Rumors in a Network: Who's the Culprit?

Devavrat Shah and Tauhid Zaman, *Member, IEEE*

***Abstract*—We provide a systematic study of the problem of finding the source of a rumor in a network. We model rumor spreading in a network with the popular susceptible-infected (SI) model and then construct an estimator for the rumor source. This estimator is based upon a novel topological quantity which we term rumor centrality. We establish that this is a maximum likelihood (ML) estimator for a class of graphs. We find the following surprising threshold phenomenon: on trees which grow faster than a line, the estimator always has nontrivial detection probability, whereas on trees that grow like a line, the detection probability will go to 0 as the network grows. Simulations performed on synthetic networks such as the popular small-world and scale-free networks, and on real networks such as an internet AS network and the U.S. electric power grid network, show that the estimator either finds the source exactly or within a few hops of the true source across different network topologies. We compare rumor centrality to another common network centrality notion known as distance centrality. We prove that on trees, the rumor center and distance center are equivalent, but on general networks, they may differ. Indeed, simulations show that rumor centrality outperforms distance centrality in finding rumor sources in networks which are not tree-like.**

***Index Terms*—Complex networks, detection algorithms, graph theory, inference algorithms, maximum likelihood (ML) detection, probability.**

I. Introduction

In the modern world the ubiquity of networks has made us vulnerable to new types of network risks. These network risks arise in many different contexts, but share a common structure: an isolated risk is amplified because it is spread by the network. For example, as we have witnessed in the recent financial crisis, the strong dependencies or 'network' between institutions have led to the situation where the failure of one (or few) institution(s) have led to global instabilities. In an electrical power grid network, an isolated failure could lead to a rolling blackout. Computer viruses utilize the Internet to infect millions of computers everyday. Finally, malicious rumors or misinformation can rapidly spread through existing social networks and lead to pernicious effects on individuals or society. In all of these situations, a policy maker, power network operator, Internet service provider or victim of a malicious rumor, would like to identify the source of the risk as quickly as possible and subsequently quarantine its effect.

In essence, all of these situations can be modeled as a rumor spreading through a network, where the goal is to find the source of the rumor in order to control and prevent these network risks based on limited information about the network structure and the "rumor infected" nodes. In this paper, we shall take important initial steps towards a systematic study of the question of identifying the rumor source based on the network structure and rumor infected nodes, as well as understand the fundamental limitations on this estimation problem.

A. Related Work

Prior work on rumor spreading has primarily focused on viral epidemics in populations. The natural (and somewhat standard) model for viral epidemics is known as the *susceptible-infected-recovered* or SIR model [1]. In this model, there are three types of nodes: (i) susceptible nodes, capable of being infected; (ii) infected nodes that can spread the virus further; and (iii) recovered nodes that are cured and can no longer become infected. Research in the SIR model has focused on understanding how the structure of the network and rates of infection/cure lead to large epidemics [2]–[5]. This motivated various researchers to propose network inference techniques for learning the relevant network parameters [6]–[10]. However, there has been little (or no) work done on inferring the source of an epidemic.

The primary reason for the lack of such work is that it is quite challenging. To substantiate this, we briefly describe a closely related (and much simpler) problem of reconstruction on trees [11], [12], or more generally, on graphs [13]. In this problem one node in the graph, call it the *root* node, starts with a value, say 0 or 1. This information is propagated to its neighbors and their neighbors recursively along a breadth-first-search (BFS) tree of the graph (when the graph is a tree, the BFS tree is the graph). Now each transmission from a node to its neighbor is noisy—a transmitted bit is flipped with a small probability. The question of interest is to estimate or reconstruct the value of the root node, based on the "noisy" information received at nodes that are far away from root. Currently, this problem is well understood only for graphs that are trees or tree-like, after a long history. Now the rumor source identification problem is, in a sense harder, as we wish to identify the location of the source among many nodes based on the infected nodes—clearly a much noisier situation than the reconstruction problem. Therefore, as the first step, we would like to understand this problem on trees.

B. Our Contributions

In this paper, we take initial steps towards understanding the question of identifying the rumor source in a network based on (rumor) infected nodes. Specifically, we start by considering a probabilistic model of rumor spreading in the network as the

Manuscript received June 24, 2009; revised February 28, 2011; accepted March 07, 2011. Date of current version July 29, 2011. This work was supported in part by the AFOSR complex networks program, in part by the NSF HSD Project, in part by the NSF EMT Project, and in part by the MIT-Shell Energy Fellowship.

The authors are with the Laboratory for Information and Decision Systems, Department of EECS, Massachusetts Institute of Technology, Cambridge, MA 02446 USA (e-mail: devavrat@mit.edu; zlisto@mit.edu).

Communicated by H.-A. Loeliger, Associate Editor for Coding Techniques.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2011.2158885

ground truth. This model is based on the SIR model which is well studied in the context of epidemiology, as mentioned earlier. It is the natural rumor spreading model with minimal side information. Therefore, such a model provides the perfect starting point to undertake the systematic study of such inference problems.

The question of interest is to identify the source of the rumor based on information about the infected nodes as well as the underlying network structure using the prior information about the probabilistic rumor spreading model. In the absence of additional information (i.e., a uniform prior), clearly the maximum likelihood (ML) estimator minimizes the estimation error. Therefore, we would like to identify: (a) a computationally tractable representation of the ML estimator if possible; and (b) evaluate the detection probability of such an estimator.

Now obtaining a succinct, useful characterization of an ML estimator for a general graph seems intractable. Therefore, following the philosophical approach of researchers working on the reconstruction problem mentioned above and on the efficient graphical model based inference algorithm (i.e., Belief Propagation), we address the above questions for tree networks.

We are able to obtain a succinct and computationally efficient characterization of the ML estimator for the rumor source when the underlying network is a regular tree. We are able to characterize the correct detection probability of this ML estimator for regular trees of any given node degree d. We find the following phase transition. For $d = 2$, i.e., when the network is a linear graph (or a path), the asymptotic detection probability is 0. For $d \geq 3$, i.e., when the network is an *expanding* tree, the asymptotic detection probability is strictly positive. For example, for $d = 3$, we identify it to be $1/4$.

As the next step, we consider nonregular trees. The ML estimator of regular trees naturally extends to provide a rumor source estimator for nonregular trees. However, it is not necessarily the ML estimator. We find that when a nonregular tree satisfies a certain *geometric* growth property (see IV-E for the precise definition), then the asymptotic detection probability of this estimator is 1. This suggests that even though this computationally simple estimator is not the ML estimator, its asymptotic performance is as good as any other (and hence the ML) estimator.

Motivated by results for trees, we develop a natural, computationally efficient heuristic estimator for general graphs based on the ML estimator for regular trees. We perform extensive simulations to show that this estimator performs quite well on a broad range of network topologies. This includes synthetic networks obtained from the small-world model and the scale-free model as well as real network topologies such as the U.S. electric power-grid and the Internet. In summary, we find that when the network structure (irrespective of being a tree) is not too *irregular*, the estimator performs well.

The estimator, which is ML for regular trees, can be thought of as assigning a nonnegative value to each node in a tree. We call this value the *rumor centrality* of the node. In essence, the estimator chooses the node with the highest rumor centrality as the estimated source, which we call the *rumor center* of the network. There are various notions of network centralities that are popular in the literature (cf. [14] and [15]). Therefore, in principle, each of these network centrality notions can act as rumor source estimators. Somewhat surprisingly, we find that the source estimator based on the popular *distance centrality* notion is identical to the rumor centrality based estimator for any tree. Therefore, in a sense our work provides theoretical justification for distance centrality in the context of rumor source detection.

Technically, the method for establishing nontrivial asymptotic detection for regular trees with $d \geq 3$ is quite different from that for geometric trees. Specifically, for regular trees with $d \geq 3$, we need to develop a refined probabilistic estimation of the rumor spreading process to establish our results. Roughly speaking, this is necessary because the rumor process exhibits high variance on expanding trees (due to exponential growth in the size of the neighborhood of a node with distance) and hence standard concentration based results are not meaningful (for establishing the result). On the other hand, for geometric trees the rumor process exhibits sharp enough concentration (due to subexponential growth in the size of the neighborhood of a node with distance) for establishing the desired result. This also allows us to deal with heterogeneity in the context of geometric trees. Similar technical contrasts between geometric and expanding structures are faced in analyzing growth processes on them. For example, in the classical percolation literature precise 'shape theorems' are known for geometric structures (e.g., d-dimensional grids) [16]–[21]. However, little is known in the context of expanding structures. Indeed, our techniques for analyzing regular (expanding) trees do overcome such challenges. Developing them further for general expanding graphs (nonregular trees and beyond) remain an important direction for future research.

Finally, we note that calculating the rumor centrality of a node is equal to computing the number of possible linear extensions of a given partial order represented by the tree structure rooted at that particular node. Subsequently, our algorithm leads to the fastest known algorithm for computing the number of possible linear extensions in this context (see [22] for the best known algorithm in the literature).

C. Organization

In Section II, the probabilistic model for rumor spreading and the derivation of the source estimator is presented. Section III studies properties of this estimator and presents an efficient algorithm for its evaluation. Section IV presents results about the effectiveness of the estimator for tree networks in terms of its asymptotic detection probability. Section V shows the effectiveness of the estimator for general networks by means of extensive simulations. Section VI provides detailed proofs of the results presented in Section IV. We conclude in Section VII with directions for future work.

II. Rumor Source Estimator

In this section we start with a description of our rumor spreading model and then we define the maximum likelihood (ML) estimator for the rumor source. For regular tree graphs, we equate the ML estimator to a novel topological quantity which we call rumor centrality. We then use rumor centrality to construct rumor source estimators for general graphs.

A. Rumor Spreading Model

We consider a network of nodes modeled as an undirected graph $G(V,E)$, where V is a countably infinite set of nodes and E is the set of edges of the form (i,j) for some i and j in V. We assume the set of nodes is countably infinite in order to avoid boundary effects. We consider the case where initially only one node v^* is the rumor source.

We use a variant of the common SIR model for the rumor spreading known as the *susceptible-infected* or SI model which does not allow for any nodes to recover, i.e., once a node has the rumor, it keeps it forever. Once a node i has the rumor, it is able to spread it to another node j if and only if there is an edge between them, i.e., if $(i,j) \in E$. Let τ_{ij} be the time it takes for node j to receive the rumor from node i once i has the rumor. In this model, τ_{ij} are independent and have exponential distribution with parameter (rate) λ. Without loss of generality, assume $\lambda = 1$.

B. Rumor Source Estimator: Maximum Likelihood (ML)

Let us suppose that the rumor starting at a node, say v^* at time 0 has spread in the network G. We observe the network at some time and find N infected nodes. By definition, these nodes must form a connected subgraph of G. We shall denote it by G_N. Our goal is to produce an estimate, which we shall denote by $\hat{v}$, of the original source v^* based on the observation G_N and the knowledge of G. To make this estimation, we know that the rumor has spread in G_N as per the SI model described above. However, a priori we do not know from which source the rumor started. Therefore, we shall assume a uniform prior probability of the source node among all nodes of G_N. With respect to this setup, the maximum likelihood (ML) estimator of v^* with respect to the SI model given G_N minimizes the error probability, i.e., maximizes the correct detection probability. By definition, the ML estimator is

$$\hat{v} \in \arg\max_{v \in G_N} \mathbf{P}(G_N|v), \tag{1}$$

where $\mathbf{P}(G_N|v)$ is the probability of observing G_N under the SI model assuming v is the source, v^*. Thus, ideally we would like to evaluate $\mathbf{P}(G_N|v)$ for all $v \in G_N$ and then select the one with the maximal value (ties broken uniformly at random).

C. Rumor Source Estimator: ML for Regular Trees

In general, evaluation of $\mathbf{P}(G_N|v)$ may not be computationally tractable. Here we shall show that for regular trees, $\mathbf{P}(G_N|v)$ becomes proportional to a quantity $R(v,G_N)$ which we define later and call rumor centrality. The $R(v,G_N)$ is a topological quantity and is intimately related to the structure of G_N.

Now to evaluate $\mathbf{P}(G_N|v)$ when the underlying graph is a tree, essentially we wish to find the probability of all possible events that result in G_N after N nodes are infected starting with v as the source under the SI model. To understand such events, let us consider a simple example as shown in Fig. 1 with $N=4$. Now, suppose node 1 was the source, i.e., we wish to calculate $\mathbf{P}(G_4|1)$. Then there are two disjoint events or node orders in which the rumor spreads that will lead to G_4 with 1 as the source: $\{1,2,3,4\}$ and $\{1,2,4,3\}$. However, due to the structure of the network, infection order $\{1,3,2,4\}$ is not possible. Therefore, in general to evaluate $\mathbf{P}(G_N|v)$, we need to find all such *permitted permutations* and their corresponding probabilities. Because permitted permutations are an important concept in this paper, we now define this notion precisely.

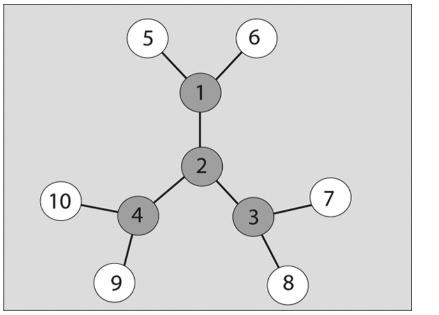

Fig. 1. Example network where the rumor graph has four nodes.

Definition 1 (Permitted Permutation): Given a connected tree $G(V,E)$ and a source node $v \in V$, consider any permution $\sigma : V \rightarrow \{1,2,\ldots,|V|\}$ of its nodes where $\sigma(u)$ denotes the position of node $u \in V$ in the permutation σ. We call σ a permitted permutation for tree $G(V,E)$ with source node v if

1) $\sigma(v) = 1$.
2) For any $(u,u') \in E$, if $d(v,u) < d(v,u')$, then $\sigma(u) < \sigma(u')$. Here $d(v,u)$ denotes the shortest path distance from v to u.

Let $\Omega(v,G_N)$ be the set of all permitted permutations starting with node v and resulting in rumor graph G_N. We wish to determine the probability $\mathbf{P}(\sigma|v)$ for each $\sigma \in \Omega(v,G_N)$. To that end, let $\sigma = \{v_1 = v, v_2, \ldots, v_N\}$. Let us define, $G_k(\sigma)$ as the subgraph (of G_N) containing nodes $\{v_1 = v, v_2, \ldots, v_k\}$ for $1 \leq k \leq N$. Then

$$\mathbf{P}\big(\sigma \mid v\big) = \prod_{k=2}^{N} \mathbf{P}\big(k^{\text{th}} \text{ infected node} = v_k \bigm| G_{k-1}(\sigma), v\big). \tag{2}$$

Each term in the product on the right-hand side (RHS) in (2), can be evaluated as follows. Given $G_{k-1}(\sigma)$ (and source v), the next infected node could be any of the neighbors of nodes in $G_{k-1}(\sigma)$ which are not yet infected. For example, in Fig. 1 G_2 is $\{1,2\}$ when the source is assumed to be 1. In that case, the next infected node could be any one of the 4 nodes: 3, 4, 5 and 6. Now due to the memoryless property of exponential random variables and since all infection times on all edges are independent and identically distributed (i.i.d.), it follows that each of these nodes is equally likely to be the next infected node. Therefore, each one of them has probability $1/4$. More generally, if $G_{k-1}(\sigma)$ has $n_{k-1}(\sigma)$ uninfected neighboring nodes, then each one of them is equally likely to be the next infected node with probability $1/n_{k-1}(\sigma)$. Therefore, (2) reduces to

$$\mathbf{P}\big(\sigma \mid v\big) = \prod_{k=2}^{N} \frac{1}{n_{k-1}(\sigma)}. \tag{3}$$

Given (3), now the problem of computing $\mathbf{P}\big(\sigma \bigm| v\big)$ boils down to evaluating the size of the *rumor boundary* $n_{k-1}(\sigma)$ for $2 \leq$

$k \leq N$. To that end, suppose the k^{th} added node to $G_{k-1}(\sigma)$ is $v_k(\sigma)$ with degree $d_k(\sigma)$. Then it contributes $d_k(\sigma) - 2$ new edges (and hence nodes in the tree) to the rumor boundary. This is because, $d_k(\sigma)$ new edges are added but we must remove the edge along which the recent infection happened, which is counted twice. That is, $n_k(\sigma) = n_{k-1}(\sigma) + d_k(\sigma) - 2$. Subsequently

$$n_k(\sigma) = d_1(\sigma) + \sum_{i=2}^{k}(d_i(\sigma) - 2). \tag{4}$$

Therefore

$$\mathbf{P}(\sigma \mid v) = \prod_{k=2}^{N} \frac{1}{d_1(\sigma) + \sum\limits_{i=2}^{k-1}(d_i(\sigma) - 2)}. \tag{5}$$

For a d regular tree, since all nodes have the same degree d, it follows from (5) that every permitted permutation σ has the same probability, independent of the source. Specifically, for any source v and permitted permutation σ

$$\begin{aligned}\mathbf{P}(\sigma \mid v) &= \prod_{k=1}^{N-1} \frac{1}{dk - 2(k-1)} \\ &\equiv p(d, N).\end{aligned}$$

From above, it follows immediately that for a d regular tree, for any G_N and candidate source v, $\mathbf{P}(G_N \mid v)$ is proportional to $|\Omega(v, G_N)|$. Formally, we shall denote the number of distinct permitted permutations $|\Omega(v, G_N)|$ by $R(v, G_N)$.

Definition 2: Given a graph G and vertex v of G, we define $R(v, G)$ as the total number of distinct permitted permutations of nodes of G that begin with node $v \in G$ and respect the graph structure of G.

In summary, the ML estimator for a regular tree becomes

$$\begin{aligned}\hat{v} &\in \arg\max_{v \in G_N} \mathbf{P}(G_N \mid v) \\ &= \arg\max_{v \in G_N} \sum_{\sigma \in \Omega(v, G_N)} \mathbf{P}(\sigma \mid v) \\ &= \arg\max_{v \in G_N} R(v, G_N) p(d, N) \\ &= \arg\max_{v \in G_N} R(v, G_N)\end{aligned} \tag{6}$$

with ties broken uniformly at random.

D. Rumor Source Estimator: General Trees

As (6) suggests, the ML estimator for a regular tree can be obtained by simply evaluating $R(v, G_N)$ for all v. However, as indicated by (5), such is not the case for a general tree with heterogeneous degree. This is because in the regular tree, all permitted permutations were equally likely, whereas in a general tree, different permitted permutations have different probabilities. To form an ML estimator for a general tree we would need to keep track of the probability of every permitted permutation. This could be computationally quite expensive because of the exponential number of terms involved. Therefore, we construct a simple heuristic to take into account the degree heterogeneity.

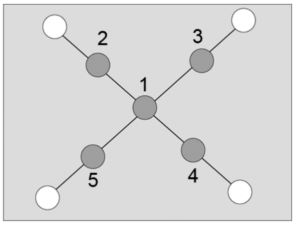

Fig. 2. Example network where rumor centrality with the BFS heuristic equals the likelihood $\mathbf{P}(G_N|v)$. The rumor infected nodes are in gray and labeled with numbers.

Our heuristic is based upon the following simple idea. The likelihood of a node is a sum of the probability of every permitted permutation for which it is the source. In general, these will have different values, but it may be that a majority of them have a common value. We then need to determine this value of the probability of the common permitted permutations. To do this, we assume the nodes receive the rumor in a breadth-first search (BFS) fashion. Roughly speaking, this corresponds to the fastest or most probable spreading of the rumor.

To calculate the BFS permitted permutation probability, we construct a sequence of nodes in a BFS fashion, with the source node fixed. For example, consider the network in Fig. 2. If we let node 2 be the source, then a BFS sequence of nodes would be $2, 1, 3, 4, 5$ and the probability of this permitted permutation is given by (5).

If we define the BFS permitted permutation with node v as the source as σ_v^*, then the rumor source estimator becomes (ties broken uniformly at random)

$$\hat{v} \in \arg\max_{v \in G_N} \mathbf{P}(\sigma_v^* \mid v) R(v, G_N). \tag{7}$$

We now consider an example to show the effect of the BFS heuristic. For the network in Fig. 2, the corresponding estimator value for node 1 is

$$\begin{aligned}\mathbf{P}(\sigma_1^* \mid v) R(1, G_N) &= \left(\frac{1}{4}\right)^4 4! \\ &= 6\left(\frac{1}{4}\right)^3\end{aligned}$$

and for node 2 it is

$$\begin{aligned}\mathbf{P}(\sigma_2^* \mid v) R(2, G_N) &= \frac{1}{2}\left(\frac{1}{4}\right)^3 3! \\ &= 3\left(\frac{1}{4}\right)^3.\end{aligned}$$

For comparison, the exact likelihood of node 1 is

$$\begin{aligned}\mathbf{P}(G_N|1) &= \mathbf{P}(\{1,2,3,4,5\}|1) + \mathbf{P}(\{1,2,3,5,4\}|1) \\ &\quad + \mathbf{P}(\{1,2,4,3,5\}|1) + \cdots + \mathbf{P}(\{1,5,4,3,2\}|1) \\ &= 24\left(\frac{1}{4}\right)^4 \\ &= 6\left(\frac{1}{4}\right)^3\end{aligned}$$

and for node 2 it is

$$\begin{aligned}\mathbf{P}(G_N|2) =& \mathbf{P}(\{2,1,3,4,5\}|2) + \mathbf{P}(\{2,1,3,5,4\}|2)\\ &+ \mathbf{P}(\{2,1,4,3,5\}|2) + \mathbf{P}(\{2,1,4,5,3\}|2)\\ &+ \mathbf{P}(\{2,1,5,3,4\}|2) + \mathbf{P}(\{2,1,5,4,3\}|2)\\ =& 6\frac{1}{2}\left(\frac{1}{4}\right)^3\\ =& 3\left(\frac{1}{4}\right)^3.\end{aligned}$$

In this case we find that the BFS heuristic equals the true likelihood for both nodes. Second, node 1 is only twice as likely as node 2 to be the source. However, if we look at the ratio of the rumor centralities of the nodes we find

$$\begin{aligned}\frac{R(1,G_N)}{R(2,G_N)} &= \frac{4!}{3!}\\ &= 4.\end{aligned}$$

Thus, the rumor centrality of node 1 is four times as large as that of node 2. What is happening is that without the BFS heuristic, rumor centrality is being fooled to always select higher degree nodes because it assumes all nodes have the same degree. Therefore, if a node only has a few infected neighbors (such as node 2), rumor centrality assumes that the node was not immediately infected and consequently did not have time to infect its neighbors. However, the BFS heuristic tries to compensate for the tendency of rumor centrality to favor higher degree nodes.

Indeed, as we shall see in Section V, this heuristic is an improvement over the naive extension of the estimator(6) for networks with very heterogeneous degree distributions. That is, biasing as per $\mathbf{P}\left(\sigma_v^* \mid v\right)$ in (7) is better than the unbiased version of it.

E. Rumor Source Estimator: General Graphs

The ML estimator for a general graph, in principle can be computed by following a similar approach as that for general trees. Specifically, it corresponds to computing the summation of the likelihoods of all possible *permitted permutations* given the network structure. This could be computationally prohibitive. Therefore, we propose a simple heuristic.

To that end, note that even in a general graph the rumor spreads along a spanning tree of the observed graph corresponding to the *first time* each node receives the rumor. Therefore, a reasonable approximation for computing the likelihood $\mathbf{P}(G_N|v)$ is as follows. First, suppose we know which spanning tree was involved in the rumor spreading. Then, using this spanning tree, we could apply the previously developed tree estimator. However, it is the lack of knowledge of the spanning tree that makes the rumor source estimation problem complicated.

We circumvent the issue of not knowing the underlying spanning tree as follows. We assume that if node $v \in G_N$ was the source, then the rumor spreads along a breadth first search (BFS) tree rooted at $v, T_{\text{bfs}}(v)$. The intuition is that if v was the source, then the BFS tree would correspond to the fastest (intuitively, most likely) spread of the rumor. Therefore, effectively we obtain the following rumor source estimator for a general rumor graph G_N:

$$\hat{v} \in \arg\max_{v \in G_N} \mathbf{P}\left(\sigma_v^* \mid v\right) R(v, T_{\text{bfs}}(v)). \quad (8)$$

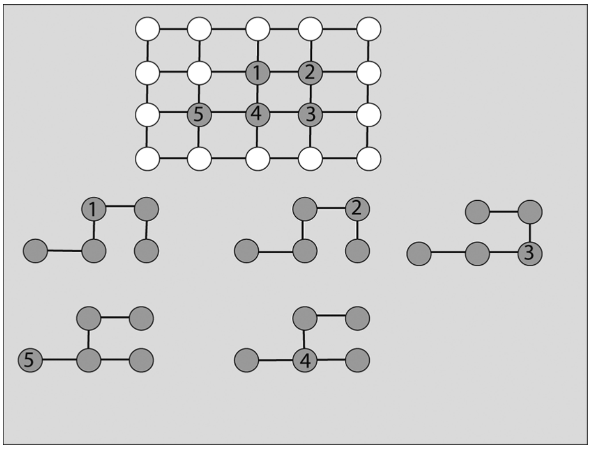

Fig. 3. Example network with a BFS tree for each node shown. The rumor infected nodes are shown in gray.

In this, ties are broken uniformly at random as before. Also, like in (7), the σ_v^* represents the BFS ordering of nodes in the tree $T_{\text{bfs}}(v)$.

For example, consider the network in Fig. 3. The BFS trees for each node are shown. Using the expression for $R(v, G_N)$ from Section III-A, the general graph estimator values for the nodes are

$$\begin{aligned}\mathbf{P}\left(\sigma_1^* \mid 1\right) R(1, T_{\text{bfs}}(1)) &= \frac{1}{4*6*8*10}\frac{5!}{20}\\ \mathbf{P}\left(\sigma_2^* \mid 2\right) R(2, T_{\text{bfs}}(2)) &= \frac{1}{4*6*8*10}\frac{5!}{30}\\ \mathbf{P}\left(\sigma_3^* \mid 3\right) R(3, T_{\text{bfs}}(3)) &= \frac{1}{4*6*8*10}\frac{5!}{20}\\ \mathbf{P}\left(\sigma_4^* \mid 4\right) R(4, T_{\text{bfs}}(4)) &= \frac{1}{4*6*8*10}\frac{5!}{10}\\ \mathbf{P}\left(\sigma_5^* \mid 5\right) R(5, T_{\text{bfs}}(5)) &= \frac{1}{4*6*8*10}\frac{5!}{40}.\end{aligned}$$

Node 4 maximizes this value and would be the estimate of the rumor source for this network. We will show with simulations that this general graph estimator performs well on different network topologies.

III. Rumor Centrality: Properties and Algorithm

The quantity $R(v, G_N)$ plays an important role in each of the rumor source estimators (6), (7), and (8). Recall that $R(v, G_N)$ counts the number of distinct ways a rumor can spread in the network G_N starting from source v. Thus, it assigns each node of G_N a nonnegative number or score. We shall call this number, $R(v, G_N)$, the **rumor centrality** of the node v with respect to G_N. The node with maximum rumor centrality will be called the **rumor center** of the network. Indeed, the rumor center is the ML estimation of the rumor source for regular trees.

This section describes ways to evaluate $R(v, G_N)$ efficiently when G_N is a tree. It also describes an important property of rumor centrality that will be useful in establishing our main results later in the paper. Further, we discuss a surprising relation

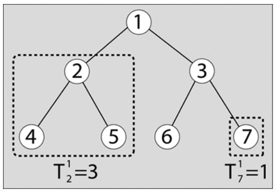

Fig. 4. Illustration of subtree variable T_u^v.

between the rumor center and the so called distance center of a tree. Finally, we remark on the relation between rumor centrality and the number of linear extensions of a partially ordered set described by a tree graph.

A. Rumor Centrality: Succinct Representation

Let G_N be a tree graph. Define T_u^v as the number of nodes in the subtree rooted at node u, with node v as the source. To illustrate this notation, a simple example is shown in Fig. 4. Here $T_2^1 = 3$ because there are 3 nodes in the subtree with node 2 as the root and node 1 as the source. Similarly, $T_7^1 = 1$ because there is only 1 node in the subtree with node 7 as the root and node 1 as the source.

We now can count the permitted permutations of G_N with v as the source. In the following analysis, we will abuse notation and use T_u^v to refer to both the subtrees and the number of nodes in the subtrees. Recall that we are looking for permitted permutations of N nodes of G_N. That is, we have N slots in a given permitted permutation, the first of which must be the source node v. The question is, how many distinct ways can we fill the remaining $N-1$ slots. The basic constraint is due to the causality induced by the tree graph that a node u must come before all the nodes in its subtree T_u^v. Given a slot assignment for all nodes in T_u^v subject to this constraint, there are $R(u, T_u^v)$ different ways in which these nodes can be ordered. This suggests a natural recursive relation between the rumor centrality $R(v, G_N)$ and the rumor centrality of its immediate children's subtrees $R(u, T_u^v)$ with $u \in \mathrm{child}(v)$. Here $\mathrm{child}(v)$ represents the set of all children of v in tree G_N assuming v as its root. Specifically, there is no constraint between the orderings of the nodes of different subtrees T_u^v with $u \in \mathrm{child}(v)$. This leads to

$$R(v, G_N) = (N-1)! \prod_{u \in \mathrm{child}(v)} \frac{R(u, T_u^v)}{T_u^v!}. \tag{9}$$

To understand the above expression, note that the number of ways to partition $N-1$ slots for different subtrees is $(N-1)! \prod_{u \in \mathrm{child}(v)} \frac{1}{T_u^v!}$, and the partition corresponding to T_u^v, $u \in \mathrm{child}(v)$ leads to $R(u, T_u^v)$ distinct orderings, thus resulting in (9).

If we expand this recursion (9) to the next level of depth in G_N we obtain

$$\begin{aligned} R(v, G_N) &= (N-1)! \prod_{u \in \mathrm{child}(v)} \frac{R(u, T_u^v)}{T_u^v!} \\ &= (N-1)! \prod_{u \in \mathrm{child}(v)} \frac{(T_u^v - 1)!}{T_u^v!} \prod_{w \in \mathrm{child}(u)} \frac{R(w, T_w^v)}{T_w^v!} \\ &= (N-1)! \prod_{u \in \mathrm{child}(v)} \frac{1}{T_u^v} \prod_{w \in \mathrm{child}(u)} \frac{R(w, T_w^v)}{T_w^v!}. \end{aligned}$$

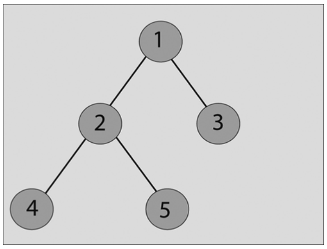

Fig. 5. Example network for calculating rumor centrality.

A leaf node l will have have 1 node and 1 permitted permutation, so $R(l, T_l^v) = 1$. If we continue this recursion until we reach the leaves of the tree, then we find that the number of permitted permutations for a given tree G_N rooted at v is

$$\begin{aligned} R(v, G_N) &= (N-1)! \prod_{u \in G_N \setminus v} \frac{1}{T_u^v} \\ &= N! \prod_{u \in G_N} \frac{1}{T_u^v}. \end{aligned} \tag{10}$$

In the last line, we have used the fact that $T_v^v = N$. We thus end up with a simple expression for rumor centrality in terms of the size of the subtrees of all nodes in G_N.

As an example of the use of rumor centrality, consider the network in Fig. 5. Using the rumor centrality formula, we find that the rumor centrality of node 1 is

$$R(1, G) = \frac{5!}{5 * 3} = 8.$$

Indeed, there are 8 permitted permutations of this network with node 1 as the source, which we list here.

$$\{1,3,2,4,5\}, \{1,2,3,4,5\}, \{1,2,4,3,5\}, \{1,2,4,5,3\},$$
$$\{1,3,2,5,4\}, \{1,2,3,5,4\}, \{1,2,5,3,4\}, \{1,2,5,4,3\}.$$

B. Rumor Centrality via Message-Passing

In order to find the rumor center of an N node tree G_N, we need to first find the rumor centrality of every node in G_N. To do this we need the size of the subtrees T_u^v for all v and u in G_N. There are N^2 of these subtrees. Therefore, a naive algorithm can lead to $\Omega(N^2)$ operations. We shall utilize a local relation between the rumor centrality of neighboring nodes in order to calculate it in $O(N)$ computation in a distributed, message-passing manner.

To this end, consider two neighboring nodes u and v in G_N. All of their subtrees will be the same size except for those rooted at u and v. In fact, there is a special relation between these two subtrees

$$T_u^v = N - T_v^u. \tag{11}$$

For example, in Fig. 4, for node 1, T_2^1 has 3 nodes, while for node 2, T_1^2 has $N - T_2^1$ or 4 nodes. Because of this relation, we can relate the rumor centralities of any two neighboring nodes

$$R(u, G_N) = R(v, G_N) \frac{T_u^v}{N - T_u^v}. \tag{12}$$

This result is the key to our algorithm for calculating the rumor centrality for all nodes in G_N. We first select any node v as the source node and calculate the size of all of its subtrees T_u^v and its rumor centrality $R(v, G_N)$. This can be done by having each node u pass two messages up to its parent. The first message is the number of nodes in u's subtree, which we call $t_{u \rightarrow \text{parent}(u)}^{up}$. The second message is the cumulative product of the size of the subtrees of all nodes in u's subtree, which we call $p_{u \rightarrow \text{parent}(u)}^{up}$. The parent node then adds the $t_{u \rightarrow \text{parent}(u)}^{up}$ messages together to obtain the size of its own subtree, and multiplies the $p_{u \rightarrow \text{parent}(u)}^{up}$ messages together to obtain its cumulative subtree product. These messages are then passed upward until the source node receives the messages. By multiplying the cumulative subtree products of its children, the source node will obtain its rumor centrality, $R(v, G_N)$.

With the rumor centrality of node v, we then evaluate the rumor centrality for the children of v using (12). Each node passes its rumor centrality to its children in a message we define as $r_{u \rightarrow u'}^{down}$ for $u' \in \text{child}(u)$. Each node u can calculate its rumor centrality using its parent's rumor centrality and its own subtree size T_u^v. We recall that the rumor centrality of a node is the number of permitted permutations that result in G_N. Thus, this message-passing algorithm is able to count the (exponential) number of permitted permutations for every node in G_N using only $O(N)$ computations. The pseudocode for this message-passing algorithm is included for completeness.

Algorithm 1 Rumor Centrality Message-Passing Algorithm

```
Choose a root node $v \in G_N$
for $u$ in $G_N$ do
    if u is a leaf then
        $t_{u \rightarrow \text{parent}(u)}^{up} = 1$
        $p_{u \rightarrow \text{parent}(u)}^{up} = 1$
    else
        if u is root v then
            $\forall\, v' \in \text{child}(v) r_{v \rightarrow v'}^{down} = \frac{N!}{N \prod_{j \in \text{child}(v)} p_{j \rightarrow v}^{up}}$
        else
            $t_{u \rightarrow \text{parent}(u)}^{up} = \sum_{j \in \text{child}(u)} t_{j \rightarrow u}^{up} + 1$
            $p_{u \rightarrow \text{parent}(u)}^{up} = t_{u \rightarrow \text{parent}(u)}^{up} \prod_{j \in \text{child}(u)} p_{j \rightarrow u}^{up}$
            $\forall\, u' \in \text{child}(u) r_{u \rightarrow u'}^{down} = r_{\text{parent}(u) \rightarrow u}^{down} \frac{t_{u \rightarrow \text{parent}(u)}^{up}}{N - t_{u \rightarrow \text{parent}(u)}^{up}}$
        end if
    end if
end for
```

C. A Property of Rumor Centrality

The following is an important characterization of the rumor center in terms of the sizes of its local subtrees. As we shall see, this will play a crucial role in establishing our main results about the performance of rumor centrality as an estimator for tree graphs.

Proposition 1: Given an N node tree, if node v^* is the rumor center, then any subtree with v^* as the source must have the following property:

$$T_v^{v^*} \leq \frac{N}{2}. \tag{13}$$

If there is a node u such that for all $v \neq u$

$$T_v^u \leq \frac{N}{2} \tag{14}$$

then u is a rumor center. Furthermore, a tree can have at most 2 rumor centers.

Proof: We showed that for a tree with N total nodes, for any neighboring nodes u and v

$$T_u^v = N - T_v^u. \tag{15}$$

For a node v one hop from v^*, we find

$$\frac{R(v,T)}{R(v^*,T)} = \frac{T_{v^*}^{v^*} T_v^{v^*}}{T_{v^*}^v T_v^v} = \frac{T_v^{v^*}}{(N - T_v^{v^*})}.$$

When v is two hops from v^*, all of the subtrees are the same except for those rooted at v, v^*, and the node in between, which we call node 1. Fig. 6 shows an example. In this case, we find

$$\frac{R(v,T)}{R(v^*,T)} = \frac{T_v^{v^*} T_1^{v^*}}{(N - T_1^{v^*})(N - T_v^{v^*})}.$$

Continuing this way, we find that in general, for any node v in T

$$\frac{R(v,T)}{R(v^*,T)} = \prod_{i \in \mathcal{P}(v^*,v)} \frac{T_i^{v^*}}{(N - T_i^{v^*})} \tag{16}$$

where $\mathcal{P}(v^*, v)$ is the set of nodes in the path between v^* and v, not including v^*. Now imagine that v^* is the rumor center. Then we have

$$\frac{R(v,T)}{R(v^*,T)} = \prod_{i \in \mathcal{P}(v^*,v)} \frac{T_i^{v^*}}{(N - T_i^{v^*})} \leq 1. \tag{17}$$

For a node v one hop from v^*, this gives us that

$$T_v^{v^*} \leq \frac{N}{2}. \tag{18}$$

For any node u in subtree $T_v^{v^*}$, we will have $T_u^{v^*} \leq T_v^{v^*} - 1$. Therefore, (18) will hold for any node $u \in T$. This proves the first part of Proposition 1.

Now assume that the node v^* satisfies (18) for all $v \neq v^*$. Then the ratios in (16) will all be less than or equal to 1. Thus, we find that

$$\frac{R(v,T)}{R(v^*,T)} = \prod_{i \in \mathcal{P}(v^*,v)} \frac{T_i^{v^*}}{(N - T_i^{v^*})} \leq 1. \tag{19}$$

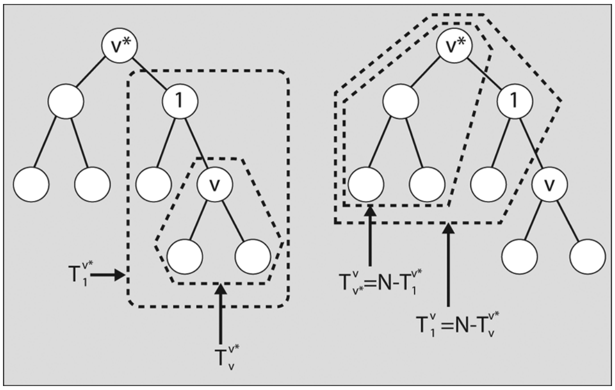

Fig. 6. T_i^j variables for source nodes 2 hops apart.

Thus, v^* is the rumor center, as claimed in the second part of Proposition 1.

Finally, assume that v^* is a rumor center and that all of its subtrees satisfy $T_v^{v^*} < N/2$. Then, any other node v will have at least one subtree that is larger than $N/2$, so v^* is the unique rumor center. Now assume that v^* has a neighbor v such that $T_v^{v^*} = N/2$. Then, $T_{v^*}^v = N/2$ also, and all other subtrees $T_u^v < N/2$, so v is also a rumor center. There can be at most 2 nodes in a tree with subtrees of size $N/2$, so a tree can have at most 2 rumor centers. ■

D. Rumor Centrality versus Distance Centrality

Here we shall compare rumor centrality with distance centrality that has become popular in the literature as a graph based score function for various other applications. To start with, we recall the definition of distance centrality. For a graph G, the distance centrality of node $v \in G$, $D(v, G)$, is defined as

$$D(v,G) = \sum_{j \in G} d(v,j) \tag{20}$$

where $d(v,j)$ is the shortest path distance from node v to node j. The distance center of a graph is the node with the smallest distance centrality. Intuitively, it is the node closest to all other nodes. On a tree, we will show the distance center is equivalent to the rumor center. Therefore, by establishing correctness of rumor centrality for tree graphs, one immediately finds that such is the case for distance centrality.

We will prove the following proposition for the distance center of a tree.

Proposition 2: On an N node tree, if v_D is the distance center, then, for all $v \neq v_D$

$$T_v^{v_D} \leq \frac{N}{2}. \tag{21}$$

Furthermore, if there is a unique rumor center on the tree, then it is equivalent to the distance center.

Proof: Assume that node v_D is the distance center of a tree T which has N nodes. The distance centrality of v_D is less than any other node. We consider a node v_ℓ which is ℓ hops from v_D, and label a node on the path between v_ℓ and v_D which is h hops from v_D by v_h. Now, because we are dealing with a tree, we have the following important property. For a node j which is in subtree $T_{v_h}^{v_D}$ but not in subtree $T_{v_{h+1}}^{v_D}$, we have $d(v_\ell, j) = d(v_D, j) + l - 2h$. Using this, we find

$$\begin{aligned}
D(v_D,T) &\leq D(v_\ell,T) \\
\sum_{j\in T} d(v_D,j) &\leq \sum_{j\in T} d(v_\ell,j) \\
&\leq \sum_{j\in T} d(v_D,j) + \ell(N - T_{v_1}^{v_D}) + \\
&\quad (\ell-2)(T_{v_1}^{v_D} - T_{v_2}^{v_D}) + \cdots + (\ell-2\ell)(T_{v_\ell}^{v_D}) \\
0 &\leq \ell N - 2T_{v_1}^{v_D} + 2T_{v_2}^{v_D} - 4T_{v_2}^{v_D} + \cdots \\
&\quad + 2(\ell-1)T_{v_l}^{v_D} - 2\ell T_{v_l}^{v_D} \\
0 &\leq (N - T_{v_1}^{v_D}) - T_{v_1}^{v_D} + (N - T_{v_2}^{v_D}) - T_{v_2}^{v_D} + \cdots \\
&\quad + (N - T_{v_l}^{v_D}) - T_{v_l}^{v_D} \\
\sum_{h=1}^{\ell} T_{v_h}^{v_D} &\leq \sum_{h=1}^{\ell} (N - T_{v_h}^{v_D}).
\end{aligned} \tag{22}$$

If we consider a node v_1 adjacent to v_D, we find the same condition we had for the rumor center. That is,

$$T_{v_1}^{v_D} \leq \frac{N}{2}. \tag{23}$$

For any node u in subtree $T_{v_1}^{v_D}$, we will have $T_u^{v_D} \leq T_{v_1}^{v_D} - 1$. Therefore, (23) will hold for any node $u \in T$. This proves the first half of Proposition 2.

If v_D is a rumor center, then, it also satisfies (23) as previously shown. Thus, when unique, the rumor center is equivalent to the distance center on a tree. This proves the second half of Proposition 2. ■

Now in contrast with trees, in a general nontree network, the rumor center and distance center need not be equivalent. Specifically, we shall define rumor centrality for a general graph to be the node with maximal value of rumor centrality on its own BFS tree. Stated more precisely, the rumor center of a general graph is the node $\hat{v}$ with the following property (ties broken uniformly at random):

$$\hat{v} \in \arg\max_{v \in G_N} R(v, T_{\text{bfs}}(v)). \tag{24}$$

In a general graph, as can be seen in Fig. 7, this general graph rumor center is not always equivalent to the distance center as it was for trees. We will see later that the general graph rumor center will be a better estimator of the rumor source than the distance center. The intuition for this is that the distance center is evaluated using only the shortest paths in the graph, whereas the general graph rumor centrality utilizes more of the network structure for estimation of the source.

E. Rumor Centrality and Linear Extensions of Posets

The rumor graph on a network can be viewed as a partially ordered set, or *poset*, of nodes if we fix a source node as the root and consider the network to be directed, with edges pointing from the node that had the rumor to the node it infected. These directed edges impose a partial order on the nodes. We have referred to any permutation of the nodes which satisfies this partial order as a permitted permutation. However, it is also known as a *linear extension* of the poset. It is known that counting the

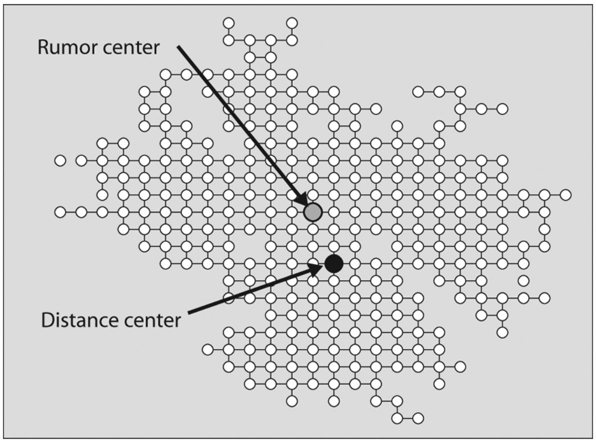

Fig. 7. A network where the distance center does not equal the general graph rumor center.

number of linear extensions of a poset is in general a very hard problem (specifically, it falls in the complexity class #P-complete [23]). However, on trees, counting linear extensions becomes computationally tractable. To the best of our knowledge, the fastest known algorithm for counting linear extensions on a tree requires $O(N^2)$ computation [22]. In contrast, the message-passing algorithm we presented in Section III-B required only $O(N)$ computation.

IV. Main Results: Theory

This section examines the behavior of the detection probability of the rumor source estimators for different graph structures. We establish that the asymptotic detection probability has a phase-transition effect: for linear graphs it is 0, while for trees which grow faster than a line it is strictly greater than 0. We will use different proof techniques to establish these results for trees with different rates of expansion. Throughout this section, we will be interested in the rumor source detection probability, which we formally define now.

Definition 3: Define $\mathcal{C}_t$ to be the event of correct rumor source detection using the rumor centrality based estimator after the rumor has spread for a time t on a graph $G(V, E)$.

A. Linear Graphs: No Detection

We first consider the detection probability for a linear graph, which is a regular tree of degree 2. We will establish the following result.

Theorem 1: Suppose the rumor starts spreading on a linear graph at time 0 as per the SI model. Then we have that

$$\mathbf{P}(\mathcal{C}_t) = O\left(\frac{1}{\sqrt{t}}\right).$$

As can be seen, the linear graph detection probability scales as $t^{-1/2}$, which goes to 0 as t goes to infinity. The intuition for this result is that the estimator provides very little information because of the linear graph's trivial structure.

B. Regular Expander Trees: Nontrivial Detection

We next consider detection on a regular degree expander tree. We assume each node has degree $d > 2$. For $d = 2$, the tree is a line, and we have seen that the detection probability goes to 0 as the network grows in size. For a regular tree with $d > 2$, we obtain the following result.

Theorem 2: Suppose the rumor starts spreading on a regular expander tree with degree $d > 2$ at time 0 as per the SI model. Then there exists a constant $\alpha_d > 0$ for all $d > 2$ so that

$$0 < \alpha_d \leq \liminf_t \mathbf{P}(\mathcal{C}_t) \leq \limsup_t \mathbf{P}(\mathcal{C}_t) \leq \frac{1}{2}.$$

Unlike linear graphs, when $d > 2$ then there is enough "complexity" in the network that allows us to perform detection of the rumor source with strictly positive probability irrespective of t (or size of the rumor network). The above result also says that the detection probability is always upper bounded by $1/2$ for any $d > 2$.

C. Degree 3 Regular Expander Trees: Exact Detection Probability

For regular trees of arbitrary degree $d > 2$, Theorem 2 states that the detection happens with strictly positive probability irrespective of the size of the network. However, we are unable to evaluate the exact asymptotic detection probability as $t \to \infty$. For $d = 3$, however we are able to obtain the exact value.

Theorem 3: Suppose the rumor starts spreading on a regular expander tree with degree $d = 3$ at time 0 as per the SI model. Then

$$\lim_t \mathbf{P}(\mathcal{C}_t) = \frac{1}{4}.$$

D. Geometric Trees: Correct Detection

The above stated results cover the case of regular trees. We now consider the detection probability of our estimator in nonregular trees. As a candidate class of such trees, we consider trees that grow polynomially. We shall call them geometric trees. These nonregular trees are parameterized by constants α, b, and c, with $0 < b \leq c$. We fix a source node v^* and consider each neighboring subtree of v^*. Let d^* be the degree of v^*. Then there are d^* subtrees of v^*, say $T_1, \ldots, T_{d^*}$. Consider the ith such subtree T_i, $1 \leq i \leq d^*$. Let v be any node in T_i and let $n^i(v, r)$ be the number of nodes in T_i at distance exactly r from the node v. Then we require that for all $1 \leq i \leq d^*$ and $v \in T_i$

$$br^\alpha \leq n^i(v, r) \leq cr^\alpha. \tag{25}$$

The condition imposed by (25) states that each of the neighboring subtrees of the source should satisfy polynomial growth (with exponent $\alpha > 0$) and regularity properties. The parameter $\alpha > 0$ characterizes the growth of the subtrees and the ratio c/b describes the regularity of the subtrees. If $c/b \approx 1$ then the subtrees are somewhat regular, whereas if the ratio is much greater than 1, there is substantial heterogeneity in the subtrees.

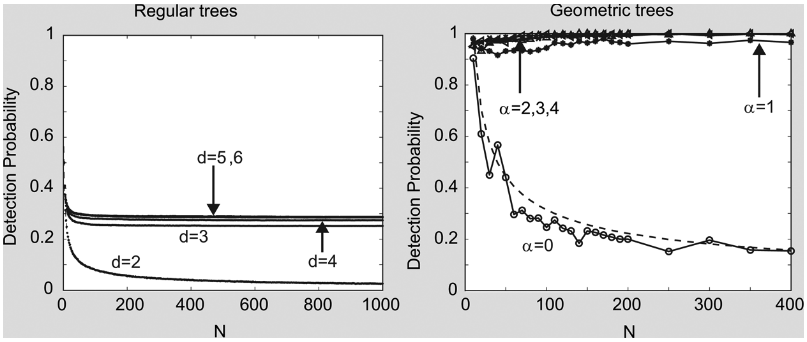

Fig. 8. Rumor centrality detection probability for regular trees (left) and geometric trees (right) versus number of nodes N. The dotted lines are plots of $N^{-1/2}$.

We note that unlike in regular trees, in a geometric tree the rumor centrality is not necessarily the ML estimator due to the heterogeneity. Nevertheless, we can use it as an estimator. Indeed, as stated below we find that the rumor centrality based estimator has an asymptotic detection probability of 1. That is, it is as good as the best possible estimator.

Theorem 4: Consider a geometric tree as described above with parameters $\alpha > 0$, $0 < b \leq c$ that satisfy (25) for a node v^* with degree $d_{v*} \geq 3$. Let the following condition be satisfied

$$d_{v^*} > \frac{c}{b} + 1.$$

Suppose the rumor starts spreading from node v^* at time 0 as per the SI model. Then

$$\liminf_t \mathbf{P}(\mathcal{C}_t) = 1.$$

This theorem says that $\alpha = 0$ and $\alpha > 0$ serve as a threshold for nontrivial detection: for $\alpha = 0$, the graph is essentially a line graph, so we would expect the detection probability to go to 0 as $t \to \infty$ based on Theorem 1, but for $\alpha > 0$ the detection probability converges to 1 as $t \to \infty$.

V. Simulation Results

This section provides simulation results for our rumor source estimators on different network topologies. These include synthetic topologies such as the popular scale-free and small-world networks, and also real topologies such as the Internet and the U.S. electric power grid.

A. Tree Networks

The detection probability of rumor centrality versus network size for different trees is show in Fig. 8. As can be seen, the detection probability decays as $N^{-1/2}$ as predicted in Theorem 1 for the graphs which grow like lines ($d = 2$ and $\alpha = 0$).

For regular degree trees we see that the detection probability is less than $1/2$ and for $d \geq 3$ it does not decay to 0, as predicted by Theorem 2. In fact, the detection probabilities appear to converge to asymptotic values. This value is $1/4$ for $d = 3$ as predicted by Theorem 3, and seems to increase by smaller amounts for $d = 4, 5, 6$.

For geometric trees with $\alpha > 0$, we see that the detection probability does not decay to 0 and is very close to 1 as predicted by Theorem 4.

B. Synthetic Networks

We performed simulations on synthetic small-world [24] and scale-free [25] networks. These are two very popular models for networks and so we would like our rumor source estimator to perform well on these topologies. For both topologies, the underlying graph contained 5000 nodes and in the simulations we let the rumor spread to 400 nodes.

Figs. 9 and 10 show an example of rumor spreading in a small-world and a scale-free network. The graphs show the rumor infected nodes in white. Also shown are the histograms of the estimator error for three different estimators. The estimators are distance centrality, rumor centrality on a BFS tree, and rumor centrality on a BFS tree with the BFS heuristic. For comparison, we also show with a dotted line a smooth fit of the histogram for the error from randomly choosing the source from the 400 node rumor network. As can be seen, for both networks, the histogram for the random guessing is shifted to the right of the estimator histograms. Thus, the centrality based estimators are a substantial improvement over random guessing for both small-world and scale-free networks.

The distance centrality estimator performs very similarly to the rumor centrality estimator. However, we see that on the small-world network, rumor centrality is better able to correctly find the source (0 error) than distance centrality (16% correct detection versus 2%). For the scale-free network used here, the average ratio of edges to nodes in the 400 node rumor graphs is 1.5 and for the small-world network used here, the average ratio is 2.5. For a tree, the ratio would be 1, so the small-world rumor graphs are less tree-like. This may explain why rumor centrality does better than distance centrality at correctly identifying the source on the small-world network.

The BFS heuristic leads to two visible effects. First, as can be seen for the scale-free network, we have a larger correct detection probability. Scale-free networks have power-law degree distributions, and thus contain many high degree *hubs*. The BFS heuristic works well in these types of networks because it was precisely designed for networks with heterogeneous degree distributions.

The second effect of the BFS heuristic is that larger errors become more likely. For both networks, the histograms spread out to higher errors. We see that for networks with less heterogeneous degree distributions, such as the small-world network, the BFS heuristic is actually degrading performance. It may be that for more regular networks, the BFS heuristic amplifies the

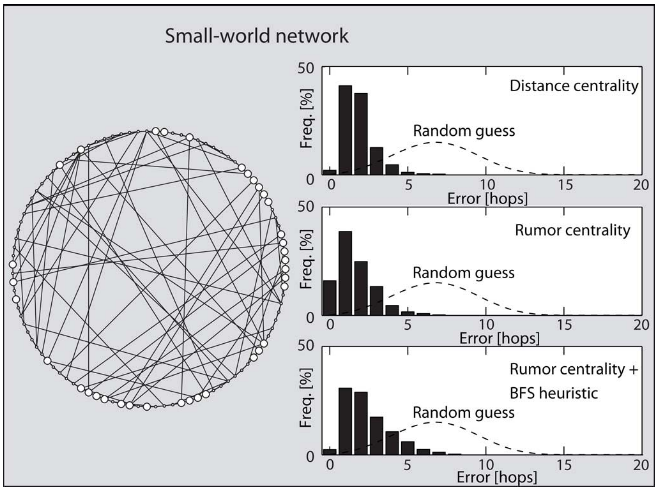

Fig. 9. Histograms of the error for distance centrality, rumor centrality, and rumor centrality with BFS heuristic estimators on a 400 node rumor network on a small-world network. The dotted line is a smooth fit of the histogram for randomly guessing the source in the rumor network. An example of a rumor graph (infected nodes in white) is shown on the right.

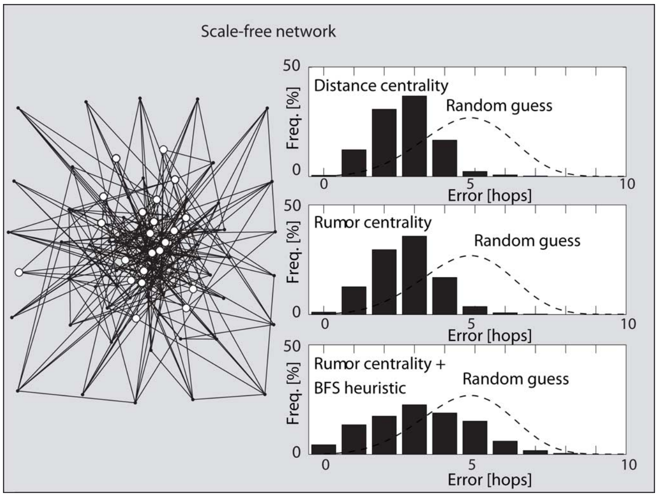

Fig. 10. Histograms of the error for distance centrality, rumor centrality, and rumor centrality with BFS heuristic estimators on a 400 node rumor network on a scale-free network. The dotted line is a smooth fit of the histogram for randomly guessing the source in the rumor network. An example of a rumor graph (infected nodes in white) is shown on the right.

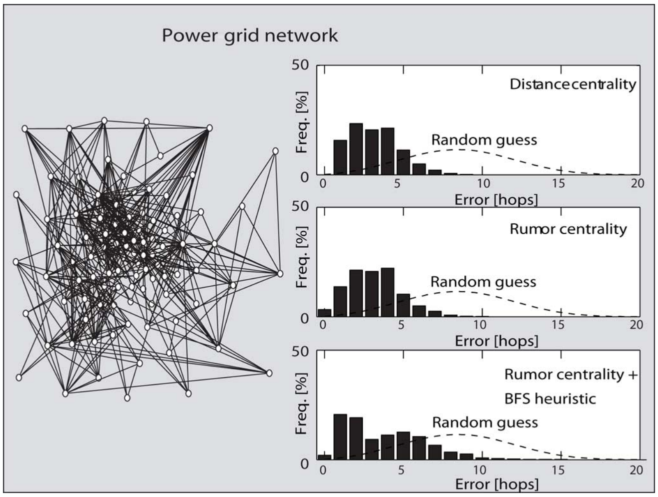

Fig. 11. Histograms of the error for distance centrality, rumor centrality, and rumor centrality with BFS heuristic estimators on a 400 node rumor network on the U.S. electric power grid network. The dotted line is a smooth fit of the histogram for randomly guessing the source in the rumor network. An example of a rumor graph (infected nodes in white) is shown on the right.

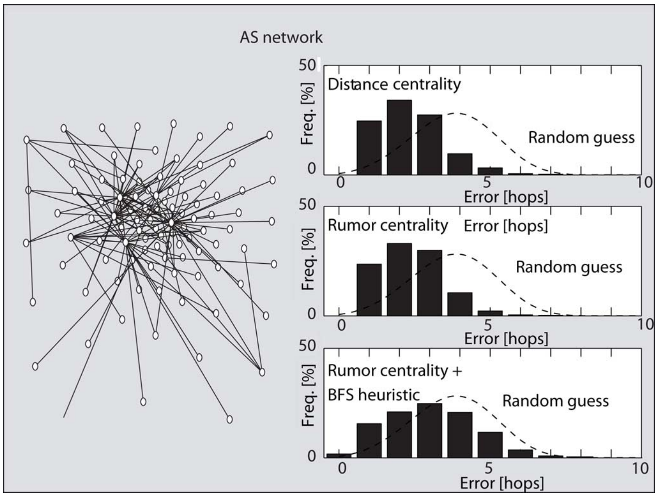

Fig. 12. Histogram of the error for distance centrality, rumor centrality, and rumor centrality with BFS heuristic estimators on a 400 node rumor network on an Internet autonomous system (AS) network. The dotted line is a smooth fit of the histogram for randomly guessing the source in the rumor network. An example of a rumor graph (infected nodes in white) is shown on the right.

effect any slight degree heterogeneity and causes detection errors.

C. Real Networks

We performed simulations on an Internet autonomous system (AS) network [26] and the U.S electric power grid network [25]. These are two important real networks so we would like our rumor source estimator to perform well on these topologies. The AS network contained 32,434 nodes and the power grid network contained 4,941 nodes. In the simulations we let the rumor spread to 400 nodes.

Figs. 11 and 12 show an example of rumor spreading in both of these networks. Also shown are the histograms of the estimator error for three different estimators. The estimators are distance centrality, rumor centrality on a BFS tree, and rumor centrality on a BFS tree with the BFS heuristic. For comparison, we also show with a dotted line a smooth fit of the histogram for the error from randomly choosing the source from the 400 node rumor network. As with the synthetic networks, the histogram for the random guessing is shifted to the right of the estimator histograms. Thus, on these real networks, the centrality based estimators are a substantial improvement over random guessing for both small-world and scale-free networks.

We see that rumor centrality and distance centrality have similar performance, but for the power grid network, rumor centrality is better able to correctly find the source than distance centrality (3% correct detection versus 0%). For the power grid network, the average ratio of edges to nodes in the 400 node rumor graphs is 4.2, and for the AS network the average ratio is 1.3. Thus, the rumor graphs on the power grid network are less tree-like. Similar to the small-world networks, this may explain why rumor centrality outperforms distance centrality on the power grid network.

The BFS heuristic improves the correct detection probability for the AS network. This is due to the fact that the AS network has many high degree hubs, similar to scale-free networks. However, in the powergrid network, the BFS heuristic spreads out the histogram to higher errors. Again, this may be due to the fact that the powergrid network does not have as much degree heterogeneity as the AS network, and the BFS heuristic is amplifying weak heterogeneities, similar to small-world networks.

VI. Proofs

This section establishes the proofs of Theorems 1–4. All of them utilize rumor centrality as the estimator to obtain the desired conclusion. To study the property of the rumor centrality, we shall utilize the property of rumor center as established in Proposition 1 crucially.

A. *Proof of Theorem 1*

We consider the spread of the rumor in a line network starting from a source, say v^*. We shall establish that for any time $t > 0$, the probability of the rumor center being equal to v^* decays as $O(1/\sqrt{t})$. Since a line is a regular graph and hence the rumor center is the ML estimator, it follows that the detection probability of any estimator decays as $O(1/\sqrt{t})$. This is because in the absence of any prior information (or uniform prior) the ML estimator minimizes the detection error (c.f see [27]).

Now rumor spreading in a line graph is equivalent to 2 independent Poisson processes with rate 1 beginning at the source and spreading in opposite directions. We refer to these processes as $N_1(t)$ and $N_2(t)$ with $N_1(0) = N_2(0) = 0$.

It follows from results of Section III -C that the rumor center of a line is the center of the line: if the line has an odd number of nodes then the rumor center is uniquely defined, else there are two rumor centers. Thus, we will correctly detect the source with probability 1 if the two Poisson processes on each side of the source have exactly the same number of arrivals and with probability $1/2$ if one of the Poisson processes is one less than the other. Then, the probability of the event of correct detection at time t, $\mathcal{C}_t$ is given by

$$\begin{aligned}\mathbf{P}(\mathcal{C}_t)=&\,\mathbf{P}(N_1(t) = N_2(t))\\ &+\frac{1}{2}\big(\mathbf{P}(N_1(t)=N_2(t)+1)+\mathbf{P}(N_1(t)+1=N_2(t))\big)\\ =&\sum_{k=0}^{\infty}\left(\mathbf{e}^{-t}\frac{t^k}{k!}\right)^2+\left(\mathbf{e}^{-t}\frac{t^k}{k!}\right)^2\frac{t}{k+1}.\end{aligned}$$

Let $a_k = e^{-2t}t^{2k}/(k!)^2$ and $b_k = a_k t/(k+1)$. Then

$$\mathbf{P}(\mathcal{C}_t) = \sum_{k=0}^{\infty} a_k + b_k.$$

We shall show that both $\sum_k a_k$ and $\sum_k b_k$ are bounded as $O(1/\sqrt{t})$. This will conclude the proof of Theorem 1.

To that end, first we consider summation of a_k. Let us consider the ratios of the successive terms:

$$\frac{a_k}{a_{k-1}} = \left(\frac{t}{k}\right)^2.$$

This ratio will be greater than 1 as long as $k \leq t$ and beyond that it will be less than 1. Thus, a_k is maximum for $k = \lfloor t \rfloor$. By Stirling's approximation, it follows that:

$$\begin{aligned}\log(a_{\lfloor t\rfloor}) =&-2\lfloor t\rfloor + 2\lfloor t\rfloor\log(\lfloor t\rfloor) - 2\log(\lfloor t\rfloor!)\\ =&-2\lfloor t\rfloor + 2\lfloor t\rfloor\log(\lfloor t\rfloor) - 2\lfloor t\rfloor\log(\lfloor t\rfloor)\\ &+2\lfloor t\rfloor - \log(\lfloor t\rfloor) + \Theta(1)\\ =&-\log(\lfloor t\rfloor) + \Theta(1).\end{aligned}$$

Therefore $a_{\lfloor t\rfloor} = \Theta\big(1/\lfloor t\rfloor\big)$. Given this, we shall bound all a_k relative to $a_{\lfloor t\rfloor}$ to obtain bound of $O(1/\sqrt{t})$ on $\sum_k a_k$.

To that end, since a_k is decreasing for $k \geq \lfloor t\rfloor$, we have that for any $k' \geq \lfloor t\rfloor$,

$$\sum_{k=k'}^{k'+\sqrt{t}} a_k \leq \sqrt{t}a_{k'}.$$

Similary, since for $k \leq \lfloor t\rfloor$, a_k is increasing we have that for any $k' \leq \lfloor t\rfloor - \sqrt{t}$,

$$\sum_{k=k'}^{k'+\sqrt{t}} a_k \leq \sqrt{t}a_{k'+\sqrt{t}}.$$

Given the above two inequalities, it will suffice to bound $a_{\lfloor t\rfloor+\ell\sqrt{t}}$ and $a_{\lfloor t\rfloor-\ell\sqrt{t}}$ for all $\ell \geq 0$ relate to $a_{\lfloor t\rfloor}$. For this, consider the following. For $k \geq \lfloor t\rfloor$

$$\begin{aligned}\frac{a_{k+\sqrt{t}}}{a_k} &= \left(\frac{t^{\sqrt{t}}}{\prod_{j=1}^{\sqrt{t}}(k+j)}\right)^2\\ &= \prod_{j=1}^{\sqrt{t}}\left(1+\frac{j+(k-t)}{t}\right)^{-2}\\ &\overset{(a)}{\leq} \prod_{j=1}^{\sqrt{t}}\left(1+\frac{j-1}{t}\right)^{-2}\\ &\overset{(b)}{\leq} \prod_{j=1}^{\sqrt{t}} e^{-\frac{(j-1)}{t}}\\ &= e^{-\frac{1}{2}+\frac{1}{2\sqrt{t}}}\\ &\leq e^{-\frac{1}{3}}\end{aligned}$$

for $t \geq 36$. In above (a) follows from $k \geq \lfloor t \rfloor \geq t-1$, (b) follows from $1+x \geq e^{x/2}$ for $x \in [0,1]$ and $j \leq \sqrt{t}$. Similarly, for $k \geq \lfloor t \rfloor - \sqrt{t}$

$$\begin{aligned}\frac{a_{k+\sqrt{t}}}{a_k} &= \left(\frac{t^{\sqrt{t}}}{\prod_{j=1}^{\sqrt{t}}(k+j)}\right)^2\\ &= \prod_{j=1}^{\sqrt{t}}\left(1+\frac{j+(k-t)}{t}\right)^{-2}\\ &\overset{(c)}{\geq} \prod_{j=1}^{\sqrt{t}}\left(1+\frac{j}{t}\right)^{-2}\\ &\overset{(d)}{\geq} \prod_{j=1}^{\sqrt{t}} e^{-\frac{j}{t}}\\ &\geq e^{-\frac{1}{2}}\end{aligned}$$

where (c) follows from $k \leq \lfloor t \rfloor \leq t$ and (b) from $1+x \leq e^x$ for $x \geq 0$. From above, it follows that

$$\begin{aligned}a_{\lfloor t \rfloor + \ell\sqrt{t}} &\leq e^{-\ell/3} a_{\lfloor t \rfloor}\\ a_{\lfloor t \rfloor - \ell\sqrt{t}} &\leq e^{-\ell/2} a_{\lfloor t \rfloor}.\end{aligned}$$

Using all of the above discussion, it follows that for $t \geq 36$

$$\begin{aligned}\sum_{k=1}^{\infty} a_k &\leq \sqrt{t} a_{\lfloor t \rfloor}\Big(\sum_{\ell=0}^{\infty} e^{-\ell/3} + e^{-\ell/2}\Big)\\ &= O\Big(a_{\lfloor t \rfloor}\sqrt{t}\Big)\\ &= O(1/\sqrt{t}).\end{aligned}$$

In a very similar manner, it can be shown that $\sum_k b_k = O(1/\sqrt{t})$. Therefore, it follows that the probability of correct detection is bounded as $O(1/\sqrt{t})$ in a line graph. This completes the proof of Theorem 1.

B. *Proof of Theorem 2*

We wish to establish that given a $d \geq 3$ regular tree, the probability of correct detection of the source using rumor centrality, irrespective of t, is uniformly lower bounded by a strictly positive constant, $\alpha_d > 0$ and upper bounded by $1/2$.

To find the lower bound on the detection probability, we shall utilize Proposition 1 which states the following. The source node has d subtrees and let the random variable $N_j(t) \geq 0$, $1 \leq j \leq d$, be the number of rumor infected nodes in the jth subtree at time t. If all $N_j(t)$ are strictly less than half the total number of rumor infected nodes at time t, then the source is the unique rumor center. Using this implication of Proposition 1, we obtain the following bound on the event of correct detection $\mathcal{C}_t$:

$$\left\{\omega \Big| \max_{1 \leq i \leq d} N_j(t,\omega) < \frac{1}{2}\left(\sum_{j=1}^{d} N_j(t,\omega) + 1\right)\right\} \subseteq \mathcal{C}_t. \quad (26)$$

The additional 1 is in the above expression to account for the source node. Therefore, by lower bounding the probability of the event on the left in (26), we shall obtain the desired lower bound $\mathbf{P}(\mathcal{C}_t)$. Since $N_j(t)$ are independent and identically distributed due to regularity of tree, we shall find the marginal distribution of $N_j(t)$. Now finding the precise "closed form" expression for the probability mass function of $N_j(t)$ for all $d \geq 3$ seems challenging. Instead, we shall obtain something almost close to that.

To that end, consider $N_1(t)$. Let T_n, $n \geq 1$, denote the time between the $n-1$st node getting infected and nth getting infected in $N_1(t)$. Since the source node is connected to the root of the first subtree via the edge along which rumor starts spreading as per an exponential distribution of rate 1, it follows that T_1 has an exponential distribution of rate 1. Once the first node gets infected in the subtree, the number of edges along which the rumor can spread further is $d-1 = 1+(d-2)$. More generally, every time a new node gets infected, it brings in new $d-1$ edges and removes 1 edge along with rumor can spread in the subtree. Now the spreading time along all edges is independent and identically distributed as per an exponential distribution of rate 1 and exponential random variables have the 'memoryless' property: if X is an exponential random variable with rate 1, then $\mathbf{P}(X > \zeta + \eta | X > \eta) = \mathbf{P}(X > \zeta)$ for any $\zeta, \eta \geq 0$. From above, we can conclude that T_n equals the minimum of $1+(d-2)(n-1)$ independent exponential random variables of rate 1. By the property of the exponential distribution, this equals an exponential random variable of rate $1+(d-2)(n-1)$. Now let S_n be the total time for n nodes to get infected in the subtree, that is

$$S_n = \sum_{i=1}^{n} T_i. \quad (27)$$

We state the following Lemma which states the precise density of S_n and some of its useful properties. It's proof is presented later in the section.

Lemma 1: The density of S_n for a degree d regular tree, $f_{S_n}(t)$ is given by

$$f_{S_n}(t) = \begin{cases} e^{-t} & \text{for } n = 1\\ \Big(\prod_{i=1}^{n-1}\Big(1+\frac{1}{ai}\Big)\Big)e^{-t}(1-e^{-at})^{n-1} & \text{for } n \geq 2, \end{cases} \quad (28)$$

where $a = d-2$. Further, let $\tau_n = \frac{1}{a}\big(\log(n) + \log(\frac{3}{4a})\big)$ and $t_n = \frac{1}{a}\log((n-1)a+1)$. Then

1. $\tau_n \leq t_{n-1} \leq t_n$ for all $n \geq 2$.
2. For all $n \geq 1$ and $t \in (0, t_n)$,

$$\frac{df_{S_n}(t)}{dt} > 0.$$

3. There exists finite constants $B_a, C_a > 0$ so that

$$\liminf_n f_{S_n}(\tau_n) \geq B_a, \text{ and } \limsup_n f_{S_n}(t_n) \leq C_a.$$

4. There exists $\gamma \in (0,1)$ so that for all $t \in (0, t_n)$

$$\limsup_n \frac{f_{S_{(d-1)n}}(t)}{f_{S_n}(t)} \leq (1-\gamma).$$

Next we use Lemma 1 to obtain a lower bound on $\mathbf{P}(\mathcal{C}_t)$. For this, define $\mathcal{D}_n(t)$ as the event under which all the d subtrees have between n and $(d-1)n$ infected nodes at time t. That is,

$$\mathcal{D}_n(t) = \bigcap_{j=1}^{d} \left\{ n \leq N_j(t) \leq (d-1)n \right\}, \ \ \text{for}n \geq 0. \tag{29}$$

Under $\mathcal{D}_n(t)$ for any $n \geq 0$, it follows that $N_j(t) \leq \frac{1}{2}\sum_{j'=1}^{d} N_{j'}(t)$ for all $1 \leq j \leq d$. That is, event $\mathcal{C}_t$ holds. Therefore

$$\mathbf{P}(\mathcal{C}_t) \geq \sup_{n \geq 0} \mathbf{P}(\mathcal{D}_n(t)). \tag{30}$$

Therefore, to uniformly lower bound $\mathbf{P}(\mathcal{C}_t)$ and establish Theorem 2, it is sufficient to find $n(t)$ so that $\mathbf{P}(\mathcal{D}_{n(t)}(t)) \geq \alpha_d > 0$ for all t large enough. To that end, we shall first study how to bound $\mathbf{P}(\mathcal{D}_n(t))$. Under $\mathcal{D}_n(t)$, we have $n \leq N_j(t) \leq (d-1)n$ for $1 \leq j \leq d$. Now consider $N_1(t)$. For $n \leq N_1(t) \leq (d-1)n$, it must be that the nth node in it must have got infected before t and the $(d-1)n+1$st node must have got infected after t. Using this along with the independent and identical distribution of $N_j(t)$ for $1 \leq j \leq d$, we obtain

$$\begin{aligned}
\mathbf{P}\big(\mathcal{D}_n(t)\big) &= \mathbf{P}\Big(\bigcap_{j=1}^{d} n \leq N_j(t) \leq (d-1)n\Big) \\
&= \mathbf{P}\Big(n \leq N_1(t) \leq (d-1)n\Big)^d \\
&= \mathbf{P}\Big(S_n \leq t < S_{(d-1)n}\Big)^d \\
&= \Big(\mathbf{P}\big(S_n \leq t\big) - \mathbf{P}\big(S_{(d-1)n} \leq t\big)\Big)^d \\
&= \left(\int_0^t \big(f_{S_n}(\tau) - f_{S_{(d-1)n}}(\tau)\, d\tau\big)\right)^d.
\end{aligned} \tag{31}$$

We shall use Lemma 1 to uniformly lower bound (31) by a strictly positive constant $\alpha_d > 0$ for all t large enough. To that end, consider a large enough t and hence $n \geq 2$ large enough so that using Lemma 1, we have for any given small enough $\delta \in (0,1)$: (a) $\tau_n \leq t_{n-1} \leq t \leq t_n$, (b) $f_{S_n}(\tau_n) \geq B_a(1-\delta)$, (c) $f_{S_n}(t_n) \leq C_a(1+\delta)$ and (d) $f_{S_{(d-1)n}}(t) \leq \big(1-\gamma(1-\delta)\big) f_{S_n}(t)$ for all $t \in (0, t_n)$. Then using $\gamma' = \gamma(1-\delta)$

$$\begin{aligned}
&\int_0^t (f_{S_n}(\tau) - f_{S_{(d-1)n}}(\tau))\, d\tau \\
&\geq \int_0^t \big(f_{S_n}(\tau) - \big(1-\gamma'\big) f_{S_n}(\tau)\big)\ d\tau \\
&\geq \gamma' \int_0^t f_{S_n}(\tau)\, d\tau \\
&\geq \gamma' \Big(\int_{\tau_n}^{t_n} f_{S_n}(\tau)\, d\tau - \int_{t_{n-1}}^{t_n} f_{S_n}(\tau)\, d\tau\Big) \\
&\geq \gamma' \Big(f_{S_n}(\tau_n)(t_n - \tau_n) - f_{S_n}(t_n)(t_n - t_{n-1})\Big) \\
&\geq \gamma(1-\delta)^2 B_a \frac{1}{a} \log\left(\frac{4}{3}\right) \\
&\quad - \gamma C_a (1-\delta^2) \frac{1}{a} \log\left(\frac{(n-1)a+1}{(n-2)a+1}\right).
\end{aligned}$$

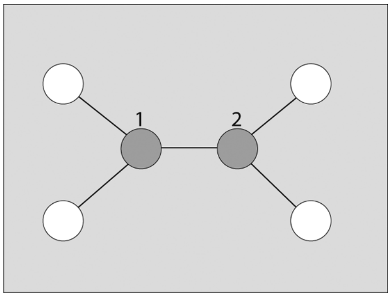

Fig. 13. Symmetric 2 node rumor network in a regular tree.

Above we have used the fact that for $n \geq 1$, $t_n - \tau_n \geq \frac{1}{a} \log \frac{4a}{3}$ and $a = d - 2 \geq 1$. As $t \to \infty$, the corresponding n so that $\tau_n \leq t_{n-1} \leq t \leq t_n$ increases to ∞ as well. Since the choice of $\delta \in (0,1)$ is arbitrary, from above along with (30) and(31), it follows that:

$$\begin{aligned}
\liminf_{n} \mathbf{P}(\mathcal{C}_t) &\geq \left(\gamma B_a \frac{1}{a} \log\left(\frac{4}{3}\right)\right)^d \\
&\triangleq \alpha_d > 0.
\end{aligned}$$

Next, we consider the upper bound of $1/2$. This bound can be obtained by using symmetry arguments. First imagine that the rumor has spread to two nodes. First, because of the memoryless property of the spreading times, the spreading process essentially resets after the second node is infected, so we can treat these two nodes as just a single, enlarged rumor source. Second, because of the regularity of the tree, the rumor boundary is symmetric about these two nodes, as shown in Fig. 13. Therefore, within this enlarged rumor source, the estimator will not be able to distinguish between these two nodes due to symmetry. For example, in Fig. 13, the estimator will select node 1 or node 2 with equal probability. In the best scenario, the estimator will detect this enlarged rumor source exactly with probability 1. This happens for example, when the rumor network only has 2 nodes as in Fig. 13. Then due to symmetry, the probability of correctly detecting the source is $1/2$ since each node in the enlarged rumor source is chosen with equal probability. The probability of the estimator detecting the enlarged source is no greater than 1 ever, so the correct detection probability can never be greater than $1/2$.

This completes the proof of Theorem 2.

Proof of Lemma 1: We derive the density by induction. For $n = 1$, we trivially have

$$f_{S_1}(t) = e^{-t}. \tag{32}$$

Now, inductively assume that f_{S_n} has the form as claimed in Lemma 1. That is

$$f_{S_n}(\tau) = C(n) e^{-\tau} \big(1 - e^{-a\tau}\big)^{n-1}, \ \text{for } \tau \geq 0$$

with $C(n) = \prod_{i=1}^{n-1}\left(1+\frac{1}{ai}\right)$. Now $S_{n+1} = S_n + T_{n+1}$; T_{n+1} is independent of S_n and has exponential distribution with rate $1+(d-2)n = 1+an$. Therefore

$$\begin{aligned}
f_{S_{n+1}}(t) &= \int_0^t f_{S_n}(\tau) f_{T_{n+1}}(t-\tau)\, d\tau \\
&= C(n)(1+an)\int_0^t e^{-\tau}(1-e^{-a\tau})^{n-1} e^{-(1+an)(t-\tau)}\, d\tau \\
&= C(n+1)ane^{-(1+an)t}\Big(\int_0^t e^{an\tau}(1-e^{-a\tau})^{n-1} d\tau\Big).
\end{aligned}$$

Expanding $(1-e^{-a\tau})^{n-1} = \sum_i \binom{n-1}{i}(-1)^i e^{-ai\tau}$ and integrating, we obtain

$$\begin{aligned}
\int_0^t e^{an\tau}(1-e^{-a\tau})^{n-1} d\tau &= \sum_{i=0}^{n-1}\binom{n-1}{i}(-1)^i \frac{e^{a(n-i)t}-1}{a(n-i)} \\
&= \frac{1}{an}\sum_{i=0}^{n-1}\binom{n}{i}(-1)^i\big(e^{a(n-i)t}-1\big).
\end{aligned}$$

From this and above, we obtain

$$\begin{aligned}
&f_{S_{n+1}}(t) \\
&= C(n+1)e^{-(1+an)t}\sum_{i=0}^{n-1}\binom{n}{i}(-1)^i\big(e^{a(n-i)t}-1\big) \\
&= C(n+1)e^{-t}\sum_{i=0}^{n-1}\binom{n}{i}(-1)^i\big(e^{-ait}-e^{-ant}\big) \\
&= C(n+1)e^{-t}\Big(\Big(\sum_{i=0}^{n}\binom{n}{i}\big(-e^{-at}\big)^i\Big)-\big(-e^{-at}\big)^n\Big) \\
&\quad - C(n+1)e^{-t}e^{-ant}\Big(\Big(\sum_{i=0}^{n}\binom{n}{i}(-1)^i\Big)-(-1)^n\Big) \\
&= C(n+1)e^{-t}\Big((1-e^{-at})^n-(-e^{-at})^n+(-e^{-at})^n\Big) \\
&= C(n+1)e^{-t}(1-e^{-at})^n.
\end{aligned}$$

This is precisely the form claimed by Lemma 1. This completes the induction step and establishes the form of density of S_n as claimed. Next we establish four properties claimed in Lemma 1.

Property 1: Let $n \geq 8$. Clearly, $t_{n-1} \leq t_n$ as $a = d-2 \geq 1$ for $d \geq 3$. Now for $n \geq 8$, we have $an/4 \geq 2a-1$. Therefore, $(n-2)a+1 \geq 3an/4$. Therefore, it follows that $\tau_n \leq t_{n-1}$.

Property 2: Consider any $n \geq 1$ and $t \in (0, t_n)$ with $t_n = \frac{1}{a}\log((n-1)a+1)$. Now

$$\begin{aligned}
&t < t_n \\
&\Rightarrow t < \frac{1}{a}\log((n-1)a+1) \\
&\Rightarrow 0 < e^{-at}((n-1)a+1)-1 \\
&\Rightarrow 0 < \frac{df_{S_n}(t)}{dt}.
\end{aligned}$$

In above, we have used the form of f_{S_n} established earlier.

Property 3: Our interest is in obtaining uniform upper bound on $f_{S_n}(t_n)$ and uniform lower bound on $f_{S_n}(\tau_n)$. To that end, let us start with the following standard inequality.

$$\begin{aligned}
1+\frac{1}{a}\log n &\geq \frac{1}{a}\sum_{i=1}^{n-1}\frac{1}{i} \\
&\geq \log\left(\prod_{i=1}^{n-1}\left(1+\frac{1}{ai}\right)\right) \\
&\geq \frac{1}{a}\sum_{i=1}^{n-1}\left(\frac{1}{i}-\frac{1}{i^2}\right) \\
&\geq \frac{1}{a}\log(n-1)-\frac{\zeta(2)}{a}
\end{aligned}$$

where $\zeta(2) = \sum_{i=1}^{\infty} i^{-2} = \pi^2/6$. Recall that $C(n) = \prod_{i=1}^{n-1}\left(1+\frac{1}{ai}\right)$. Therefore, from above we have

$$en^{\frac{1}{a}} \geq C(n) \geq (n-1)^{\frac{1}{a}}e^{-\frac{\zeta(2)}{a}}. \tag{33}$$

Recalling $\tau_n = \frac{1}{a}\log\frac{3n}{4a}$ and from above, $f_{S_n}(\tau_n)$ can be lower bounded as

$$\begin{aligned}
f_{S_n}(\tau_n) &= C(n)e^{-\tau_n}\big(1-e^{-a\tau_n}\big)^{n-1} \\
&= C(n)\Big(\frac{4a}{3n}\Big)^{\frac{1}{a}}\Big(1-\frac{4a}{3n}\Big)^{n-1} \\
&\geq e^{-\frac{\zeta(2)}{a}}\Big(\frac{4a(n-1)}{3n}\Big)^{\frac{1}{a}}\Big(1-\frac{4a}{3n}\Big)^{n-1}.
\end{aligned}$$

Now the term $e^{-\frac{\zeta(2)}{a}}$ is strictly positive constant; $(\frac{n-1}{n})^{\frac{1}{a}} \to 1$ as $n \to \infty$; and $\big(1-\frac{4a}{3n}\big)^{n-1} \to e^{-\frac{4a}{3}}$ as $n \to \infty$. Therefore, it follows that

$$\begin{aligned}
\liminf_{n\to\infty} f_{S_n}(\tau_n) &\geq e^{-\frac{\zeta(2)}{a}}\Big(\frac{4a}{3}\Big)^{\frac{1}{a}}e^{-\frac{4a}{3}} \\
&\triangleq B_a > 0.
\end{aligned}$$

Similarly, for $t_n = \frac{1}{a}\log((n-1)a+1)$

$$\begin{aligned}
f_{S_n}(t_n) &= C(n)e^{-t_n}\big(1-e^{-at_n}\big)^{n-1} \\
&= C(n)\Big(\frac{1}{(n-1)a+1}\Big)^{\frac{1}{a}}\Big(1-\frac{1}{(n-1)a+1}\Big)^{n-1} \\
&\leq e\Big(\frac{n}{(n-1)a+1}\Big)^{\frac{1}{a}}\Big(1-\frac{1}{(n-1)a+1}\Big)^{n-1} \\
&\xrightarrow{n\to\infty} ea^{-\frac{1}{a}}e^{-\frac{1}{a}} \\
&\triangleq C_a < \infty.
\end{aligned}$$

Thus

$$\limsup_{n} f_{S_n}(t_n) \leq C_a < \infty. \tag{34}$$

Property 4: To establish this property, recall that for $m \geq 2$

$$\begin{aligned}
\log f_{S_m}(t) = \sum_{i=1}^{m-1}\log\Big(1+\frac{1}{ai}\Big)-t \\
+(m-1)\log\Big(1-e^{-at}\Big).
\end{aligned}$$

Therefore (using $a = d - 2$)

$$\begin{aligned}\log\left(\frac{f_{S_{(d-1)n}}(t)}{f_{S_n}(t)}\right)\\ &= \sum_{i=n}^{(a+1)n-1} \log\left(1+\frac{1}{ai}\right) + an\log(1-e^{-at})\\ &\le \frac{1}{a}\Big(\sum_{i=n}^{(a+1)n-1}\frac{1}{i}\Big) + an\log(1-e^{-at})\\ &\le \frac{1}{a}\Big(\log\Big(\frac{(a+1)n}{n}\Big)\Big) + an\log(1-e^{-at}).\\ &= \frac{1}{a}\log(a+1) + an\log(1-e^{-at}).\end{aligned}$$

For $t \in (0, t_n)$, $e^{-at} \ge e^{-at_n}$ and hence

$$e^{-at} \ge e^{-at_n} = \frac{1}{(n-1)a+1} \ge \frac{1}{na},$$

since $a \ge 1$. Therefore, for $t \in (0, t_n)$ we have

$$\begin{aligned}\log\left(\frac{f_{S_{(d-1)n}}(t)}{f_{S_n}(t)}\right)\\ &\le \frac{1}{a}\log(a+1) + an\log\Big(1-\frac{1}{an}\Big)\\ &\overset{(x)}{\le} \frac{1}{a}\log(a+1) + an\Big(-\frac{1}{an}+\frac{1}{2a^2n^2}\Big)\\ &\le \frac{1}{a}\log(a+1) - 1 + \frac{1}{2an},\end{aligned}$$

where (x) follows from $\log(1-x) \le -x + x^2/2$ for all $x \in (0,1)$ and $an \ge 1$. Now $\frac{1}{2an} \to 0$ as $n \to \infty$; $\log(a+1) - a$ is a decreasing function in $a \ge 0$ and for $a = 1$, it is $\log 2 - 1 \le 1/6$. Therefore, it follows that there exists a $\gamma \in (0,1)$ so that for any $a \ge 1$

$$\begin{aligned}\limsup_{n\to\infty}\log\left(\frac{f_{S_{(d-1)n}}(t)}{f_{S_n}(t)}\right) &\le \frac{1}{a}\log(a+1) - 1\\ &\le \log(1-\gamma).\end{aligned}$$

This completes the proof of Lemma 1. ■

C. Proof of Theorem 3

We are interested in degree $d = 3$ regular tree. By Proposition 1, the event of correct detection at time t, $\mathcal{C}_t$ is such that

$$\left\{\omega\Big|\max_{j\in 1,2,3} N_j(t,\omega) < \frac{1}{2}\Big(\sum_{j=1}^{3} N_j(t,\omega)+1\Big)\right\} \subseteq \mathcal{C}_t \quad (35)$$

and

$$\mathcal{C}_t \subseteq \left\{\omega\Big|\max_{j\in 1,2,3} N_j(t,\omega) \le \frac{1}{2}\Big(\sum_{j=1}^{3} N_j(t,\omega)+1\Big)+1\right\}.$$

If we knew the exact form for the distribution of the number of arrivals at time t, $N_j(t)$ for $1 \le j \le 3$, then we could bound the probability of $\mathcal{C}_t$ explicitly. For regular trees with degree 3, this is indeed possible. To that end, by Lemma 1 we find that the distribution of the time for n nodes to get rumor infected in the jth subtree, for $1 \le j \le 3$ is given by

$$f_{S_n}(t) = ne^{-t}(1-e^{-t})^{n-1}, \text{ for } n \ge 1.$$

Now in order for there to be exactly n rumor infected nodes in the jth subtree by time t, the nth node must get infected before t and $n+1$st node must get infected after t. Therefore, the distribution of $N_j(t)$, for $1 \le j \le 3$, is

$$\begin{aligned}\mathbf{P}\big(N_j(t)=n\big) &= \mathbf{P}\big(S_n \le t\big) - \mathbf{P}\big(S_{n+1} \le t\big)\\ &= \int_0^t ne^{-\tau}(1-e^{-\tau})^{n-1}d\tau\\ &\quad - \int_0^t (n+1)e^{-\tau}(1-e^{-\tau})^{n}d\tau\\ &= e^{-t}(1-e^{-t})^{n}\end{aligned}$$

for $n \ge 1$. Indeed,

$$\mathbf{P}\big(N_j(t)=0\big) = \mathbf{P}(S_1 > t) = e^{-t}.$$

If we denote e^{-t} by p then the above becomes

$$\mathbf{P}\big(N_j(t)=n\big) = p(1-p)^n, \text{ for } n \ge 0. \quad (36)$$

That is, $N_i(t)$ has a geometric distribution with parameter $p = e^{-t}$. Next we evaluate lower bound on $\mathbf{P}(\mathcal{C}_t)$ using (35). To that end define $\mathcal{H}_t^n$ as

$$\mathcal{H}_t^n = \left\{\omega\Big|\sum_{j=1}^{3} N_j(t,\omega) = n-1 \bigcap \max_{j=1,2,3} N_j(t,\omega) < \frac{n}{2}\right\}.$$

The elements of $\mathcal{H}_t^n$ correspond to the event where there are n infected nodes, and each subtree has less than $n/2$ nodes. Also, the sets $\mathcal{H}_t^n$ are disjoint because if one occurs for a specific t and n, then the others did not occur. We will also define the set H^n as

$$H^n = \left\{n_1, n_2, n_3 \in \mathbf{N}\Big|\sum_{j=1}^{3} n_j = n-1 \bigcap \max_{j=1,2,3} n_j < \frac{n}{2}\right\}.$$

The elements of H^n correspond to nonnegative integers n_1, n_2, and n_3 that sum to $n-1$ and are less than $n/2$. Then from (35)

$$\bigcup_{n=1}^{\infty} \mathcal{H}_t^n \subseteq \mathcal{C}_t. \quad (37)$$

Therefore, using the fact that the spreading on subtrees happens independently, we obtain

$$\begin{aligned}\mathbf{P}(\mathcal{C}_t) &\ge \sum_{n=1}^{\infty} \mathbf{P}(\mathcal{H}_t^n)\\ &\ge \sum_{n=1}^{\infty} \sum_{n_1,n_2,n_3 \in H^n} \prod_{j=1}^{3} P(N_j(t)=n_j)\\ &= \sum_{n=1}^{\infty} \sum_{n_1,n_2,n_3 \in H^n} p^3(1-p)^{n_1+n_2+n_3}\\ &= p^3 \sum_{n=0}^{\infty} (1-p)^n \sum_{n_1,n_2,n_3 \in H^n} 1. \quad (38)\end{aligned}$$

The sum over the n_j's require us to count the number of states in H^n. It follows that:

$$\begin{aligned}|H^n| &\ge 3! \times \Big|\Big\{(n_1,n_2,n_3) : n_3 < n_2 < n_1 < \frac{n}{2}\Big\}\Big|\\ &= 6\Big|\Big\{(n_1,n_2,n_3) : n_3 < n_2 < n_1 < \frac{n}{2}\Big\}\Big|.\end{aligned}$$

Now when $n_3 < n_2 < n_1 < n/2$, it must be that $n/3 < n_1 < n/2$; for a given such n_1, $(n - n_1)/2 < n_2 < n_1$ and $n_3 = n - 1 - n_1 - n_2$. Using these relations, it follows that the number of such (n_1, n_2, n_3) triples are $\frac{n^2}{48} + O(n)$. Therefore, $|H^n|$ is at least $\frac{n^2}{8} + O(n)$. Using this, we obtain

$$\begin{aligned}\mathbf{P}(\mathcal{C}_t) &\geq p^3\Big(\sum_{n=0}^{\infty}(1-p)^n\Big(\frac{n^2}{8} + O(n)\Big)\Big)\\ &= \frac{1}{8}p^3(1-p)^2\Big(\sum_{n=0}^{\infty}(n+2)(n+1)(1-p)^n\Big)\\ &\quad + p^3(1-p)O\Big(\sum_{n=0}^{\infty}n(1-p)^{n-1}\Big)\\ &\quad + p^3 O\Big(\sum_{n=0}^{\infty}(1-p)^n\Big)\\ &= \frac{p^3(1-p)^2}{8}\frac{2}{p^3} + p^3(1-p)O(p^{-2}) + p^3 O(p^{-1})\\ &= \frac{(1-p)^2}{4} + O\big(p(1-p) + p^2\big).\end{aligned}$$

Now since $p = e^{-t}$, as $t \to \infty$, $p \to 0$. Therefore, we obtain that

$$\liminf_{t\to\infty} \mathbf{P}(\mathcal{C}_t) \geq \frac{1}{4}.$$

In a very similar manner, using (36) it follows that

$$\begin{aligned}\mathbf{P}(\mathcal{C}_t) &\leq p^3\Big(\sum_{n=0}^{\infty}(1-p)^n\Big(\frac{n^2}{8} + O(n)\Big)\Big)\\ &= \frac{(1-p)^2}{4} + O\big(p(1-p) + p^2\big)\end{aligned}$$

where $p = e^{-t}$. Therefore

$$\limsup_{t\to\infty} \mathbf{P}(\mathcal{C}_t) \leq \frac{1}{4}.$$

This concludes the proof of Theorem 3.

D. Proof of Theorem 4

The proof of Theorem 4, as before, uses the characterization of rumor center provided by Proposition 1. That is, we wish to show that for all t large enough, the probability of the event that the size of the d^* rumor infected subtrees of the source v^* are essentially "balanced" [c.f (26)] with high enough probability. To establish this, we shall use coarse estimations on the size of each of these subtrees using the standard concentration property of the Poisson process along with geometric growth. This will be unlike the proof for regular trees where we had to necessarily delve into very fine detailed probabilistic estimates of the size of the subtrees to establish the result. This relatively easier proof for geometric trees (despite heterogeneity) brings out the fact that it is fundamentally much more difficult to analyze expanding trees than geometric structure as expanding trees do not yield to generic concentration based estimations as they necessarily have very high variances.

To that end, we shall start by obtaining sharp estimations on the size of each of the rumor infected d^* subtrees of v^* for any given time t. Now initially, at time 0 the source node v^* has the rumor. It starts spreading along its d^* children (neighbors). Let $N_i(t)$ denote the size of the rumor infected subtree, denoted by $G_i(t)$, rooted at the ith child (or neighbor) of node v^*. Initially, $N_i(0) = 0$. The $N_i(\cdot)$ is a Poisson process with time-varying rate: the rate at time t depends on the 'boundary' of the tree as discussed earlier. Due to the balanced and geometric growth conditions assumed in Theorem 4, the following will be satisfied: for small enough $\epsilon > 0$: (a) every node within a distance $t(1-\epsilon)$ of v^* is in one of the $G_i(t)$; and (b) no node beyond distance $t(1+\epsilon)$ of v^* is in any of the $G_i(t)$. Such a tight characterization of the "shape" of $G_i(t)$ along with the polynomial growth will provide sharp enough bound on $N_i(t)$ that will result in establishing Theorem 4. This result is summarized below with its proof in the Appendix.

Theorem 5: Consider a geometric tree with parameters $\alpha > 0$ and $0 < b \leq c$ as assumed in Theorem 4 and let the rumor spread from source v^* starting at time 0. Define $\epsilon = t^{-1/2+\delta}$ for any small $0 < \delta < 0.1$. Let $G(t)$ be the set of all rumor infected nodes in the tree at time t. Let $\mathcal{G}_t$ be the set of all subtrees rooted at v^* (rumor graphs) such that all nodes within distance $t(1-\epsilon)$ from v^* are in the tree and no node beyond distance $t(1+\epsilon)$ from v^* is in the tree. Then

$$\begin{aligned}\mathbf{P}(G_t \in \mathcal{G}_t) &= 1 - O\big(e^{-t^{\delta}}\big)\\ &\xrightarrow{t\to\infty} 1.\end{aligned}$$

Define $\mathcal{E}_t$ as the event that $G_t \in \mathcal{G}_t$. Under event $\mathcal{E}_t$, consider the sizes of the subtrees $N_i(t)$ for $1 \leq i \leq d_{v^*}$. Due to the polynomial growth condition and $\mathcal{E}_t$, we obtain the following bounds on each $N_i(t)$ for all $1 \leq i \leq d_{v^*}$:

$$\begin{aligned}\sum_{r=1}^{t(1-\epsilon)-1} br^{\alpha} &\leq N_i(t)\\ &\leq \sum_{r=1}^{t(1+\epsilon)-1} cr^{\alpha}.\end{aligned}$$

Now bounding the summations by Reimann's integrals, we have

$$\int_0^{L-1} r^{\alpha}dr \leq \sum_{r=1}^{L} r^{\alpha} \leq \int_0^{L+1} r^{\alpha}dr.$$

Therefore, it follows that under event $\mathcal{E}_t$, for all $1 \leq i \leq d_{v^*}$

$$\frac{b}{1+\alpha}\big(t(1-\epsilon)-2\big)^{\alpha+1} \leq N_i(t) \leq \frac{c}{1+\alpha}\big(t(1+\epsilon)\big)^{\alpha+1}.$$

In the most "unbalanced" situation, $d_{v^*} - 1$ of these subtrees have minimal size $N_{\min}(t)$ and the remaining one subtree has size $N_{\max}(t)$ where

$$\begin{aligned}N_{\min}(t) &= \frac{b}{1+\alpha}\big(t(1-\epsilon)-2\big)^{\alpha+1},\\ N_{\max}(t) &= \frac{c}{1+\alpha}\big(t(1+\epsilon)\big)^{\alpha+1}.\end{aligned}$$

Since by assumption $c < b(d_{v^*} - 1)$, there exists $\gamma > 0$ such that $(1+\gamma)c < b(d_{v^*} - 1)$. Therefore, for any choice of $\epsilon = t^{-1/2+\delta}$ for some $\delta \in (0, 0.1)$, we have

$$\begin{aligned}\frac{(d^*-1)N_{\min}(t)+1}{N_{\max}(t)} &= \frac{b(d_{v^*}-1)}{c}\Big(\frac{t - t^{\frac{1}{2}+\delta} - 2}{t + t^{\frac{1}{2}+\delta}}\Big)^{\alpha+1} \\ &\quad + \frac{1+\alpha}{c}\Big(\frac{1}{t+t^{\frac{1}{2}+\delta}}\Big)^{\alpha+1} \\ &\overset{(i)}{>} (1+\gamma)\Big(\frac{1 - t^{-\frac{1}{2}+\delta} - \frac{2}{t}}{1 + t^{-\frac{1}{2}+\delta}}\Big)^{\alpha+1} \\ &\quad + \frac{1+\alpha}{c}\Big(\frac{1}{t+t^{\frac{1}{2}+\delta}}\Big)^{\alpha+1} \\ &> 1+\gamma \\ &> 1\end{aligned}$$

for t large enough since as $t \to \infty$ the second term in inequality (i) goes to 1 and the third term goes to 0. From this, it immediately follows that under event $\mathcal{E}_t$ for t large enough

$$\max_{1 \le i \le d_{v^*}} N_i(t) < \frac{1}{2}\left(\sum_{i=1}^{d_{v^*}} N_i(t) + 1\right).$$

Therefore, by Proposition 1 it follows that the rumor center is unique and equals v^*. Therefore, for t large enough $\mathcal{E}_t \subset \mathcal{C}_t$. From above and Theorem 5

$$\begin{aligned}\liminf_t \mathbf{P}(\mathcal{C}_t) &\ge \lim_t \mathbf{P}(\mathcal{E}_t) \\ &= 1.\end{aligned}$$

VII. Conclusion and Future Work

This paper has provided, to the best of the authors' knowledge, the first systematic study of the problem of finding rumor sources in networks. Using the well known SI model, we constructed an estimator for the rumor source in regular trees, general trees, and general graphs. We defined the ML estimator for a regular tree to be a new notion of network centrality which we called rumor centrality and used this as the basis for estimators for general trees and general graphs.

We analyzed the asymptotic behavior of the rumor source estimator for regular trees and geometric trees. For linear graphs, it was shown that the detection probability goes to 0 as the network grows in size. However, for trees which grew faster than lines, it was shown that there was always nontrivial detection probability. This analysis highlighted the different techniques which must be used for networks with expansion versus those with only polynomial growth. Simulations performed on synthetic graphs agreed with these tree results and also demonstrated that the general graph estimator performed well in different network topologies, both synthetic (small-world, scale-free) and real (AS, power grid).

On trees, we showed that the rumor center is equivalent to the distance center. However, these were not equivalent in a general network. Also, it was seen that in networks which are not tree-like, rumor centrality is a better rumor source estimator than distance centrality.

The next step of this work would be to better understand the effect of the BFS heuristic on the estimation error and under what precise conditions it improves or degrades performance. Another future direction would be to generalize the estimator to networks with a heterogeneous rumor spreading rate.

Appendix A
Proof of Theorem 5

We recall that Theorem 5 stated that the rumor graph on a geometric tree is full up to a distance $t(1-\epsilon)$ and does not extend beyond $t(1+\epsilon)$, for $\epsilon = t^{-1/2+\delta}$ for some positive $\delta \in (0, 0.1)$. To establish this, we shall use the following well known concentration property of the unit rate Poisson process. We provide its proof later for completeness.

Theorem 6: Consider a unit rate Poisson process $P(\cdot)$ with rate 1. Then for any $t > 0$ and for any $\gamma \in (0, 0.25)$

$$\mathbf{P}\big(\big|P(t) - t\big| \ge \gamma t\big) \le 2e^{-\frac{1}{4}t\gamma^2}.$$

Now we use Theorem 6 to establish Theorem 5. Recall that the spreading time along each edge is an independent and identically distributed exponential random variable with parameter 1. Now the underlying network graph is a tree. Therefore for any node v at distance r from source node v^*, there is a unique path (of length r) connecting v and v^*. Then, the spread of the rumor along this path can be thought of as a unit rate Poisson process, say $P(t)$, and node v is infected by time t if and only if $P(t) \ge r$. Therefore, from Theorem 6 it follows that for any node v that is at distance $t(1-\epsilon)$ for $\epsilon = t^{-\frac{1}{2}+\delta}$ for some $\delta \in (0, 0.1)$

$$\begin{aligned}\mathbf{P}\big(v \text{ is not rumor infected}\big) &\le 2e^{-\frac{1}{4}t\epsilon^2} \\ &= 2e^{-\frac{1}{4}t^{2\delta}}.\end{aligned}$$

Now the number of such nodes at distance $t(1-\epsilon)$ from v^* is at most $O(t^{1+\alpha})$ (follows from arguments similar to those in the proof of Theorem 4). Therefore, by an application of union bound it follows that:

$$\begin{aligned}&\mathbf{P}\big(\text{ a node at distance t}(1-\epsilon) \text{ from } v^* \text{ isn't infected}\big) \\ &\quad = O\Big(t^{\alpha+1}e^{-\frac{1}{4}t^{2\delta}}\Big) \\ &\quad = O\Big(e^{-t^{\delta}}\Big).\end{aligned}$$

Using similar argument and another application of Theorem 6, it can be argued that

$$\begin{aligned}&\mathbf{P}\big(\text{ a node at distance t}(1+\epsilon) \text{ from v}^* \text{ is infected}\big) \\ &\quad = O\Big(e^{-t^{\delta}}\Big).\end{aligned}$$

Since the rumor is a 'spreading' process, if all nodes at distance r from v^* are infected, then so are all nodes at distance $r' < r$ from v^*; if all nodes at distance r from v^* are not infected then so are all nodes at distance $r' > r$ from v^*. Therefore, it follows that with probability $1 - O(e^{-t^{\delta}})$, all nodes at distance up to $t(1-\epsilon)$ from v^* are infected and all nodes beyond distance $t(1+\epsilon)$ from v^* are not infected. This completes the proof of Theorem 5.

Appendix B
Proof of Theorem 6

We wish to prove bounds on the probability of $P(t) \leq t(1-\gamma)$ and $P(t) \geq t(1+\gamma)$ for a unit rate Poisson process $P(\cdot)$. To that end, for $\theta > 0$ it follows that:

$$\begin{aligned}\mathbf{P}\big(P(t) \leq t(1-\gamma)\big) &= \mathbf{P}\big(-\theta P(t) \geq -\theta t(1-\gamma)\big)\\ &= \mathbf{P}\big(e^{-\theta P(t)} \geq e^{-\theta t(1-\gamma)}\big)\\ &\leq e^{\theta t(1-\gamma)}\mathbf{E}\big[e^{-\theta P(t)}\big]\\ &= e^{\theta t(1-\gamma)}e^{t(e^{-\theta}-1)}\end{aligned}$$

where the last equality follows from the fact that $P(t)$ is a Poisson random variable with parameter t. That is

$$\mathbf{P}\big(P(t) \leq t(1-\gamma)\big) \leq \inf_{\theta>0} e^{\theta t(1-\gamma)+te^{-\theta}-t}.$$

The minimal value of the exponent in the RHS above is achieved for value of $\theta = -\log(1-\gamma)$. For this value of θ, using the fact that $\gamma \in (0, 0.25)$ and $\log(1-\gamma) \geq -\gamma - 3\gamma^2/4$ for $\gamma < 1/3$, it follows that:

$$\mathbf{P}\big(P(t) \leq t(1-\gamma)\big) \leq e^{-\frac{1}{4}t\gamma^2}.$$

Next, to establish the bound on the probability of $P(t) \geq t(1+\gamma)$, using similar argument it follows that:

$$\mathbf{P}\big(P(t) \geq t(1+\gamma)\big) \leq \inf_{\theta>0} e^{t\left(-\theta(1+\gamma)+(e^{\theta}-1)\right)}.$$

The RHS is minimized for $\theta = \log(1+\gamma)$. Using $\log(1+\gamma) \geq \gamma - \gamma^2/2$ for $\gamma \leq 0.5$ it follows that:

$$\mathbf{P}\big(P(t) \geq t(1+\gamma)\big) \leq e^{-\frac{1}{4}t\gamma^2}.$$

This completes the proof of Theorem 6.

Acknowledgment

One of the authors (D. Shah) would like to acknowledge a stimulating conversation with D. Gamarnik and A. Montanari at the Banff International Research Station (BIRS) in summer 2008 that seeded this work and would like to thank the program at BIRS.

References

[1] N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases and its Applications*. London, U.K.: Griffin, 1975.

[2] C. Moore and M. E. J. Newman, "Epidemics and percolation in in small-world networks," *Phys. Rev. E*, vol. 61, pp. 5678–5682, 2000.

[3] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, pp. 3200–3203, 2001.

[4] M. E. J. Newman, "The spread of epidemic disease on networks," *Phys. Rev. E*, vol. 66, p. 016128, 2002.

[5] A. Ganesh, L. Massoulie, and D. Towsley, "The effect of network topology on the spread of epidemics," in *Proc. 24th Ann. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, 2005, vol. 2, pp. 1455–1466.

[6] N. Demiris and P. D. O'Neill, "Bayesian inference for epidemics with two levels of mixing," *Scand. J. Statist.*, vol. 32, pp. 265–280, 2005.

[7] G. Streftaris and G. J. Gibson, "Statistical inference for stochastic epidemic models," in *Proc. 17th Int. Workshop on Statist. Model.*, 2002, pp. 609–616.

[8] P. D. O'Neill, "A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods," *Math. Biosci.*, vol. 180, pp. 103–114, 2002.

[9] N. Demiris and P. D. O'Neill, "Bayesian inference for stochastic multitype epidemics in structured populations via random graphs," *J. Roy. Statist. Soc. B*, vol. 67, pp. 731–745, 2005.

[10] H. Okamura, K. Tateishi, and T. Doshi, "Statistical inference of computer virus propagation using nonhomogeneous poisson processes," in *Proc. 18th IEEE Int. Symp. Software Reliabil.*, 2007, vol. 5, pp. 149–158.

[11] W. Evans, C. Kenyon, Y. Peres, and L. Schulman, "Broadcasting on trees and the Ising model," *Ann. Appl. Prob.*, vol. 10, pp. 410–433, 2000.

[12] E. Mossel, "Reconstruction on trees: beating the second eigenvalue," *Ann. Appl. Prob.*, vol. 11, pp. 285–300, 2001.

[13] A. Gerschenfeld and A. Montanari, "Reconstruction for models on random graphs," in *Proc. 48th IEEE Symp. Found. Comp. Sci.*, 2007, pp. 194–204.

[14] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, pp. 581–603, 1966.

[15] L. C. Freeman, "A set of measure of centrality based on betweenness," *Sociometry*, vol. 40, pp. 35–41, 1977.

[16] J. M. Hammersley and D. J. A. Welsh, "First-passage percolation, subadditive processes, stochastic networks, and generalized renewal theory," in *Bernoulli-Bayes-Laplace Anniversary Volume*. Berlin: Springer, 1965.

[17] R. Smythe and J. C. Wierman*, First Passage Percolation on the Square Lattice, Lecture Notes in Math.* Berlin: Springer, 1978.

[18] J. T. Cox and R. Durrett, "Some limit theorems for percolation processes with necessary and sufficient conditions," *Ann. Appl. Prob.*, vol. 9, pp. 583–603, 1981.

[19] H. Kesten*, Aspects of First Passage Percolation. Lecture Notes in Math.* Berlin: Springer, 1986, vol. 1180, pp. 125–264.

[20] H. Kesten, "Percolation theory and first-passage percolation," *Ann. Appl. Prob.*, vol. 15, pp. 1231–1271, 1987.

[21] H. Kesten, "On the speed of convergence in first-passage percolation," *Ann. Appl. Prob.*, vol. 3, pp. 296–338, 1993.

[22] M. D. Atkinson, "On computing the number of linear extensions of a tree," *Order*, vol. 7, pp. 23–25, 1990.

[23] G. Brightwell and P. Winkler, "Counting linear extensions is #P-complete," in *JSTOC '91: Proc. 23rd Ann. ACM Symp. Theory of Comput.*, 1991, pp. 175–181.

[24] D. J. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[25] A. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[26] The CAIDA AS relationships dataset Aug. 30, 2009 [Online]. Available: http://www.caida.org/data/active/as-relationships/

[27] D. Betsekas and J. N. Tsitsiklis*, Probabilistic System Analysis*, 2nd ed. New York: Aetna, 2009.

Devavrat Shah received the B.Tech. degree in 1999 and the Ph.D. degree in 2004).

He is currently a Jamieson Career Development Associate Professor with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Canbridge. He is a member of the Laboratory for Information and Decision Systems (LIDS) and Operations Research Center (ORC). His research focus is on the theory of large complex networks, which includes network algorithms, stochastic networks, network information theory, and large scale statistical inference.

Professor Shah was co-awarded the Best Paper awards at the IEEE INFOCOM '04 and ACM SIGMETRICS/Performance '06 and supervised Best Student Paper awards at Neural Information Processing Systems'08 and ACM SIGMETRICS/Performance '09. He received the 2005 George B. Dantzig Best Disseration award from the INFORMS. He received the first ACM SIGMETRICS Rising Star Award 2008 for his work on network scheduling algorithms. He was awarded the 2010 Erlang Prize from INFORMS which is given to a young researcher for outstanding contributions to applied probability. He is currently an Associate Editor of Operations Research.

Tauhid Zaman (M'10) receved the B.S. degree in 2004 and the M.Eng degree in 2006. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge.

His research focuses on techniques for knowledge extraction and inference in large-scale networks, especially social networks. He is currently an MIT-Shell Energy Fellow.