Information Source Detection in the SIR Model: A Sample Path Based Approach

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

***Abstract*—This paper studies the problem of detecting the information source in a network in which the spread of information follows the popular Susceptible-Infected-Recovered (SIR) model. We assume all nodes in the network are in the susceptible state initially except the information source which is in the infected state. Susceptible nodes may then be infected by infected nodes, and infected nodes may recover and will not be infected again after recovery. Given a snapshot of the network, from which we know all infected nodes but cannot distinguish susceptible nodes and recovered nodes, the problem is to find the information source based on the snapshot and the network topology. We develop a sample path based approach where the estimator of the information source is chosen to be the root node associated with the sample path that most likely leads to the observed snapshot. We prove for infinite-trees, the estimator is a node that minimizes the maximum distance to the infected nodes. A reverse-infection algorithm is proposed to find such an estimator in general graphs. We prove that for g-regular trees such that $gq > 1$, where g is the node degree and q is the infection probability, the estimator is within a constant distance from the actual source with high probability, independent of the number of infected nodes and the time the snapshot is taken. Our simulation results show that for tree networks, the estimator produced by the reverse-infection algorithm is closer to the actual source than the one identified by the closeness centrality heuristic.**

I. Introduction

Diffusion processes in networks refer to the spread of information throughout the networks, and have been widely used to model many real-world phenomena such as the outbreak of epidemics, the spreading of gossips over online social networks, the spreading of computer virus over the Internet, and the adoption of innovations. Important properties of diffusion processes such as the outbreak thresholds [1] and the impact of network topologies [2] have been intensively studied.

In this paper, we are interested in the reverse of the diffusion problem: given a snapshot of the diffusion process at time t, can we tell which node is the source of the diffusion? The answer to this problem has many important applications, and can help us answer the following questions: who is the rumor source in online social networks? which computer is the first one infected by a computer virus? who is the one who uploaded contraband materials to the Internet? and where is the source of an epidemic?

We call this problem information source detection problem. This information source detection problem has been studied in [3]–[5] under the Susceptible-Infected (SI) model, in which susceptible nodes may be infected but infected nodes cannot recover. The authors formulated the problem as a maximum likelihood estimation (MLE) problem, and developed novel algorithms to detect the source.

In this paper, we adopt the Susceptible-Infected-Recovered (SIR) model, a standard model of epidemics [6], [7]. The network is assumed to be an undirected graph and each node in the network has three possible states: susceptible (S), infected (I), and recovered (R). Nodes in state S can be infected and change to state I, and nodes in state I can recover and change to state R. Recovered nodes cannot be infected again. We assume that initially all nodes are in the susceptible state except one infected node (called the information source). The information source then infects its neighbors, and the information starts to spread in the network. Now given a snapshot of the network, in which we can identify infected nodes and healthy (susceptible and recovered) nodes (we assume susceptible nodes and recovered nodes are indistinguishable), the question is which node is the information source.

We remark that it is very important to take recovery into consideration since recovery can happen due to various reasons in practice. For example, a contraband material uploader may delete the file, a computer may recover from a virus attack after anti-virus software removes the virus, and a user may delete the rumor from her/his blog. In order to solve the information source detection problem in these scenarios, we study the SIR model in this paper, which makes the problem significantly more challenging than that in the SI model as we will explain in the related work section.

A. Main Results

The main results of this paper are summarized below.

- Similar to the SI model, the information source detection problem can be formalized as an MLE problem. Unfortunately, to solve the MLE problem, we need to consider all possible infection sample paths, and for each sample path, we need to specify the infection time and recovery time for each healthy node and the infection time for each infected node, so the number of possible sample paths is at the order of $\Omega(t^N)$, where N is the network size and t is the time the snapshot is obtained. Therefore, the MLE problem is difficult to solve even when t is known.

The problem becomes much harder when t is unknown, which is the assumption of this paper. To overcome this difficulty, we propose a sample path based approach. We propose to find the sample path which most likely leads to the observed snapshot and view the source associated with that sample path as the information source. We call this problem optimal sample path detection problem. We investigate the structure properties of the optimal sample path in trees. Defining the infection eccentricity of a node to be the maximum distance from the node to infected nodes, we prove that the source node of the optimal sample path is the node with the minimum infection eccentricity. Since a node with the minimum eccentricity in a graph is called the Jordan center, we call the nodes with the minimum infection eccentricity the Jordan infection centers. Therefore, the sample path based estimator is one of the Jordan infection centers.

- We propose a low complexity algorithm, called reverse infection algorithm, to find the sample path based estimator in general graphs. In the algorithm, each infected node broadcasts its identity in the network, the node who first collect all identities of infected nodes declares itself as the information source, breaking ties based on the sum of distances to infected nodes. The running time of this algorithm is equal to the minimum infection eccentricity, and the number of messages each node receives/sends is bounded by the degree of the node.
- We analyze the performance of the reverse infection algorithm on g-regular trees, and show that the algorithm can output a node within a constant distance from the actual source with high probability, independent of the number of infected nodes and the time the snapshot is taken.
- We conduct simulations over tree networks to verify the performance of the reverse infection algorithm. The detection rate over regular trees is found to be around 60%, and is higher than that of the infection closeness centrality (or called distance centrality) heuristic. The infection closeness of a node is defined to be the inverse of the sum of distances to infected nodes and the infection closeness centrality heuristic is to claim the node with the maximum infection closeness as the source. Note that in [3]–[5], the authors proved the node with the maximum infection closeness is the MLE on regular trees.

B. Related Work

There have been extensive studies on the spread of epidemics in networks based on the SIR model (see [1], [2], [8], [9] and references within). The work most related to this paper is [3]–[5], in which the information source detection problem was studied under the SI model. This paper considers the SIR model, where infection nodes may recover, which can occur in many practical scenarios as we have explained. Because of node recovery, the information source detection problem under the SIR model differs significantly from that under the SI model. The differences are summarized below.

- The set of possible sources in the SI model [3]–[5] is restricted to the set of infected nodes. In the SIR model, all nodes are possible information sources because we assume susceptible nodes and recovered nodes are indistinguishable and a healthy node may be a recovered node so can be the information source. Therefore, the number of candidate sources is much larger in the SIR model than that in the SI model.
- A key observation in [3]–[5] is that on regular trees, all permitted permutations of infection sequences (a infection sequence specifies the order at which nodes are infected) are equally likely under the SI model. The number of possible permutations from a fixed root node, therefore, decides the likelihood of the root node being the source. However, under the SIR model, different infection sequences are associated with different probabilities, so counting the number of permutations are not sufficient.
- [3]–[5] proved that the node with the minimum closeness centrality is the an MLE on regular-trees. We define the infection closeness centrality to be the inverse of the sum of distances to infected nodes. Our simulations show that the sample path based estimator is closer to the actual source than the nodes with the maximum infection closeness.

Other related works include: (1) detecting the first adopter of innovations based on a game theoretical model [10] in which the authors derived the MLE but the computational complexity is exponential in the number of nodes, (2) network forensics under the SI model [11], where the goal is to distinguish an epidemic infection from a random infection, and (3) geospatial abduction problems [12], [13].

II. Problem Formulation

A. The SIR Model for Information Propagation

Consider an undirected graph $G = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of (undirected) edges. Each node $v \in \mathcal{V}$ has three possible states: susceptible (S), infected (I), and recovered (R). We assume a time slotted system. Nodes change their states at the beginning of each time slot, and the state of node v in time slot t is denoted by $X_v(t)$. Initially, all nodes are in state S except node v^* which is in state I and is the information source. At the beginning of each time slot, each infected node infects each of its susceptible neighbors with probability q, independent of other nodes, i.e., a susceptible node is infected with probability $1-(1-q)^n$ if it has n infected neighbors. Each infected node recovers with probability p, i.e., its state changes from I to R with probability p. In addition, we assume a recovered node cannot be infected again. Since whether a node gets infected only depends on the states of its neighbors and whether a node becomes a recovered node only depends on its own state in the previous time slot, the infection process can be modeled as a discrete time Markov chain $\mathbf{X}(t)$ where $\mathbf{X}(t) = \{X_v(t), v \in \mathcal{V}\}$ is the states of all the nodes at time

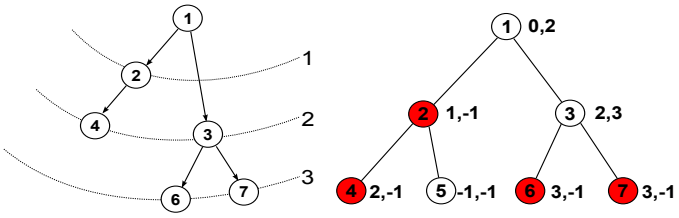

Figure 1. An Example of Information Propagation

slot t. The initial state of this Markov chain is $X_v(0) = S$ for $v \neq v^*$ and $X_{v^*}(0) = I$.

B. Information Source Detection

We assume $\mathbf{X}(t)$ is not fully observable since we cannot distinguish susceptible nodes and recovered ones. So at time t, we observe $\mathbf{Y} = \{Y_v, v \in \mathcal{V}\}$ such that

$$Y_v = \begin{cases} 1, & \text{if } v \text{ is in state } I; \\ 0, & \text{if } v \text{ is in state } S \text{ or } R. \end{cases}$$

The information source detection problem is to identify v^* given the graph G and $\mathbf{Y}$, where t is an unknown parameter.

Figure 1 is an example of the infection process. The left figure shows the information propagation over time. The nodes on each dotted line are the nodes which are infected at that time slot, and the arrows indicate where the infection comes from (e.g., node 4 is infected by node 2).

The figure on the right is the network we observe, where the shaded nodes are infected nodes and others are susceptible or recovered nodes. The pair of numbers next to each node are the corresponding infection time and recovery time. For example, node 3 was infected at time slot 2 and recovered at time slot 3. -1 indicates that the infection or recovery has yet occurred. Note that these two pieces of information are not available to us, and we include them in the figure to illustrate the infection and recovery processes. If we observe the network at the end of time slot 3, then the snapshot of the network is $\mathbf{Y} = \{0, 1, 0, 1, 0, 1, 1\}$, where the states are ordered according to the indices of the nodes.

C. Maximum Likelihood Detection

We define $\mathbf{X}[0, t] = \{\mathbf{X}(\tau) : 0 < \tau \leq t\}$ to be a sample path of the infection process from 0 to t. In addition, we define function $F(\cdot)$ such that

$$F(X_v(t)) = \begin{cases} 1, & \text{if } X_v(t) = I; \\ 0, & \text{otherwise.} \end{cases}$$

We say $\mathbf{F}(\mathbf{X}[t]) = \mathbf{Y}$ if $F(X_v(t)) = Y_v$ for all v. Identifying the information source can be formulated as a maximum likelihood detection problem as follows:

$$v^\dagger \in \arg\max_{v \in \mathcal{V}} \sum_{\mathbf{X}[0,t]:\mathbf{F}(\mathbf{X}(t))=\mathbf{Y}} \Pr(\mathbf{X}[0, t]|v^* = v),$$

where $\Pr(\mathbf{X}[0, t]|v^* = v)$ is the probability to obtain sample path $\mathbf{X}[0, t]$ given the information source is node v.

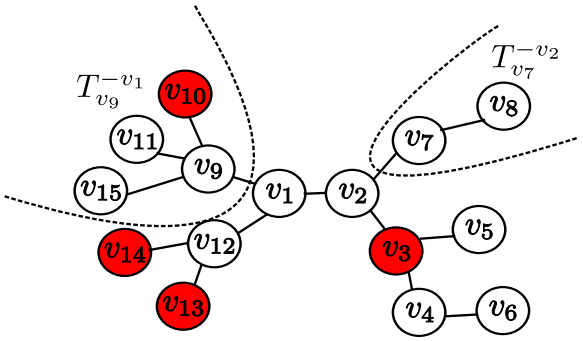

Figure 2. An Example Illustrating the Infection Eccentricity

We note the difficulty of solving this maximum likelihood problem is the curse of dimensionality. For each v such that $Y_v = 0$, we need to decide its infection time and recovered time (the node is in susceptible state if the infection time is $> t$), i.e., $O(t^2)$ possible choices; for each v such that $Y_v = 1$, we need to decide the infection time, i.e., $O(t)$ possible choices. Therefore, even for a fixed t, the number of possible sample paths is at least at the order of t^N, where N is the number of nodes in the network. This curse of dimensionality makes it computational expensive, if not impossible, to solve the maximum likelihood problem. To overcome this difficulty, we propose a sample path based approach which is discussed below.

D. Sample Path Based Detection

Instead of computing the marginal probability, we propose to identify the sample path $\mathbf{X}^*[0, t^*]$ that most likely leads to $\mathbf{Y}$, i.e.,

$$\mathbf{X}^*[0, t^*] = \arg \max_{t, \mathbf{X}[0,t] \in \mathcal{X}(t)} \Pr(\mathbf{X}[0, t]), \tag{1}$$

where $\mathcal{X}(t) = \{\mathbf{X}[0, t]|\mathbf{F}(\mathbf{X}(t)) = \mathbf{Y}\}$. The source node associated with $\mathbf{X}^*[0, t^*]$ is viewed as the information source.

III. Sample Path Based Detection On Tree Networks

The optimal sample paths for general graphs are still difficult to obtain. In this section, we focus on tree networks and derive structure properties of the optimal sample paths.

First, we introduce the definition of eccentricity in graph theory [14]. The eccentricity $e(v)$ of a vertex v is the maximum distance between v and any other vertex in the graph. The Jordan centers of a graph are the nodes which have the minimum eccentricity. For example, in Figure 2, the eccentricity of node v_1 is 4 and the Jordan center is v_2, whose eccentricity is 3.

Following a similar terminology, we define the infection eccentricity $\tilde{e}(v)$ given $\mathbf{Y}$ as the maximum distance between v and any infected nodes in the graph. Define the Jordan infection centers of a graph to be the nodes with the minimum infection eccentricity given $\mathbf{Y}$. In Figure 2, nodes v_3, v_{10}, v_{13} and v_{14} are observed to be infected. The infection eccentricities of v_1, v_2, v_3, v_4 are $2, 3, 4, 5$, respectively, and the Jordan infection center is v_1.

We will show that the source associated with the optimal sample path is a node with the minimum infection eccentricity. We derive this result using three steps: first, assuming the information source is v_r, we analyze $t^*_{v_r}$ such that

$$t^*_{v_r} = \arg_t \max_{t,\mathbf{X}[0,t]} \Pr(\mathbf{X}[0,t]|v^* = v_r),$$

i.e., $t^*_{v_r}$ is the time duration of the optimal sample path in which v_r is the information source. It turns out that $t^*_{v_r}$ equals to the infection eccentricity of node v_r. Considering Figure 2 if the source is v_1, then the time duration of the optimal sample path is 2.

In the second step, we consider two neighboring nodes, say nodes v_1 and v_2. We will prove that if $\tilde{e}(v_1) < \tilde{e}(v_2)$, then the optimal sample path rooted at v_1 occurs with a higher probability than the optimal sample path rooted at v_2.

Finally, at the third step, we will show that given any two nodes u and v, if v has the minimum infection eccentricity and u has a larger infection eccentricity, then there exists a path from u to v along which the infection eccentricity monotonically decreases, which implies that the source of the optimal sample path must be a Jordan infection center. For example, in Figure 2, node v_4 has a larger infection eccentricity than v_1 and $v_4 \to v_3 \to v_2 \to v_1$ is the path along which the infection eccentricity monotonically decreases from 5 to 2.

A. The Optimal Time

Lemma 1. *Consider a tree network rooted at v_r and with infinitely many levels. Assume the information source is the root, and the observed infection topology is $\mathbf{Y}$ which contains at least one infected node. If $\tilde{e}(v_r) \leq t_1 < t_2$, then the following inequality holds*

$$\max_{\mathbf{X}[0,t_1]\in\mathcal{X}(t_1)} \Pr(\mathbf{X}[0,t_1]) > \max_{\mathbf{X}[0,t_2]\in\mathcal{X}(t_2)} \Pr(\mathbf{X}[0,t_2]),$$

where $\mathcal{X}(t) = \{\mathbf{X}[0,t]|\mathbf{F}(\mathbf{X}(t)) = \mathbf{Y}\}$. In addition,

$$t^*_{v_r} = \tilde{e}(v_r) = \max_{u\in\mathcal{I}} d(v_r, u),$$

where $d(v_r, u)$ is the length of the shortest path between v_r and u and also called the distance between v_r and u, and $\mathcal{I}$ is the set of infected nodes. □

The proof of Lemma 1 can be found in [15].

B. The Sample Path Based Estimator

After deriving t^*_v, we have a unique t^*_v for each $v \in \mathcal{V}$. The next lemma states that the optimal sample path rooted a node with a smaller infection eccentricity is more likely to occur.

Lemma 2. *Consider a tree network with infinitely many levels. Assume the information source is the root, and the observed infection topology is $\mathbf{Y}$ which contains at least one infected node. For $u, v \in \mathcal{V}$ such that $(u, v) \in \mathcal{E}$, if $t^*_u > t^*_v$, then*

$$\Pr(\mathbf{X}^*_u([0, t^*_u])) < \Pr(\mathbf{X}^*_v([0, t^*_v])),$$

where $\mathbf{X}^*_u[0, t^*_u]$ is the optimal sample path starting from node u.

Proof. Denote by T_v the tree rooted in v and T_u^{-v} the tree rooted at u but without the branch from v. See $T_{v_9}^{-v_1}$ and $T_{v_7}^{-v_2}$ in Figure 2. Furthermore, denote by $\mathcal{C}(v)$ the set of children of v. The sample path $\mathbf{X}[0,t]$ restricted to T_u^{-v} is defined to be $\mathbf{X}([0,t], T_u^{-v})$.

Step 1: The first step is to show $t^*_u = t^*_v + 1$. First we claim $T_v^{-u} \cap \mathcal{I} \neq \emptyset$. Otherwise, all infected node are on T_u^{-v}. Since on a tree, v can only reach nodes in T_u^{-v} through edge (u, v), $t^*_v = t^*_u + 1$, which contradicts $t^*_u > t^*_v$.

If $T_u^{-v} \cap \mathcal{I} \neq \emptyset$, $\forall a \in T_u^{-v} \cap \mathcal{I}$, we have

$$d(u, a) = d(v, a) - 1 \leq t^*_v - 1,$$

and $\forall b \in T_v^{-u} \cap \mathcal{I}$,

$$d(u, b) = d(v, b) + 1 \leq t^*_v + 1.$$

Hence,

$$t^*_u \leq t^*_v + 1,$$

which implies that

$$t^*_v < t^*_u \leq t^*_v + 1,$$

i.e., $t^*_u = t^*_v + 1$.

If $T_u^{-v} \cap \mathcal{I} = \emptyset$, all infected nodes are in T_v^{-u}, so it is obvious $t^*_u = t^*_v + 1$.

Step 2: In this step, we will prove that $t^I_v = 1$ on the sample path $\mathbf{X}^*_u[0, t^*_u]$. If $t^I_v > 1$ on $\mathbf{X}^*_u([0, t^*_u])$, then

$$t^*_u - t^I_v = t^*_v + 1 - t^I_v < t^*_v.$$

Note that according to the definition of t^*_u and t^I_v, within $t^*_u - t^I_v$ time slots, node v can infect all infected nodes on T_v^{-u}. Since $t^*_u = t^*_v + 1$, the infected node farthest from node u must be on T_v^{-u}, which implies that there exists a node $a \in T_v^{-u}$ such that $d(u, a) = t^*_u = t^*_v + 1$ and $d(v, a) = t^*_v$. So node v cannot reach a within $t^*_u - t^I_v$ time slots, which contradicts the fact that the infection can spread from node v to a within $t^*_u - t^I_v$ time slots along the sample path $X^*_u[0, t^*_u]$. Therefore, $t^I_v = 1$.

Step 3: Now given sample path $\mathbf{X}^*_u[0, t^*_u]$, we construct $\mathbf{X}_v[0, t^*_v]$ which occurs with a higher probability. We divide the sample path $\mathbf{X}^*_u[0, t^*_u]$ into two parts along subtrees T_u^{-v} and T_v^{-u}. Since $t^I_v = 1$, we have

$$\begin{aligned}&\Pr(\mathbf{X}^*_u[0, t^*_u])\\ &= q \Pr\left(\mathbf{X}^*_u\left([0, t^*_u], T_v^{-u}\right) \middle| t^I_v = 1\right) \Pr\left(\mathbf{X}^*_u\left([0, t^*_u], T_u^{-v}\right)\right),\end{aligned}$$

where q is the probability that v is infected at the first time slot. Suppose in $\mathbf{X}_v[0, t^*_v]$, node u was infected at the first time slot, then

$$\begin{aligned}&\Pr(\mathbf{X}_v[0, t^*_v]) =\\ &q \Pr\left(\mathbf{X}_v\left([0, t^*_v], T_v^{-u}\right)\right) \Pr\left(\mathbf{X}_v\left([0, t^*_v], T_u^{-v}\right) \middle| t^I_u = 1\right).\end{aligned}$$

For the subtree T_v^{-u}, given $\mathbf{X}^*_u\left([0, t^*_u], T_v^{-u}\right)$, in which $t^I_v = 1$, we construct the partial sample path $\mathbf{X}_v\left([0, t^*_v], T_v^{-u}\right)$ to be identical to $\mathbf{X}^*_u\left([0, t^*_u], T_v^{-u}\right)$ except that all events occur one time slot earlier, i.e.,

$$\mathbf{X}_v\left([0, t^*_v], T_v^{-u}\right) = \mathbf{X}^*_u\left([1, t^*_u], T_v^{-u}\right).$$

This is feasible because $t_v^* = t_u^* - 1$. Then

$$\Pr\left(\mathbf{X}_u^*\left([0,t_u^*],T_v^{-u}\right)\Big|t_v^I=1\right)=\Pr\left(\mathbf{X}_v\left([0,t_v^*],T_v^{-u}\right)\right).$$

For the subtree T_u^{-v}, we construct $\mathbf{X}_v([0,t_v^*],T_u^{-v})$ such that

$$\mathbf{X}_v([0,t_v^*],T_u^{-v}) \in$$
$$\arg\max_{\tilde{\mathbf{X}}([0,t_v^*],T_u^{-v})\in\mathcal{X}(t_v^*,T_u^{-v})}\Pr\left(\tilde{\mathbf{X}}\left([0,t_v^*],T_u^{-v}\right)\Big|t_u^I=1\right).$$

Based on Lemma 1, we have

$$\max_{\tilde{\mathbf{X}}([0,t_v^*],T_u^{-v})\in\mathcal{X}(t_v^*,T_u^{-v})}\Pr\left(\tilde{\mathbf{X}}\left([0,t_v^*],T_u^{-v}\right)\Big|t_u^I=1\right)=$$
$$\max_{\tilde{\mathbf{X}}([0,t_u^*-1],T_u^{-v})\in\mathcal{X}(t_u^*-1,T_u^{-v})}\Pr\left(\tilde{\mathbf{X}}\left([0,t_u^*-1],T_u^{-v}\right)\Big|t_u^I=1\right)$$
$$>\max_{\mathbf{X}([0,t_u^*],T_u^{-v})\in\mathcal{X}(t_u^*,T_u^{-v})}\Pr\left(\mathbf{X}\left([0,t_u^*],T_u^{-v}\right)\right).$$

Therefore, given the optimal sample path rooted at u, we have constructed a sample path rooted at v which occurs with a higher probability. The lemma holds. □

Next, we give a useful property of the Jordan infection centers in the following lemma.

Lemma 3. *On a tree network with at least one infected node, there exist at most two Jordan infection centers. When the network has two Jordan infection centers, the two must be neighbors.* □

The proof of Lemma 3 can be found in [15].

Based on Lemma 2 and Lemma 3, we finish this section with the following theorem.

Theorem 4. *Consider a tree network with infinitely many levels. Assume that the observed infection topology* $\mathbf{Y}$ *contains at least one infected node. Then the source node associated with* $\mathbf{X}^*[0,t^*]$ *(the solution to the optimization problem (1)) is a Jordan infection center, i.e.,*

$$v^\dagger = \arg\min_{v\in\mathcal{V}}\tilde{e}(v).$$

Proof. We assume the network has two Jordan infection centers: w and u, and assume $\tilde{e}(w)=\tilde{e}(u)=\lambda$. The same argument works for the case where the network has only one Jordan infection center.

Based on Lemma 3, w and u must be adjacent. We will show for any $a\in\mathcal{V}\backslash\{w,u\}$, there exists a path from a to u (or w) along which the infection eccentricity strictly decreases.

Step 1: First, it is easy to see from Figure 3 that $d(\gamma,w)\le\lambda-1\ \forall\gamma\in T_w^{-u}\cap\mathcal{I}$. We next show that there exists a node ξ such that the equality holds.

Suppose that $d(\gamma,w)\le\lambda-2$ for any $\gamma\in T_w^{-u}\cap\mathcal{I}$, which implies

$$d(\gamma,u)\le\lambda-1\quad\forall\gamma\in T_w^{-u}\cap\mathcal{I}.$$

Since w and u are both Jordan infection centers, we have $\forall\gamma\in T_u^{-w}\cap\mathcal{I}$,

$$d(\gamma,w)\le\lambda$$
$$d(\gamma,u)\le\lambda-1.$$

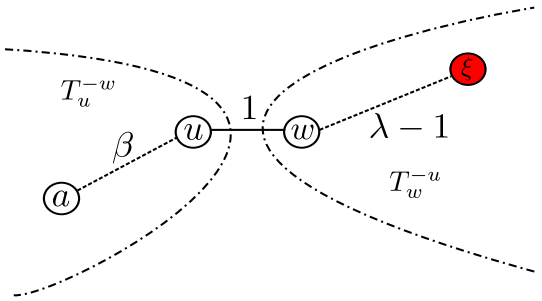

Figure 3. A Pictorial Description of the Positions of Nodes a, u, w and ξ.

In a summary, $\forall\gamma\in\mathcal{I}$,

$$d(\gamma,u)\le\lambda-1.$$

This contradicts the fact that $\tilde{e}(w)=\tilde{e}(u)=\lambda$. Therefore, there exists $\xi\in T_w^{-u}\cap\mathcal{I}$ such that

$$d(\xi,w)=\lambda-1.$$

Step 2: Similarly, $\forall\gamma\in T_u^{-w}\cap\mathcal{I}$,

$$d(\gamma,u)\le\lambda-1,$$

and there exists a node such that the equality holds.

Step 3: Next we consider $a\in\mathcal{V}\backslash\{w,u\}$, and assume $a\in T_u^{-w}$ and $d(a,u)=\beta$. Then for any $\gamma\in T_w^{-u}\cap\mathcal{I}$, we have

$$\begin{aligned}d(a,\gamma)&=d(a,u)+d(u,w)+d(w,\gamma)\\&\le\beta+1+\lambda-1\\&=\lambda+\beta,\end{aligned}$$

and there exists $\xi\in T_w^{-u}\cap\mathcal{I}$ such that the equality holds. On the other hand, $\forall\gamma\in T_u^{-w}\cap\mathcal{I}$.

$$\begin{aligned}d(a,\gamma)&\le d(a,u)+d(u,\gamma)\\&\le\beta+\lambda-1.\end{aligned}$$

Therefore, we conclude that

$$\tilde{e}(a)=\lambda+\beta,$$

so the infection eccentricity decreases along the path from a to u.

Step 4: Repeatedly applying Lemma 2 along the path from node a to u, we can conclude that the optimal sample path rooted at node u is more likely to occur than the optimal sample path rooted at node a. Therefore, the root node associated with the optimal sample path $\mathbf{X}^*[0,t^*]$ must be a Jordan infection center, and the theorem holds. □

IV. Reverse Infection Algorithm

Since in tree networks with infinitely many levels, the estimator based on the sample path approach is a Jordan infection center, we view the Jordan infection centers as possible candidates of the information source. We next present a simple algorithm to find the information source in general networks. The algorithm is to first identify the Jordan infection

centers, and then break ties based on the sum of distances to infected nodes.

The key idea of the algorithm is to let every infected node broadcast a message containing its identity (ID) to its neighbors. Each node, after receiving messages from its neighbors, checks whether the ID in the message has been received. If not, the node records the ID (say v), the time at which the message is received (say t_v), and then broadcasts the ID to its neighbors. When a node receives the IDs of all infected nodes, it claims itself as the information source and the algorithm terminates. If there are multiple nodes receiving all IDs at the same time, the tie is broken by selecting the node with the smallest $\sum t_v$.

The tie-breaking rule we proposed is to choose the node with the maximum infection closeness [16]. The closeness measures the efficiency of a node to spread information to all other nodes. The closeness of a node is the inverse of the sum of distances from the node to any other nodes. In our model, we define the *infection closeness* as the inverse of the sum of distances from a node to all infected nodes, which reflects the efficiency to spread information to infected nodes. We select a Jordan infection center with the largest infection closeness, breaking ties at random.

Algorithm 1 Reverse Infection Algorithm

for $i \in \mathcal{I}$ **do**
  i sends its ID ω_i to its neighbors.
end for
while $t \geq 1$ and STOP$== 0$ **do**
  for $u \in \mathcal{V}$ **do**
    if u receives ω_i for the first time **then**
      Set $t_{ui} = t$ and then broadcast the message ω_i to its neighbors.
      If there exists a node who received $|\mathcal{I}|$ distinct messages, then set $STOP == 1$.
    end if
  end for
end while
return $u^\dagger = \arg\min_{u\in\mathcal{S}} \sum_{i\in\mathcal{I}} t_{ui}$, where $\mathcal{S}$ is the set of nodes who receive $|\mathcal{I}|$ distinct messages when the algorithm terminates. Ties are broken at random.

It is easy to verify that the set $\mathcal{S}$ is the set of the Jordan infection centers. The running time of the algorithm is equal to the minimum infection eccentricity and the number of messages each node receives/sends during each time slot is bounded by its degree.

V. Performance Analysis

The reverse infection algorithm is based on the structure properties of the optimal sample paths on trees. While the MLE is the node that maximizes the likelihood of the snapshot among all possible nodes, the sample path based estimator does not have such a guarantee. To demonstrate the effectiveness of the sample path based approach, we next show that on $(g+1)$-regular trees where each node has $g+1$ neighbors, the information source generated by the reverse infection algorithm is within a constant distance from the actual source with high probability, independent of the number of infected nodes and the time at which the snapshot $\mathbf{Y}$ was taken.

Theorem 5. *Consider a $(g+1)$-regular tree with infinitely many levels where $g > 2$ and $gq > 1$. Assume that the observed infection topology $\mathbf{Y}$ contains at least one infected node. Given $\epsilon > 0$, there exists d_ϵ such that the distance between the optimal sample path estimator and the actual source is d_ϵ with probability $1-\epsilon$, where d_ϵ is independent of the number of infected nodes and the time the snapshot $\mathbf{Y}$ was taken.*

Proof. Consider the tree rooted at the information source v^*. We say v^* is at level 0. We denote by $\mathcal{Z}_l$ the set of infected and recovered nodes at level l. Furthermore, we define $\mathcal{Z}_l^\tau$ to be the set of infected and recovered nodes at level l whose parents are in set $\mathcal{Z}_{l-1}^\tau$ and who were infected within τ time slots after their parents were infected. We assume $\mathcal{Z}_0^\tau = \{v^*\}$. In addition, let $Z_l = |\mathcal{Z}_l|$ and $Z_l^\tau = |\mathcal{Z}_l^\tau|$.

Note

$$\lim_{\tau\to\infty} Z_l^\tau = Z_l,$$

and given v and $u \in \mathcal{Z}_l^\tau$,

$$|t_v^I - t_u^I| \leq l(\tau-1),$$

i.e., the infection times of nodes in $\mathcal{Z}_l^\tau$ differ by at most $l(\tau-1)$ (note that the difference is not $\tau-1$ since the parents of u and v may be infected at different times). Our proof is based on the Galton Watson (GW) branching process [17]. A GW branching process is a stochastic process $B(l)$ which evolves according to the recurrence formula $B(0)=1$ and

$$B(l) = \sum_{i=1}^{B(l-1)} \zeta_i,$$

where $\{\zeta_i\}$ is a set of random variables, taking values from nonnegative integers. The distribution of ζ_i is called the offspring distribution of the branching process. In a $(g+1)$-regular tree, the evolution of $\mathcal{Z}_l^\tau$ is a branching process, where the offspring distribution is a function of τ. We use B^τ to denote the corresponding branching process, and $B^\tau(l)$ to denote the number of offsprings at level l, i.e., $B^\tau(l) = Z_l^\tau$ (we use these two notations interchangeably). Given a node is in the infected state for t time slots, the number of infected offsprings follows a binomial distribution. Note the following two facts:

- The number of time slots at which a node is in the infected state follows a geometric distribution with parameter p.
- A child remains to be susceptible with probability $(1-q)^\tau$ when the parent has been in the infected state for τ time slot.

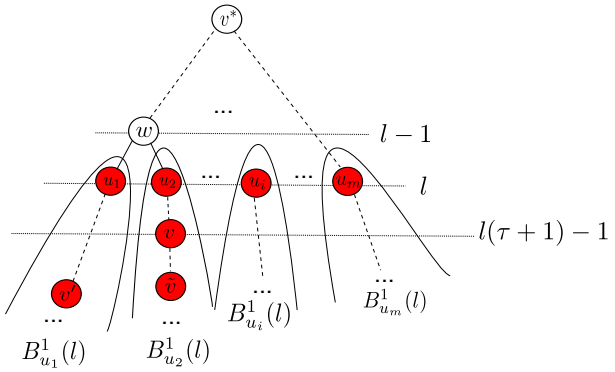

Figure 4. A pictorial description of the positions of v', $\tilde{v}$, u_1, and w.

Therefore, the offspring distribution of the branching process B^τ at level ≥ 1[1] is

$$\begin{aligned}&\Pr(\gamma = i)\\ &= \sum_{t=1}^{\tau-1}(1-p)^{t-1}p\binom{g}{i}\left(1-(1-q)^t\right)^i(1-q)^{t(g-i)}\\ &+\left(1-\sum_{t=1}^{\tau-1}(1-p)^{t-1}p\right)\binom{g}{i}\left(1-(1-q)^\tau\right)^i(1-q)^{\tau(g-i)},\end{aligned}$$

where γ is the number of offsprings of a node. The offspring distribution of branching process B^∞ is

$$\begin{aligned}&\Pr(\gamma' = i)\\ &= \sum_{t=1}^{\infty}(1-p)^{t-1}p\binom{g}{i}\left(1-(1-q)^t\right)^i(1-q)^{t(g-i)}.\end{aligned}$$

Each infected node can be viewed the source of branching processes on the subtree rooted at the node. We define K_l to be the number of survived B^1 branching processes whose roots are in set $\mathcal{Z}_l^\tau$, where a branching process survives if it never dies out.

Now given $L \geq 2$, we consider the following events:

- Event 1: $Z_L = 0$
- Event 2: $K_l \geq 2$ for some $l \leq L$. In other words, at least two B^1 branching processes starting from $\mathcal{Z}_l^\tau$ survive for some $l \leq L$.

We note that these two are disjoint events.

When $Z_l = 0$, no node at level L is infected and the infection process terminates at level $L-1$. When there is at least one infected node in $\mathbf{Y}$, since $\tilde{e}(v^*) \leq L-1$, the minimum infection eccentricity is at most $L-1$. Therefore, the distance between v^* and $v^\dagger$ is no more than $2(L-1)$.

Given $K_l \geq 2$ for some $l \leq L$, we will argue that the distance between the sample path based estimator and the actual one is upper bounded by $(\tau+1)L-1$. Consider Figure 4, where the shaded nodes are infected and recovered nodes. We will show that if two B^1 branching processes starting from $l \leq L$ survive, a node at level $\geq (\tau+1)L-1$ cannot be a Jordan infection center. Recall that at time t, the distance between any infected node and the actual source is no more than t, which implies the eccentricity of a Jordan infection center is $\leq t$. Now consider a node $\tilde{v}$ at level $\geq (\tau+1)l-1$. Recall that at least two B^1 branching processes starting from level l survive. Let $u_1 \in \mathcal{Z}_l^\tau$ be the root of a survived B^1 branching process, and assume node $\tilde{v}$ is not on the subtree rooted at u_1. Further, assume v' is an infected node at the lowest level on sub-tree $T_{u_1}^{-w}$. Since the branching process $B^1_{u_1}$ survives, the infection process propagates one level lower at each time slot and node v' is at level $l + t - t^I_{u_1}$.

From Figure 4, it is easy to see that the distance between v' and $\tilde{v}$ is at least

$$t - t^I_{u_1} + 2 + (\tau+1)l - 1 - l = t - t^I_{u_1} + \tau l + 1,$$

which occurs when the first common predecessor of nodes v' and $\tilde{v}$ is at $l-1$ level. Note that the common predecessor cannot appear at level $\geq l$ since $\tilde{v}$ is not on $T_{u_1}^{-w}$. Since $u_1 \in \mathcal{Z}_l^\tau$, the infection time of node u_1 is no later than τl, i.e., $t^I_{u_1} \leq \tau l$. Therefore, the distance between v' and $\tilde{v}$ is at least $t+1$, which is larger than t. Hence, v' cannot be a Jordan infection center. Since $l \leq L$, any node at or below level $(\tau+1)L-1$ cannot be a Jordan infection center. In a summary, if event 2 occurs, then we have

$$d(v^*, v^\dagger) \leq (\tau+1)L - 1.$$

We next show that given any ϵ, we can find sufficiently large τ and L, independent of t and the number of infected nodes, such that the probability that either event 1 or event 2 occurs is at least $1-\epsilon$.

Given $n_0 > 0$ and $\tau > 0$, we define

$$l^\dagger = \min\left\{l : Z_l^\tau > n_0\right\},$$

i.e., $l^\dagger$ is the first level at which B^τ has more than n_0 nodes. We first have

$$\begin{aligned}&\Pr(Z_L = 0) + \Pr\left(K_l \geq 2 \text{ for some } l \leq L\right)\\ &\geq \Pr(Z_L = 0) + \Pr\left(K_{l^\dagger} \geq 2 \text{ and } l^\dagger \leq L\right)\\ &= \Pr(Z_L = 0) + \Pr\left(l^\dagger \leq L\right)\Pr\left(K_{l^\dagger} \geq 2 \middle| l^\dagger \leq L\right)\\ &= \Pr(Z_L = 0) + \Pr\left(\bigcup_{i=1}^{L}\{Z_i^\tau > n_0\}\right)\Pr\left(K_{l^\dagger} \geq 2 \middle| l^\dagger \leq L\right)\\ &\geq \left(1 - \Pr\left(\bigcap_{i=1}^{L}\{0 < Z_i^\tau \leq n_0\}\right) - \Pr\left(\bigcup_{i=1}^{L}\{Z_i^\tau = 0\}\right)\right)\\ &\quad\times \Pr\left(K_{l^\dagger} \geq 2 \middle| l^\dagger \leq L\right) + \Pr(Z_L = 0).\end{aligned}$$

Note that we have

$$\begin{aligned}&\Pr(K_{l^\dagger} \geq 2 | l^\dagger \leq L)\\ &= \sum_{l=1}^{L}\Pr(K_{l^\dagger} \geq 2 | l^\dagger = l)\Pr(l^\dagger = l | l^\dagger \leq L).\end{aligned} \tag{2}$$

According to Lemma 6, given any $\epsilon_1 > 0$, we can find a sufficiently large n_0 such that

$$\Pr(K_{l^\dagger} \geq 2 | l^\dagger = l) \geq (1-\epsilon_1),$$

[1]The source node has $g+1$ children while other nodes have g children

which implies that for sufficiently large n_0,

$$\Pr(K_{l^\dagger} \geq 2 | l^\dagger \leq L) \geq 1 - \epsilon_1.$$

We can then conclude

$$\begin{aligned}
&\Pr(Z_L = 0) + \Pr\left(K_l \geq 2 \text{ for some } l \leq L\right) \\
\geq &\left(1 - \Pr\left(\bigcap_{i=1}^{L} \{0 < Z_i^\tau \leq n_0\}\right)\right)(1 - \epsilon_1) \\
&- \Pr\left(\bigcup_{i=1}^{L} \{Z_i^\tau = 0\}\right) + \Pr(Z_L = 0) \\
= &\left(1 - \Pr\left(\bigcap_{i=1}^{L} \{0 < Z_i^\tau \leq n_0\}\right)\right)(1 - \epsilon_1) \\
&+ \Pr(Z_L = 0) - \Pr(Z_L^\tau = 0),
\end{aligned}$$

where $\Pr(\cup_{i=1}^{L} \{Z_i^\tau = 0\}) = \Pr(Z_L^\tau = 0)$ because $Z_l^\tau = 0$ implies that $Z_L^\tau = 0$ for $l \leq L$.

According to Lemma 7 and Lemma 8, given any $\epsilon_2 > 0$ and $\epsilon_3 > 0$, there exist sufficiently large τ and L such that

$$\left(1 - \Pr\left(\bigcap_{i=1}^{L} \{0 < Z_i^\tau \leq n_0\}\right)\right) > 1 - \epsilon_2,$$

and

$$\Pr(Z_L = 0) - \Pr(Z_L^\tau = 0) \geq -\epsilon_3.$$

Hence, we have

$$\begin{aligned}
&\Pr(Z_L = 0) + \Pr\left(K_l \geq 2 \text{ for some } l \leq L\right) \\
\geq \quad &(1 - \epsilon_1)(1 - \epsilon_2) - \epsilon_3.
\end{aligned}$$

Now choosing $\epsilon_1 = \epsilon_2 = \epsilon_3 = \epsilon_4/3$ for some $\epsilon_4 > 0$, we have

$$\begin{aligned}
&\Pr(Z_L = 0) + \Pr\left(K_l \geq 2 \text{ for some } l \leq L\right) \\
\geq \quad &1 - \epsilon_4.
\end{aligned}$$

Now let $|\mathbf{Y}|$ denote the number of infected nodes in the observation $\mathbf{Y}$. Define events $E_1 = \{Z_L = 0\}$ and $E_2 = \{K_l \geq 2 \text{ for some } l \leq L\}$. We have

$$\begin{aligned}
&\Pr(E_1 || \mathbf{Y}| = 1) + \Pr\left(E_2 || \mathbf{Y}| = 1\right) \\
= &\frac{1}{\Pr(|\mathbf{Y}| = 1)} \left(\Pr(E_1 \cap \{|\mathbf{Y}| = 1\}) + \Pr\left(E_2 \cap \{|\mathbf{Y}| = 1\}\right)\right).
\end{aligned}$$

Since E_2 implies that $|\mathbf{Y}| = 1$, we have

$$\begin{aligned}
&\Pr(E_1 || \mathbf{Y}| = 1) + \Pr\left(E_2 || \mathbf{Y}| = 1\right) \\
= &\frac{1}{\Pr(|\mathbf{Y}| = 1)} \left(\Pr(E_1 \cap \{|\mathbf{Y}| = 1\}) + \Pr\left(E_2\right)\right) \\
= &\frac{1}{\Pr(|\mathbf{Y}| = 1)} \left(\Pr(E_1) - \Pr(E_1 \cap \{|\mathbf{Y}| = 0\}) + \Pr\left(E_2\right)\right) \\
\geq &\frac{1}{\Pr(|\mathbf{Y}| = 1)} \left(\Pr(E_1) - \Pr(\{|\mathbf{Y}| = 0\}) + \Pr\left(E_2\right)\right) \\
\geq &\frac{1}{\Pr(|\mathbf{Y}| = 1)} \left(\Pr(\{|\mathbf{Y}| = 1\}) - \epsilon_4\right) \\
= &1 - \frac{\epsilon_4}{\Pr(|\mathbf{Y}| = 1)}.
\end{aligned}$$

Note that $\Pr(|\mathbf{Y}| = 1)$ is a positive constant since the B^1 branching process starting from the information source survives with non-zero probability. The theorem holds by choosing $\epsilon_4 = \epsilon \Pr(|\mathbf{Y}| = 1)$. □

The lemmas used in the proof are listed below and the detailed proofs can be found in [15].

Lemma 6. *Consider n_0 i.i.d GW branching processes with a binomial offspring distribution with parameters g and q such that $gq > 1$. Denote by K the number of branching processes that survive. Given any $\epsilon > 0$, if*

$$n_0 \geq \frac{8 \log \frac{1}{\epsilon}}{1 - \rho},$$

then

$$\Pr(K \geq 2) \geq 1 - \epsilon,$$

where ρ is the extinction probability of the GW branching process. In the binomial case, ρ is the smallest non-negative root of equation $\rho = (1 - q + q\rho)^g$. □

Lemma 7. *Given any $\epsilon > 0$, there exists a constant L' such that for any $L \geq L'$,*

$$\Pr\left(\bigcap_{i=1}^{L} \{0 < Z_i^\tau \leq n_0\}\right) \leq \epsilon.$$

□

Lemma 8. *Given any ϵ, there exist τ' and L' such that for any $\tau > \tau'$ and $L > L'$*

$$\Pr(Z_L = 0) - \Pr(Z_L^\tau = 0) \geq -\epsilon.$$

□

VI. Simulations

In this section, we evaluate the performance of the reverse infection algorithm on tree networks. We compare the reverse infection algorithm with the closeness centrality heuristic, which selects the node with the maximum infection closeness as the information source. Note that the node with the maximum closeness is the maximum likelihood estimator of the information source on regular trees under the SI model [3]–[5].

We first studied the performance on small-size trees. The infection probability q was chosen uniformly from $(0, 1)$ and the recovery probability p was chosen uniformly from $(0, q)$. The infection process propagates t time slots where t was uniformly chosen from $[3, 5]$. To keep the size of infection topology small, we restricted the total number of infected and recovered nodes to be no more than 100. For small-size trees, we first calculated the MLE using dynamic programming for fixed t and then searching over $t \in [0, t_{\max}]$ for a large value of $t_{\max}$ to find the optimal estimator.

The detection rate is defined to be the fraction of experiments in which the estimator coincides with the actual source. We varied g from 2 to 10 and the results are shown in Figure 5. We can see that the detection rate of the reverse infection

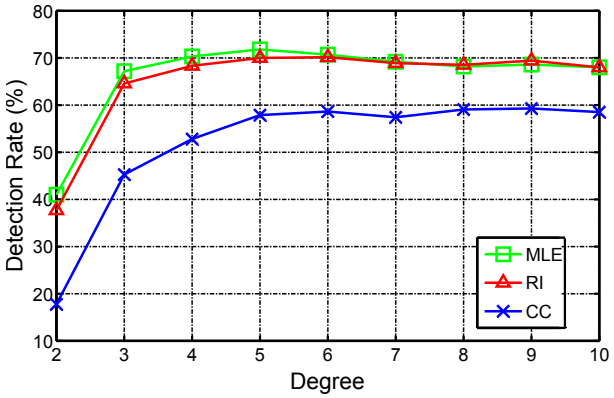

Figure 5. The Detection Rates of the Maximum Likelihood Estimator (MLE), Reverse Infection (RI) and Closeness Centrality (CC) on Regular Trees

algorithm is almost the same as that of the MLE, and is higher than that of the closeness centrality heuristic by approximately 20% when the degree is small and by 10% when the degree is large.

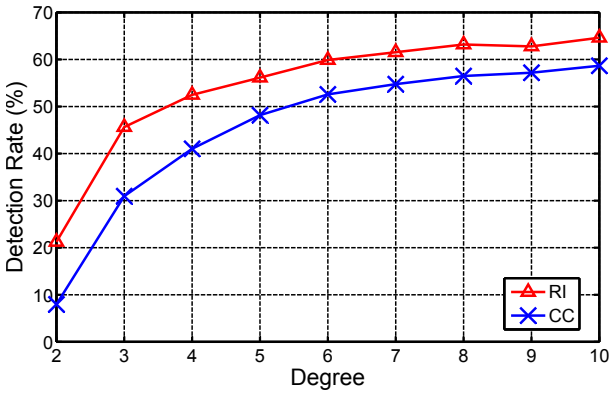

Figure 6. The Detection Rates of the Reverse Infection (RI) and Closeness Centrality (CC) Algorithms on Regular Trees

We further conducted our simulations on large-size g-regular trees. The infection probability q was chosen uniformly from $(0, 1)$ and the recovery probability p was chosen uniformly from $(0, q)$. The infection process propagates t time slots where t was uniformly chosen from $[3, 20]$. We selected the networks in which the total number of infected and recovered nodes is no more than 500.

We varied g from 2 to 10. Figure 6 shows the detection rate as a function of g. We can see the detection rates of both the reverse infection and closeness centrality algorithms increase as the degree increases and is higher than 60% when $g > 6$. However, he detection rate of the reverse infection algorithm is higher than that of the closeness centrality algorithm, and the average difference is 8.86%.

VII. Conclusion

In this paper, we developed a sample path based approach to find the information source under the SIR model. We proved that the sample path based estimator is a node with the minimum infection eccentricity. Based on that, a reverse infection algorithm has been proposed. We analyzed the performance of the reverse infection algorithm on regular trees, and showed that with high probability the distance between the estimator and actual source is a constant, independent of the number of infected nodes and the time the network was observed. We evaluated the performance of the proposed reverse infection algorithm on tree networks.

References

[1] C. Moore and M. E. J. Newman, "Epidemics and percolation in small-world networks," *Phys. Rev. E*, vol. 61, no. 5, pp. 5678–5682, 2000.

[2] A. Ganesh, L. Massoulie, and D. Towsley, "The effect of network topology on the spread of epidemics," in *Proc. IEEE Int. Conf. Computer Communications (INFOCOM)*, Miami, FL, Mar. 2005, pp. 1455–1466.

[3] 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.

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

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

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

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

[8] M. E. J. Newman, "The spread of epidemic disease on networks," *Phys. Rev. E*, vol. 66, no. 1, p. 016128, Jul. 2002.

[9] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, no. 14, pp. 3200–3203, 2001.

[10] V. G. Subramanian and R. Berry, "Spotting trendsetters: Inference for network games," in *Proc. Ann. Allerton Conf. Commununication, Control and Computing*, 2012.

[11] C. Milling, C. Caramanis, S. Mannor, and S. Shakkottai, "Network forensics: Random infection vs spreading epidemic," in *Proc. Ann. ACM SIGMETRICS Conf.*, 2012, pp. 223–234.

[12] P. Shakarian, V. S. Subrahmanian, and M. L. Sapino, "GAPs: Geospatial abduction problems," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 1, pp. 1–27, Oct. 2011.

[13] P. Shakarian and V. S. Subrahmanian, *Geospatial Abduction: Principles and Practice*. Springer, 2011.

[14] F. Harary, *Graph theory*. Addison-Wesley, 1991.

[15] K. Zhu and L. Ying, "Information source detection in the SIR model: A sample path based approach," *Arxiv preprint arXiv:1206.5421*, 2012.

[16] D. Koschutzki, K. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl, and O. Zlotowski, "Centrality indices," in *Network Analysis*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2005, vol. 3418, pp. 16–61.

[17] P. Haccou, P. Jagers, and V. A. Vatutin, *Branching processes: Variation, growth, and extinction of populations*. Cambridge University Press, 2005.