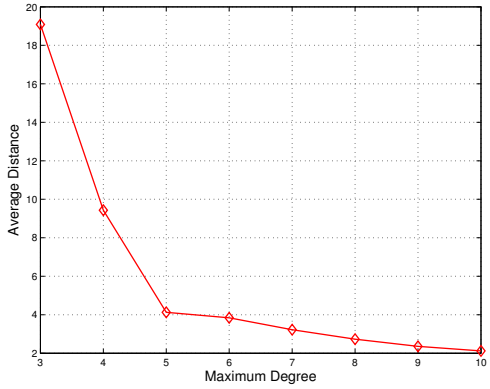

Figure 2. The Average Distance from Sources to the Estimators on Binomial Random Trees

B. The Power Grid Network

In this set of simulations, we tested the performance of our algorithms on the power grid network [19], which has $4,941$ nodes and $6,594$ edges. For each simulation, we randomly selected four nodes as the sources. q was uniformly chosen from $(0, 0.4)$, and p was chosen uniformly from $(0, 0.2)$.

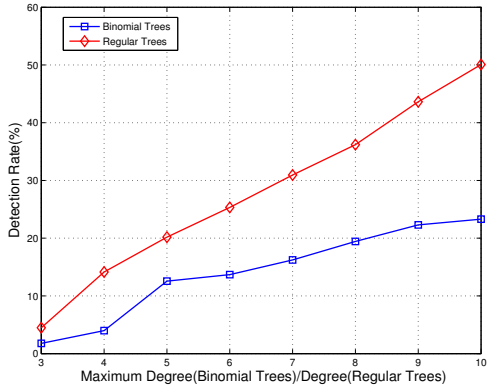

Figure 3. The Detection Rates on Regular Trees and Binomial Trees

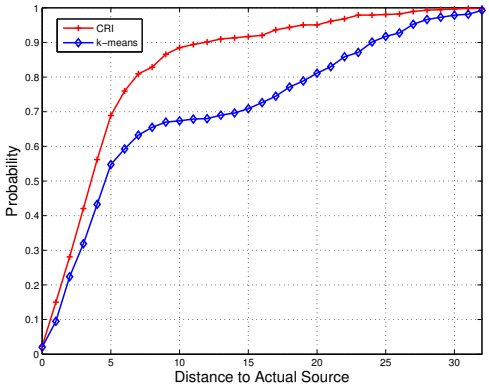

Figure 4. The CDF of CRI and k-means on the Power Grid Network (Assign Estimators to Sources)

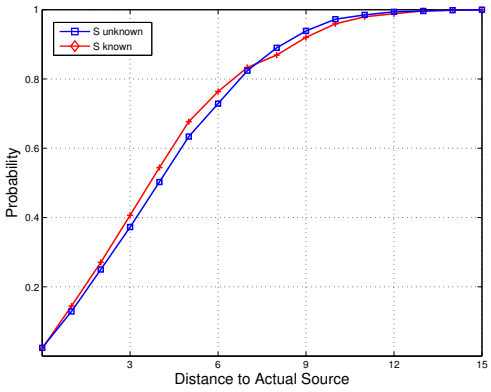

Figure 5. The CDF of CRI when the nubmer of sources, S, is known and unknown

We compared our algorithms with a heuristic algorithm based on k-means. In the k-means heuristic, the initial centroids were the nodes in set $\mathcal{B}$ obtained by using steps 1-2 of Algorithm 1. During the clustering step of each iteration in the k-means heuristic, we used distance centrality to select the centroid of each cluster. In this simulation, we assumed that the number of sources is unknown so we used Algorithm 2 to estimate the number of sources. The results are shown in Figure 4. We can see that our algorithm performs better than the k-means heuristic.

We further compared the performance of our algorithm with or without the knowledge of the number of real sources. The results are shown in Figure 5. We can see that the joint detection and estimation algorithm has a similar performance as the detection algorithm that knows the number of real sources.

V. Conclusion

In this paper, we studied the problem of detecting multiple information sources under the SIR model. We developed an algorithm for tree network when the number of sources is known, and proved that under fairly general conditions, each estimator is within a constant distance to the closest real source with a high probability on g-regular tree networks. Then we proposed a heuristic algorithm for general networks and a heuristic algorithm to estimate the number of real sources when it is unknown.

Acknowledgment: Research supported in part by ARO grant W911NF-13-1-0279.

References

[1] D. Shah and T. Zaman, "Detecting sources of computer viruses in networks: Theory and experiment," in *Proc. Ann. ACM SIGMETRICS Conf.*, New York, NY, 2010, pp. 203–214.

[2] ——, "Rumors in a network: Who's the culprit?" *IEEE Trans. Inf. Theory*, vol. 57, pp. 5163–5181, Aug. 2011.

[3] ——, "Rumor centrality: a universal source detector," in *Proc. Ann. ACM SIGMETRICS Conf.*, London, England, UK, 2012, pp. 199–210.

[4] W. Luo and W. P. Tay, "Identifying multiple infection sources in a network," in *Proc. Asilomar Conf. Signals, Systems and Computers*, 2012.

[5] W. Luo, W. P. Tay, and M. Leng, "Identifying infection sources and regions in large networks," *IEEE Trans. Signal Process.*, vol. 61, pp. 2850–2865.

[6] N. Karamchandani and M. Franceschetti, "Rumor source detection under probabilistic sampling," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Istanbul, Turkey, July 2013.

[7] W. Dong, W. Zhang, and C. W. Tan, "Rooting out the rumor culprit from suspects," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Istanbul, Turkey, 2013, pp. 2671–2675.

[8] K. Zhu and L. Ying, "Information source detection in the SIR model: A sample path based approach," in *Proc. Information Theory and Applications Workshop (ITA)*, Feb. 2013.

[9] ——, "A robust information source estimator with sparse observations," in *Proc. IEEE Int. Conf. Computer Communications (INFOCOM)*, Toronto, Canada, April-May 2014.

[10] W. Luo and W. P. Tay, "Finding an infection source under the SIS model," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, BC, May 2013.

[11] ——, "Estimating infection sources in a network with incomplete observations," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Austin, TX, 2013, pp. 301–304.

[12] B. A. Prakash, J. Vreeken, and C. Faloutsos, "Spotting culprits in epidemics: How many and which ones?" in *IEEE Int. Conf. Data Mining (ICDM)*, Brussels, Belgium, 2012, pp. 11–20.

[13] A. Y. Lokhov, M. Mezard, H. Ohta, and L. Zdeborova, "Inferring the origin of an epidemy with dynamic message-passing algorithm," *arXiv preprint arXiv:1303.5315*, 2013.

[14] A. Agaskar and Y. M. Lu, "A fast monte carlo algorithm for source localization on graphs," in *SPIE Optical Engineering and Applications*, 2013.

[15] E. Seo, P. Mohapatra, and T. Abdelzaher, "Identifying rumors and their sources in social networks," in *SPIE Defense, Security, and Sensing*, 2012.

[16] N. T. J. Bailey, *The mathematical theory of infectious diseases and its applications*. Hafner Press, 1975.

[17] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[18] Z. Chen, K. Zhu, and L. Ying, "Detecting multiple information sources in networks under the sir model," *Technical Report, Arizona State University*, 2014.

[19] "Western states power grid of the united states [online]," 1998, available: http://www-personal.umich.edu/~mejn/netdata/.