MCMC Approaches to Rumor Source Inference using Pairwise Information

Anand Kalvit
Dept. of Electrical Engineering,
IIT Bombay
anandiitb12@gmail.com

Vivek S. Borkar
Dept. of Electrical Engineering,
IIT Bombay
borkar.vs@gmail.com

Nikhil Karamchandani
Dept. of Electrical Engineering,
IIT Bombay
nikhilk@ee.iitb.ac.in

ABSTRACT

In this work, we examine the problem of rumor source inference on a network whose topology is known, given infected nodes and pairwise information in the form of pairwise partial orders on the set of nodes of the underlying graph based on the order in which they were infected. We analyze the Maximum Likelihood Estimator (MLE) of the rumor source, assumed unique, and compare it with other estimators popular in literature, e.g., rumor center, distance center and Jordan center. We propose an approximation to the MLE and a class of estimators based on this approximation that is agnostic to the underlying rumor model. We also propose MCMC algorithms to implement them. Further, we assess the robustness of the proposed estimators to different graph topologies via extensive simulations on the Erdős-Rényi and Barabási-Albert models.

CCS CONCEPTS

•**Computing methodologies** →**Network science;** *Modeling methodologies;* •**Networks** →*Network simulations;*

KEYWORDS

source detection, inference on networks, MCMC, epidemic models

ACM Reference format:
Anand Kalvit, Vivek S. Borkar, and Nikhil Karamchandani. 2017. MCMC Approaches to Rumor Source Inference using Pairwise Information. In *Proceedings of 11th EAI International Conference on Performance Evaluation Methodologies and Tools, Venice, Italy, December 5–7, 2017 (VALUETOOLS 2017),* 8 pages.
DOI: 10.1145/3150928.3150936

1 INTRODUCTION

Locating the source of a branching process from its observed spread on a network is a very relevant contemporary problem that is receiving a lot of attention. From locating faults in electricity grids to triangulating locations of contagions in networks, its applications are widespread. A particularly interesting case is when the source of such a process is unique. The source node infects its neighbor(s) who in turn infect their neighbor(s) and so on. The problem is to infer the source given a snapshot of infected nodes. Distance center and Jordan center are two estimators popular in literature that use such information alone to estimate the source node. These depend purely on the topological properties of the infected network and are agnostic to the underlying stochastic model of rumor spread. There is also a wide body of literature that uses information about the underlying rumor spread model in addition to such snapshot information (set of infected nodes at time t) to estimate the source (see [6, 8–10, 12, 13] etc). In this work, we investigate and propose a class of rumor source estimators that is agnostic to the underlying stochastic rumor model and robust to different graph topologies, assumimg additional availability of information in the form of pairwise partial orders on the set of nodes of the underlying graph based on the order in which they were infected. This scenario is typical to ranking problems. The rumor source inference problem has a natural interpretation as a raking problem and vice-versa. We however refrain from elaborating on the connection between the two in this paper. In our knowledge, our work is the first systematic study of rumor source inference using pairwise information after [7] and follows in spirit the ideas introduced in [7]. We propose a class of model free approximations to the Maximum Likelihood Estimator (MLE) of the rumor source that is agnostic to the underlying rumor model and propose MCMC algorithms to implement them. The novelty in our approach lies in the generality of our MCMC setup which can be extended to the case of multiple rumor sources and also, to the problem of rank aggregation from pairwise comparisons. We however do not pursue these extensions in this paper.

Our model is as follows. Let $G(V_0, E_0)$ denote the graph underlying the network where V_0 := the set of nodes connected by bidirectional edges in E_0. A unique source node v^* on $G(V_0, E_0)$ starts a rumor at time $t = 0$ which spreads along the edges E_0 of the graph. An infected node can spread the rumor to its hitherto uninfected neighbors and the propagation times along the various edges are random and unknown. A node once infected stays as such thereafter, i.e., there is no recovery. After sufficient time has elapsed, we observe a set of infected vertices V. Let $G(V, E)$ denote the subgraph induced by V on $G(V_0, E_0)$. We call $G(V, E)$ the rumor infected graph. We define a pairwise partial order on V as $\{(a, b) : a \prec b \mid a, b \in V\}$; where $a \prec b$ indicates that a was infected before b. The pairwise information is modeled by the matrix $I_{|I|\times 2}$ where $|I|$ is the number of available pairwise comparisons and the i-th row of I is (a, b) if $a \prec b$. The problem is to estimate the root of this order given $G(V, E)$ and $I_{|I|\times 2}$. It is easy to see that if $|V| = n$, then there can be at most $\binom{n}{2}$ pairwise comparisons available. If all $\binom{n}{2}$ comparisons are available, identifying the rumor source is trivial. We restrict our attention to the interesting case when only a fraction $f < 1$ of the $\binom{n}{2}$ possible pairwise comparisons is available.

Recall that an arborescence is a directed tree with at most one outgoing edge for each node and maximal w.r.t. this property. It

VALUETOOLS 2017, Venice, Italy

DOI: 10.1145/3150928.3150936

perforce has a unique node with no outgoing edge, called its root. A *rumor tree* or a *spreading pattern* on $G(V,E)$ rooted at $v \in V$ is defined as an arborescence of $G(V,E)$ rooted at v with the directions of all edges reversed. A *rumor permutation rooted at* v or *starting from* v on $G(V,E)$ is defined as an ordering (permutation) of the nodes in V with v as the starting element. We also define the notion of *compatibility* for rumor trees and rumor permutations. Let $[a\ b]$ denote any row of I. A rumor tree is said to be compatible with this row if there does not exist any directed path from b to a on it, and with I if it is compatible with all rows of I. A rumor permutation is said to be compatible with $[a\ b]$ if a features before b in the permutation and compatible with I if it is compatible with all rows of I. It is compatible with $G(V,E)$ if it respects the graph structure of $G(V,E)$. Let $\sigma = \{v_1, v_2, ..., v_{|V|}\}$ be a rumor permutation on $G(V,E)$ where $\{v_i\}_{i=1,...,|V|} \in V$ and v_1 is the root of the permutation. If for each k-length ($k = \{1,2,...,|V|-1\}$) subsequence $\sigma_k = \{v_1, ...v_k\}$ of σ, v_{k+1} is a direct neighbor of a node in σ_k, then σ is said to be compatible with $G(V,E)$.

We will use the following notation throughout:

1. $G(V,E)$: A connected graph on the set of vertices V with E as a bidirectional edge set.
2. $V(G)$: Set of vertices of G.
3. $E(G)$: Set of edges of G.
4. $E_1(V)$: Set of edges with only one end vertex in V.
5. $E_2(V)$: Set of edges with both end vertices in V.
6. $\mathbb{P}(G, I|v^* = v)$: Probability of observing the rumor infected graph G and pairwise information I, given that the rumor spread started from $v \in V(G)$. All other conditional probabilities $\mathbb{P}(\mathbf{y}|\mathbf{x})$ are to be interpreted in similar spirit, treating $\mathbf{x}$ and $\mathbf{y}$ as random variables.
7. $S_{\{V(G),v\}}$: Set of rumor permutations on $V(G)$ starting from $v \in V(G)$.
8. $\Omega(G,v)$: Set of rumor permutations on $V(G)$ starting from $v \in V(G)$ that are compatible with the graph topology G $\left(\Omega(G,v) \subset S_{\{V(G),v\}}\right)$.
9. $\Omega(G,I,v)$: Set of rumor permutations starting from $v \in V(G)$ that are compatible with G and I both $\left(\Omega(G,I,v) \subset \Omega(G,v) \subset S_{\{V(G),v\}}\right)$.
10. $\mathbb{T}(G,v)$: Set of rumor trees of G rooted at $v \in V(G)$.
11. $\mathbb{T}(G,I,v)$: Set of rumor trees of G rooted at $v \in V(G)$ that are compatible with I $(\mathbb{T}(G,I,v) \subset \mathbb{T}(G,v))$.

The rest of the paper is organized as follows. In section 2, we analyze the Maximum Likelihood Estimator (MLE) and investigate different approximations leading to other estimators popular in literature. We also propose an approximation to the MLE and a class of estimators based on this approximation. In sections 3,4, we propose algorithms for implementing these estimators. In section 5, we discuss a graph sparsification technique to speed up our algorithms. Finally, in section 6, we present the results of simulations on the Erdős-Rényi and Barabási-Albert models.

2 DIFFERENT ESTIMATORS

Given the infected graph $G(V,E)$ and the additional pairwise information I, our goal in this paper is to design estimators for recovering the rumor source v^*. We assume that a priori each node is equally likely to be the source and hence the best estimator in terms of the estimation error probability is the Maximum Likelihood Estimator (MLE) given by

$$\begin{aligned} v_{ML} &\in arg \max_{v \in V(G)} \mathbb{P}(G, I|v) \\ &= arg \max_{v \in V(G)} \sum_{\sigma \in S_{\{V(G),v\}}} \mathbb{P}(G, I, \sigma|v) \\ &= arg \max_{v \in V(G)} \sum_{\sigma \in S_{\{V(G),v\}}} \mathbb{P}(G, I|\sigma)\mathbb{P}(\sigma|v) \\ &= arg \max_{v \in V(G)} \sum_{\sigma \in \Omega(G,I,v)} \mathbb{P}(\sigma|v). \end{aligned} \tag{1}$$

In [8], the authors proposed an estimator specific to the Susceptible Infected (SI) model of rumor spread with the propagation time along the edges of the graph assumed to be IID exponential random variables. Their work, however, does not assume availability of any pairwise information. It can be observed from equation (1) that the MLE in absence of pairwise information is $v_{ML} \in arg \max_{v \in V(G)} \sum_{\sigma \in \Omega(G,v)} \mathbb{P}(\sigma|v)$. It was shown in [8] that the MLE in this case is simply the node with the maximum value of $R(v,G) := |\Omega(G,v)|$, called the *rumor centrality* of v in $G(V,E)$. The node with the maximum rumor centrality is called the *rumor center* of the network. The authors in [8] also proposed an $O(|V(G)|)$ message passing algorithm to evaluate $R(v,G)$ on a tree graph G. However, the MLE is computationally infeasible on arbitrary graphs. For general graphs, [8] proposed an approximate estimator that uses a Breadth First Search (BFS) heuristic whereby only the BFS rumor propagation trees are considered. The reasoning behind this heuristic is that an SI infection is likely to spread out from the source in a BFS manner. The rumor centrality estimator in [8] for general graphs is given by:

$$\hat{v} \in arg \max_{v \in V(G)} R(v, T_{bfs}(v)) \tag{2}$$

where $R(v, T_{bfs}(v))$ is the rumor centrality of v with respect to a BFS tree $T_{bfs}(v)$ of G rooted at v.

Unlike [8], we will propose an approximation to the ML estimator which scores each node based on the number of compatible rumor propagation trees rooted at the node. Note that the ML estimator from equation (1) can alternately be written as:

$$\begin{aligned} v_{ML} &\in arg \max_{v \in V(G)} \mathbb{P}(G, I|v) \\ &= arg \max_{v \in V(G)} \sum_{\tau \in \mathbb{T}(G,I,v)} \mathbb{P}(G, I, \tau|v) \\ &= arg \max_{v \in V(G)} \sum_{\tau \in \mathbb{T}(G,I,v)} \sum_{\sigma \in \Omega(\tau,v)} \mathbb{P}(G, I|\sigma)\mathbb{P}(\sigma|\tau)\mathbb{P}(\tau|v) \\ &= arg \max_{v \in V(G)} \sum_{\tau \in \mathbb{T}(G,I,v)} \left(\sum_{\sigma \in \Omega(\tau,I,v)} \mathbb{P}(\sigma|\tau) \right) \mathbb{P}(\tau|v). \end{aligned} \tag{3}$$

Our approximation to the MLE is based on the following assumption in equation (3):

$$\sum_{\sigma \in \Omega(\tau,I,v)} \mathbb{P}(\sigma|\tau) \approx 1 \quad \forall\ \tau \in \mathbb{T}(G,I,v). \tag{4}$$

This would mean that on a rumor tree τ that is compatible with the pairwise information I, *almost all* permissible rumor permutations

also comply with I. This simplifies the estimator to:

$$\hat{v} \in arg \max_{v \in V(G)} \sum_{\tau \in \mathbb{T}(G,I,v)} \mathbb{P}(\tau|v) \tag{5}$$

Equation (5) represents the form of the general rumor source estimator we study in this work. The probability $\mathbb{P}(\tau|v)$ represents the likelihood of the rumor propagating along a given tree τ given the source $v^* = v$. This probability can be computed if the underlying rumor spreading model is known. Since we do not assume this information to be available, a key objective of this work is to study and identify good proxies for the true probability $\mathbb{P}(\tau|v)$ to enable robust estimation of the rumor source.

To complete the discussion, we discuss below two popular rumor source estimators that are agnostic to the stochastic model of rumor spread: Jordan center and distance center. These depend on the network's topological properties alone and are defined as:

$$\hat{v}_{JC} \in arg \min_{v \in V(G)} \max_{u \in V(G)} d(v,u) \tag{6}$$

$$\hat{v}_{DC} \in arg \min_{v \in V(G)} \sum_{u \in V(G)} d(v,u) \tag{7}$$

where $d(v,u)$ is the shortest distance between v and u on G. As we will discuss later in the section, both of these estimators are in fact special cases of our general estimator in equation (5). In the absence of pairwise information, the Jordan center is the root of the smallest diameter[1] BFS tree of the rumor infected graph and the distance center is the root of the BFS tree on which the sum-of-distances metric ($\sum_{u \in V(G)} d(v,u)$) is minimized. These estimators therefore have fast polynomial-time algorithms for implementation in the absence of pairwise information. When pairwise information is available, there might not exist a BFS tree of the rumor infected graph which is compatible with the available information. In this case, the problem of finding the Jordan center (distance center) requires finding an acceptable arborescence of the rumor graph on which the diameter (the sum-of-distances metric) is minimized. This is a hard combinatorial problem. We therefore do not consider these estimators under a pairwise information setting. Our work instead focuses on the study of estimators that are agnostic to the stochastic model of rumor spread but at the same time are also amenable to implementation under a pairwise information setting.

For an estimator of the form (5), we need good proxies for $\mathbb{P}(\tau|v)$. This prior probability associated with each potential rumor propagation tree of $G(V,E)$ will be a function of its topological parameters. Let τ_v denote a rumor propagation tree rooted at v. We consider the *diameter* $D(\tau_v)$ (distance of the farthermost node on τ_v from v) and *size of rumor boundary* $L(\tau_v)$ (number of leaf nodes on τ_v) as two important topological parameters of τ_v and define $\mathbb{P}(\tau_v|v)$ in terms of $D(\tau_v)$ and $L(\tau_v)$ in two ways:

(1) Assign equal prior probabilities to all rumor trees rooted at a node v, independent of their topological parameters, i.e.,

$$\mathbb{P}(\tau_v|v) = \frac{1}{|\mathbb{T}(G,v)|} \tag{8}$$

where $\mathbb{T}(G,v)$ is the set of all rumor trees of $G(V,E)$ rooted at v. This choice of our proxy for $\mathbb{P}(\tau|v)$ simplifies the estimator in equation (5) to:

$$\hat{v} \in arg \max_{v \in V(G)} \sum_{\tau \in \mathbb{T}(G,I,v)} \frac{1}{|\mathbb{T}(G,v)|} = arg \max_{v \in V(G)} \frac{|\mathbb{T}(G,I,v)|}{|\mathbb{T}(G,v)|} \tag{9}$$

(2) Assign prior probabilities to rumor trees based on their topological parameters as:

$$\mathbb{P}(\tau_v|v) \propto \alpha(D_0 - D(\tau_v)) + (1-\alpha)L(\tau_v) \tag{10}$$

where $0 \leq \alpha \leq 1$ and D_0 is a graph constant that corresponds to the largest diameter of any arborescence of G. Such a definition of $\mathbb{P}(\tau_v|v)$ assigns higher probabilities to rumor trees with a short diameter and a long rumor boundary, i.e., short and fat trees. This is a reasonable assignment since most epidemic models lead to higher likelihoods for the rumor to spread along propagation trees that broadly fall in this category. Hence this proxy for $\mathbb{P}(\tau_v|v)$ defined is a natural choice and we explore the performance of estimators based on this assignment.

In addition, we also consider threshold versions of our general estimator defined in equation (5):

$$\hat{v} \in arg \max_{v \in V(G)} \sum_{\tau \in \mathbb{T}(G,I,v)} \mathbb{P}(\tau|v) \mathbb{1}_{\{\mathbb{P}(\tau|v) > \delta\}} \tag{11}$$

$$\hat{v} \in arg \max_{v \in V(G)} \sum_{\tau \in \mathbb{T}(G,I,v)} \mathbb{1}_{\{\mathbb{P}(\tau|v) > \delta\}} \tag{12}$$

where δ is the threshold parameter. Notice that (11) is a threshold version of (5) and (12) is a threshold version of (9) with $\mathbb{P}(\tau|v)$ defined as per (10).

Depending on the particular choice we make, the general estimator defined in equation (5) can lead to several different estimators:

(1) **E1** - node with the maximum fraction of compatible rumor trees rooted at it i.e.,

$$\hat{v}_{E1} \in arg \max_{v \in V} \frac{|\mathbb{T}(G,I,v)|}{|\mathbb{T}(G,v)|}$$

(2) **E2** - node with the maximum average weight of compatible arborescences rooted at it i.e.,

$$\hat{v}_{E2} \in arg \max_{v \in V(G)} \frac{\sum_{\tau \in \mathbb{T}(G,I,v)} \mathbb{P}(\tau|v)}{|\mathbb{T}(G,I,v)|}$$

where $\mathbb{P}(\tau|v)$ is set according to (10).

(3) **E3** - root of the shortest diameter compatible rumor tree (*Jordan center*) i.e.,

$$\hat{v}_{E3} \in arg \min_{v \in V(G)} \max_{u \in V(G)} d(v,u).$$

As discussed before, implementing the Jordan center under a pairwise information setting is a hard combinatorial problem and hence we will not pursue this in the course of this paper. We however keep E3 in the discussion here since it is a special case of estimators E5 and E6 (see below).

(4) **E4** - node with the maximum measure ($\sum_{\tau \in \mathbb{T}(G,I,v)} \mathbb{P}(\tau|v)$) on the set of compatible rumor trees $\mathbb{T}(G,I,v)$ rooted at it i.e.,

$$\hat{v}_{E4} \in arg \max_{v \in V} \sum_{\tau \in \mathbb{T}(G,I,v)} \mathbb{P}(\tau|v)$$

[1] distance of the farthermost node from the root

where $\mathbb{P}(\tau|v)$ is set according to (10).

(5) **E5** - E4 with a threshold parameter δ i.e.,

$$\hat{v}_{E5} \in arg \max_{v \in V(G)} \sum_{\tau \in \mathbb{T}(G,I,v)} \mathbb{P}(\tau|v) \mathbb{1}_{\{\mathbb{P}(\tau|v) > \delta\}}$$

Observe that for $\delta = 0$, E5 = E4. Also observe that if $\delta = \max_{v \in V} \max_{\tau \in \mathbb{T}(G,I,v)} \mathbb{P}(\tau|v)$ where $\mathbb{P}(\tau|v)$ is set according to (10) with $\alpha = 1$, then estimator E5 reduces to E3 (Jordan center). It is also noteworthy that if the definition of $\mathbb{P}(\tau|v)$ is altered to $\mathbb{P}(\tau|v) \propto |V|^2 - \sum_{u \in V} d_\tau(v,u)$ where $d_\tau(v,u)$ is the distance between v and u on τ, with δ set as before, E5 reduces to the distance center.

(6) **E6** - E1 with a threshold parameter δ i.e.,

$$\hat{v}_{E6} \in arg \max_{v \in V(G)} \sum_{\tau \in \mathbb{T}(G,I,v)} \mathbb{1}_{\{\mathbb{P}(\tau|v) > \delta\}}$$

where $\mathbb{P}(\tau|v)$ is set according to (10). Observe that for $\delta = 0$, E6 = E1. With δ and $\mathbb{P}(\tau|v)$ set as per the preceding discussion for E5, E6 also reduces to E3 (Jordan center) or the distance center.

The rest of paper is devoted to describing schemes for implementing the various estimators discussed above.

3 REJECTION SAMPLING ON RUMOR TREES

In this section, we first present an efficient way of checking compatibility of rumor trees with the given pairwise information I. We next propose an MCMC-based algorithm to implement E1-E6 that invokes this compatibility check procedure.

3.1 Checking compatibility of rumor trees

Recall that each rumor tree corresponds to an arborescence with the directions of the edges reversed. Given this equivalence, we will describe a rejection sampling algorithm for arborescences. Recall that any row of the information matrix I looks like $[a\ b]$, with the interpretation that a was infected before b. We reject an arborescence if there exists a directed path from a to b on it since this would imply that the rumor reached b before a on the propagation tree. If the arborescence survives compatibility checks with all rows of I, it is accepted. A brute force way of checking compatibility is to check the existence of b on the directed path from a to the root, for all rows of I. This is computationally intensive. The problem is aggravated by the high sample correlation of the MCMC processes, as successively generated arborescences differ only in a few edges (see section 3.2). This leads to a large number of samples in a rejection episode, slowing the algorithm down. We present a scheme to make the compatibility checks faster and computationally cheaper to implement.

Let P denote the adjacency matrix of an arborescence τ. Let $R = (I - P)^{-1}$. For the i^{th} row of R, all columns with a value > 0 indicate the nodes that are visited on the directed path from node i to the root. Thus if $R(a,b) = 1$, τ is rejected. Matrix inversion, however, is a costly operation. We leverage the fact that successively generated arborescences differ only in a few edges (see section 3.2), so the adjacency matrix of the $(k+1)$-th arborescence P_{k+1} is a perturbation of P_k by a low rank matrix and we can use the *Sherman-Morrison-Woodbury* formula ([4], section 2.1.3) to evaluate $(I - P_{k+1})^{-1}$ from $(I - P_k)^{-1}$ given $(P_{k+1} - P_k)$. This drastically speeds up the runtime.

3.2 MCMC-based inference algorithm

Here, we describe an MCMC-based algorithm to implement E1-E6. We begin with an MCMC scheme to generate an arborescence-valued Markov chain with a desired stationary distribution, based on a simple random walk on $G(V,E)$ biased using a suitable Metropolis-Hastings [5] filter $H(\cdot)$ to improve the sampling process. This is then used to implement the estimators E1-E6 using the procedure described in section 3.1.

For any arborescence τ_r rooted at node r, let $h(\tau_r) := \mathbb{P}(\tau_r|r)$ denote the score of τ_r, where $\mathbb{P}(\tau_r|r)$ is defined as per (8) or (10). We also set a parameter $T > 0$ which controls the mixing rate of the underlying Markov chain and the Metropolis-Hastings filter $H(\tau_r)$ which is supposed to be representative of the 'energy' of arborescence τ_r. As shown below, the sampling process favours low energy arborescences. We do not explicitly define $H(\tau_r)$, but it can conveniently be set as a function of the topological parameters or the score of τ_r. For instance, a natural choice is $H(\tau_r) = -\mathbb{P}(\tau_r|r)$ so that the arborescences with a higher score will be sampled more often.

The algorithm goes as follows. Generate an arborescence τ_r of G rooted at r. Initialize a vector of counts $\{c\}$ and scores $\{s_c\}$ of nodes in V to all zeros.

(1) Select a neighbor r' of r on $G(V,E)$ uniformly at random. Add the directed edge $r \longrightarrow r'$ to τ_r and remove the unique outgoing edge from r' in the resulting graph. Denote the arborescence so formed by $\tau'_{r'}$.

(2) Do:

$$(\tau_r, r) \leftarrow \begin{cases} (\tau'_{r'}, r') & \text{with prob. } \min\left\{1, e^{-\left(\frac{H(\tau'_{r'}) - H(\tau_r)}{T}\right)}\right\} \\ (\tau_r, r) & \text{with prob. } 1 - \min\left\{1, e^{-\left(\frac{H(\tau'_{r'}) - H(\tau_r)}{T}\right)}\right\} \end{cases}$$

(3) Do: $c(r) \leftarrow c(r) + 1$.

(4) Check compatibility of τ_r as described in section 3.1. If τ_r is accepted, do: $s_c(r) \leftarrow s_c(r) + h(\tau_r) \frac{e^{\frac{H(\tau_r)}{T}}}{d(r)}$.

(5) Go to (1).

It can be easily proved that this arborescence-valued Markov chain is reversible and has a unique stationary distribution over the set of arborescences of $G(V,E)$ given by $\pi(\tau_r) = \frac{d(r)}{2|E(G)|} \frac{e^{-\frac{H(\tau_r)}{T}}}{Z}$, where $d(r)$ denotes the degree of r on $G(V,E)$, $Z = \sum_{\tau \in \mathbb{T}(G)} e^{-\frac{H(\tau)}{T}}$ and $\mathbb{T}(G)$ is the set of all arborescences of $G(V,E)$. Notice that a correction factor of $e^{\frac{H(\tau_r)}{T}}/d(r)$ is introduced in step (4) to account for the non-uniformity introduced in the sampling process. At stationarity, the unbiased score of a node v is given by $s(v) = s_c(v)/c(v)$. The estimated rumor source is given by $\hat{v} = arg \max_{v \in V(G)} s(v)$ with ties broken uniformly at random.

Another way of generating an arborescence-valued Markov chain is following. At each node v in V, do:

(1) Generate a spanning tree τ of $G(V,E)$.

(2) Sample an edge e uniformly at random from $E(G)\backslash E(\tau)$ and add to τ. The resulting graph $G(V, E(\tau) \cup e)$ has a unique loop.
(3) Select and remove one of the edges e' of the loop so formed in $G(V, E(\tau) \cup e)$ (other than the edge just added) uniformly at random. The resulting graph $G(V, E(\tau) \cup e \backslash e')$ is a tree. Denote it by τ'.
(4) $\tau \leftarrow \tau'$.
(5) Go to (2).

It can be easily proved that the ensuing Markov chain is reversible and has a unique stationary distribution that is uniform over the set of spanning trees of $G(V, E)$. To any tree in this process, there is a unique way of assigning directions to edges so to turn it into an arborescence rooted at v. Consequently, the stationary distribution over the set of arborescences rooted at v is also uniform. A total of $|V|$ such chains is simulated, one for each $v \in V$ and all chains are run for the same number of steps. This MCMC algorithm therefore can be useful when parallel computing resources are available.

4 SIMULATING RUMOR GROWTHS

The algorithm we present in this section can be used to implement all estimators E1-E6 and the MLE as defined in section 2, as well as the rumor center as defined in [8] (without the BFS heuristic) which counts the number of compatible rumor permutations starting from each node v. This algorithm is different from the one discussed in the previous section in that the samples generated are IID. In this method, we simulate a rumor growth starting from each $v \in V(G)$ according to the Susceptible-Infected (SI) model with IID exponential clocks. This results in a rumor permutation σ_v of $V(G)$ on a rumor tree τ_v of G rooted at v with probabilities $p(\sigma_v)$ and $p(\tau_v)$ respectively. We generate an IID sequence of rumor permutations and rumor trees and use the samples to compute the measure on the set of acceptable rumor permutations and rumor trees respectively.

The following is a $|V(G)|$-step rumor growth algorithm:

(1) Initialize as: $U = \{v\}$, $p(\tau_v \cap \sigma_v) = 1$, $p(\sigma_v) = 1$.
(2) Sample an edge e uniformly at random from $E_1(U)$ (recall definition from section 1). Let v_e denote the vertex at the end of e that is outside U. Add v_e to U. Also do:

$$p(\tau_v \cap \sigma_v) = p(\tau_v \cap \sigma_v) \times \frac{1}{|E_1(U)|}$$

$$p(\sigma_v) = p(\sigma_v) \times \frac{\sum_{e' \in E_1(U)} \mathbb{1}_{\{v_{e'} = v_e\}}}{|E_1(U)|}$$

(3) **If** $|U| = |V(G)|$, STOP; **Else** go to (2).

This results in a rumor permutation σ_v on a rumor tree τ_v rooted at v with $p(\tau_v \cap \sigma_v)$ and $p(\sigma_v|G) \equiv p(\sigma_v)$ computable in runtime. The conditional $p(\sigma_v|\tau_v)$ is required to compute the marginal $p(\tau_v)$. $p(\sigma_v|\tau_v)$ can be computed using the same growth algorithm on τ instead of G. The marginal $p(\tau_v)$ is then given by Bayes' theorem as $p(\tau_v) = \frac{p(\tau_v \cap \sigma_v)}{p(\sigma_v|\tau_v)}$. Once the marginal probabilities associated with σ_v and τ_v have been calculated, what remains is to use these to assign scores to the various nodes based on the particular estimator amongst E1-E6, MLE, or rumor center we are implementing. This is done using the inference algorithm we describe next.

Inference algorithm

At each node $v \in V$, do the following:

(1) Initialize the score $s(v)$ zero. Let $h(\tau_v) := \mathbb{P}(\tau_v|v)$ denote the score of τ_v, where $\mathbb{P}(\tau_v|v)$ is defined as per (8) or (10).
(2) Simulate a rumor spread starting from v. Check for its compatibility[2]. If the rumor spread is accepted, go to (3); if rejected, repeat (2).
(3) **For E1-E6, do:**
$s(v) \leftarrow s(v) + \frac{h(\tau_v)}{p(\tau_v)}$
For rumor center (w/o BFS heuristic), do:
$s(v) \leftarrow s(v) + \frac{1}{p(\sigma_v)}$
For MLE, do: $s(v) \leftarrow s(v) + 1$.
(4) Go to (2).

Equal number of rumor spreads are simulated from each $v \in V$. The estimated rumor source is given by $\hat{v} \in arg\max_{v \in V} s(v)$ with ties broken uniformly at random. Furthermore, this method of generating arborescences/permutations is amenable to on-the-go compatibility checks, i.e., a rumor subtree (and a permutation thereon) need not have spanned the entire infected graph before it can be checked for compatibility. This offers huge computational as well as runtime savings.

5 GRAPH SPARSIFICATION BY EFFECTIVE RESISTANCES

The number of spanning trees of a graph G is given by $\mu_2\mu_3...\mu_n/n$ [1] where $\{\mu_i\}_{i \geq 2}$ is the set of non-zero eigenvalues of the Laplacian of G. For a complete graph on n nodes, this value is n^{n-2}. Even for graphs that are far from complete, the size of the set of spanning trees/arborescences scales exponentially with the number of nodes in the graph. This is clearly a point of concern and graph sparsification helps address this issue. In particular, we sparsify the graph based on the effective resistances R_{eff} of its edges. The effective resistance of an edge is known to be equal to the probability that the edge appears in a random spanning tree of G [11]. The main idea in sparsification by effective resistances is to include each edge of G in the sparsifier with a probability proportional to its effective resistance. This is known to maintain the essential structure and connectivity of the graph. An infection spread over a network strongly follows its connectivity and link structure. We therefore expect this sparsification scheme to retain a significant amount of information about the underlying rumor spread.

We use the tree-valued MCMC algorithm (described towards the end of section 3) to estimate edge resistances. In particular, we simulate a Markov chain with a uniform stationary distribution over the set of spanning trees of G. The effective resistance of an edge is then simply the long run fraction of times it featured in a random spanning tree generated by the Markov chain. In order to sparsify the graph, we only retain edges occurring more than a threshold fraction p_t of times. In section 6, we examine the effect of this threshold p_t on the detection rate of the different estimators via simulations.

[2] section 3.1 describes the rejection sampling procedure for arborescences which can be used for estimators E1-E6; for the MLE and rumor center, a similar procedure has to be conducted for rumor permutations

6 SIMULATION RESULTS

The rumor spread is simulated from a single source node using the Susceptible-Infected (SI) model with IID exponential clocks. The rumor starts from the source node and propagates along edges of the underlying graph with propagation times along edges modeled as IID exponential random variables. We simulate this on Erdős-Rényi (ER) graphs [3] and scale-free networks (SFN) generated using the Barabási-Albert model [2] with 1000 nodes and stop when $n = 400$ nodes get infected. The pairwise information matrix $I_{\lfloor f\cdot\binom{n}{2}\rfloor\times 2}$ consists of rows sampled uniformly at random without replacement from all $\binom{n}{2}$ possible pairwise comparisons, where $0 \leq f < 1$. We perform experiments on ER graphs $G(1000, p)$ over a range of values of the parameter $p :=$ the probability of any given edge existing in the graph. We also perform experiments on SFNs over a range of values of the parameters $m_0 :=$ the starting number of nodes in the graph and $m :=$ the number of nodes that is added during each successive iteration of the generative model. The performance measure used is the *detection rate*, which is the fraction of experiments in which the correct rumor source is detected.

We compare the performance of the various estimators defined in section 2. We also compare the performance of all our estimators with *random guessing*. At an available fraction f of the pairwise information, the nodes that feature in the second column of $I_{\lfloor f\cdot\binom{n}{2}\rfloor\times 2}$ are disqualified as potential rumor source. Using this observation, it is easy to verify that the detection rate for a scheme which randomly guesses the source from amongst the set $V\backslash V_2(f)$ is asymptotically equal to the fraction of available information f, as the number of infected nodes becomes large.

Figures 1,2 show the performance of E1 on ER graphs and SFNs with different graph parameters. On the ER graph, the performance hardly changes with the p-value, possibly because of the growth isotropy inherent to the ER model. Figures 3,4 compare the performance of estimator E1 when the detection rate computed is based on exact detection of the rumor source and when the criteria is relaxed to detection within a 1 or 2 hop neighborhood of the rumor source. Figures 5,6 show the performance of E1 on sparsified ER graphs and SFNs for different sparsification thresholds p_t, as discussed in section 5. Thus $p_t = 0$ corresponds to the original rumor infected graph. As p_t is increased, performance of the estimator degrades. A comparison of the performance of E1, E2 and E4 on ER graphs and SFNs can be seen from figures 7,8. E4 expectedly outperforms E1. Perhaps surprisingly, E2 is outperformed by E1 and E4 both, and is only slightly better than guessing randomly. This could possibly be because of a high variance in the scores of compatible arborescences that the expectation does not take into account. This variance decreases as the available fraction f of pairwise information I increases since a lesser number of arborescences survive compatibility checks. Thus, the performance of the estimator E2 is expected to improve with increasing f. This is reflected in figures 7,8. Figures 9,10 show the performance of E6 with different score thresholds s_t. Recall from section 2 that E6 is based on using a threshold parameter δ on the value of the chosen proxy for $\mathbb{P}(\tau_v|v)$. We set s_t as a suitably chosen monotonically increasing function of the threshold parameter δ that maps $[0, 1]$ to $(-\infty, \infty)$. This is done in order to increase the dynamic range of the threshold parameter to enhance the observability of performance trends with varying

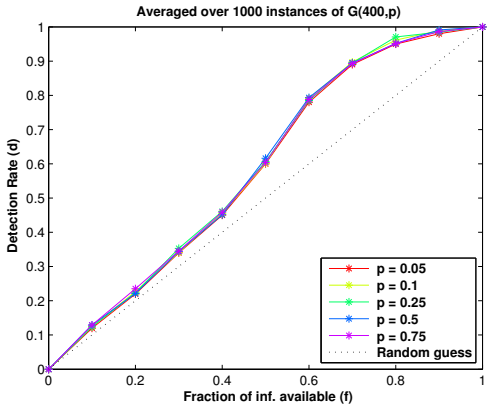

Figure 1: Performance of the estimator E1 on ER graphs with different p.

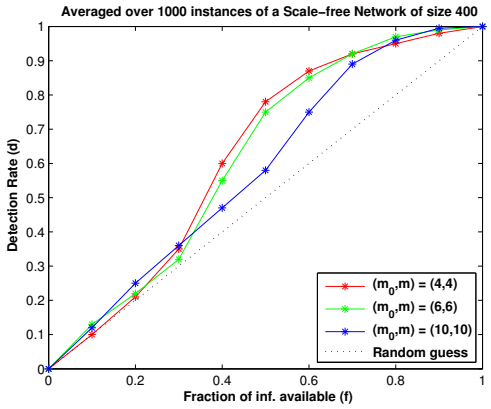

Figure 2: Performance of the estimator E1 on SFNs with different graph parameters m_0, m.

thresholds. At $s_t = -\infty$, E6 = E1. The performance improves with increasing thresholds up to a certain value and starts degrading as the threshold is increased beyond this optimal value. The set of arborescences with scores above this optimal threshold is the most informative set about the rumor spread. As the score threshold is lowered from this optimal value, arborescences that are unlikely to have been the propagation tree start contributing to the estimator value. This amounts to noise and hence the performance degrades. The performance also degrades as the threshold is increased beyond the optimal value as arborescences that well qualify to be the propagation tree start getting ignored. A similar trend is expected for estimator E5 as well. Figures 11,12 show the performance comparison of E5 and E6 with the best score thresholds s_t against the rumor source MLE. As expected, E5 outperforms E6 and the MLE outperforms them both. It is however very important to note that the performance disparity between E5,E6, and the MLE is quite limited, especially considering the huge computational as well as runtime savings that E5,E6 offer over the MLE.

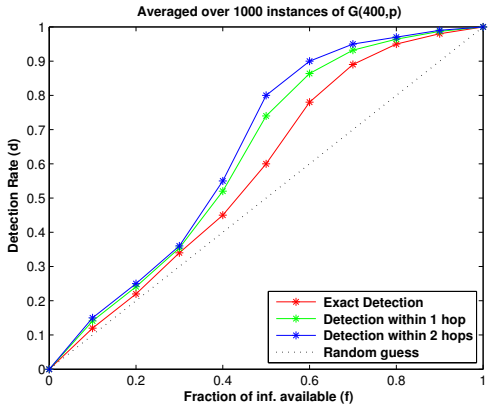

Figure 3: Performance of the estimator E1 on ER graph.

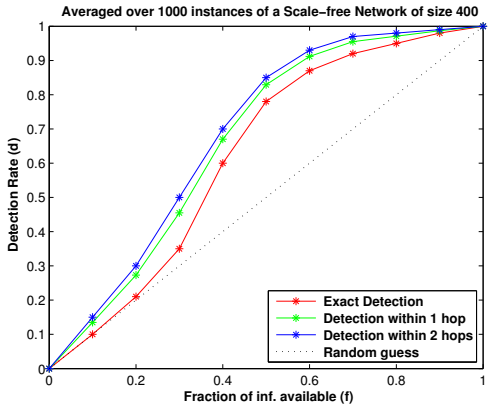

Figure 4: Performance of the estimator E1 on SFN.

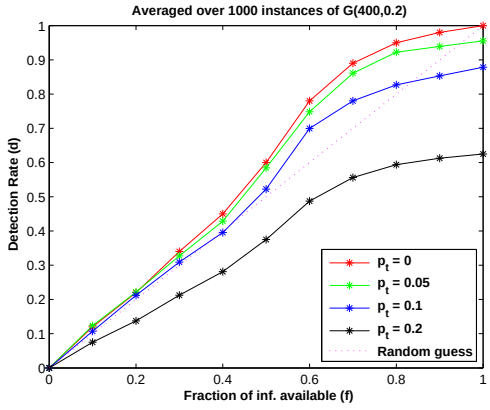

Figure 5: Performance of the estimator E1 on sparsified ER graphs. Edges with R_{eff} less than p_t are dropped off.

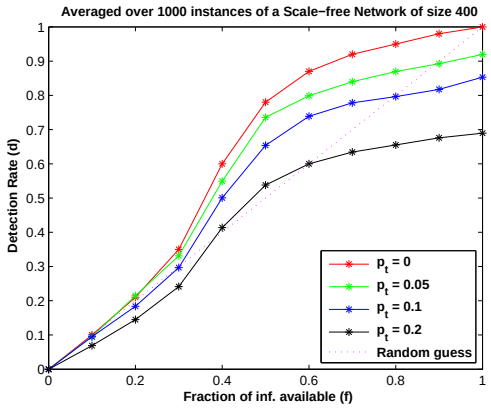

Figure 6: Performance of the estimator E1 on sparsified SFNs. Edges with R_{eff} less than p_t are dropped off.

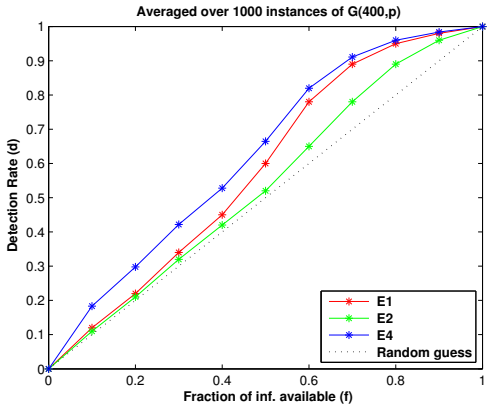

Figure 7: Comparison of E1, E2 and E4 on ER graph.

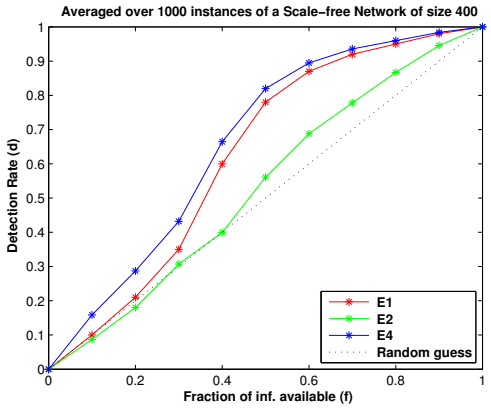

Figure 8: Comparison of E1, E2 and E4 on SFN.

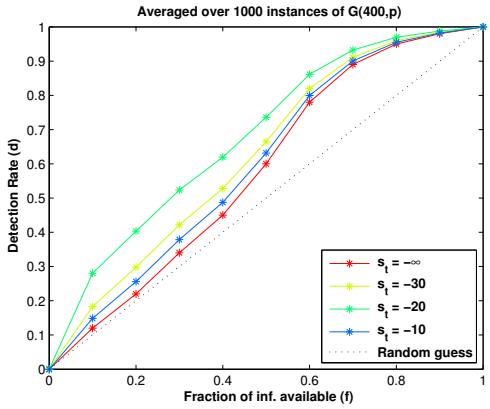

Figure 9: Performance of E6 with different thresholds on ER graph.

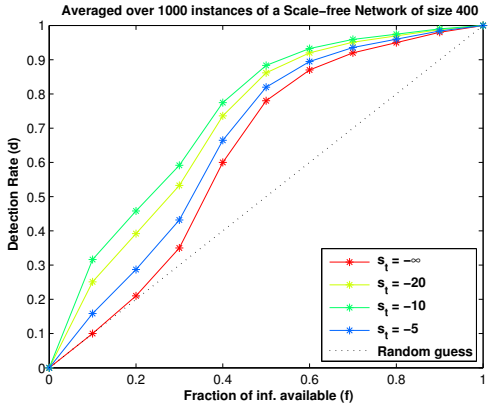

Figure 10: Performance of E6 with different thresholds on SFN.

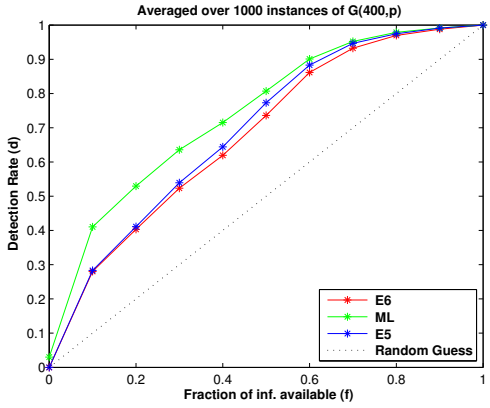

Figure 11: Comparison of E5 and E6 with the best thresholds against the MLE on ER graph.

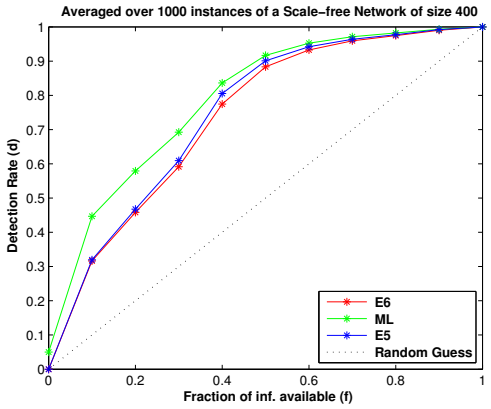

Figure 12: Comparison of E5 and E6 with the best thresholds against the MLE on SFN.

To conclude, we comment on the convergence rates of our MCMC schemes. Note that our MCMC scheme can be viewed as a random walk on the set of arborescences (or spanning trees) of the underlying graph. Using results on the mixing time of a simple random walk [1], it can be verified that the mixing times of our MCMC algorithms are at most polynomial in $n \log n$.

Acknowledgements Work of VSB supported in part by a J. C. Bose Fellowship. Work of VSB and NK supported in part by an Indo-French grant No. IFC/DST-Inria-2016-01/448 "Machine Learning for Network Analytics". NK also acknowledges the support from DST INSPIRE Faculty Fellowship and a seed grant from IIT Bombay.

REFERENCES

[1] David Aldous and Jim Fill. 2002. Reversible Markov chains and random walks on graphs. (2002).
[2] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.
[3] P. Erdős and A. Rényi. 1959. On random graphs. I. *Publ. Math. Debrecen* 6 (1959), 290–297.
[4] GH Golub and CF Van Loan. 1996. Matrix computations 3rd edition The John Hopkins University Press. *Baltimore, MD* (1996).
[5] W Keith Hastings. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 1 (1970), 97–109.
[6] Nikhil Karamchandani and Massimo Franceschetti. 2013. Rumor source detection under probabilistic sampling. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE.
[7] Ankit Kumar, Vivek Borkar, and Nikhil Karamchandani. 2017. Temporally Agnostic Rumor Source Detection. *IEEE Transactions on Signal and Information Processing over Networks* (2017).
[8] Devavrat Shah and Tauhid Zaman. 2011. Rumors in a network: who's the culprit? *IEEE Transactions on Information Theory* (2011).
[9] Devavrat Shah and Tauhid Zaman. 2012. Finding rumor sources on random graphs. (2012).
[10] Devavrat Shah and Tauhid Zaman. 2012. Rumor centrality: a universal source detector. In *ACM SIGMETRICS Performance Evaluation Review*, Vol. 40. ACM, 199–210.
[11] Daniel A Spielman and Nikhil Srivastava. 2011. Graph sparsification by effective resistances. *SIAM J. Comput.* 40, 6 (2011), 1913–1926.
[12] Kai Zhu and Lei Ying. 2015. Source localization in networks: Trees and beyond. *arXiv preprint arXiv:1510.01814* (2015).
[13] Kai Zhu and Lei Ying. 2016. Information source detection in the SIR model: A sample-path-based approach. *IEEE/ACM Transactions on Networking (TON)* 24, 1 (2016), 408–421.