A fast Monte Carlo algorithm for source localization on graphs

Ameya Agaskar[a,b] and Yue M. Lu[a]
[a]Harvard School of Engineering and Applied Sciences, 33 Oxford Street, Cambridge, MA 02138
[b]MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA 02420

ABSTRACT

Epidemic models on networks have long been studied by biologists and social sciences to determine the steady state levels of an infection on a network. Recently, however, several authors have begun considering the more difficult problem of estimating the source of an infection given information about its behavior some time after the initial infection. In this paper, we describe a technique to estimate the source of an infection on a general graph based on observations from a small set of observers during a fixed time window at some unknown time after the initial infection. We describe an alternate representation for the susceptible-infected (SI) infection model based on geodesic distances on a randomly-weighted version of the graph; this representation allows us to exploit fast algorithms to compute geodesic distances to estimate the marginal distributions for each observer and compute a pseudo-likelihood function that is maximized to find the source.

1. INTRODUCTION

Epidemic models have long been studied by biologists and social scientists to study the spread of contagion on networks of varying scales and geometries.[1] These models range from the early, very simple models on (implicit) complete graphs,[2] to very recent models on sophisticated random graphs.[3] This standard line of work has been mainly focused on the problem of determining the steady-state behavior of the model: will an epidemic die out, or will it remain active, with some fraction of the population always infected?

Recently, however, a new line of work has emerged to address the problem of detecting the source of an epidemic.[4–11] In many ways, this is a much harder problem, which explains why it has only recently begun to receive attention. To determine the steady-state behavior, it is often enough to solve a system of differential equations. However, the source localization problem requires either high computational complexity to find near-optimal solutions, or simplified heuristics to achieve suboptimal performance. Often, algorithms are designed to work on trees, whose properties can greatly simplify the problem, and extended in an *ad hoc* manner to general graphs.[4–6]

In this work, we describe an algorithm that operates on a sequence of observations made on a small subset of the nodes; our observation window begins at some unknown time after the initial infection. This algorithm is designed to function on general graphs, not just trees, and uses an initial Monte Carlo stage to determine a pseudo-likelihood function for sources on a particular graph.

1.1 Related Work

Serious investigation of the epidemic source localization problem began with Shah and Zaman[4] in 2011. They considered a continuous-time SI epidemic model on a graph, and sought to find the most likely source given the full set of infected nodes at some time. The infected nodes under this model form a connected subgraph, and they developed a metric called the *rumor centrality* for each node in the infection subgraph that served as a proxy for the likelihood of a particular node being the source. They provided an efficient message passing algorithm for calculating the rumor centrality on a tree, and developed the breadth-first search (BFS) heuristic to approximate the rumor centrality on more general graphs. The epidemic source was chosen to be the member of the infected subgraph with the highest rumor centrality.

Email: {aagaskar, yuelu}@seas.harvard.edu. The Lincoln Laboratory portion of this work is sponsored by the Department of the Air Force under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Wavelets and Sparsity XV, edited by Dimitri Van De Ville, Vivek K. Goyal, Manos Papadakis,
Proc. of SPIE Vol. 8858, 88581N · © 2013 SPIE · CCC code: 0277-786X/13/$18
doi: 10.1117/12.2023039

Others soon considered spectral techniques for source localization.[7,10] Given the infected subgraph at some time, information about the likely source of infection can be gained from the eigendecomposition of the adjacency or Laplacian matrices of the subgraph. Luo and Tay[8] considered the case of multiple infection sources and developed an algorithm to detect these sources and determine the original infection source associated with each infected vertex. Seo *et al.* considered the case where only a subset of the vertices are monitored and developed four metrics most likely to be associated with the source; each successive metric breaks ties in the previous metrics.[11] Zhu and Ying[9] first considered the susceptible-infected-recovered (SIR) model in which nodes can recover from infection after some period of time and are then immune. They developed a message passing algorithm called the *reverse infection algorithm* that choose a source. More recently, Lokhov *et al.* have provided the statistical physics community's answer to the problem: a message passing algorithm on trees that computes the exact marginal probability for a vertex to be infected at some time given a source.[5] This facilitates the computation of a mean-field approximation to the likelihood of an observation. Though not exact on general graphs, it seems to give reasonable results. Meanwhile, Pinto *et al.* considered the problem of locating the source given the exact time of infection for a set of observers. Their estimator is linear on trees, and is extended to general graphs using the BFS heuristic.[6]

1.2 Contribution

Our approach is the first to use multiple snapshots in a fixed, small time window for a sparse set of observers to estimate the source. Since the infection process we consider is Markov, if we could measure all the vertices, a snapshot of the state at a single time would be sufficient and any more data collection would be superfluous. However, because we only observe a small subset of the vertices, we can gain additional information by observing them over a period of time. Assuming that at least one observer sees a transition from susceptible to infected, our observation can be transformed into bounds on the difference between that observer's transition and all others. We introduce an alternate representation for the infection process in terms of the infection times for each vertex. The alternate representation allows us to quickly sample infection times conditioned on each source. This allows us to estimate the marginals of the observed relative infection times and compute a pseudolikelihood that is the product of those marginals, conditioned on each possible source.

The rest of the paper is organized as follows. In Section 2, we formalize the model and develop an alternate representation for the model that will be useful for developing a source localization algorithm. In Section 3, we present a Monte Carlo technique for computing parameters of the observations and describe how to use them for the source localization problem. In Section 4 we discuss the performance of the algorithm in numerical experiments. Finally, in Section 5 we conclude with some remarks.

2. EPIDEMIC MODEL

2.1 SI Model

We start with a graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of N vertices and $\mathcal{E}$ is the set of M edges, each of which is an unordered pair of vertices. If $\{v_i, v_j\} \in \mathcal{E}$, then v_i and v_j are in contact and capable of transmitting the infection to each other. We will use the notation $i \sim j$ to mean that v_i and v_j are neighbors on the graph. We consider a discrete-time version of the susceptible-infected (SI) model on G. At any time, each vertex is in one of two states, susceptible or infected. To model this mathematically, we will give every susceptible vertex the value 0 and every infected vertex the value 1. The epidemic model, illustrated in Figure 1, is a random process $\mathcal{X} = (\boldsymbol{x}(t))_{t=0}^{\infty}$ whose sample paths are sequences of binary vectors in $\{0,1\}^N$. We will assume that initially, a vertex $s \in \mathcal{V}$ is chosen uniformly at random to be the infection source, and that the initial state vector $\boldsymbol{x}(0) = \boldsymbol{e}_s$ is 1 at s and 0 elsewhere.

At each time $t \in \{0, 1, \ldots\}$, any vertex that is infected remains so at time $t+1$, while any vertex that is susceptible becomes infected at time $t+1$ if it receives an infection signal from one of its neighbors on G. All infection signals are independent, and at each time step an infected vertex v_j sends an infection signal to susceptible neighbor v_i with probability λ_{ji}. Although the graph is undirected, we allow λ_{ji} and λ_{ij} to differ. The infection process is Markov; the vector $\boldsymbol{x}(t+1)$ depends on the previous states only through $\boldsymbol{x}(t)$. To

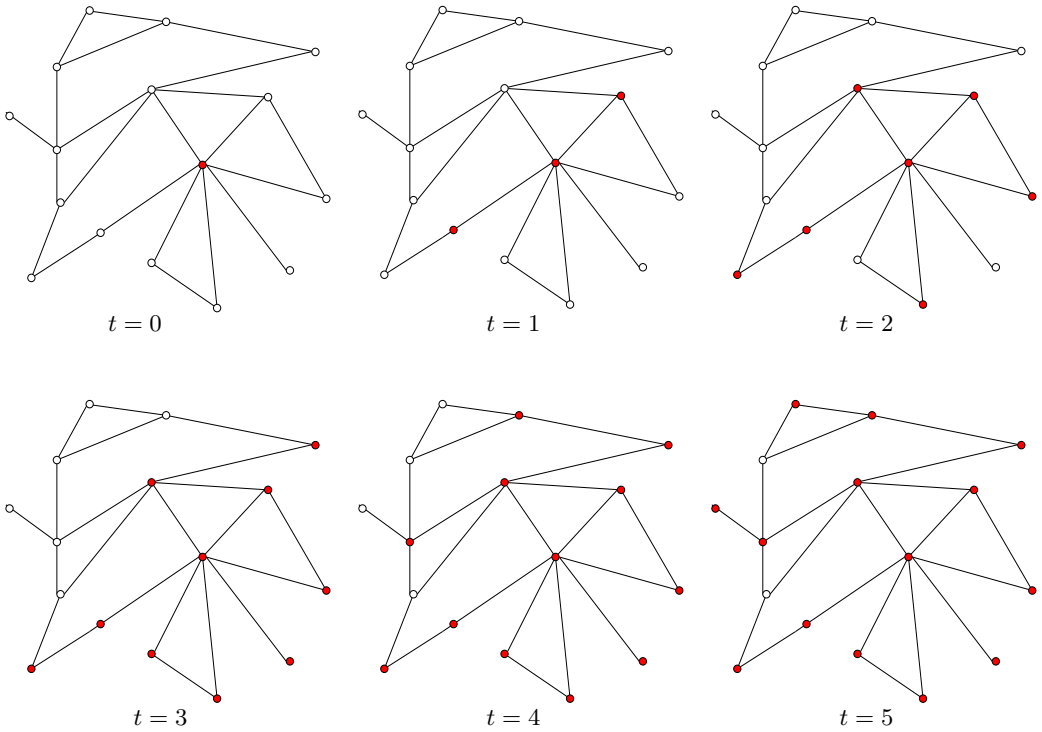

Figure 1. At time $t = 0$, a single vertex is infected (filled red circle) and the remaining ones are not. At each time step, an infected vertex has some probability of infecting each of its neighbors.

simplify the notation, we will define $\mathcal{I}_i(t) = \{j : v_j \sim v_i, x_j(t) = 1\}$ to be the set of neighbors of i that are infected at time t. Then the transition rule can be written as

$$\Pr\{x_i(t+1) = 1 | \boldsymbol{x}(t)\} = \begin{cases} 1 & \text{if } x_i(t) = 1 \\ 1 - \prod_{j \in \mathcal{I}_i(t)} (1 - \lambda_{ji}) & \text{if } x_i(t) = 0, \end{cases} \tag{1}$$

where the elements of $\boldsymbol{x}(t+1)$ are independent conditioned on $\boldsymbol{x}(t)$. $\mathcal{X}$ can be seen as a Markov chain with $2^N - 1$ states (if we ignore the all-susceptible state that is inaccessible from any other state); the distribution of $\boldsymbol{x}(t+1)$ conditioned on its history depends only on $\boldsymbol{x}(t)$, its state at time t. The all-infected state is an absorbing state, and will be reached in finite time with probability 1. The source localization problem we are interested in amounts to inferring the initial state of the Markov chain given partial measurements.

2.2 Alternate Representation

We can use an alternate representation of the process $\mathcal{X}$. Let $\tau_i = \min_t x_i(t) = 1$ be the time at which vertex i first becomes infected. Then $(\tau_1, \ldots, \tau_N)$ contains all of the information of $\mathcal{X}$, since we can reconstruct the sequence $(\boldsymbol{x}(0), \boldsymbol{x}(1), \ldots)$ from it. Using a randomly weighted version of the graph G, we can directly sample $(\tau_1, \ldots, \tau_N)$ given the source s.

Let H be the random, directed weighted graph $H = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where $\mathcal{V}$ and $\mathcal{E}$ are the vertex and edge sets for the graph G, and $\mathcal{W}$ consists of random weights w_{ij} and w_{ji} associated with each edge $\{i, j\}$ in $\mathcal{E}_G$. H is directed because the weights for two different directions of an edge in $\mathcal{E}$ can be different. Each weight w_{ij} is a geometric random variable taking values in $\{1, 2, \ldots\}$ with parameter λ_{ij}. If v_i became infected at some time, and v_j were connected to no other vertices, then w_{ij} represents the number of time steps it would take for v_j to become infected.

We define the following geodesic quasimetric on H:

$$d_H(v_i, v_j) = \min_{P \in \mathcal{P}_{ij}} \sum_{k=1}^{\text{len}(P)-1} w_{p_k, p_{k+1}}, \tag{2}$$

where $\mathcal{P}_{ij}$ is the set of paths on the graph from v_i to v_j, and $P = (p_1 = v_i, p_2, \ldots, p_{\ell(P)} = v_j)$. $d_H(v_i, v_j)$ is the length of the shortest possible path from v_i to v_j, where traversing an edge from v_k to v_l costs w_{kl}. It is a quasimetric and not a metric because it is not symmetric, due to the different weights on different directions of each edge. Now define $\mathcal{Y} = (\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots)$ as follows:

$$s \sim \text{Unif}(1, \ldots, N); \quad y_i(t) = \begin{cases} 0 & \text{if } t < d_H(s,i) \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

$\mathcal{Y}$ starts with only the source s, chosen uniformly at random as in $\mathcal{X}$, infected; the infection time of each remaining vertex determined by its distance from the source on H. We have the following proposition:

PROPOSITION 1. $\mathcal{Y} \stackrel{d}{=} \mathcal{X}$, i.e. $\Pr\{\boldsymbol{x}(0), \boldsymbol{x}(1), \ldots, \boldsymbol{x}(t)\} = \Pr\{\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(t)\}$ *for every* t.

Proof. First, it is clear that $\boldsymbol{y}(0) \stackrel{\text{d}}{=} \boldsymbol{x}(0)$ since each is a standard basis vector chosen uniformly at random. It remains to be shown that for each t, the distribution of $\boldsymbol{y}(t+1)$ given the history $(\boldsymbol{y}(t), \ldots, \boldsymbol{y}(0))$ is identical to that for $\boldsymbol{x}(t+1)$ given $(\boldsymbol{x}(t), \ldots, \boldsymbol{x}(0))$. Along the way, we will show that $\mathcal{Y}$ is a Markov process.

First, we note that if $y_i(t) = 1$, then $\Pr\{y_i(t+1) = 1 | \boldsymbol{y}(t), \ldots, \boldsymbol{y}(0)\} = 1$, regardless of any of the other current or previous values in $\boldsymbol{y}$, since if $t \geq d_H(s,i)$, then $t+1 \geq d_H(s,i)$. On the other hand, if $y_i(t) = 0$, then

$$\begin{aligned} &\Pr\{y_i(t+1) = 1 | \boldsymbol{y}(t), \ldots, \boldsymbol{y}(0), \text{with } y_i(t) = 0\} \\ &\quad = 1 - \Pr\{d_H(s,i) > t+1 | d_H(s,i) > t, d_H(s,j) > t \text{ for all } y_j(t) = 0, d_H(s,j) = k_j \text{ for all } y_j(t) = 1\}, \end{aligned} \tag{4}$$

by the definition of $\mathcal{Y}$, where k_j is the time at which vertex j switched from 0 to 1 in the sequence $\boldsymbol{y}(0), \ldots, \boldsymbol{y}(t)$. From the definition of geodesic distance, we can rewrite $d_H(s,i) = \min_{j \sim i}(d_H(s,j) + w_{ji})$. Since $d_H(s,j) + w_{ji} > t+1$ automatically if $y_j(t) = 0$, we can consider in the minimum only those vertices that are infected at time t. So we get

$$\begin{aligned} &\Pr\{y_i(t+1) = 1 | \boldsymbol{y}(t), \ldots, \boldsymbol{y}(0), \text{with } y_i(t) = 0\} \\ &\quad = 1 - \Pr\{w_{ji} + d_H(s,j) > t+1 \text{ for all } j \in \mathcal{I}_i(t) | w_{ji} + d_H(s,j) > t \, \forall j \in \mathcal{I}_i(t), d_H(s,j) = k_j \text{ for all } j \in \mathcal{I}_i(t)\}, \end{aligned} \tag{5}$$

$$= 1 - \Pr\{w_{ji} > t+1-k_j \text{ for all } j \in \mathcal{I}_i(t) | w_{ji} > t - k_j, d_H(s,j) = k_j \text{ for all } j \in \mathcal{I}_i(t)\}. \tag{6}$$

Since the w_{ji} are geometric random variables, the conditional probability in (6) does not depend on the k_j's. Combined with the fact that the w_{ji}'s are independent, we arrive at

$$\Pr\{y_i(t+1) = 1 | \boldsymbol{y}(t), \ldots, \boldsymbol{y}(0), \text{with } y_i(t) = 0\} = \Pr\{y_i(t+1) = 1 | \boldsymbol{y}(t)\} = 1 - \prod_{j \in \mathcal{I}_i(t)} (1 - \lambda_{ji}),$$

the same as the Markov transition rule for $\mathcal{X}$. Thus $\mathcal{Y} \stackrel{\text{d}}{=} \mathcal{X}$. □

Thus, to sample $(\tau_1, \ldots, \tau_N)$ given a source s, we simply draw random weights w_{ij} and w_{ji} for each edge $\{i,j\} \in \mathcal{E}$, and let $\tau_i = d_H(s,i)$. The distances can be computed using the Dijkstra shortest paths algorithm.

3. SOURCE LOCALIZATION

We now formally define the source localization problem. Suppose we can observe the states of L vertices $\mathcal{O} \subset \mathcal{V}$ during an observation window $\{t_0, t_0+1, \ldots, t_0+T-1\}$. Without loss of generality, we assume $\mathcal{O} = \{1, 2, \ldots, L\}$. We do not know t_0, so we have no knowledge of how long ago the infection began spreading from the source. Given the observations and knowledge of the graph and the infection parameters λ_{ji} we would like to estimate the source. If all vertices are observed, so that $L = N$, then $\boldsymbol{x}(t_0)$ is sufficient for estimation and we do not need to use the whole sequence of observations, since the infection sequence before t_0 is independent of the infection sequence after t_0 given the realization at t_0 thanks to the Markov property. This is not true if $L < N$; in that case our observation sequence is a hidden Markov process (HMP) and does not satisfy the Markov property.

Observations of a single vertex during our window may fall into three categories: we may observe its transition from susceptible to infected during our observation window, it may be susceptible during the entire window, or

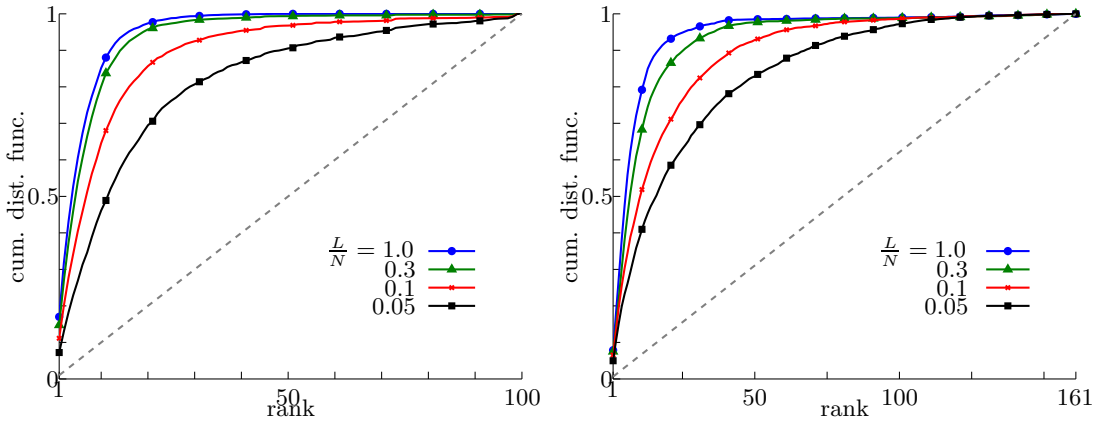

Figure 2. On the left, results for a random geometric graph with 100 vertices; on the right, a 4-regular tree truncated 5 levels from a root vertex. For each graph, a series of trials was run in which a random source was chosen and a realization of $\mathcal{X}$ was generated. Our algorithm was given the realization from $t = 4$ to $t = 23$, for twenty observations, and computed pseudolikelihoods for each potential source. For the random geometric graph, the infection rate was 0.1 for every edge; for the tree it was 0.4. The CDF of the rank of the true source in the ordered list of pseudolikelihoods is shown here. Curves closer to the top right indicate better performance. On both graphs, the CDFs presented are for sampling factors of 1, 0.3, 0.1, and 0.05. It is evident that a lot of decimation is needed before the performance degrades appreciably. A lower bound (uniform source estimator) is given by the dotted line.

it may be infected during the entire window. This defines a partition of the set of observers $\mathcal{O} = \mathcal{O}_T \cup \mathcal{O}_S \cup \mathcal{O}_I$, where $\mathcal{O}_T$, $\mathcal{O}_S$, and $\mathcal{O}_I$ represent these categories, respectively. Suppose that $\mathcal{O}_T$ is nonempty, and without loss of generality, that $1 \in \mathcal{O}_T$. For each $i \in \mathcal{O}_T$, we let $m_i \in \{1, 2, \ldots, T-1\}$ be the index of the first observation for which the vertex transitions to the infected state (where our first observation has index 0). Because we do not have an absolute time reference, we have knowledge only about the relative infection times $\tau_i - \tau_1$ for $i \in \mathcal{O} \backslash \{i\}$.

Then we define a log-pseudolikelihood function ℓ by assuming that the relative delays of the observed nodes are independent:

$$\ell(s) = \sum_{i \in \mathcal{O}_T \backslash \{1\}} \log \Pr\{\tau_i - \tau_1 = m_i - m_1 | s\} + \sum_{i \in \mathcal{O}_I} \log \Pr\{\tau_i - \tau_1 \leq -m_1 | s\} + \sum_{i \in \mathcal{O}_S} \log \Pr\{\tau_i - \tau_1 \geq T - m_1 | s\}. \quad (7)$$

To estimate the marginals for the relative times, we will use Monte Carlo simulations to approximate the means and variances of the τ_i. We will use the procedure described in Section 2.2 to sample $(\tau_1, \ldots, \tau_L)$ for several iterations for each source s, then set $\mu_i(s)$ and $\sigma_i^2(s)$ to be the mean and variance of those samples. By approximating the infection times as Gaussian, the pseudo-likelihood function (7) becomes

$$\begin{aligned} \ell(s) = -\frac{1}{2} \sum_{i \in \mathcal{O}_T \backslash \{1\}} \left(\log 2\pi \left(\sigma_i^2(s) + \sigma_1^2(s) \right) + \frac{(m_i - m_1 - \mu_i(s) + \mu_1(s))^2}{\sigma_i^2(s) + \sigma_1^2(s)} \right) \\ + \sum_{i \in \mathcal{O}_I} \log \Phi \left(\frac{-m_1 - \mu_i + \mu_1}{\sqrt{\sigma_i^2(s) + \sigma_1^2(s)}} \right) + \sum_{i \in \mathcal{O}_S} \log \left(1 - \Phi \left(\frac{T - m_1 - \tau_i + \tau_1}{\sqrt{\sigma_i^2(s) + \sigma_1^2(s)}} \right) \right), \end{aligned} \quad (8)$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution.

The computation is dominated by the Dijkstra shortest path algorithm in the sampling procedure for the infection times. By the central limit theorem, $O(1/\epsilon^2)$ samples are needed to achieve an error $o(\epsilon)$. The Dijkstra algorithm requires $O(M + N \log N)$ time to find the lengths of the shortest path to a single vertex from every other vertex. Thus, it takes $O(L(M + N \log N)/\epsilon^2)$ time to estimate $\mu_i(s)$ and $\sigma_i^2(s)$ for every $i \in \mathcal{O}$ and $s \in \mathcal{V}$. In typical applications, graphs tend to be sparse, so we may take $M = o(N \log N)$, giving us a computational complexity of $O(LN \log(N)/\epsilon^2)$.

4. NUMERICAL RESULTS

To analyze the performance of this algorithm, we simulated the epidemic process on several graphs and computed the pseudolikelihoods for various observer fractions. The choice of performance metric is not entirely obvious: the error probability is too pessimistic, since choosing, say, a neighbor of the true source is much better than choosing a source at the other end of the graph; yet average geodesic (or some other) distance is too optimistic, since many realistic graphs have small diameter. Most authors have instead used the following metric: the vertices are sorted by the value of the function to be maximized, and the rank of the true source is considered.

In Figure 2, we show the cumulative distribution functions (CDFs) of the source's rank in our various experiments. We considered two graphs. First, a random geometric graph with 100 nodes randomly placed on the unit square and two nodes connected if they are within a radius of 0.3 of one another; on this graph, an infection rate $\lambda = 0.1$ is used for every edge. Second, a 4-regular tree truncated after 5 levels; on this graph, an infection rate $\lambda = 0.4$ is used for every edge. In both cases, observations start at $t = 4$ and continue for 20 time steps. And in both cases, CDFs are computed when the fraction of observers is 1, 0.3, 0.1, and 0.05.

As expected, the performance of the estimator degrades as fewer vertices are observed. Interestingly, however, the degradation begins very slowly as the observer fraction drops from 1 down to 0.3, then speeds up as the fraction continues to fall. Note that on the 100-vertex random geometric graph, and observer fraction of 0.05 means that only 5 random vertices are observed. Meanwhile, it is interesting to see that for both graphs, observing only 30% of the vertices gives almost as good results as observing them all.

5. CONCLUSION

We introduced a technique to estimate the source of an infection on a graph given observations of the states of a subset of vertices during a time window some time after the infection begins. The technique uses a Monte Carlo algorithm that uses a representation of the SI model based on geodesic distances on a randomly weighted version of the graph; these distances can be computed efficiently using the Dijkstra algorithm. Numerical results show that the fraction of observed vertices can be reduced significantly without performance degradation.

REFERENCES

1. A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks*, Cambridge University Press, 2008.
2. W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London. Series A* **115**, pp. 700–721, Aug. 1927.
3. E. Volz, "SIR dynamics in random networks with heterogeneous connectivity," *Journal of Mathematical Biology* **56**, pp. 293–310, Mar. 2008.
4. D. Shah and T. Zaman, "Rumors in a network: Who's the culprit?," *IEEE Transactions on Information Theory* **57**(8), pp. 5163–5181, 2011.
5. A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, "Inferring the origin of an epidemy with dynamic message-passing algorithm," *arXiv:1303.5315* , Mar. 2013.
6. P. C. Pinto, P. Thiran, and M. Vetterli, "Locating the source of diffusion in large-scale networks," *Physical Review Letters* **109**, p. 068702, Aug. 2012.
7. V. Fioriti and M. Chinnici, "Predicting the sources of an outbreak with a spectral technique," *arXiv preprint arXiv:1211.2333* , 2012.
8. W. Luo and W. P. Tay, "Identifying multiple infection sources in a network," in *Proc. Asilomar*, pp. 1483–1489, 2012.
9. K. Zhu and L. Ying, "Information source detection in the SIR model: A sample path based approach," *arXiv:1206.5421* , June 2012.
10. B. Prakash, J. Vreeken, and C. Faloutsos, "Spotting culprits in epidemics: How many and which ones?," in *2012 IEEE 12th International Conference on Data Mining (ICDM)*, pp. 11–20, 2012.
11. E. Seo, P. Mohapatra, and T. Abdelzaher, "Identifying rumors and their sources in social networks," in *Proc. SPIE 8389, Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR III*, May 2012.