Zhu and Ying *Computational Social Networks* 2014, **1**:3
http://www.computationalsocialnetworks.com/content/1/1/3

Computational Social Networks
a SpringerOpen Journal

RESEARCH Open Access

A robust information source estimator with sparse observations

Kai Zhu* and Lei Ying

*Correspondence: kzhu17@asu.edu
School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287, USA

Abstract

Purpose/Background: In this paper, we consider the problem of locating the information source with sparse observations. We assume that a piece of information spreads in a network following a heterogeneous susceptible-infected-recovered (SIR) model, where a node is said to be *infected* when it receives the information and *recovered* when it removes or hides the information. We further assume that a small subset of infected nodes are reported, from which we need to find the source of the information.

Methods: We adopt the sample path-based estimator developed in the work of Zhu and Ying (arXiv:1206.5421, 2012) and prove that on infinite trees, the sample path-based estimator is a Jordan infection center with respect to the set of observed infected nodes. In other words, the sample path-based estimator minimizes the maximum distance to observed infected nodes. We further prove that the distance between the estimator and the actual source is upper bounded by a constant independent of the number of infected nodes with a high probability on infinite trees.

Results: Our simulations on tree networks and real-world networks show that the sample path-based estimator is closer to the actual source than several other algorithms.

Conclusions: In this paper, we proposed the sample path-based estimator for information source localization. Both theoretic analysis and numerical evaluations showed that the sample path-based estimator is robust and close to the real source.

Keywords: Information source detection; Heterogeneous SIR model; Sparse observation

Background

In this paper, we are interested in locating the source of information that spreads in a network by using sparse observations. The solution to this problem has important applications such as locating the sources of epidemics, the sources of news/rumors in social networks, or the sources of online computer virus. The problem has been studied in [1-5] under a homogeneous susceptible-infected (SI) model for information diffusion and in [6] under a homogeneous susceptible-infected-recovered (SIR) model for information diffusion, assuming that a complete snapshot of the network is given.

While [1-6] answered some basic questions about information source detection in large-scale networks, a complete snapshot of a real-world network, which may have

hundreds of millions of nodes, is expensive to obtain. Furthermore, these works assume homogeneous infection across links and homogeneous recovery across nodes, but in reality, most networks are heterogeneous. For example, people close to each other are more likely to share rumors, and epidemics are more infectious in the regions with poor medical care systems. Therefore, it is important to take sparse observations and network heterogeneity into account when locating information sources. In this paper, we assume that the information spreads in the network following a heterogeneous SIR model and assume that only a small subset of infected nodes are reported to us. The goal is to identify the information source in a heterogeneous network by using sparse observations.

We use the sample path-based approach developed in [6] for locating the information source with sparse observations. Surprisingly, we find that the sample path-based estimator is robust to network heterogeneity and the number of observed infected nodes. In particular, our results show that even under a heterogeneous SIR model and with sparse observations, the sample path-based estimator remains to be a Jordan infection center in infinite trees, where the Jordan infection centers with a partial observation are the nodes that minimize the maximum distance to observed infected nodes. We further show that in an infinite tree, the distance between a Jordan infection center and the actual source can be bounded by a value independent of the size of an infected subnetwork with a high probability, where the infected subnetwork is the subnetwork consisting of nodes which are either infected or recovered, and is a connected component. Assume that the size of the infected subnetwork is n, and the result says that a Jordan infection center is a distance of $O(1)$ from the actual source.

We remark that the locations of the Jordan centers only depend on the network topology and are independent of the infection and recovery probabilities, so the sample path-based estimators (or the Jordan infection centers) are also robust to the information diffusion model, which makes it very appealing in practice since the accurate knowledge of the SIR parameters can be difficult to measure in reality.

Related works

Other than [1-6], there are several related works in this area including the following: (1) detecting the first adopter of an innovation based on game theory [7], in which the maximum likelihood estimator is derived but the computational complexity of finding the estimator is exponential in the number of nodes; (2) distinguishing epidemic infection from random infection under the SI model [8]; and (3) geospatial abduction which deals with reasoning certain locations in a two-dimensional geographical area that can explain observed phenomena [9,10]. A recent paper [11] also proposed a dynamic message passing algorithm (DMP) to detect the information source under a general SIR model with complete or partial observations. However, the algorithm needs the complete information of infection and recovery probabilities. In addition, the complexity of DMP is very high under partial observations since almost all nodes in the network are candidates of the source, and the calculation needs to be repeated for every possible candidate. In the simulations, we will show that our algorithm significantly outperforms DMP in terms of both accuracy and speed. We will see that our algorithm is 400 times faster even when we limit the DMP algorithm to a subnetwork.

Methods

A heterogeneous SIR model

In this section, we introduce the heterogeneous SIR model for information propagation. Different from the homogeneous SIR model in which infection and recovery probabilities are both homogeneous [6], the heterogeneous SIR model we consider allows different infection probabilities at different links and different recovery probabilities at different nodes.

Consider an undirected graph $G = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges. Denote by $(u, v) \in \mathcal{E}$ the edge between node u and node v. Each node $v \in \mathcal{V}$ has three states: susceptible (S), infected (I), and recovered (R). A node is said to be susceptible if it has not received the information, infected after it receives the information, and recovered if the node removes or hides the information. Time is slotted. At the beginning of each time slot, each infected node attempts to contact all its susceptible neighbors. A contact from node u to node v *succeeds* with probability q_{uv}. A susceptible node becomes infected after being *successfully* contacted by one of its infected neighbors. At the middle of each time slot, an infected node, *if it is infected before the current time slot,* recovers with probability p_v. A recovered node cannot be infected again. We assume that contacts succeed independently across links and time slots and that nodes recover independently across nodes and time slots.

Consider a network shown in Figure 1, where node e is in the susceptible state, nodes a and c are in the infected state, and nodes b and d are in the recovered state. Then, at the next time slot, node e becomes infected with probability

$$1 - (1 - q_{ae})(1 - q_{ce}),$$

and nodes a and c recover with probability p_a and p_c, respectively.

Problem formulation

In this section, we formally define the problem of information source detection. Table 1 summarizes the notations used in the paper. Adopting the notation in [6], we define $X_v(t)$

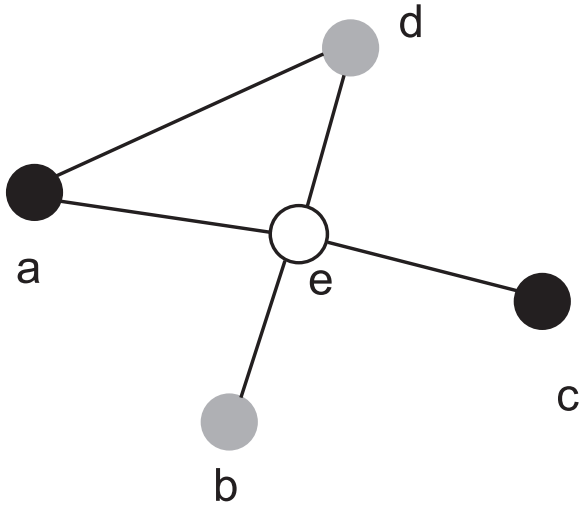

Figure 1 An example for illustrating the heterogeneous SIR model.

Table 1 Notation table

	Description
q_{uv}	The probability an infected node u infects its neighbor node v
p_v	The probability an infected node v recovers
$\mathbf{Y}$	The partial snapshot
$X_v(t)$	The state of node v at time t
$\mathbf{X}(t)$	The states of all nodes at time t
$\mathbf{X}[0,t]$	The sample path from 0 to t
$\mathcal{X}(t)$	The set of all valid sample paths from time slot 0 to t
$\mathcal{I}_{\mathbf{Y}}$	The set of the observed infected nodes
$\mathcal{H}_{\mathbf{Y}}$	The set of the unobserved nodes
$\tilde{e}(v,\mathcal{I}_{\mathbf{Y}})$	The observed infection eccentricity of node v
$v^{\dagger}$	The estimator of the information source
v^*	The actual information source
t_v^*	The time duration associated to the optimal sample path in which node v is the information source
$\mathcal{C}(v)$	The set of children of v
$\phi(v)$	The parent of node v
$\mathcal{Y}^k$	The set of infection topologies where the maximum distance from the source to an infected node is k
T_v	The tree rooted in v
T_v^{-u}	The tree rooted in v without the branch from its neighbor u
$\mathbf{X}\left([0,t],T_v^{-u}\right)$	The sample path restricted to topology T_v^{-u}
t_v^I, t_v^R	The infection time and recovery time of node v
$d(v,u)$	The length of the shortest path between node v and node u

to be the states of node v at the end of time slot t such that

$$X_v(t) = \begin{cases} \text{S}, & \text{if } v \text{ is in state S at timet;} \\ \text{I}, & \text{if } v \text{ is in state I at timet;} \\ \text{R}, & \text{if } v \text{ is in state R at timet.} \end{cases}$$

Let $\mathbf{X}(t) = \{X_v(t) : \forall v \in \mathcal{V}\}$ denote the states of all nodes at time instant t.

In this paper, we assume that we only have *one partial snapshot* of the network, which is *a subset of the infected nodes.* This observation can be sparse, and details will be given in the next section. We assume that the states of other nodes are unknown. We let Y_v denote the state of node v in the snapshot such that

$$Y_v = \begin{cases} 1, & \text{if node } v \text{ is observed to be infected;} \\ 0, & \text{otherwise.} \end{cases}$$

Let $\mathbf{Y} = \{Y_v : \forall v \in \mathcal{V}\}$. We denote by v^* the information source. The problem of information source detection is to locate v^* based on the partial observation $\mathbf{Y}$ and the network topology G.

Due to recovery and partial observations, all nodes in the network are potential candidates of the information source. The maximum likelihood estimator of the problem is therefore computationally expensive to find as pointed out in [6]. In this paper, we follow the sample path-based approach proposed in [6] to find an estimator of v^*.

Since $\mathbf{X}(t)$ is the state of the network at time t, the sequence $\{\mathbf{X}(\tau)\}_{0\le\tau\le t}$ specifies the complete infection process. Therefore, we call $\mathbf{X}[0,t] = \{\mathbf{X}(\tau) : 0 \le \tau \le t\}$ a sample path

which is the states of all nodes from time 0 to time t. We further define a function $F(\cdot)$ such that

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

This function maps the actual state of a node to the observed state of the node. $\mathbf{F}(\mathbf{X}(t)) = \mathbf{Y}$ if and only if $F(X_v(t)) = Y_v, \forall v \in \mathcal{V}$. The optimal sample path $\mathbf{X}^*[0, t^*]$ is defined to be the most likely sample path that results in the observed snapshot, i.e., it solves the following optimization problem:

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

where $\mathcal{X}(t) = \{\mathbf{X}[0, t] \,|\mathbf{F}(\mathbf{X}(t)) = \mathbf{Y}\}$ and $\Pr(\mathbf{X}[0, t])$ is the probability that the sample path $\mathbf{X}[0, t]$ occurs. The source that associates with $\mathbf{X}^*[0, t^*]$ is called *the sample path-based estimator*. It is proved in [6] that the sample path-based estimator on an infinite tree is a Jordan infection center under the homogeneous SIR model with a complete snapshot. The focus of this paper is to identify the sample path-based estimator under the heterogeneous SIR model with sparse observations.

Main results

In this section, we summarize the main results of this paper.

Main result 1: the Jordan infection centers as the sample path-based estimators

In our theoretical analysis, we consider tree networks with infinitely many levels (or called infinite trees) to derive the sample path-based estimator under the heterogeneous SIR model with a partial snapshot. Let $\mathcal{I}_\mathbf{Y}$ denote the set of observed infected nodes. We define the observed infection eccentricity $\tilde{e}(v, \mathcal{I}_\mathbf{Y})$ of node v to be the maximum distance between v and any observed infected node where the distance is defined to be the shortest distance between two nodes. The Jordan infection centers of the partial snapshot are then defined to be the nodes with the minimum observed infection eccentricity. The following theorem states that on an infinite tree, the sample path-based estimator is a Jordan infection center of the partial snapshot.

Theorem 1. *Consider an infinite tree and assume that the partial snapshot* $\mathbf{Y}$ *contains at least one infected node. The sample path-based estimator, denoted by* $v^\dagger$*, is a Jordan infection center, i.e.,*

$$v^\dagger \in \arg\min_{v \in \mathcal{V}} \tilde{e}(v, \mathcal{I}_\mathbf{Y}). \tag{2}$$

□

The proof of this theorem consists of the following key steps.

1. In the first step, we focus on the sample paths originated from node v (i.e., we assume node v is the source). We consider two groups of sample paths: $\mathcal{X}_v(t)$ and $\mathcal{X}_v(t+1)$, where $\mathcal{X}_v(t)$ is the set of the sample paths that are *originated from* v, *have time duration* t, and *are consistent with the partial snapshot*, i.e., $\mathbf{F}(\mathbf{X}(t)) = \mathbf{Y}$ for any $\mathbf{X}[0, t] \in \mathcal{X}_v(t)$. The set $\mathcal{X}_v(t+1)$ is similarly defined. We show that for any

$t \geq \tilde{e}(v, \mathcal{I}_{\mathbf{Y}})$, the sample path with the highest probability in $\mathcal{X}_v(t)$ occurs more likely than the one in $\mathcal{X}_v(t+1)$. In other words,

$$\max_{\mathbf{X}[0,t] \in \mathcal{X}_v(t)} \Pr(\mathbf{X}[0,t]) > \max_{\mathbf{X}[0,t+1] \in \mathcal{X}_v(t+1)} \Pr(\mathbf{X}[0,t+1]).$$

As a consequence of this result, we conclude that *the sample path that has the highest probability among those originated from node v has a duration of $\tilde{e}(v, \mathcal{I}_{\mathbf{Y}})$ (the observed infection eccentricity of node v).* This result will be proved in Lemma 1 in the 'Proofs' section.

2. In the second step, we consider two neighboring nodes, say nodes u and v, and assume node v has a smaller observed infection eccentricity than node u. Based on Lemma 1, we will prove that the optimal sample path associated with node v occurs with a higher probability than that of node u. The key idea is to construct a sample path originated from node v based on the optimal sample path originated from node u and show that it occurs with a higher probability. This result will be proved in Lemma 2 in the 'Proofs' section.
3. We will finally prove that starting from any node, there exists a path from the node to a Jordan infection center such that the observed infection eccentricity strictly decreases along the path. Consider an example in Figure 2. Nodes b and f are two observed infected nodes. So node a is a Jordan infection center with observed infection eccentricity 1. The path from node e to node a is

   $$e \to d \to c \to b \to a,$$

   along which the observed infection eccentricity decreases as

   $$5 \to 4 \to 3 \to 2 \to 1.$$

   By repeatedly using Lemma 2, it can be shown that the optimal sample path originated from a Jordan infection center occurs with a higher probability than the optimal sample path originated from a node which is not a Jordan infection center, which implies that the sample path-based estimator must be a Jordan infection center.

Main result 2: an $O(1)$ bound on the distance between a Jordan infection center and the actual information source

Unlike the maximum likelihood estimator, the sample path estimator does not guarantee that the estimator is the node that most likely leads to the observation. It has been shown in [6] that on tree networks and under the homogeneous SIR model, the distance between

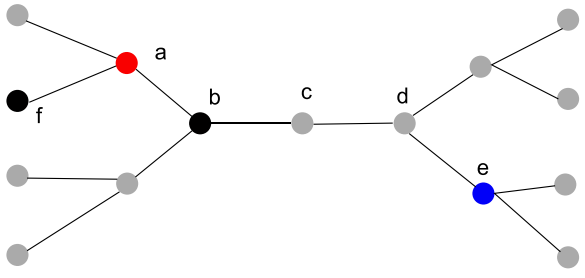

Figure 2 The key intuition behind Theorem 1.

the estimator and the actual source is a constant with a high probability. It is easy to see that with a partial observation, the distance between the estimator and the actual source cannot be bounded if the observed infection nodes are arbitrarily chosen. In this paper, we consider a class of fairly general sampling algorithms that generate the partial observation (and maybe sparse). The sampling algorithms have the following property: *for any set of M infected nodes, the probability that at least one node in the set is reported approaches to 1 as M goes to infinity.* We call such a sampling algorithm *unbiased*; in other words, any subset of infected nodes is likely to contain an observed infected node when the size of the subset is large enough. Note that if an infected node is reported with probability at least δ for some $\delta > 0$, independent of other nodes, then it satisfies the property above. Our second main result is that the sample path estimator is within a constant distance from the actual source independent of the size of the infected subnetwork if the sampling algorithm is unbiased. *We also emphasize that the observation generated by an unbiased sampling algorithm can be very sparse since we only require that one observed infected node is reported with a high probability among M nodes when M is sufficiently large.*

Theorem 2. *Consider an infinite tree. Let* $g_{\min}$ *be the lower bound on the number of children and* $q_{\min} > 0$ *be the lower bound on q. Assume* $g_{\min} > 1$, $g_{\min}q_{\min} > 1$, *and the observed infection topology* $\boldsymbol{Y}$ *contains at least one infected node and is generated by an unbiased sampling algorithm. Then given* $\epsilon > 0$, *the distance between the sample path estimator and the actual source is* d_ϵ *with probability* $1 - \epsilon$, *where* d_ϵ *is independent of the size of the infected subnetwork. In other words, the distance is* $O(1)$ *with a high probability.* □

The idea of the proof is illustrated using Figure 3, which consists of the following key steps:

1. We first define a one-time-slot infection subtree to be a subtree of the infected subnetwork such that each node on the subtree is infected in the next time slot after the parent is infected, except the source node. Note that the depth of a one-time-slot infection subtree grows by 1 deterministically until it terminates. We further say a node survives at time t if it is the root of a one-time-slot infection subtree which has not terminated by time t.
2. In the first step, we will prove that there exist at least two survived nodes within a distance L from the information source. In Figure 3, node a is the information source, and nodes b and c are two survived nodes.

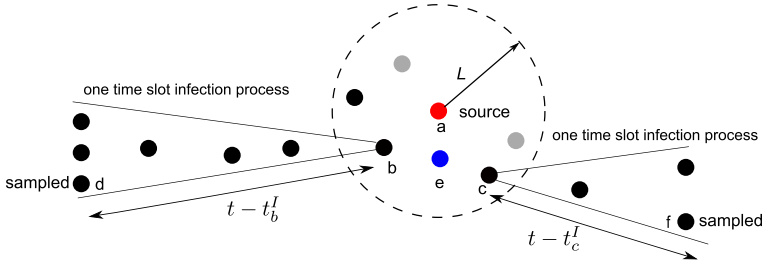

Figure 3 The key intuition behind Theorem 2.

3. In the second step, we will show that with a high probability, at least one infected node at the bottom of a one-time-slot infection subtree, which has not terminated, is observed under an unbiased sampling algorithm. In Figure 3, nodes d and f are two sampled nodes corresponding to the two one-time-slot infection subtrees starting from nodes b and c, respectively.
4. Since a one-time-slot infection subtree grows by 1 deterministically at each time slot, the depth of a one-time-slot infection subtree is $t - t_k^I$, where k is the root node of the one-time-slot infection subtree. Recall that the Jordan infection centers minimize the maximum distance to observed infected nodes, so a Jordan infection center must be within a $O(1)$ distance from the two survived nodes (nodes b and c). Considering Figure 3, we know that the actual source (node a) has an infection eccentricity $\leq t$ since the information can propagate at most t hops at time t. So the infection eccentricity of the Jordan infection centers is no more than t according to the definition. Assume node e in Figure 3 is a Jordan infection center, then it is within a distance of $O(t)$ from nodes d and f, and so is within a distance of $O(1)$ from nodes b and c. Since nodes b and c are no more than L hops from the actual source a, we can conclude that the distance between the actual source a and the estimator e is $O(1)$.

Reverse infection algorithm

The Jordan infection centers for general graphs can be identified by the reverse infection algorithm proposed in [6]. In the algorithm, each observed infected node broadcasts its identity (ID) to its neighbors. All nodes in the network record the distinct IDs they received. When a node receives a new distinct ID, it records it and then broadcasts it to its neighbors. This process stops when there is a node which receives the IDs from all observed infected nodes. It is easy to verify that the set of nodes which first receive all infected IDs is the set of Jordan infection centers. When there are multiple Jordan infection centers in the graph, we select the one with the maximum infection closeness centrality as the information center. The infection closeness centrality is defined as the inverse of the sum of the distances from one node to all observed infected nodes.

We explain the reverse infection algorithm using an example in Figure 4. The red nodes are the observed infected nodes, and the black nodes are the unobserved nodes. The array next to each node records the IDs that the node has received. When an ID is received, it

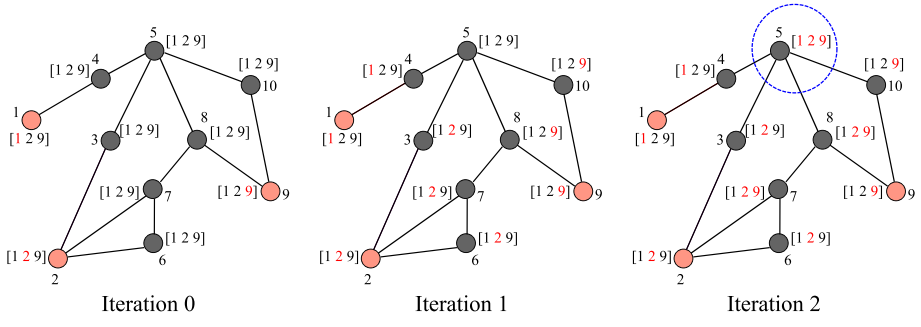

Figure 4 An example of the reverse infection algorithm.

is colored in red. For example, node 7 in iteration 1 has received the ID of node 2 which is colored in red and has not received the ID of nodes 1 and 9 which are in black. At each iteration, each node broadcasts its newly received IDs to its neighbors. For example, node 4 just received the ID of node 1 in iteration 1 so it will broadcast node 1's ID to its neighbors in iteration 2. The algorithm terminates when some nodes receive the IDs of all observed infected nodes, and this node is the Jordan infection center. In iteration 3, node 5 received all IDs and so node 5 is the Jordan infection center in the example.

Discussion: robustness

According to the two main results above, we know that the sample path-based estimator remains to be a Jordan infection center. This is a somewhat surprising result since the locations of the Jordan infection centers are determined by the topology of the network and are *independent of the parameters of the heterogeneous SIR model.* In other words, the locations of the Jordan infection centers remain the same for different SIR processes as long as the set of observed infected nodes is the same. This property suggests that the sample path-based estimator is a robust estimator and can be used in the case when the parameters of the SIR model are unknown, which is a very desirable property since knowing these parameters can be difficult in practice.

In the simulations, we also consider a weighted graph with the link weights chosen proportionally according to the SIR parameters and use the weighted Jordan infection centers as the estimator. Interestingly, we will see that the performance is worse than the unweighted Jordan infection centers, which again demonstrates the robustness of the sample path-based estimator.

Furthermore, the main results hold as long as the sampling algorithm is unbiased and are independent of the number of samples. So the results are valid for sparse observations and are robust to the number of observations.

Results and discussion

Simulations

In this section, we evaluate the performance of the reverse infection algorithm for the heterogeneous SIR model on different networks including tree networks and real-world networks.

We first describe the heterogeneous SIR model we used in the simulation. Each edge $e \in \mathcal{E}$ is assigned with a weight q_e which is uniformly distributed over $(0, 1)$. The infection time over each edge $e \in \mathcal{E}$ is geometrically distributed with mean $1/q_e$. Similarly, each node $v \in \mathcal{V}$ is assigned with a weight p_v generated by a uniform distribution over $(0, 1)$, and the recovery time is geometrically distributed with mean $1/p_v$. The information source is randomly selected. The total number of infected and recovered nodes in each infection graph is within the range of $[100, 300]$. Each infected node v in the infection graph reports with probability σ, independently. The snapshots used in the simulations have at least one infected node. We changed σ and evaluated the performance on different networks.

We briefly introduce the three main algorithms which were used to compare with the reverse infection algorithm (RI).

1. *Closeness centrality algorithm (CC)*: The closeness centrality algorithm selects the node with the maximum infection closeness as the information source.
2. *Weighted reverse infection algorithm (wRI)*: The weighted reverse infection algorithm selects the node with the minimum weighted infection eccentricity as the information source where the weighted infection eccentricity is similar to the infection eccentricity except that the length of a path is defined to be the sum of the link weights instead of the number of hops, and the link weight is the average time it takes to spread the information over the link, i.e., $\lfloor 1/q_e \rfloor$ on edge e.
3. *Weighted closeness centrality algorithm (wCC)*: The weighted closeness centrality algorithm selects the node with the maximum weighted infection closeness as the information source.

Tree networks

We first evaluated the performance of the RI algorithm on tree networks.

Regular trees A g-regular tree is a tree where each node has g neighbors. We set the degree $g = 5$ in our simulations.

We varied the sample probability σ from 0.01 to 0.1. The simulation results are summarized in Figure 5a, which shows the average distance between the estimator and the actual information source versus the sampling probability. When the sample probability increases, the performance of all algorithms improves. When the sample probability is

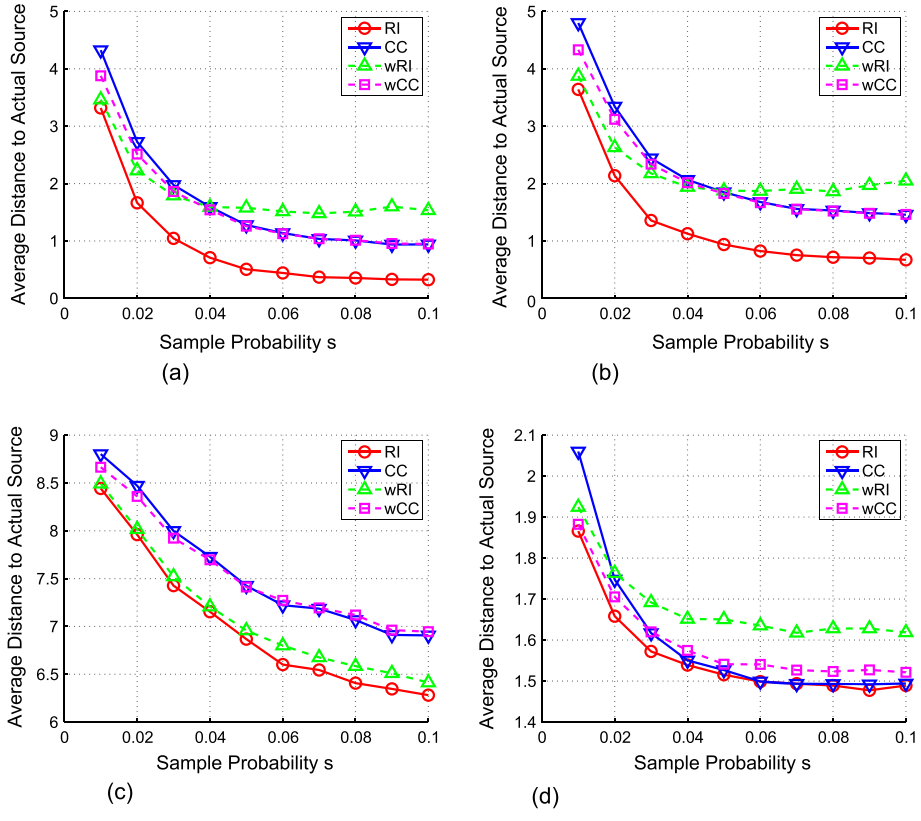

Figure 5 The Performance of RI, CC, wRI, and wCC on different graphs. (a) Regular tree. **(b)** Binomial tree. **(c)** The power grid network. **(d)** The Internet autonomous systems network.

larger than 6%, the average distance becomes stable which means that a small number of infected nodes is enough to obtain a good estimator. We also notice that the average distance of RI is smaller than all other algorithms and is less than one hop when $\sigma \geq 0.04$. wRI has a similar performance with RI when the sample probability is small (=0.01) but becomes much worse when the sample probability increases.

Binomial trees We further evaluated the performance of RI and other algorithms on binomial trees $T(\xi, \beta)$ where the number of children of each node follows a binomial distribution such that ξ is the number of trials and β is the success probability of each trial. In the simulations, we selected $\xi = 10$ and $\beta = 0.4$. Again, we varied σ from 0.01 to 0.1. The results are shown in Figure 5b. Similar to the regular trees, the performance of RI dominates CC, wRI, and wCC, and the difference in terms of the average number of hops is approximately 1 when $\sigma \geq 0.03$.

Real-world networks

In this section, we conducted experiments on two real-world networks: the Internet autonomous systems (IAS) network which is available at http://snap.stanford.edu/data/index.html and the power grid (PG) network which is available at http://www-personal.umich.edu/~mejn/netdata/.

The power grid network The power grid network has 4,941 nodes and 6,594 edges. On average, each node has 1.33 edges. So the power grid network is a sparse network. The simulation results are shown in Figure 5c. In the power grid network, we can see that RI and wRI have similar performance, and both outperform CC and wCC by at least one hop when $\sigma \geq 0.04$.

The internet autonomous systems network The Internet autonomous systems network is the data collected on 31 March 2001. There are 10,670 nodes and 22,002 edges in the network. The simulation results are shown in Figure 5d. wRI and wCC always perform worse than RI. Although RI and CC have similar performance when the sample probability is large, RI outperforms CC when $\sigma \leq 0.03$.

RI versus DMP

We finally compared the performance of RI and DMP. We conducted the simulation on the power grid network and fixed the sample probability to be 10%. Under this setting, the complexity of DMP is very high since the DMP computation needs to be repeated for every node in the network. Since nodes far away from the observed infected nodes are not likely to be the information source, we ran DMP over a small subset of nodes close to the Jordan infection centers (roughly 10%) to reduce the complexity of the algorithm.

We tested the speed of RI and DMP on a machine with 1.8 GB memory, 4 cores 2.4 GHz Intel i5 CPU and Ubuntu 12.10. The algorithms are implemented in Python 2.7. On average, it took RI 0.57 s to locate the estimator for one snapshot and took DMP 229.12 s. So RI is much faster than DMP.

Figure 6 shows the cumulative distribution function (CDF) of the distance from the estimator to the actual source under DMP and RI. We can see that RI dominates DMP; in particular, 71% of the estimators under RI are no more than seven hops from the actual

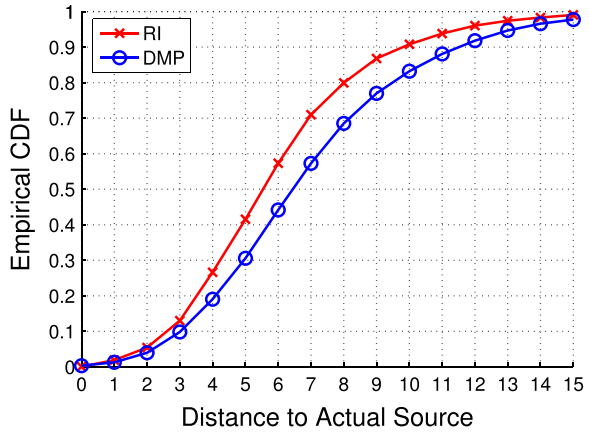

Figure 6 The CDF of RI and DMP on the power grid network.

source compared to 57% under DMP. Therefore, RI outperforms DMP in terms of both speed and accuracy. We remark that we did not compare the performance of RI and DMP on the IAS network because the complexity of running DMP on a large-sized network like the IAS network is prohibitively high.

Proofs

In this section, we present the proofs of the main results.

Proof of Theorem 1

Denote by $\mathcal{I}_{\mathbf{Y}} = \{v | Y_v = 1\}$ the set of observed infected nodes and $\mathcal{H}_{\mathbf{Y}} = \{v | Y_v = 0\}$ the set of unobserved nodes. Given a node v, define the optimal time t_v^* to be

$$t_v^* \triangleq \arg_t \max_{t, \mathbf{X}[0,t] \in \mathcal{X}(t)} \Pr\left(\mathbf{X}[\,0, t] \,| v \text{ is information source}\right),$$

i.e., it is the duration of the optimal sample path with node v as the information source.

Lemma 1 (**Time Inequality**). *Consider an infinite tree rooted at v_r. Assume that v_r is the information source and the observed snapshot $\boldsymbol{Y}$ contains at least one infected node. If $\tilde{e}(v_r, \mathcal{I}_{\boldsymbol{Y}}) \leq t_1 < t_2$, the following inequality holds:*

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

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

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

i.e., $t_{v_r}^*$ is equal to the observed infection eccentricity of v_r with respect to $\mathcal{I}_{\mathbf{Y}}$.

Proof. We adopt the notations defined in [6], which are listed below:

- $\mathcal{C}(v)$ is the set of children of v.
- $\phi(v)$ is the parent of node v.

- $\mathcal{Y}^k$ is the set of infection topologies where the maximum distance from v_r to an infected node is k. All possible infection topologies are then partitioned into countable subsets $\{\mathcal{Y}^k\}$.
- T_v is the tree rooted in v.
- T_v^{-u} is the tree rooted in v without the branch from its neighbor u.
- $\mathbf{X}([0,t]\,, T_v^{-u})$ is the sample path restricted to topology T_v^{-u}.
- t_v^I, t_v^R are the infection time and recovery time of node v.

Considering the case where the time difference of two sample paths is 1, we will show that

$$\max_{\mathbf{X}[0,t]\in\tilde{\mathcal{X}}(t)} \Pr(\mathbf{X}[0,t]) > \max_{\mathbf{X}[0,t+1]\in\tilde{\mathcal{X}}(t+1)} \Pr(\mathbf{X}[0,t+1]).$$

Next, we use induction over $\mathcal{Y}^k$.

Step 1 $k=0$ v_r is the only observed infected node in this case. Given a sample path $\mathbf{X}[0,t+1]\in\tilde{\mathcal{X}}(t+1)$, the probability of the sample path can be written as

$$\Pr\left(\mathbf{X}[0,t+1]\right) = \Pr\left(\mathbf{X}[0,t]\right)\Pr(\mathbf{X}(t+1)|\mathbf{X}[0,t]).$$

Since v_r is the only observed infected node and all other nodes' states are unknown, we assign $\mathbf{X}'[0,t]\in\tilde{\mathcal{X}}(t)$ to be same as the first t time slots in $\mathbf{X}[0,t+1]\,$, i.e., $\mathbf{X}'[0,t]=\mathbf{X}[0,t]\,$. Hence, we obtain that

$$\Pr\left(\mathbf{X}'[0,t]\right) = \Pr\left(\mathbf{X}[0,t]\right) > \Pr\left(\mathbf{X}[0,t+1]\right).$$

Therefore, the case $k=0$ is proved.

Step 2 Assume the inequality holds for $k\le n$ and consider $k=n+1$, i.e., $\mathbf{Y}\in\mathcal{Y}^{n+1}$. Clearly, $t\ge n+1\ge 1$ for each $\mathbf{X}[0,t]$. Furthermore, the set of subtrees $\mathcal{T}=\{T_u^{-v_r}|u\in\mathcal{C}(v_r)\}$ are divided into two subsets:

$$\mathcal{T}^h=\{T_u^{-v_r}|u\in\mathcal{C}(v_r), T_u^{-v_r}\cap\mathcal{I}_{\mathbf{Y}}=\emptyset\}$$

and

$$\mathcal{T}^i=\mathcal{T}\backslash\mathcal{T}^h.$$

Given $t_{v_r}^R$, the infection processes on the subtrees are mutually independent.

We construct $\mathbf{X}'[0,t]$ which occurs more likely than $\mathbf{X}^*[0,t+1]$ according to the following steps, where $\mathbf{X}^*[0,t+1]=\arg\max_{\mathbf{X}[0,t+1]\in\tilde{\mathcal{X}}(t+1)}\Pr(\mathbf{X}[0,t+1])$.

Part 1 $\mathcal{T}^i$. For a subtree in $\boldsymbol{\mathcal{T}}^i$ the proof follows Step 2.b and Step 2.c of Lemma 1 in [6]. The intuition is as follows: Consider a subtree and a sample path on it with duration $t+1$. If u is not infected at the first time slot, we can construct a sample path with duration t by moving the events one time slot earlier. The new sample path (with duration t) has a higher probability to occur than the original one. If u is infected in the first time slot, we can invoke the induction assumption to the subtree rooted at u, which belongs to $\mathcal{Y}^n$.

Part 2 v_r. In this part, we have the freedom to assign the unobserved node as infected or healthy. In part 1, the infection time of each root u in subtrees $\mathcal{T}^i$ of $\mathbf{X}'[0,t]$ is either the same as or one time slot earlier than its infection time in $\mathbf{X}^*[0,t+1]$. Therefore, if $t_{v_r}^R\le t$, the recovery time of the source v_r in $\mathbf{X}'[0,t]$ can be assigned the same as that in $\mathbf{X}^*[0,t+1]$.

If $t_{v_r}^R=t+1$, the source v_r recovers at time slot $t+1$ which means v_r is not observed since the observation set only contains infected nodes. Therefore, in $\mathbf{X}'[0,t]$ we assign the source to be in state I at time t, which is the same as the state of v_r at time t in $\mathbf{X}^*[0,t+1]$.

If $t^R_{v_r} > t+1$, v_r remains infected in the sample path $\mathbf{X}^*[0, t+1]$. We assign the source to be in state I in $\mathbf{X}'[0, t]$.

As a summary, according to the assignment above, the states of the source v_r in $\mathbf{X}'[0, t]$ are the same as those of the first t time slots in $\mathbf{X}^*[0, t+1]$.

Part 3 $\mathcal{T}^{\mathbf{h}}$. Based on the conclusion of part 2, the subtrees belonging to $\mathcal{T}^h$ in $\mathbf{X}'[0, t]$ mimic the behaviors of the first t time slots in $\mathbf{X}^*[0, t+1]$.

Since $\mathbf{X}^*[0, t+1]$ has one extra time slot during which some extra events occur, $\mathbf{X}'[0, t]$ occurs with a higher probability on the subtrees in $\mathcal{T}^h$.

According to the discussion above, we conclude that time inequality holds for $k = n+1$ and hence for any k according to the principle of induction. Therefore, the lemma holds. □

Lemma 2 (**Adjacent nodes inequality**). *Consider an infinite tree with partial observation* $\boldsymbol{Y}$ *which contains at least one infected node. For* $u, v \in \mathcal{V}$ *such that* $(u, v) \in \mathcal{E}$, *if* $t^*_u > t^*_v$

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

where $\boldsymbol{X}^*_u[0, t^*_u]$ *is the optimal sample path associated with root* u.

Proof. The proof of the lemma follows the proof of Lemma 2 in [6]. The key idea is to construct a sample path rooted at v, which has a higher probability than the optimal sample path rooted at u. It is not hard to see that $t^*_u = t^*_v + 1$ based on the definition of the infection eccentricity. The graph is partitioned into T_v^{-u} and T_u^{-v} which are mutually independent after the infection of v and u. With this observation, we construct $\tilde{\mathbf{X}}_v[0, t^*_v]$ which infects u at the first time slot. $\tilde{\mathbf{X}}_v\left([0, t^*_v], T_v^{-u}\right)$ then mimics the behavior of $\mathbf{X}^*_u\left([0, t^*_u], T_v^{-u}\right)$, and $\tilde{\mathbf{X}}_v\left([0, t^*_v - 1], T_u^{-v}\right)$ has a higher probability than $\mathbf{X}^*_u\left([0, t^*_u], T_u^{-v}\right)$ based on Lemma 1. □

The adjacent nodes inequality results in partial orders in the tree and makes it possible to compare the likelihood of optimal sample paths associated with adjacent nodes without knowing the actual probability of the optimal sample path. Following the proof of Theorem 4 in [6], it can be shown that in tree networks, from any node, there exists a path from the node to a Jordan infection center such that the observed infection eccentricity strictly decreases along the path. By repeatedly using Lemma 2, we can then prove that the source of the optimal sample path must be a Jordan infection center.

Proof of Theorem 2

In this subsection, we present the proof that shows that the sample path estimator is within a constant distance from the actual source independent of the size of the infected subnetwork. Given a tree rooted in v^* where the information starts from v^* following the general SIR model, we define the following three branching processes:

1. $\mathcal{Z}_l(T_{v^*})$ denotes the set of nodes which are in infected or recovered states at level l on tree T_{v^*}. Let $Z_l(T_{v^*})$ denote the cardinality of $\mathcal{Z}_l(T_{v^*})$. Note that $\mathcal{Z}_0(T_{v^*}) = \{v^*\}$. We call this process the *original infection process.*
2. $\mathcal{Z}^\tau_l(T_{v^*})$ denotes the set of infected and recovered nodes at level l whose parents are in set $\mathcal{Z}^\tau_{l-1}(T_{v^*})$ and who were infected within τ time slots after their parents

were infected. This process adds a deadline τ on infection. If a node is not infected within τ time slots after its parent is infected, it is not included in this branching process. This process is called τ*-deadline infection process.* From the definition, if $u, v \in \mathcal{Z}_l^\tau(T_{v^*})$, then

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

For $\tau = 1$, we call $\mathcal{Z}_l^1(T_{v^*})$ the *one-time-slot infection process*. The extinction probability of a branching process is the probability that there is no offspring at a certain level of the branching process, i.e., $Z_l^1(T_{v^*}) = 0$ for some l. Denote by ρ_v the extinction probability of $Z_l^1\left(T_v^{-\phi(v)}\right)$.

3. We define the *binomial branching process* as a branching process whose offspring distribution follows binomial distribution $B(g, \varphi)$ where g is the number of trials and φ is the success probability. Denote by ρ the extinction probability of the binomial branching process.

   The following notations will be used in later analysis:

   - $v^\dagger$ denotes the optimal sample path estimator.
   - $g_{\min}$ is the lower bound on the number of children, i.e.,

     $$\min_v |\mathcal{C}(v)| \geq g_{\min}, \forall v \in \mathcal{V}.$$

   - $q_{\min}$ is the lower bound on the infection probability, i.e.,

     $$q_{\min} = \min_e q_e, \forall e \in \mathcal{E}.$$

   - σ_v^τ is the probability that a node v infects at least one of its children within τ time slot after v is infected.

Given $n_0 > 0$ and $\tau > 0$, define $l^\dagger = \min l'$ where $Z_{l'}^\tau(T_{v^*}) > n_0$, i.e., $l^\dagger$ is the first level where the τ-deadline infection process has more than n_0 offsprings.

Given τ and level $L \geq 2$, we consider the following two events:

Event 1: $Z_L(T_{v^*}) = 0$.

Event 2: $l^\dagger \leq L$ and at least two one-time-slot infection processes starting from level $l^\dagger$ survive, i.e., $\exists u, v \in \mathcal{Z}_{l^\dagger}^\tau(T_{v^*})$ such that $\forall l$, $Z_l^1\left(T_u^{-\phi(u)}\right) \neq 0$ and $Z_l^1\left(T_v^{-\phi(v)}\right) \neq 0$. In addition, at least one infected node at the bottom of each survived one-time-slot infection process is observed.

For event 1, no node at level L gets infected and the infection process terminates at level $L - 1$. So the infection eccentricity of v^* is at most $L - 1$, and the minimum infection eccentricity of the network is at most $L - 1$. Therefore, the distance between v^* and $v^\dagger$ is no more than $2(L - 1)$.

Considering event 2, we assume that the information propagates for t time slots. The deadline property of the τ-deadline infection process indicates $t_{u_1}^I \leq \tau l^\dagger$ and $t_{u_2}^I \leq \tau l^\dagger$. Given a node $\tilde{v}$ at level $(\tau + 1)l^\dagger - 1$ where $\tilde{v} \in T_{u_2}^{-\phi(u_2)}$ and a node $v' \in T_{u_1}^{-\phi(u_1)}$ which is an observed infected node at the bottom of the infection tree, from Figure 7, we obtain

$$d\left(\tilde{v}, v'\right) = t - t_{u_1}^I + \tau l^\dagger + 1 \geq t + 1.$$

Note that $\forall u \in \mathcal{I}$,

$$d\left(v^*, u\right) \leq t < d\left(\tilde{v}, v'\right).$$

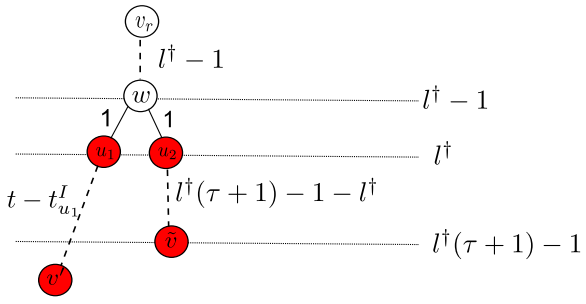

Figure 7 A pictorial description of the distance relations in Theorem 2.

Since $l^\dagger \leq L$, any node at or below level $L(\tau+1)-1$ has an infection eccentricity larger than that of v^*. Hence, $v^\dagger$ cannot be at or below level $L(\tau+1)-1$. Therefore,

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

Next, we prove the probability that either event 1 or event 2 happens goes asymptotically to 1. Denote by $K_{l^\dagger}$ the number of one-time-slot infection processes which start from level $l^\dagger$ and survive. Denote by E the event that a survived one-time-slot infection process has at least one observed infected node at its lowest level.

According to the discussion above, the probability that the distance between the estimator and the actual source is no more than $(\tau+1)L-1$ is at least

$$\begin{aligned}
&\Pr\left(Z_L\left(T_{v^*}\right)=0\right) + \Pr\left(K_{l^\dagger} \geq 2, l^\dagger \leq L\right) \Pr(E)^2 \\
&\quad \geq \Pr\left(Z_L\left(T_{v^*}\right)=0\right) + \Pr\left(l^\dagger \leq L\right) \Pr\left(K_{l^\dagger} \geq 2 \middle| l^\dagger \leq L\right) \Pr(E)^2 \\
&\quad = \Pr\left(Z_L\left(T_{v^*}\right)=0\right) + \Pr\left(\bigcup_{i=1}^{L} Z_i^\tau > n_0\right) \\
&\qquad \times \Pr(K_{l^\dagger} \geq 2 | l^\dagger \leq L) \Pr(E)^2 \\
&\quad = \left(1 - \Pr\left(\bigcap_{i=1}^{L} 0 < Z_i^\tau(T_{v^*}) \leq n_0\right) - \Pr\left(\bigcup_{i=1}^{L} Z_i^\tau(T_{v^*}) = 0\right)\right) \\
&\qquad \times \Pr\left(K_{l^\dagger} \geq 2 | l^\dagger \leq L\right) \Pr(E)^2 + \Pr\left(Z_L\left(T_{v^*}\right)=0\right).
\end{aligned}$$

In addition, we have

$$\Pr\left(K_{l^\dagger} \geq 2 | l^\dagger \leq L\right) = \sum_{l=1}^{L} \Pr\left(K_{l^\dagger} \geq 2, l^\dagger = l | l^\dagger \leq L\right) \tag{3}$$

$$= \sum_{l=1}^{L} \Pr\left(K_{l^\dagger} \geq 2 | l^\dagger = l\right) \Pr\left(l^\dagger = l | l^\dagger \leq L\right). \tag{4}$$

In Lemma 3, we prove that the extinction probability of each branching process from level $l^\dagger$ is upper bounded by the extinction probability ρ of the binomial infection process $B(g_{\min}, q_{\min})$. Therefore, at level $l^\dagger$, we have n_0 i.i.d one-time infection processes whose extinction probabilities are upper bounded by ρ. The probability that at least two of them survive goes asymptotically to 1 when n_0 increases. Therefore, $\forall \epsilon_1 > 0$, we have enough large n_0, such that

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

Therefore, Equation 4 becomes

$$\begin{aligned}\Pr\left(K_{l^\dagger} \geq 2 | l^\dagger \leq L\right) &\geq (1-\epsilon_1) \sum_{l=1}^{L} \Pr\left(l^\dagger = l | l^\dagger \leq L\right) \\ &= (1-\epsilon_1).\end{aligned}$$

We show in Lemma 4 that $\Pr(E) \geq 1-\epsilon_2$ given $\epsilon_2 > 0$. If n_0 and t are sufficiently large, we have

$$\Pr\left(K_{l^\dagger} \geq 2 | l^\dagger \leq L\right) \Pr(E)^2 \geq (1-\epsilon_1)\,(1-\epsilon_2)^2\,.$$

Therefore,

$$\begin{aligned}&\Pr(Z_L(T_{v^*}) = 0) + \Pr\left(K_{l^\dagger} \geq 2, l^\dagger \leq L\right) \Pr(E)^2 \\ &\quad \geq \left(1 - \Pr\left(\bigcap_{i=1}^{L} 0 < Z_i^\tau(T_{v^*}) \leq n_0\right)\right)(1-\epsilon_1)(1-\epsilon_2)^2 \\ &\qquad - \Pr\left(\bigcup_{i=1}^{L} Z_i^\tau(T_{v^*}) = 0\right) + \Pr(Z_L(T_{v^*}) = 0) \\ &\quad = \underbrace{\left(1 - \Pr\left(\bigcap_{i=1}^{L} 0 < Z_i^\tau(T_{v^*}) \leq n_0\right)\right)}_{\text{Part 1}}(1-\epsilon_1)(1-\epsilon_2)^2 \\ &\qquad + \underbrace{\Pr(Z_L(T_{v^*}) = 0) - \Pr\left(Z_L^\tau(T_{v^*}) = 0\right)}_{\text{Part 2}},\end{aligned} \tag{5}$$

where Equation 5 holds since $Z_l^\tau(T_{v^*}) = 0$ implies that $Z_L^\tau(T_{v^*}) = 0$ for $l \leq L$.

For part 1 in Equation 5, we prove in Lemma 4, given $\epsilon_3 > 0$, when τ and L are sufficiently large,

$$1 - \Pr\left(\bigcap_{i=1}^{L} 0 < Z_i^\tau(T_{v^*}) \leq n_0\right) > 1 - \epsilon_3.$$

For part 2 in Equation 5, we have

$$\lim_{\tau \to \infty} \Pr(Z_L^\tau(T_{v^*}) = 0) = \Pr(Z_L(T_{v^*}) = 0).$$

Therefore, given $\epsilon_4 > 0$, when τ is sufficiently large,

$$\Pr(Z_L(T_{v^*}) = 0) - \Pr\left(Z_L^\tau\,(T_{v^*}) = 0\right) \geq -\epsilon_4.$$

Hence, we have

$$\begin{aligned}&\Pr\left(Z_L\,(T_{v^*}) = 0\right) + \Pr\left(K_{l^\dagger} \geq 2, l^\dagger \leq L\right) \Pr(E)^2 \\ &\qquad \geq (1-\epsilon_1)\,(1-\epsilon_2)^2\,(1-\epsilon_3) - \epsilon_4.\end{aligned}$$

Now choosing $\epsilon_1 = \epsilon_2 = \epsilon_3 = \epsilon_4 = \epsilon_5/5$ for some $\epsilon_4 > 0$, we have

$$\Pr\left(Z_L\,(T_{v^*}) = 0\right) + \Pr\left(K_{l^\dagger} \geq 2, l^\dagger \leq L\right) \Pr(E)^2 \geq 1 - \epsilon_5.$$

Now let $|\mathbf{Y}|$ denote the number of infected nodes in the observation $\mathbf{Y}$. Define events $E_1 = \{Z_L = 0\}$ and $E_2 = \{K_l \geq 2 \text{for some } l \leq L\}$, and E_3 is the event that two of the survived one-time-slot infection processes have at least one observed infected node each

at their bottoms. We have

$$\begin{aligned}&\Pr(E_1||\mathbf{Y}|\geq 1)+\Pr\left(E_2\cap E_3||\mathbf{Y}|\geq 1\right)\\&\quad=\frac{1}{\Pr(|\mathbf{Y}|\geq 1)}\left(\Pr(E_1\cap\{|\mathbf{Y}|\geq 1\})\right.\\&\qquad\left.+\Pr\left(E_2\cap E_3\cap\{|\mathbf{Y}|\geq 1\}\right)\right).\end{aligned}$$

Since $E_2\cap E_3$ implies that $|\mathbf{Y}|\geq 1$, we have

$$\begin{aligned}&\Pr(E_1||\mathbf{Y}|\geq 1)+\Pr\left(E_2\cap E_3||\mathbf{Y}|\geq 1\right)\\&\qquad=\frac{1}{\Pr(|\mathbf{Y}|\geq 1)}\left(\Pr(E_1\cap\{|\mathbf{Y}|\geq 1\})+\Pr\left(E_2\cap E_3\right)\right)\\&\qquad=\frac{1}{\Pr(|\mathbf{Y}|\geq 1)}\left(\Pr(E_1)-\Pr(E_1\cap\{|\mathbf{Y}|=0\})\right.\\&\qquad\quad\left.+\Pr\left(E_2\cap E_3\right)\right)\\&\qquad\geq\frac{1}{\Pr(|\mathbf{Y}|\geq 1)}\left(\Pr(E_1)-\Pr(\{|\mathbf{Y}|=0\})+\Pr\left(E_2\cap E_3\right)\right)\\&\qquad\geq\frac{1}{\Pr(|\mathbf{Y}|\geq 1)}\left(\Pr(\{|\mathbf{Y}|\geq 1\})-\epsilon_5\right)=1-\frac{\epsilon_5}{\Pr(|\mathbf{Y}|\geq 1)}.\end{aligned}\tag{6}$$

Note that $\Pr(|\mathbf{Y}|\geq 1)$ is a positive constant since the one-time-slot infection process starting from the information source survives with non-zero probability. The theorem holds by choosing $\epsilon_5=\epsilon\Pr(|\mathbf{Y}|\geq 1)$.

Lemma 3. *The extinction probability of a one-time-slot infection process is smaller than the extinction probability of a binomial branching process* $B(g_{\min},q_{\min})$*, i.e.,* $\forall v\in\mathcal{V}$*,*

$$\rho_v<\rho.$$

Proof. As shown in Figure 8, we construct a *virtual source process* $Z_l^{(vs)}\left(T_v^{-\phi(v)}\right)$ and a *min-infection process* $Z_l^{(mi)}\left(T_v^{-\phi(v)}\right)$ as auxiliary processes over the same tree topology where $Y_v^{(vs)}$ and $Y_v^{(mi)}$ are the binary numbers indicating whether node v has been infected. Denote by $\rho_v^{(vs)}$ and $\rho_v^{(mi)}$ the extinction probabilities, respectively.

In the min-infection process, infection spreads over edges with probability $q_{\min}$. In the virtual source process, the probability that a node gets infected is

$$\Pr\left(Y_v^{(vs)}=1\right)=\Pr\left(Y_v^{(mi)}=1\right)+\Pr\left(Y_v^{(mi)}=0\right)\cdot\frac{q_{uv}-q_{\min}}{1-q_{\min}}=q_{uv},$$

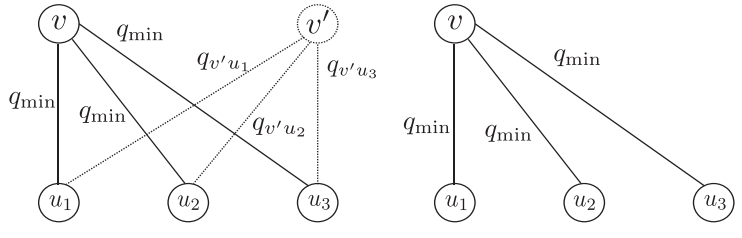

Figure 8 A pictorial description of the two auxiliary processes in Lemma 3.

i.e., for each node $u \in \mathcal{C}(v)$, v tries to infect u with probability $q_{\min}$. If v fails to infect u, a *virtual source* v' tries to infect u with probability $\frac{q_{vu}-q_{\min}}{1-q_{\min}}$. Therefore, the virtual source process has the same distribution with the one-time-slot infection process.

We now couple the min-infection process and the virtual source infection process as follows:

- If $Y_v^{(mi)} = 1$, then $Y_v^{(vs)} = 1$.
- If $Y_v^{(mi)} = 0$, then $Y_v^{(vs)} = 1$ with probability $\frac{q_{uv}-q_{\min}}{1-q_{\min}}$.

Since a node is more likely to get infected in the virtual source infection process, we obtain

$$\rho_v^{(vs)} \leq \rho_v^{(mi)}.$$

Recalling that the one-time-slot infection process has the same distribution with the virtual source branching process, we obtain $\rho_v \leq \rho_v^{(mi)}, \forall v$.

In addition, the min-infection process has more children than the binomial branching process with the same infection probability for each child. It is obvious that the binomial branching process is more likely to die out, i.e., $\rho_v^{(mi)} < \rho$.

As a summary, we prove

$$\rho_v < \rho.$$

□

Lemma 4. *Assume* $\exists \xi > 0$ *such that* $\sigma_v^\tau < 1 - \xi, \forall v \in \mathcal{V}$. *Given any* $\epsilon > 0$, *there exists a constant* L' *such that for any* $L \geq L'$,

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

Proof. Follows the same argument of Lemma 7 in [6], and by choosing

$$L' = \left\lceil \frac{\log \epsilon}{\log(1 - \xi^{n_0})} \right\rceil,$$

we obtain for any $L \geq L', \epsilon > 0$

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

□

Lemma 5. *For any* $\epsilon > 0$, *there exists a sufficiently large t such that*

$$\Pr(E) \geq 1 - \epsilon.$$

Proof. Note that the binomial branching process $B(g_{\min}, q_{\min})$ is a Galton-Watson (GW) process [12] which requires each node to have an i.i.d offspring distribution. The previous result about the instability of the Galton-Watson process in Theorem 6.2 in [12] proves that the GW process either goes to infinity or goes to 0. If the GW process survives, the number of offsprings goes to infinity as the level increases. Therefore, for a sufficiently long time, the survived binomial branching process will have a sufficiently large number of offsprings at the lowest level. Since the one-time-slot infection process always has at

least the same number of children as the binomial branching process, the survived one-time-slot infection process will have enough number of infected nodes at the lowest level as time increases. According to the unbiased property of the partial observation, after a sufficiently long time, the probability that at least one infected node in the lowest level is observed goes to 1 asymptotically, i.e.,

$$\Pr(E) \geq 1 - \epsilon.$$

□

Conclusions

In this paper, we studied the problem of detecting the information source in a heterogeneous SIR model with sparse observations. We proved that the optimal sample path estimator on an infinite tree is a node with the minimum infection eccentricity with partial observations. With a fairly general condition, we proved that the estimator is within constant distance from the actual information source with a high probability with a sparse observation. Extensive simulation results showed that our estimator outperforms other algorithms significantly.

Abbreviations
CC: closeness centrality; DMP: dynamic message passing; RI: reverse infection; SI: susceptible-infected; SIR: susceptible-infected-recovered; wCC: weighted closeness centrality; wRI: weighted reverse infection.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
KZ and LY contributed equally to this work. Both authors read and approved the final manuscript.

Authors' information
KZ received his B.E. degree in Electronics Engineering from Tsinghua University, Beijing, China, in 2010. He is currently working towards a Ph.D. degree at the School of Electrical, Computer and Energy Engineering at Arizona State University. His research interest is in social networks.
LY received his B.E. degree from Tsinghua University, Beijing, in 2001 and his M.S. and Ph.D in Electrical Engineering from the University of Illinois at Urbana-Champaign in 2003 and 2007, respectively. During Fall 2007, he worked as a postdoctoral fellow in the University of Texas at Austin. He was an assistant professor at the Department of Electrical and Computer Engineering at Iowa State University from January 2008 to August 2012. He currently is an associate professor at the School of Electrical, Computer and Energy Engineering at Arizona State University and an associate editor of the IEEE/ACM Transactions on Networking. His research interest is broadly in the area of stochastic networks, including big data and cloud computing, cyber security, P2P networks, social networks, and wireless networks. He won the Young Investigator Award from the Defense Threat Reduction Agency (DTRA) in 2009 and NSF CAREER Award in 2010. He was the Northrop Grumman Assistant Professor (formerly the Litton Industries Assistant Professor) in the Department of Electrical and Computer Engineering at Iowa State University from 2010 to 2012.

Acknowledgements
This research was supported in part by ARO grant W911NF-13-1-0279.

Received: 12 May 2014 Accepted: 14 July 2014
Published online: 15 October 2014

References

1. Shah, D, Zaman, T: Detecting sources of computer viruses in networks: theory and experiment. In: Proc. Ann. ACM SIGMETRICS Conf., pp. 203–214. ACM, New York, NY (2010)
2. Shah, D, Zaman, T: Rumors in a network: who's the culprit? IEEE Trans. Inf. Theory **57**, 5163–5181 (2011)
3. Shah, D, Zaman, T: Rumor centrality: a universal source detector. In: Proc. Ann. ACM SIGMETRICS Conf., pp. 199–210. ACM, London, England, UK (2012)
4. Luo, W, Tay, WP, Leng, M: Identifying infection sources and regions in large networks. Arxiv preprint arXiv:1204.0354 (2012)
5. Nguyen, DT, Nguyen, NP, Thai, MT: Sources of misinformation in online social networks: who to suspect? In: Military Communications Conference, 2012-MILCOM 2012, Orlando, FL, USA, 29 Oct 2012, pp. 1–6. IEEE (2012)
6. Zhu, K, Ying, L: Information source detection in the SIR model: a sample path based approach. Arxiv preprint arXiv:1206.5421 (2012)

7. Subramanian, VG, Berry, R: Spotting trendsetters: inference for network games. In: Proc. Annu. Allerton Conf. Communication, Control and Computing, Monticello, IL, USA, 1 Oct 2012, (2012)
8. Milling, C, Caramanis, C, Mannor, S, Shakkottai, S: Network forensics: random infection vs spreading epidemic. In: Proc. Ann. ACM SIGMETRICS Conf., London, England, UK, 11 Jun 2012, pp. 223–234. (2012)
9. Shakarian, P, Subrahmanian, VS, Sapino, ML: GAPs: geospatial abduction problems. ACM Trans. Intell. Syst. Technol. **3**(1), 1–27 (2011)
10. Shakarian, P, Subrahmanian, VS: Geospatial Abduction: Principles and Practice. Springer, New York (2011)
11. Lokhov, AY, Mezard, M, Ohta, H, Zdeborova, L: Inferring the origin of an epidemy with dynamic message-passing algorithm. arXiv preprint arXiv:1303.5315 (2013)
12. Harris, TE: The Theory of Branching Processes. Dover Pubns, New York (1963)

doi:10.1186/s40649-014-0003-2
Cite this article as: Zhu and Ying: **A robust information source estimator with sparse observations.** *Computational Social Networks* 2014 **1**:3.