Rumor Source Detection under Probabilistic Sampling

Nikhil Karamchandani
Dept. of EE,
University of California Los Angeles,
Los Angeles, USA
Email: nikhil@ee.ucla.edu

Massimo Franceschetti
Dep. of Electrical & Computer Engineering
University of California San Diego
La Jolla, USA
Email: massimo@ece.ucsd.edu

***Abstract*—Consider a network where an unidentified source starts a rumor. The rumor spreads along the edges of the network to other nodes in the network. After a sufficiently long amount of time, we observe a subset of the nodes that have heard the rumor, and using this information wish to identify the source. Optimal estimators were recently proposed for regular (exponential growth) and irregular geometric (polynomial growth) trees when all nodes that heard the rumor reveal themselves. We provide the extension to the case in which nodes reveal whether they have heard the rumor with probability p, independent of each other. For geometric trees and $p > 0$, we achieve the same performance as the optimal estimator with $p = 1$. For regular trees, the estimator can achieve performance within ϵ of the optimal, provided that p is larger than a threshold.**

I. Introduction

Consider a network where an unidentified source starts a rumor. The rumor spreads along the edges of the network and after a sufficiently long amount of time, we observe a subset of the nodes that have heard the rumor. Using only this information, we wish to identify the source. This problem arises in many different contexts including, for example, identifying the malicious device propagating a virus on a computer network, the first infected person spreading a contagious disease in a community, or a fault in an electrical power network leading to a large-scale blackout.

The problem was recently studied analytically in [1], where the rumor-spreading process was modeled according to the *Susceptible-Infected (SI)* model with homogenous exponentially distributed spreading times on the individual links of the network. The authors proposed a *rumor centrality estimator* which takes as input the rumor-infected subgraph of the network at some time and then assigns a score to each infected node which reflects the probability of the node being the rumor source. They showed that the node with the highest score in such an estimator is the Maximum Likelihood estimate of the rumor source when the underlying rumor spreading graph is a regular tree. The detection performance of the estimator was also studied for irregular geometric trees which have polynomial growth, unlike regular trees. These results were improved in [2], [3] where exact expressions for the probability of correct detection were derived and extensions to other spreading time distributions and random irregular tree graphs were studied.

All the papers mentioned above assume that the infection status of all the nodes in the network is known. However, this is not always possible in large networks because of the prohibitive cost of data collection. Our contribution is the extension to the case of partial observation where the infection status of some nodes in the network is not known. We assume that each node reports its infection status with probability $p > 0$, independently of all other nodes. We apply the rumor centrality estimator to the reported rumor subgraph and study the impact of such missing information on its performance when compared to the case of $p = 1$ studied in [1]–[3]. For regular trees, we show that for any $\epsilon > 0$, the rumor centrality estimator achieves a probability of correct detection within ϵ of the optimal, obtained when all nodes reveal their infection status, as long as the sampling probability p is higher than a fixed threshold. On the other hand, for geometric trees, we show that for any $p > 0$, the estimator achieves essentially the same performance as in the case of $p = 1$.

Following [1], there has been further work on the problem of rumor source detection. The works [4], [5] study the scenario of multiple sources, whereas [6] considers the *Susceptible-Infected-Recovered (SIR)* model of rumor spreading. The work in [7] also allows for the possibility that not all nodes report their infection status, however it introduces the additional assumption that timing and direction of spreading information is available at each reporting node. Other prior work on rumor spreading has focused on the *forward problem* where given a rumor source, the objective is to identify conditions on network topology and spreading rates which can lead to large-scale spreading. See for example, [8] and references therein. There has been recent work on other *inverse problems* as well beside rumor source detection, for example [9]–[11] infer the underlying network graph after observing the results of rumor spreading on the network.

We formally describe our problem setup in Section II. Regular trees and geometric trees are studied in Sections III and IV respectively. Finally, Section V concludes the paper.

The work of N. Karamchandani was supported by the Information Theory and Applications Center at the University of California San Diego and AFOSR MURI award FA9550-09-064: ”Information Dynamics as Foundation for Network Management”.

II. Problem setup and notation

We model a network as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where the set of vertices $\mathcal{V}$ represents the nodes in the network and the set of edges $\mathcal{E}$ represents the links between the network nodes. Throughout this paper, we will let $\mathcal{V}$ be a countably infinite set to avoid boundary effects. Further, similar to [1]–[3] we will focus attention on networks where the underlying graph $\mathcal{G}$ is a tree.

A. Rumor Spread Model

We model the spread of the rumor using the *Susceptible-Infected (SI)* model. In this model, once a node is infected with the rumor, it retains it forever. Let a node $v^* \in \mathcal{G}$ be the rumor source. Once a node i gets the rumor, it can pass it on to (or infect) any node j with which it shares a link, i.e., $(i,j) \in \mathcal{E}$. The time τ_{ij} that it takes for this to happen is modeled as an exponential random variable with parameter λ. We assume that the random variables $\{\tau_{ij}\}_{(i,j)\in\mathcal{E}}$ are independent and identically distributed. Without loss of generality, let $\lambda = 1$.

B. Rumor Centrality Estimator

Let a node v^* start the spread of a rumor in the network $\mathcal{G}$ at time 0. After a sufficiently long time, say N nodes are infected with the rumor and let $\mathcal{G}_N$ denote the corresponding connected subgraph of $\mathcal{G}$ induced by the infected nodes. [1] presented an estimator which assigns a score, called *rumor centrality*, to each node in the infected subgraph G_N. The estimator then chooses the node with the highest score to be the estimated rumor source, which we will refer to as the *rumor center* henceforth. Let T_u^v be the number of nodes in the sub-tree of $\mathcal{G}_N$ rooted at node u when node v is the source. See Figure 1 for an illustration. A key property of the rumor center, the node with the highest rumor centrality, is given by the following result [1, Proposition 1].

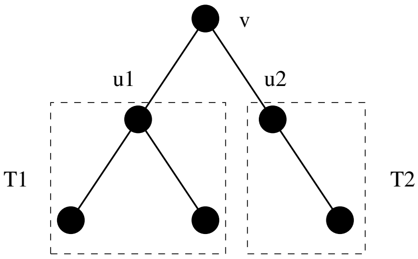

Fig. 1. Illustration of the subtree T_u^v

Proposition II.1. *Given an N node tree graph $\mathcal{G}_N$, a node $v \in \mathcal{G}_N$ is a rumor center if and only if*

$$T_u^v \leq \frac{1}{2}\left(\sum_{w:(v,w)\in\mathcal{E}} T_w^v\right) \quad \textit{for all } u \textit{ such that } (v,u) \in \mathcal{E}.$$

Furthermore, the node v is the unique rumor center if the above inequality is strict.

C. Rumor Centrality Estimator under Probabilistic Sampling

Unlike [1], in this work we allow for incomplete information about the infection state of the network. In particular, we assume that when we observe the infected subgraph $\mathcal{G}_N$, each node only reveals its infection state with probability $p \in (0,1)$. Let $\mathcal{G}_N^p$ denote the minimum subgraph connecting all the nodes which reveal their state to be infected. Hereafter, we will refer to this infected subgraph as the *reported rumor subgraph*. In what follows, we will analyze the performance of the rumor centrality estimator when applied to the reported rumor subgraph $\mathcal{G}_N^p$ instead of the true infection graph $\mathcal{G}_N$. In other words, for every node $v \in \mathcal{G}_N^p$, the new estimator would pick an estimate for the rumor source by assigning a rumor centrality score of $R(v, \mathcal{G}_N^p)$.

III. Regular trees

Consider a network where the underlying graph $\mathcal{G}$ is an infinite d-regular tree. Suppose a node v^* starts a rumor on this graph at time 0. Let $\mathcal{G}_{N(t)}$ denote the rumor infected subgraph at time t and let $\mathcal{G}_{N(t)}^p$ denote the corresponding reported rumor subgraph. Let $\mathcal{C}_t^p$ denote the event that the rumor centrality estimator applied to $\mathcal{G}_{N(t)}^p$ correctly identifies the rumor source v^*. We will be interested in studying $\mathbb{P}(\mathcal{C}_t^p)$ as t grows large, or equivalently after the rumor has spread for a sufficiently long time. For any $d \geq 3$, [2, Theorem 3.1] characterizes the probability of correct rumor source detection $\alpha_d^1 \triangleq \lim_{t\to\infty} \mathbb{P}(\mathcal{C}_t^1)$ when complete information about the infection states of the network nodes is available. We generalize this result to the case of probabilistic sampling as follows.

Theorem III.1. *Let $\mathcal{G}$ be an infinite d-regular tree with $d \geq 3$ and let the rumor spread on $\mathcal{G}$ as per the SI model with exponential distribution. Assume that each infected node reveals its state with probability p, independent of all other nodes. Then for any $0 < \beta, \epsilon < 1$, there exists a constant $0 < \delta(\epsilon) < 1/2$ such that if*

$$p > \frac{7 - 14\delta(\epsilon)}{7 + 2\delta(\epsilon)} \ ,$$

then we have

$$\alpha_d^p \triangleq \lim_{t\to\infty} \mathbb{P}(\mathcal{C}_t^p) \geq \alpha_d^1 - \epsilon \ .$$

Proof. Let the rumor start at a node v^*. At time 0, only v^* is infected and each of its d neighbors $u_1, u_2, \ldots, u_d$ is susceptible to getting the rumor. At any time $t > 0$, let $T_{u_i}^{v^*}(t)$ denote the infected subtree rooted at u_i with v^* as the source. To simplify notation, we will use $T_i(t)$ to denote this subtree as well as the number of infected nodes in the subtree. [2, Theorem 3.1] characterizes the probability of correct rumor source detection α_d^1 as follows.

$$\begin{aligned}\alpha_d^1 = \lim_{t\to\infty} \mathbb{P}(\mathcal{C}_t^1) &\overset{(a)}{=} \lim_{t\to\infty} \mathbb{P}\left(\frac{T_i(t)}{\sum_{j=1}^d T_j(t)} < \frac{1}{2}\right) \\ &= d I_{1/2}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right) - (d-1) \ , \qquad (1)\end{aligned}$$

where (a) follows from Proposition II.1 and the regularized incomplete beta function $I_x(a,b)$ denotes the probability that a random variable with the Beta distribution with parameters a and b is less than $x \in [0,1]$. Then it follows from (1) that for any $0 < \epsilon < 1$, there exists $0 < \delta(\epsilon) < 1$ such that

$$\lim_{t\to\infty} \mathbb{P}\left(\mathcal{E}_1(t)\right) \triangleq \lim_{t\to\infty} \mathbb{P}\left(\frac{T_i(t)}{\sum_{j=1}^{d} T_j(t)} < \frac{1}{2} - \delta(\epsilon)\right) \geq \alpha_d^1 - \epsilon. \tag{2}$$

For simplicity, we will refer to $\delta(\epsilon)$ as δ hereafter. Next, denote the total number of infected nodes at time t by $N(t) = 1 + \sum_{i=1}^{d} T_i(t)$ and let the infected subgraph be $\mathcal{G}_{N(t)}$. Next, each infected node in this subgraph reveals its status with probability p. Let $\mathcal{G}^p_{N(t)}$ denote the reported rumor subgraph, the minimum subgraph connecting all the nodes which reveal their state to be infected. Let $T_i^p(t)$ denote the number of infected nodes in the i^{th} subtree $T_i(t)$ which also belong to the reported subgraph $\mathcal{G}^p_{N(t)}$. Let $H_i^p(t) = T_i(t) - T_i^p(t)$ denote the number of infected nodes in the i^{th} subtree which are not included in the reported rumor subgraph. Proposition II.1 implies that the true source v^* is correctly identified by the rumor centrality estimator applied to $\mathcal{G}^p_{N(t)}$ only if for every $i \in \{1, 2, \ldots, d\}$

$$\frac{T_i^p(t)}{\sum_{j=1}^{d} T_j^p(t)} = \frac{T_i(t) - H_i^p(t)}{\sum_{j=1}^{d} T_j^p(t)} < \frac{1}{2} \; . \tag{3}$$

The rest of the proof focuses on bounding $H_i^p(t)$. Note that an infected node w in $T_i(t)$ belongs to the set of $H_i^p(t)$ unreported nodes only if no infected nodes in the subtree $T_w^{v^*}(t)$ rooted at w reveal their state. This suggests that infected nodes which are not part of the reported rumor subgraph $\mathcal{G}^p_{N(t)}$ will most likely be close to the boundary of the infected subgraph $\mathcal{G}_{N(t)}$. This intuition will prove useful for the proofs in this paper.

Clearly, each infected node in $T_i(t)$ which reports its status is included in the reported rumor subgraph. Then its follows that the probability that $T_i^p(t)$ is at most K is at most the probability that a binomial random variable $B \sim \text{Bin}(T_i(t), p)$ is at most K. Then for any $0 < \beta < 1$, we have from the Chernoff bound that

$$\mathbb{P}\big(T_i^p(t) \leq (1-\beta)pT_i(t)\big) \leq \exp\left(-\beta^2 p T_i(t)/2\right) .$$

It follows from the union bound that

$$\mathbb{P}(\mathcal{E}_2(t)) \triangleq \mathbb{P}\Big(\sum_{j=1}^{d} T_j^p(t) \geq (1-\beta)p\sum_{j=1}^{d} T_i(t)\Big) \geq 1 - \sum_{i=1}^{d} \exp\left(-\beta^2 p T_i(t)/2\right) . \tag{4}$$

On the other hand, any infected node which is a leaf in the sub-tree $T_i(t)$ is absent from the reported rumor subgraph with probability $q = 1-p$. Thus, $H_i^p(t)$ is at least the number of leaves in the sub-tree $T_i(t)$ which do not report their infection status. Let $L_i(t)$ denote the number of leaves in the sub-tree $T_i(t)$. Then it follows from the Chernoff bound that for any $0 < \gamma < 1$,

$$\mathbb{P}\left(H_i^p(t) \leq (1-\gamma)qL_i(t)\right) \leq \exp\left(-\gamma^2 q L_i(t)/2\right) \tag{5}$$

For $0 < \eta < 1$, let $\theta = 1 - (1-\gamma)(1-\eta)$. Then we have

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_3(t)\right) &\triangleq \mathbb{P}\left(H_i^p(t) \geq (1-\theta)q/8)\, T_i(t) \;\forall\; i \in \{1,\ldots,d\}\right) \\ &\geq \mathbb{P}\left(H_i^p(t) \geq (1-\gamma)qL_i(t), L_i(t) \geq (1-\eta)T_i(t)/8 \;\forall\; i\right) \\ &\geq 1 - \sum_{i=1}^{d} \left(\exp\left(-\gamma^2 q L_i(t)/2\right) + \exp\left(-\eta^2 T_i(t)/16\right)\right) \end{aligned}\tag{6}$$

where the last inequality follows from (5), Lemma III.2, and the union bound. Next, note that if the events $\mathcal{E}_1(t), \mathcal{E}_2(t), \mathcal{E}_3(t)$ hold true, then for each $i \in \{1, 2, \ldots, d\}$ we have

$$\frac{T_i^p(t)}{\sum_{j=1}^{d} T_j^p(t)} \leq \frac{(1-(1-\theta)q/8)T_i(t)}{(1-\beta)p\sum_{j=1}^{d} T_i(t)} < \frac{7+p+\theta(1-p)}{8p(1-\beta)} \cdot \left(\frac{1}{2} - \delta\right) .$$

Then the condition for the estimator to correctly identify the rumor source as specified in (3) is satisfied if

$$\frac{7+p+\theta(1-p)}{8p(1-\beta)} \cdot \left(\frac{1}{2} - \delta\right) \leq \frac{1}{2} \implies p > \frac{7-14\delta}{7+2\delta} \; ,$$

where the last equation follows since β, θ can be made arbitrarily small. Then if p satisfies the condition above and as time t grows large, the probability that the true source v^* is correctly identified by the rumor centrality estimator applied to $\mathcal{G}^p_{N(t)}$ is given by

$$\begin{aligned}\lim_{t\to\infty} \mathbb{P}\left(\mathcal{C}_t^p\right) &\geq 1 - \lim_{t\to\infty}\Big(\mathbb{P}\left(\overline{\mathcal{E}_1(t)}\right) + \mathbb{P}\left(\overline{\mathcal{E}_2(t)}\right) + \mathbb{P}\left(\overline{\mathcal{E}_3(t)}\right)\Big) \\ &\geq \alpha_d^1 - \epsilon \; .\end{aligned}$$

where $\overline{\mathcal{E}}$ denotes the complement of an event $\mathcal{E}$. The last equation follows from (2), (4), (6) and noting that $\{T_i(t), L_i(t)\}_{i=1}^{d}$ tend to infinity as time t grows large. ■

Lemma III.2. *Let $\mathcal{G}$ be an infinite d-regular tree with $d \geq 3$ and let the rumor spread on $\mathcal{G}$ as per the SI model with exponential distribution. At time t, let $T_i(t)$ denote the i^{th} subtree of the infected rumor subgraph and let $L_i(t)$ denote the number of leaves in $T_i(t)$. Then for any $0 < \theta < 1$, we have*

$$L_i(t) \geq (1-\theta)T_i(t)/8$$

with probability at least

$$1 - \exp\left(-\theta^2 T_i(t)/16\right) .$$

Proof. Consider the i^{th} subtree $T_i(t)$, we will denote the set of nodes which are not yet infected but are neighbors of rumor infected nodes in $T_i(t)$ as the *rumor boundary* $Z_i(t)$. We will abuse notation to use $Z_i(t)$ to also denote the number of nodes in the rumor boundary of the i^{th} subtree at time t, with $Z_i(0) = 1$. By the memoryless property of the exponential distribution, one of the $Z_i(t)$ nodes on the rumor boundary is chosen uniformly at random to be rumor infected next in the i^{th} subtree. This node is removed from the boundary and its non-infected $(d-1)$ neighbors are added to the rumor boundary. In other words, each infection adds $(d-2)$ new nodes to the rumor boundary. This implies that the number of infected nodes $T_i(t)$ in the i^{th} rumor subtree and the number of nodes $Z_i(t)$ in the corresponding rumor boundary are related as $Z_i(t) = (d-2)T_i(t) + 1$.

Consider the rumor boundary $Z_i(t)$ at time t. Let t_0 denote the time when the rumor boundary of $T_i(t_0)$ consisted of $Z_i(t_0) = a \cdot Z_i(t)$ nodes, where $a = d/(2(d-1))$. Let $Z_i^t(t_0)$ denote the subset of nodes in $Z_i(t_0)$ which get infected before time t. As before, we will use $Z_i^t(t_0)$ to also denote the number of such nodes. Then it is easy to verify that for every node $z \in Z_i^t(t_0)$, there is a distinct leaf node ℓ of $T_i(t)$ such that ℓ is a descendant of z in $T_i(t)$. This implies that the number of leaves in the i^{th} subtree at time t, $L_i(t) \geq Z_i^t(t_0)$.

As mentioned before, at each infection time for the i^{th} subtree in the interval (t_0, t), a node from the rumor boundary is chosen uniformly at random to be infected. Since each infection adds $(d-2)$ new nodes to the boundary, there are

$$\frac{Z_i(t) - Z_i(t_0)}{d-2} = \frac{Z_i(t) - aZ_i(t)}{d-2} = \frac{Z_i(t)}{2(d-1)} \triangleq M(t)$$

new infections in the time interval (t_0, t). Say $k \in [1, M(t)]$ new nodes in $T_i(t)$ have been infected till some time $\tilde{t} \in (t_0, t)$. Then the probability that the next infected node will belong to $Z_i^t(t_0)$ is at least

$$\frac{Z_i(t_0) - k}{Z_i(t_0) + k(d-2)} = \frac{aZ_i(t) - k}{aZ_i(t) + k(d-2)} \ .$$

Then the expected number of nodes belonging to the infected subset $Z_i^t(t_0)$ is at least

$$\sum_{k=0}^{M(t)-1} \frac{aZ_i(t) - k}{aZ_i(t) + k(d-2)} \geq M(t) \cdot \frac{aZ_i(t) - M(t)}{aZ_i(t) + M(t)(d-2)}$$
$$= Z_i(t) \cdot \frac{1}{4(d-1)} \overset{(a)}{\geq} T_i(t) \cdot \frac{d-2}{4(d-1)} \overset{(b)}{\geq} \frac{T_i(t)}{8} \ ,$$

where (a) follows since $Z_i(t) = (d-2)T_i(t) + 1$ and (b) follows since $(d-2)/(d-1)$ is increasing in d and $d \geq 3$. The result then follows by applying the Chernoff bound. ■

IV. Geometric trees

The previous section considers regular trees. In this section, we will study a particular class of irregular trees that grow polynomially, namely geometric trees. We borrow the following definition from [1], [2].

Definition IV.1. *Geometric trees:* This class of tree graphs is parameterized by constants a, b, and c, with $a > 0$, $0 < b \leq c$, and a root node v^*. Let d^* denote the degree of the root node v^* and let $u_1, u_2, \ldots, u_{d^*}$ denote the neighboring nodes. Let T_i denote the subtree rooted at u_i and consisting of nodes away from v^*. Consider any node $v \in T_i$ and let $n^i(v, r)$ be the number of nodes in T_i at distance exactly r from v. Then we require that for all $i \in \{1, 2, \ldots, d^*\}$, $v \in T_i$, and $r \geq 1$,

$$br^a \leq n^i(v, r) \leq cr^a. \tag{7}$$

The above equation specifies that geometric trees demonstrate polynomial growth with exponent being the parameter a. They also satisfy certain regularity properties as indicated by the ratio c/b. Note that if the ratio is close to 1, then the subtrees are somewhat regular. The further it deviates from 1, the greater the heterogeneity between the subtrees.

Consider a network where the underlying graph $\mathcal{G}$ is an infinite geometric tree. Suppose the root node v^* starts a rumor on this graph at time 0. As before, let $\mathcal{C}_t^p$ denote the event that the rumor centrality estimator applied to the reported rumor subgraph $\mathcal{G}_{N(t)}^p$ at time t correctly identifies the rumor source v^*. We will be interested in studying $\mathbb{P}(\mathcal{C}_t^p)$ as t grows large, or equivalently after the rumor has spread for a sufficiently long time. [1, Theorem 4] studies the problem of rumor source detection in geometric trees when complete information about the infection states of the network nodes is available. In particular, for $\alpha > 0$ and a few technical conditions, it is shown that the probability of correct detection achieved by the rumor centrality estimator is given by $\lim_{t\to\infty} \mathbb{P}(\mathcal{C}_t^1) = 1$. We generalize this result to the case of probabilistic sampling.

Theorem IV.2. *Let $\mathcal{G}$ be a rooted geometric tree with parameters $a > 0$, $0 < b \leq c$ and a root node v^* with degree d^* such that $d^* > \max(2, c/b + 1)$. Let the rumor spread on $\mathcal{G}$ starting from the root node v^* as per the SI model with exponential distribution. Assume that each infected node reveals its state with probability $p > 0$, independent of all other nodes. Then we have*

$$\lim_{t\to\infty} \mathbb{P}(\mathcal{C}_t^p) = 1 \ .$$

Proof. Let the rumor start at the root node v^* of a geometric tree. As before, at time t let $T_i(t)$ denote the i^{th} infected subtree rooted at a neighbor u_i of the root node v^*, with v^* as the source. Then a few useful properties of the the rumor infection process are derived in the proof of [1, Theorem 4] and we restate them here.

(P1) Let $\epsilon = t^{-1/2+\delta}$ for any small $0 < \delta < 0.1$. Then with probability approaching 1 as time t grows large, all nodes within distance $t(1-\epsilon)$ of v^* are infected and no node beyond distance $t(1+\epsilon)$ of v^* is infected.

(P2) For $c < b(d^*-1)$, there exists $\gamma > 0$ such that for any $i \in \{1, 2, \ldots, d^*\}$ and time t large enough,

$$\frac{\sum_{j=1}^{d^*} T_j(t)}{T_i(t)} > 2 + \gamma \ .$$

Next, each infected node reveals its state with probability $p > 0$. Let $T_i^p(t)$ denote the number of infected nodes in the i^{th} subtree $T_i(t)$ which also belong to the reported rumor subgraph and let $H_i^p(t) = T_i(t) - T_i^p(t)$. Proposition II.1 implies that at time t, the true source v^* is correctly identified by the rumor centrality estimator applied to the reported rumor subgraph if for every $i \in \{1, 2, \ldots, d^*\}$

$$\frac{T_i^p(t)}{\sum_{j=1}^{d^*} T_j^p(t)} = \frac{T_i(t) - H_i^p(t)}{\sum_{j=1}^{d^*}(T_j(t) - H_j^p(t))} < \frac{1}{2}$$

$$\text{or, } \frac{\sum_{j=1}^{d^*} H_j^p(t)}{T_i(t)} < \frac{\sum_{j=1}^{d^*} T_{j(t)}}{T_i(t)} - 2$$

$$\text{or, } \frac{\sum_{j=1}^{d^*} H_j^p(t)}{T_i(t)} < \gamma \ , \tag{8}$$

where the last condition follows from property (P2) above. The rest of the proof focuses on bounding $H_i^p(t)$. For

$$\Delta \triangleq 2 \left(\frac{(1+a)(2+a)}{b} \log_{\frac{1}{1-p}} t \right)^{\frac{1}{1+a}} + 4 \ , \tag{9}$$

let w be a node in $T_i(t)$ which is at distance at most $(t(1-\epsilon) - \Delta)$ from the root v^*. Then it follows from the observation (P1) above that with probability approaching 1 as t grows large, all descendants of w within distance Δ are also infected by the rumor. To lower bound the number of such descendants of w, consider one such descendent x at distance $\Delta/2$ from w. Then any node within distance $\Delta/2$ of x will be a descendant of w which is also infected. From (7), the number of such nodes is at least

$$\sum_{r=1}^{\Delta/2-1} br^a \geq b \int_0^{\Delta/2-2} r^a \, \mathrm{d}r = \frac{b}{1+a} \left(\frac{\Delta}{2} - 2 \right)^{1+a} = \log_{\frac{1}{1-p}} t^{2+a},$$

where the last equality follows from (9). Note that w will not be included in the reported rumor subgraph only if all of its infected descendants do not report their status. Since each node fails to report its infection status independently with probability $1-p$, the probability that node w contributes to $H_i^p(t)$ is at most $(1-p)^{(2+a)\log_{\frac{1}{1-p}} t} = t^{-(2+a)}$. From (7), the number of such nodes w within distance $t(1-\epsilon) - \Delta$ from the root v^* is at most

$$\sum_{r=1}^{t(1-\epsilon)-\Delta-1} cr^a \leq c \int_0^{t(1-\epsilon)-\Delta} r^a \, \mathrm{d}r = \frac{c}{1+a} \left(t(1-\epsilon) - \Delta \right)^{1+a} .$$

By the union bound, the probability that some node in $T_i(t)$ within distance $t(1-\epsilon) - \Delta$ from the root v^* does not belong to the reported rumor subgraph is at most

$$\frac{c}{1+a} \left(t(1-\epsilon) - \Delta \right)^{1+a} \cdot \frac{1}{t^{2+a}}.$$

The above probability goes to 0 as t grows large. Thus, with probability approaching 1 as $t \to \infty$, an infected node in $T_i(t)$ contributes to $H_i^p(t)$ only if its distance from the root node v^* is at least $t(1-\epsilon) - \Delta$. Hereafter, we will assume this to be true. From property (P1), all infected nodes in $T_i(t)$ are at distance at most $t(1+\epsilon)$ from v^*. Then we have from (7) that with high probability as $t \to \infty$, for every $i \in \{1, 2, \ldots, d^*\}$

$$\begin{aligned} H_i^p(t) &\leq \sum_{r=t(1-\epsilon)-\Delta+1}^{t(1+\epsilon)-1} cr^a \leq c \int_{t(1-\epsilon)-\Delta}^{t(1+\epsilon)} r^a \, \mathrm{d}r \\ &= \frac{c}{1+a} \Big(\left(t(1+\epsilon)\right)^{1+a} - \left(t(1+\epsilon) - (2\epsilon t + \Delta)\right)^{1+a} \Big) \\ &\overset{(a)}{\leq} \frac{c}{1+a} (1+a)(2\epsilon t + \Delta) \left(t(1+\epsilon)\right)^a \overset{(b)}{\leq} \kappa t^{a+1/2+\delta} \ , \end{aligned} \tag{10}$$

where κ is some positive constant; (a) follows since $\left(x^{1+\beta} - (x-\gamma)^{1+\beta} \right) \leq (1+\beta)\gamma x^\beta$ for any $\beta > 0, 0 < \gamma < x$; and (b) follows from the definitions of ϵ and Δ in observation (P1) and (9) respectively.

On the other hand, property (P1) states that with high probability as $t \to \infty$, all nodes in the i^{th} subtree within distance $t(1-\epsilon)$ of the root v^* are infected. Then we have

$$T_i(t) \geq \sum_{r=1}^{t(1-\epsilon)-1} br^a \geq \frac{b}{1+a} \left(t(1-\epsilon) - 2 \right)^{1+a} .$$

Combining with (10), we have that for every $i \in \{1, 2, \ldots, d^*\}$ and time t large enough,

$$\frac{H_i^p(t)}{T_i(t)} \leq \frac{\kappa t^{a+1/2+\delta}}{\frac{b}{1+a} \left(t(1-\epsilon) - 2 \right)^{1+a}} < \frac{\gamma}{d^*} \ , \tag{11}$$

where the last inequality follows since $\delta < 0.1$ from the definition in (P1) and hence the LHS in the the above equation goes to zero as $t \to \infty$. Thus, the condition in (8) is satisfied and the proof is complete. ■

V. Conclusions and future work

In this paper, we have extended the setup of rumor source detection to the case of incomplete infection status information. For regular trees, we showed that performance close to the optimal can be achieved if p is greater than a threshold. Future work will involve improving this threshold as well as proving converse results.

References

[1] T. Zaman and D. Shah, "Rumors in a network: Who's the culprit?" *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5163–5181, Aug. 2011.

[2] D. Shah and T. Zaman, "Finding rumor sources on random graphs," *preprint*, 2011. [Online]. Available: http://arxiv.org/abs/1110.6230

[3] ——, "Rumor centrality: a universal source detector," in *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*, Jun. 2012, pp. 199–210.

[4] W. Luo and W. Tay, "Identifying infection sources in large tree networks," in *Proceedings of the 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, Jun. 2012, pp. 281–289.

[5] B. Prakash, J. Vrekeen, and C. Faloutsos, "Spotting culprits in epidemics: How many and which ones?" in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, vol. 12, Dec. 2012.

[6] K. Zhu and L. Ying, "Information source detection in the SIR model: A sample path based approach," *preprint*, 2012. [Online]. Available: http://arxiv.org/abs/1206.5421

[7] P. Pinto, P. Thiran, and M. Vetterli, "Locating the source of diffusion in large-scale networks," *Physical Review Letters*, vol. 109, no. 6, Aug. 2012.

[8] M. Draief and L. Massoulié, *Epidemics and Rumours in Complex Networks*, ser. London Mathematical Society Lecture Note Series. Cambridge University Press, 2009.

[9] P. Netrapalli and S. Sanghavi, "Learning the graph of epidemic cascades," in *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*, Jun. 2012, pp. 211–222.

[10] S. M. C. Milling, C. Caramanis and S. Shakkottai, "On identifying the causative network of an epidemic," in *Proceedings of the 50th Annual Allerton Conference on Communications, Control and Computing*, Oct. 2012.

[11] C. Milling, C. Caramanis, S. Mannor, and S. Shakkottai, "Network forensics: random infection vs spreading epidemic," in *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*, Jun. 2012, pp. 223–234.