FINDING RUMOR SOURCES ON RANDOM GRAPHS

BY DEVAVRAT SHAH AND TAUHID ZAMAN

Massachusetts Institute of Technology

We consider the problem of detecting the source of a rumor (information diffusion) in a network based on observations about which set of nodes posses the rumor. In a recent work [10] by the authors, this question was introduced and studied. The authors proposed *rumor centrality* as an estimator for detecting the source. They establish it to be the maximum likelihood estimator with respect to the popular Susceptible Infected (SI) model with exponential spreading time for regular trees. They showed that as the size of infected graph increases, for a line (2-regular tree) graph, the probability of source detection goes to 0 while for d-regular trees with $d \geq 3$ the probability of detection, say α_d, remains bounded away from 0 and is less than $1/2$. Their results, however stop short of providing insights for the heterogeneous setting such as irregular trees or the SI model with non-exponential spreading times.

This paper overcomes this limitation and establishes the effectiveness of rumor centrality for source detection for generic random trees and the SI model with a generic spreading time distribution. The key result is an interesting connection between a multi-type continuous time branching processes (an equivalent representation of a generalized Polya's urn, cf. [1]) and the effectiveness of rumor centrality. Through this, it is possible to quantify the detection probability precisely. As a consequence, we recover all the results of [10] as a special case and more importantly, we obtain a variety of results establishing the *universality* of rumor centrality in the context of tree-like graphs and the SI model with a generic spreading time distribution.

1. Introduction. Imagine a rumor spreads through a network, and after a certain amount of time, we only know who has heard the rumor and the underlying network structure. With only this information, is it possible to determine who was the rumor source? Finding rumor sources is a very general type of problem which arises in many different contexts. For example, the rumor could be a computer virus on the Internet, a contagious disease in a human population, or some trend in a social network. In each of these scenarios, detection of the source is of interest as this source may be a malicious agent, patient zero, or an influential person.

There is limited prior work on the question of finding the source of a rumor. However, there has been much work done to understand conditions under which a rumor becomes an epidemic and how to use these insights to stop the spread [7],[9],[4]. In a thematically related problem of reconstruction, the interest is in predicting the state of the source based on the noisy observations about this state that are available in the network. Like the reconstruction problem, the source de-

tection problem is quite complex: the signal of interest, the state of source for reconstruction problem and the rumor source here is extremely 'low-dimensional' while the observations, the noisy versions of the state in reconstruction problem and infected nodes here, lie in a very 'high-dimensional' setting. This makes the estimation or detection quite challenging[1]. It is not surprising that even to obtain meaningful answers for the reconstruction problem for tree or tree-like graphs, sophisticated techniques have been required, cf.[3],[8], [5].

There are two key challenges that need to be addressed to resolve the rumor source detection problem. First, how does one actually construct the rumor source estimator? The estimator would naturally need to incorporate the topology of the underlying network, but it is not obvious in what manner. Second, what are the fundamental limits to this rumor source detection problem? In particular, how well can one find the rumor source, what is the distribution of the error, and how does the network structure affect one's ability to find the rumor source?

In a recent work [10], the authors introduced and studied the problem of rumor source detection in networks. They proposed *rumor centrality*, a graph-score function, for ranking the importance of nodes as the source. They showed that the node with maximal rumor centrality is the maximum likelihood (ML) estimation in the context of regular trees and the SI model with homogeneous exponential spreading times. They showed the effectiveness of this estimator by establishing that the rumor source is found with non-trivial probability for regular trees and geometric trees under this setting. The model and precise results from [10] are described in Section 2.

The key limitations of this prior work are : (i) the results do not quantify the exact detection probability, say α_d, for d-regular graphs, under the proposed maximum likelihood estimator other than $\alpha_2 = 0$, $\alpha_3 = 0.25$ and $0 < \alpha_d \leq 0.5$ for $d \geq 4$ for the SI model with exponential spreading times; (ii) the results do not quantify the magnitude of the error in the event of not being able to identify the source; and more generally, (iii) the results do not provide any insights into how the estimator behaves for generic heterogeneous tree (or tree-like) graphs under the SI model with a generic spreading time distribution.

1.1. *Summary of results.* The primary reason behind the limitations of the results in [10] is the fact that the analytic method employed there is quite specific to regular trees with homogeneous exponential spreading times. To overcome these limitations, as the main contribution of this work we introduce a novel analysis method that utilizes connections to the classical multi-class Markov branching process (MCMBP) (equivalently, generalized Polya's urn (GPU)). As a consequence of this, we are able to quantify the probability of the error event precisely and thus

[1]The phrase *finding needle in a haystack* seems quite appropriate.

eliminate the shortcomings of the prior work. The following is a summary of the key results (see Section 3 for precise statements):

1. *Regular tree, SI model with exponential spreading time*: we characterize α_d, the detection probability for d-regular trees, for all d. Specifically, for $d \geq 3$

$$\alpha_d = dI_{1/2}\Big(\frac{1}{d-2}, \frac{d-1}{d-2}\Big) - (d-1).$$

   In above $I_x(\alpha, \beta)$ is the incomplete beta function with parameters α, β evaluated at $x \in [0,1]$ (see (3.1)). This implies that $\alpha_d > 0$ for $d \geq 3$ with $\alpha_3 = 0.25$ and $\alpha_d \to 1 - \ln 2$ as $d \to \infty$. Further, we show that the probability of rumor centrality estimating the k^{th} infected node as the source decays as $\exp(-\Theta(k))$. The precise results are stated as Theorem 3.1, Corollaries 1 and 2.
2. *Generic random tree, SI model with exponential spreading time*: for generic random trees (see Section 3.2 for precise definition) which are expanding, we establish that there is strictly positive probability of correct detection using rumor centrality. Furthermore, the probability of rumor centrality estimating the k^{th} infected node as the source decays as $\Theta(1/k)$. The precise results are stated as Theorem 3.2 and Theorem 3.3.
3. *Geometric tree, SI model with spreading time with finite moment generating function*: for any geometric tree (see Section 3.2.2 for precise definition), we establish that the probability of correct detection goes to 1 as the number of infected nodes increases. The precise result is stated as Theorem 3.4.
4. *Generic random tree, SI model with generic spreading time*: for generic expanding random tree with generic spreading time (see Section 3.2 for definition), we establish that the probability of correct source detection remains bounded away from 0. The precise result is stated as Theorem 3.2.

The above results collectively establish that, even though, rumor centrality is an ML estimator only for regular tree and the SI model with exponential spreading times, it is universally effective with respect to heterogeneity in the tree structure and spreading time distributions. It's effectiveness for generic random trees immediately implies its utility for finding sources in sparse random graphs that are locally tree-like. Examples include Erdos-Renyi and random regular graphs. A brief discussion to this effect can be found in Section 3.3.

2. Model, problem statement and rumor centrality. We start by describing the model and problem statement followed by a quick recall of the precise results from [10]. In the process, we shall recall the definition of rumor centrality and source estimation based as introduced in [10].

2.1. *Model.* Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a possibly infinite connected graph. Let $v \in \mathcal{V}$ be a rumor source from which a rumor starts spreading at time, say 0. As per the classical Susceptible Infected (SI) model the rumor spreads in the graph. Specifically, each edge say $e = (u_1, u_2)$, has a spreading time, say S_e, associated with it. If node u_1 gets infected at time t_1, then at time $t_1 + S_e$ the infection spreads from u_1 to u_2. A node, once becoming infected, remains infected. The spreading times associated with edges are independent random variables with identical distribution. Let $F : \mathbb{R} \to [0, 1]$ denote the cumulative density function of the spreading time distribution. We shall assume that the distribution is non-negative valued, i.e. $F(0) = 0$ and it is non-atomic at 0, i.e. $F(0^+) = 0$. Since it is a cumulative density function, it is non-decreasing and $\lim_{x\to\infty} F(x) = 1$. The simplest, homogeneous SI model has exponential spreading times with parameter $\lambda > 0$ with $F(x) = 1 - \exp(-\lambda x)$ for $x \geq 0$. In [10], the results were restricted to this homogeneous exponential spreading time setting. In this paper, we shall develop results for arbitrary spreading time distributions consistent with the above assumptions.

Problem statement. Given the above spreading model, we observe the rumor infected graph $G = (V, E)$ at some time $t > 0$. We do not know the value of t or the realization of the spreading times on edges $e \in E$; we only know the rumor infected nodes $V \subset \mathcal{V}$ and edges between them $E = V \times V \cap \mathcal{E}$. The goal is to find the rumor source (among V) given G.

2.2. *Rumor centrality: an estimator.* To solve this problem, the notion of rumor centrality was introduced in [10]. Rumor centrality is a 'graph score' function. That is, it takes $G = (V, E)$ as input and assigns a positive number or score to each of the vertices. Then the estimated source is the one with maximal (ties broken uniformly at random) score or rumor centrality. The estimated source is called the 'rumor center'. We start with the precise description of rumor centrality for a tree[2] graph G: the rumor centrality of node $u \in V$ with respect to $G = (V, E)$ is

$$R(u, G) = \frac{|V|!}{\prod_{w\in V} T_w^u}, \tag{2.1}$$

where T_w^u is the size of the subtree of G that is rooted at w and points away from u. For example, in Figure 1, let u be node 1. Then $|V| = 5$; the subtree sizes are $T_1^1 = 5, T_2^1 = 3, T_3^1 = T_4^1 = T_5^1 = 1$ and hence $R(1, G) = 8$: exactly equal to the number of distinct spreading orders starting from 1. In [10], a linear time algorithm is described to compute the rumor centrality of all nodes building on the relation $R(u, G)/R(v, G) = T_u^v/T_v^u$ for neighboring nodes $u, v \in V$ ($(u, v) \in E$).

The rumor centrality of a given node $u \in V$ for a tree given by (2.1) is precisely the number of distinct spreading orders that could lead to the rumor infected graph

[2] We shall call an undirected graph a *tree*, if it is connected and it does not have any cycles.

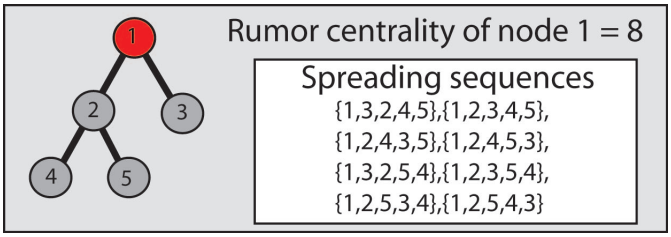

FIG 1. *Example of rumor centrality calculation for a* 5 *node network. The rumor centrality of node* 1 *is* 8 *because there are* 8 *spreading orders that it can originate, which are shown in the figure.*

G starting from u. This is equivalent to computing the number of linear extensions of the partial order imposed by the graph G due to causality constraints of rumor spreading. Under the SI model with homogeneous exponential spreading times and a regular tree, it turns out that each of the spreading orders is equally likely. Therefore, the rumor centrality turns out to be the Maximum Likelihood (ML) estimator for the source in this specific setting (cf. [10]). In general, the likelihood of each node $u \in V$ being the source given G is proportional to the weighted summation of the number of distinct spreading orders starting from u, where weight of a spreading order could depend on the details of the graph structure and spreading time distribution of the SI model. Now for a tree graph and SI model with homogeneous exponential spreading times, as mentioned above, such a quantity can be computed in linear time. But in general, this could be complicated. For example, computing the number of linear extensions of a given partial order is known to be #P-complete [2]. While there are algorithms for approximately sampling linear extensions given a partial order [6], [10] proposed a simpler alternative for general graphs.

DEFINITION 1 (Rumor Centrality). *Given node $u \in V$ in graph $G = (V, E)$, let $T \subset G$ denote the breadth-first search tree of u with respect to G. Then, the rumor centrality of u with respect to G is obtained by computing it as per* (2.1) *with respect to T. The estimated rumor source is the one with maximal rumor centrality (ties broken uniformly at random).*

2.3. *Prior results.* In [10], the authors established that rumor centrality is the maximum-likelihood estimator for the rumor source when the underlying graph $\mathcal{G}$ is a regular tree. They studied the effectiveness of this ML estimator for such regular trees. Specifically, suppose we observe the $n(t)$ node rumor infected graph G after time t, which is a subgraph of $\mathcal{G}$. Let $C^k_{n(t)}$ be the event that the source estimated as per rumor centrality is the kth infected node, and thus $C^1_{n(t)}$ corresponds to the event of correct detection. The following are key results from [10]:

THEOREM 2.1 ([10]). *Let $\mathcal{G}$ be a d-regular infinite tree with $d \geq 2$. Let*

$$\alpha_d^L = \liminf_{t\to\infty} \mathbf{P}\left(C^1_{n(t)}\right) \leq \limsup_{t\to\infty} \mathbf{P}\left(C^1_{n(t)}\right) = \alpha_d^U. \tag{2.2}$$

Then,

$$\alpha_2^L = \alpha_2^U = 0, \quad \alpha_3^L = \alpha_3^U = \frac{1}{4}, \quad \textit{and} \quad 0 < \alpha_d^L \leq \alpha_d^U \leq \frac{1}{2}, \ \forall\, d \geq 4. \tag{2.3}$$

3. Main results. We state the main results of this paper. In a nutshell, our results concern the characterization of the probability of $C^k_{n(t)}$ for any $k \geq 1$ for large t when $\mathcal{G}$ is a generic tree. As a consequence, it provides a characterization of the performance for sparse random graphs.

3.1. *Regular tree, SI model with exponential spreading time.* We first look at rumor source detection on regular trees with degree $d \geq 3$, where rumor centrality is an exact ML estimator when the spreading times are exponentially distributed. Our results will utilize properties of Beta random variables. We recall that the regularized incomplete Beta function $I_x(a,b)$ is the probability that a Beta random variable with parameters a and b is less than $x \in [0,1]$,

$$I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x t^{a-1}(1-t)^{b-1} dt, \tag{3.1}$$

where $\Gamma(\cdot)$ is the standard Gamma function. For regular trees of degree ≥ 3 we obtain the following result.

THEOREM 3.1. *Let $\mathcal{G}$ be d-regular infinite tree with $d \geq 3$. Assume a rumor spreads on $\mathcal{G}$ as per the SI model with exponential distribution with rate λ. Then, for any $k \geq 1$,*

$$\begin{aligned}\lim_{t\to\infty} \mathbf{P}\left(C^k_{n(t)}\right) =& I_{1/2}\left(k-1+\frac{1}{d-2}, 1+\frac{1}{d-2}\right) \\ &+ (d-1)\left(I_{1/2}\left(\frac{1}{d-2}, k+\frac{1}{d-2}\right) - 1\right).\end{aligned} \tag{3.2}$$

For $k = 1$, Theorem 3.1 yields that $\alpha_d^L = \alpha_d^U = \alpha_d$ for all $d \geq 3$ where

$$\alpha_d = dI_{1/2}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right) - (d-1). \tag{3.3}$$

More interestingly,

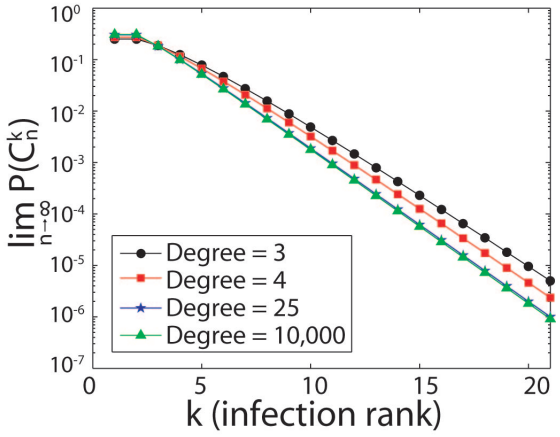

FIG 2. *(left)* $\lim_{t\to\infty} \mathbf{P}\left(C^k_{n(t)}\right)$ *versus* k *for regular trees of different degree.*

COROLLARY 1.

$$\lim_{d\to\infty} \alpha_d = 1 - \ln 2 \;\approx\; 0.307. \tag{3.4}$$

For any $d \geq 3$, we can obtain a simple upper bound for Theorem 3.1 which provides the insight that the probability of error in the estimation decays exponentially with error distance (not number of hops in graph, but based on chronological order of infection) from the true source.

COROLLARY 2. *When* $\mathcal{G}$ *is a* d*-regular infinite tree, for any* $k \geq 1$*,*

$$\lim_{t\to\infty} \mathbf{P}\Big(C^k_{n(t)}\Big) \leq k\,(k+1)\left(\frac{1}{2}\right)^{k-1} \;\sim\; \exp\Big(-\Theta(k)\Big).$$

To provide intuition, we plot the asymptotic error distribution $\lim_{t\to\infty} \mathbf{P}\left(C^k_{n(t)}\right)$ for different degree regular trees in Figure 2. As can be seen, for degrees greater than 4, all the error distributions fall on top of each other, and the probability of detecting the k^{th} infection as the source decays exponentially in k. We also plot the upper bound from Corollary 2. As can be seen, this upper bound captures the rate of decay of the error probability. Thus we see tight concentration of the error for this class of graphs. Figure 3 plots the asymptotic correct detection probability α_d versus degree d for these regular trees. It can be seen that the detection probability starts at $1/4$ for degree 3 and rapidly converges to $1 - \ln(2)$ as the degree goes to infinity.

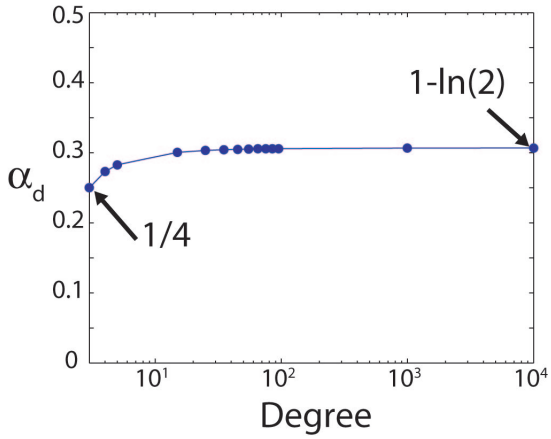

FIG 3. *α_d versus degree d for regular trees.*

3.2. *Generic random tree, SI model with general spreading time distribution.* The above precise results were obtained using the memoryless property of the exponential distribution and the regularity of the trees. Next, we wish to look at a more general setting both in terms of tree structures and spreading time distributions. In this more general setting, while we cannot obtain precise values for the detection and error probabilities, we are able to make statements about the non-triviality of the detection probability of rumor centrality. When restricted to exponential spreading times for generic trees, we can identify bounds on the error probability as well. Let us start by defining what we mean by generic random trees through a generative model.

DEFINITION 2 (Random Trees). *A random tree is generated as follows. Starting with a root node, add a random number of children, say η_0 to the root with $\eta_0 \in \{0, 1 \dots\}$ having some distribution $\mathcal{D}_0$. If $\eta_0 \neq 0$, then to each child of the root, add a random number of children chosen as per distribution $\mathcal{D}$ over $\{0, 1, \dots\}$. Recursively, to each newly added node, add independently a random number of nodes as per distribution $\mathcal{D}$. When 0 children are added to a node, that branch of the tree terminates there.*

The generative model described above is precisely the standard Galton-Watson branching process if $\mathcal{D}_0 = \mathcal{D}$. If we take $\mathcal{D}_0$ and $\mathcal{D}$ to be deterministic distributions with support on d and $d-1$ respectively, then it gives the d-regular tree. For a random d-regular graph on n nodes, as n grows the neighborhood of a randomly chosen node in the graph converges (in distribution, locally) to such a d-regular tree. If we take $\mathcal{D}_0 = \mathcal{D}$ as a Poisson distribution with mean $c > 0$, then it asymptotically

equals (in distribution) to the local neighborhood of a randomly chosen node in a sparse Erdos-Renyi graph as the number of nodes grows. Recall that a (sparse) Erdos-Renyi graph on n nodes with parameter c is generated by selecting each of the $\binom{n}{2}$ edges to be present with probability c/n independently. Effectively, random trees as described above captures the local structure for sparse random graphs reasonably well. For that reason, establishing the effectiveness of rumor centrality for source detection for such trees provide insights into its effectiveness for sparse random graph models.

We shall consider spreading time distributions to be generic: let $F : \mathbb{R} \to [0, 1]$ be the cumulative density function of spreading time distribution. Then $F(0) = 0$, $F(0^+) = 0$ and $\lim_{t\to\infty} F(t) = 1$. We state the following result about the effectiveness of rumor centrality with a generic spreading time distribution.

THEOREM 3.2. *Let η_0, distributed as $\mathcal{D}_0$, be such that $\Pr(\eta_0 \geq 3) > 0$ and let η, distributed as per $\mathcal{D}$, be such that $\mathbf{E}[\eta^2] < \infty$, $\mathbf{E}[\eta] > 1$. Suppose the rumor starts from the root of the random tree generated as per distributions $\mathcal{D}_0$ and $\mathcal{D}$ as described above and spreads as per the SI model with generic spreading time distribution as discussed above. Then,*

$$\liminf_{t\to\infty} \mathbf{P}\left(C^1_{n(t)}\right) > 0.$$

The above result says that irrespective of the structure of the random trees, spreading time distribution and time elapsed, there is non-trivial probability of detecting the root as the source by rumor centrality. The interesting aspect of the result is that this non-trivial detection probability is established by studying events when the tree grows without bound. In specific models (e.g. Poisson distribution of $\mathcal{D}_0,\ \mathcal{D}$), one may derive such a bound by identifying the probability of having an empty tree to be non-trivial. But, indeed, such events are trivial and are not of much interest to us (neither mathematically, nor motivationally).

3.2.1. *Generic random tree, SI model with exponential spreading time distribution.* Extending the results of Theorem 3.2 for explicitly bounding the probability of error event, $\mathbf{P}\left(C^k_{n(t)}\right)$, for generic spreading time distribution seems rather challenging. Here we provide a result for generic random trees with exponential spreading times.

THEOREM 3.3. *Consider the setup of Theorem 3.2 with spreading times being homogeneous exponential distributions with (unknown, but fixed) parameter $\lambda > 0$. In addition, let $\mathcal{D}_0 = \mathcal{D}$. Let (with some abuse of notation) η_i denote the number of children of i^{th} infected node. Then conditioned on the event that, (i) $\eta_k \geq 2$, and*

(ii) $\sum_{i=1}^{k-1} \eta_i \geq ck\eta_k$ *for some* $c > 1$,

$$\mathbf{P}\left(C^k_{n(t)}\right) \leq \frac{1}{k}. \tag{3.5}$$

The above result establishes an explicit upper bound on the error event under the occurrence of a specific event. The bound is relatively weaker ($1/k$ versus $\exp(-\Theta(k))$) and holds under specific conditions. However, it applies to essentially any generic random tree and demonstrates that the probability of (mis-)estimating later infected nodes decreases.

3.2.2. *Geometric trees.* The trees considered thus far, d-regular tree with $d \geq 3$ or random tree with $\mathbf{E}[\eta] > 1$, grow exponentially in size with the diameter of the tree. This is, in contrast with line graphs or d-regular trees with $d = 2$ which grow only linearly in diameter. It can be easily seen that the probability of correct detection, $\mathbf{P}(C^1_{n(t)})$ will scale as $\Theta(1/\sqrt{t})$ for line graphs as long as the spreading time distribution has non-trivial variance (see [10] for proof of this statement for SI model with exponential spreading times). In contrast, the results of this paper stated thus far suggest that the expanding trees allow for non-trivial detection as $t \to \infty$. Thus, qualitatively line (tree) graphs and expanding trees are quite different – one does not allow detection while the other does. To understand where the precise detectability threshold lies, here we study the polynomially growing *geometric* trees.

DEFINITION 3 (Geometric Tree). *This family of rooted, non-regular trees are parameterized by constants* α*,* b*, and* c*, with* $0 < b \leq c$ *and a root* v^**. Let* d^* *be the degree of this root* v^* *and let the* d^* *neighboring subtree of* v^**, be denoted by* $T_1, \ldots, T_{d^*}$*. Consider the* i*th 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* $v \in T_i$

$$br^\alpha \leq n^i(v, r) \leq cr^\alpha. \tag{3.6}$$

The condition imposed by (3.6) states that each of the neighboring subtrees of the root 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. Note that the line graph is a geometric tree with $\alpha = 0$, $b = 1$, and $c = 2$.

We shall consider the scenario where rumor starts from the root node of a rooted geometric tree. We shall show that rumor centrality detects the root as the source

with asymptotic probability of 1 for a generic spreading time distribution with exponential tails. This is quite interesting given the fact that rumor centrality is an ML estimator only for regular tree with exponential spreading time. The precise result is stated next.

THEOREM 3.4. *Let $\mathcal{G}$ be a rooted geometric tree as described above with parameters $\alpha > 0$, $0 < b \leq c$ and root node v^* with degree d^* such that*

$$d_{v^*} > \max\left(2, \frac{c}{b} + 1\right).$$

Suppose the rumor starts spreading on $\mathcal{G}$ starting from v^* as per th e SI model with generic spreading time distribution whose cumulative density function $F : \mathbb{R} \to [0, 1]$ is such that (a) $F(0) = 0$, (b) $F(0^+) = 0$, and (c) if X is a random variable distributed as per F then $\mathbf{E}[\exp(\theta X)] < \infty$ for $\theta \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$. Then

$$\lim_t \mathbf{P}(C^1_{n(t)}) = 1.$$

A similar theorem was proven in [10], but only for the SI model with an exponential distribution. We have now extended this result to arbitrarily distributed spreading times. Theorem 3.4 says that $\alpha = 0$ and $\alpha > 0$ serve as a threshold for non-trivial detection: for $\alpha = 0$, the graph is essentially a linear graph, so we would expect the detection probability to go to 0 as $t \to \infty$ as discussed above, but for $\alpha > 0$ the detection probability converges to 1 as $t \to \infty$.

3.3. *Locally tree-like graphs: discussion.* The results of the paper stated are primarily for all sorts of tree structured graphs. On one hand, they are specialized. On the other hand, they do serve as local approximations for a variety of sparse random graph models. As discussed earlier, for a random d-regular graph over m nodes, a randomly chosen node's local neighborhood (say up to distance $o(\log m)$) is a tree with high probability. Similarly, consider an Erdos-Renyi graph over m nodes with each edge being present with probability $p = c/m$ independently for any $c > 0$ ($c > 1$ is an interesting regime due to existence the of a giant component). Then again, a randomly chosen node's local neighborhood (up to distance $o(\log m)$) is a tree and distributionally equivalent (in the large m limit) to a random tree with Poisson degree distribution.

Given such 'locally tree-like' structural properties, if a rumor spreads on a random d-regular graph or sparse Erdos-Renyi graph for time $o(\log m)$ starting from a random node, then rumor centrality can detect the source with guarantees given by Theorems 3.1 and 3.2. Thus, effectively though the results of this paper are for tree structured graphs, they do have meaningful implications for tree-like sparse graphs.

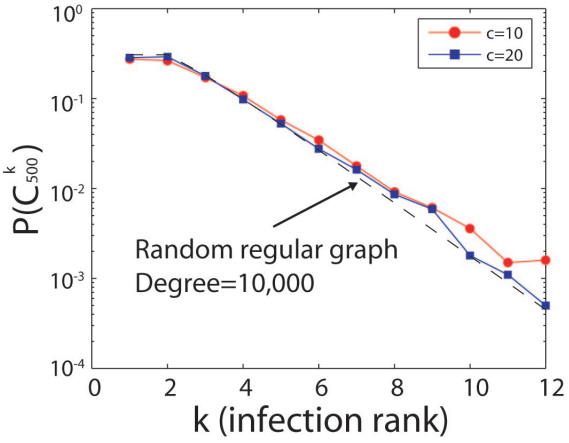

FIG 4. $\mathbf{P}\left(C_{500}^{k}\right)$ *versus* k *for Erdos-Renyi graphs with mean degree* 10 *and* 20. *Also shown with a black dashed line is* $\lim_{t\to\infty}\mathbf{P}\left(C_{n(t)}^{k}\right)$ *for a degree* $10,000$ *random regular graph.*

For the purpose of illustration, we conducted some simulations for Erdos-Renyi graphs that are reported in Figure 4. We generated graphs with $m = 50,000$ nodes and edge probabilities $p = c/m$ for $c = 10$ and $c = 20$. The rumor graph contained $n = 500$ nodes. We ran $10,000$ rumor spreading simulations to obtain the empirical error distributions plotted in Figure 4. As can be seen, the error drops of exponentially, very similar to the regular tree error distribution. In fact, we also plot the distribution for regular tree of degree $10,000$ and it can be seen that the error decays at similar, exponential rates. This indicates that even though there is substantial randomness in the graph, the asymptotic rumor source detection error distribution behaves as though it were a regular tree graph. This result also suggests that the bounds in Theorem 3.3 are loose for this graph.

4. Proofs. Here proofs of the results stated in Section 3 are presented. We establish results for d-regular trees by connecting rumor spreading with Polya urn models and branching processes. Later we extend this novel method to establish results for generic random trees under arbitrary spreading time distributions. After this, we prove Theorem 3.4 using standard Chernoff's bound and the polynomial growth property of geometric trees.

4.1. *Proof of Theorem 3.1: d-regular trees.*

4.1.1. *Setup, notations.* Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an infinite d-regular tree and let the rumor start spreading from a node, say v_1. Without loss of generality, we view it as a randomly generated tree, as described in Section 3, with v_1 being the root with

d children and all the subsequent nodes with $d-1$ children (hence each node has degree d). We shall be interested in $d \geq 3$. Now suppose the rumor is spread on this tree starting from v_1 as per the SI model with exponential distribution with rate $\lambda > 0$.

Initially, node v_1 is the only rumor infected node and its d neighbors are potential nodes that can receive the rumor. We will denote the set of nodes that are not yet rumor infected but neighbors of rumor infected nodes as the *rumor boundary*. Thus, initially the rumor boundary consists of the d neighbors of v_1. Under the SI model, each edge has an independent exponential clock of mean $1/\lambda$. The minimum of d independent exponentials of mean $1/\lambda$ is an exponential random variable (of mean $1/(d\lambda)$) and hence one of the d nodes (chosen uniformly at random) in the rumor boundary gets infected after an exponential amount of time (of mean $1/(d\lambda)$). Upon this infection, this node gets removed from the boundary but adds it's $d-1$ children to the rumor boundary. That is, each infection effectively adds $d-2$ new nodes to the rumor boundary. In summary, let $Z(t)$ denote the number of nodes in the rumor boundary at time t, then $Z(0) = d$ and $Z(t)$ evolves as follows: each of the $Z(t)$ nodes has an exponential clock of mean $1/\lambda$; when it ticks, it dies and replaces itself with $d-2$ new nodes which in turn start their own independent exponential clocks of mean $1/\lambda$ and so on. The thus described $Z(t)$ is precisely the multi-class Markov branching process (MCMBP): the multi-class comes when we think of the contributions of each of the d sub-tress of the root v_1 to the rumor boundary separately and hence effectively having d branching processes each starting at time 0 with initial value equal to 1: let $u_1, \ldots, u_d$ be the children of v_1; let $Z_i(t)$ denote the number of nodes in the rumor boundary that belong to the subtree $T_i(t)$ that is rooted at u_i with $Z_i(0) = 1$ for $1 \leq i \leq d$; $Z(t) = \sum_{i=1}^{d} Z_i(t)$. With an abuse of notation, let $T_i(t)$ also denote the total number of nodes infected in the subtree rooted at u_i at time t; initially $T_i(0) = 0$ for $1 \leq i \leq d$. Since each infected node add $d-2$ nodes to the rumor boundary, it can be easily checked that $Z_i(t) = (d-2)T_i(t) + 1$ and hence $Z(t) = (d-2)T(t) + d$ with $T(t)$ being the total number of infected nodes at time t (excluding v_1, i.e. $T(0) = 0$).

4.1.2. *Probability of correct detection* $\mathbf{P}\left(C^1_{n(t)}\right)$. Suppose we observe the rumor infected nodes at some time t which we do not know. That is, we observe the rumor infected graph $G(t)$ which contains the root v_1 and its d infected subtrees $T_i(t)$ for $1 \leq i \leq d$. We recall the following result of [10] that characterizes the rumor center.

LEMMA 1 ([10]). *Given a rumor infected tree $G = (V, E)$, there can be at*

most two rumor centers. Specifically, a node $v \in V$ is a rumor center if and only if

$$T_i^v \leq \frac{1}{2}\Big(\sum_{j\in\mathcal{N}(v)} T_j^v\Big), \quad \forall\, i \in \mathcal{N}(v), \tag{4.1}$$

where $\mathcal{N}(v) = \{u \in V : (u,v) \in E\}$ are neighbors of v in G and T_j^v denotes the sub-tree of G that is rooted at node $j \in \mathcal{N}(v)$ that includes all nodes that are away from node v (i.e. the subtree does not include v). The node v is the unique rumor center if the inequality in (4.1) is strict for all $i \in \mathcal{N}(v)$.

This immediately suggests the characterization of the event that node v_1, the true source, is identified by rumor centrality at time t: v_1 is a rumor center only if $2T_i(t) \leq \sum_{j=1}^d T_j(t)$ for all $1 \leq i \leq d$, and if the inequality is strict then it is indeed the rumor center. Now if $n(t)$ is the total number of infected nodes at time t, then as per our earlier notation, $C^1_{n(t)}$ is the event of correct detection at time t. Let $E_i = \{2T_i(t) < \sum_{j=1}^d T_j(t)\}$ and $F_i = \{2T_i(t) \leq \sum_{j=1}^d T_j(t)\}$. Then,

(4.2)

$$\mathbf{P}\Big(C^1_{n(t)}\Big) \geq \mathbf{P}\Big(\cap_{i=1}^d E_i\Big) = 1 - \mathbf{P}\Big(\cup_{i=1}^d E_i^c\Big) \overset{(a)}{\geq} 1 - \sum_{i=1}^d \mathbf{P}\Big(E_i^c\Big) \overset{(b)}{=} 1 - d\mathbf{P}\Big(E_1^c\Big).$$

Above, (a) follows from the union bound of events and (b) from symmetry. Similarly, we have

(4.3)

$$\mathbf{P}\Big(C^1_{n(t)}\Big) \leq \mathbf{P}\Big(\cap_{i=1}^d F_i\Big) = 1 - \mathbf{P}\Big(\cup_{i=1}^d F_i^c\Big) \overset{(a)}{=} 1 - \sum_{i=1}^d \mathbf{P}\Big(F_i^c\Big) \overset{(b)}{=} 1 - d\mathbf{P}\Big(F_1^c\Big).$$

Above, (a) follows because events $F_1^c, \ldots, F_d^c$ are disjoint and (b) from symmetry. Therefore, the probability of correct detection boils down to evaluating $\mathbf{P}(E_1^c)$ and $\mathbf{P}(F_1^c)$ which, as we shall see, will coincide with each other as $t \to \infty$ (equivalently, $n(t) \to \infty$). Therefore, the bounds of (4.2) and (4.3) will provide the exact evaluation of the correct detection probability as $t \to \infty$.

4.1.3. $\mathbf{P}(E_1^c)$, $\mathbf{P}(F_1^c)$ *and Polya's urn.* Effectively, the interest is in the ratio $T_1(t)/(\sum_{i=1}^d T_i(t))$ especially as $t \to \infty$ (implicitly we are assuming that this ratio is well defined for a given t or else by definition there is only one node infected which will be v_1, the true source). It can be easily verified that as $t \to \infty$, $T_i(t) \to \infty$ for all i almost surely and hence $Z_i(t) = (d-2)T_i(t) + 1$ goes to ∞ as well. Therefore, it is sufficient to study the ratio $Z_1(t)/(\sum_{j=1}^d Z_j(t))$ as $t \to \infty$ since we shall find that this ratio converges to a random variable with density on

$[0,1]$. In summary, if we establish that the ratio $Z_1(t)/(\sum_{j=1}^d Z_j(t))$ converges in distribution on $[0,1]$ with a well defined density, then it immediately follows that $\mathbf{P}(E_1^c) = \mathbf{P}(F_1^c)$ as $t \to \infty$ and we can use $Z_1(t)/(\sum_{j=1}^d Z_j(t))$ in place of $T_1(t)/(\sum_{j=1}^d T_j(t))$ to evaluate $\mathbf{P}(E_1^c)$ as $t \to \infty$.

With these in mind, let us study the ratio $Z_1(t)/(\sum_{j=1}^d Z_j(t))$. For this, it is instructive to view the simultaneous evolution of $(Z_1(t), Z_{\neq 1}(t))$ $(Z_{\neq 1}(t) \stackrel{\triangle}{=} \sum_{j=2}^d Z_j(t))$ as that induced by the standard, discrete time, Polya's urn: initially, $\tau_0 = 0$ and there is one ball of type 1 representing $Z_1(\tau_0) = 1$ and $d-1$ balls of type 2 representing $Z_{\neq 1}(\tau_0) = d-1$ in a given *urn*; the n^{th} event happens at time τ_n when one of the $Z_1(\tau_{n-1}) + Z_{\neq 1}(\tau_{n-1})$ $(= d + (n-1)(d-2))$ balls chosen uniformly at random is thrown out of the urn and new $d-2$ balls of its type are added to the urn. If we set $\tau_n - \tau_{n-1}$ equal to an exponential random variable with mean $1/(\lambda(d+(n-1)(d-2)))$, then it is easy to check that the fraction of balls of type 1 is identical in law to that of $Z_1(t)/(\sum_{j=1}^d Z_j(t))$ (here we are using the *memoryless* property of exponential random variables crucially). Therefore, for our purposes, it is sufficient to study the limit law of fraction of balls of type 1 under this Polya's urn model.

It is well understood that the fraction of balls of type 1 at time τ_n, which is equal to $Z_1(\tau_n)/(Z_1(\tau_n) + Z_{\neq 1}(\tau_n))$, is a martingale with value in $[0,1]$. By the standard martingale convergence theorem, it converges to a well defined random variable almost surely. Further, the law of this limiting random variable in our particular case turns out to be the Beta distribution with parameters $a = 1/(d-2)$ and $b = (d-1)/(d-2)$. (See chapter on generalized Polya's urn in [1], for example, for proof details of this statement; a more general version of this result will be utilized later in the context of general random trees).

From the above discussion, we conclude that the ratio $Z_1(t)/(\sum_{i=1}^d Z_i(t))$ converges to a Beta random variable with parameters $a = 1/(d-2)$ and $b = (d-1)/(d-2)$. Since the Beta distribution has density on $[0,1]$, from the above discussion it follows that as $t \to \infty$ (equivalently, $n(t) \to \infty$), $\mathbf{P}(E_1^c) = \mathbf{P}(F_1^c)$ and hence from (4.2), (4.3)

$$\lim_{t\to\infty} \mathbf{P}\Big(C^1_{n(t)}\Big) = 1 - d\Big(1 - I_{1/2}\Big(\frac{1}{d-2}, 1 + \frac{1}{d-2}\Big)\Big), \tag{4.4}$$

where recall that $I_{1/2}(a,b)$ is the probability that a Beta random variable with parameters a and b takes value in $[0,1/2]$. Note that this establishes the result of Theorem 3.1 for $k=1$ in (3.2).

4.1.4. *Probability of $C^k_{n(t)}$.* Thus far we have established Theorem 3.1 for $k = 1$, the probability of the rumor center being the true source. The probability of the event $C^k_{n(t)}$ (the kth infected node being the rumor center) is evaluated in an almost

identical manner with a minor difference. For this reason, we present an abridged version of the proof.

Let $v_k, k \geq 2$ be the k^{th} infected node when the rumor starts from v_1. We will evaluate the probability of identifying v_k as the rumor center. As before, let us suppose we observe the infected tree G at time t with $n(t) \geq k$ nodes. Let $w_1, \ldots, w_d$ be the d neighbors of v_k with respect to $\mathcal{G}$. Note that one of the neighbors of v_k is infected before it. Equivalently, viewing the tree being rooted at v_1, this neighbor is the 'parent' of v_k. We shall denote it by w_1 and let $w_2, \ldots, w_d$ be the $d-1$ 'children' of v_k. Let $T_i^k(t)$ be the subtrees of G rooted at w_i if we imagine v_k as the root of G: therefore, $T_1^k(t)$ is rooted at w_1 and includes $v_1, \ldots, v_{k-1}$; $T_i^k(t)$ are rooted at w_i for $2 \leq i \leq d$ and contain nodes in G that are away from v_k; none of the $T_i^k(t)$ for $1 \leq i \leq d$ include v_k. By definition $T_1^k(t)$ is never empty, but $T_i^k(t)$ can be empty if w_i is not infected, for $2 \leq i \leq d$. As before, with abuse of notation, we shall denote $T_i^k(t)$ as the size of the subtree as well. As per Lemma 1, v_k is identified as a rumor center if and only if all of its d subtrees are balanced, i.e.

$$2T_i^k(t) \leq \sum_{j=1}^{d} T_j^k(t), \ \ \forall\, 1 \leq i \leq d. \tag{4.5}$$

Therefore, as before,

$$\mathbf{P}\Big(C_{n(t)}^k\Big) \geq \mathbf{P}\Big(\cap_{i=1}^d E_i\Big) = 1 - \mathbf{P}\Big(\cup_{i=1}^d E_i^c\Big) \geq 1 - \sum_{i=1}^{d} \mathbf{P}\Big(E_i^c\Big), \tag{4.6}$$

and

$$\mathbf{P}\Big(C_{n(t)}^k\Big) \leq \mathbf{P}\Big(\cap_{i=1}^d F_i\Big) = 1 - \mathbf{P}\Big(\cup_{i=1}^d F_i^c\Big) = 1 - \sum_{i=1}^{d} \mathbf{P}\Big(F_i^c\Big). \tag{4.7}$$

Above, $E_i = \{2T_i^k(t) < \sum_{j=1}^d T_j^k(t)\}$ and $F_i = \{2T_i^k(t) \leq \sum_{j=1}^d T_j^k(t)\}$.

To evaluate these probabilities, as before, we shall study the evolution of the rumor boundaries in each tree. Unlike the earlier situation where all events had the same probability, we need to be a bit careful for $k \geq 2$. Specifically, note that the time when the kth node, v_k gets infected, the tree $T_1^k(\cdot)$ has size $k-1$ and it's rumor boundary, $Z_1^k(\cdot)$, is of size $(d-2)(k-1)+1$. But for $2 \leq i \leq d$, $T_i^k(\cdot)$ is empty and has its rumor boundary, $Z_i^k(\cdot)$, of size 1. Beyond this difference in initial values, the evolution is the same as before and therefore the limiting laws of the ratios of the sizes of rumor boundaries can be evaluated as before as the limit of the fraction of balls of a given type in a Polya's urn model with different initial number of balls of two types. Specifically, to evaluate E_1^c (and F_1^c), we consider a Polya's urn in

which we start with $(d-2)(k-1)+1$ balls of type 1 (corresponding to $Z_1^k(\cdot)$) and $d-1$ balls of type 2 (corresponding to $\sum_{j=2}^d Z_j^k(\cdot)$). With these initial conditions, the limit law of fraction of balls of type 1 turns out to be (see [1] for details) a Beta distribution with parameters $a = ((d-2)(k-1)+1)/(d-2) = (k-1) + 1/(d-2)$ and $b = (d-1)/(d-2) = 1 + 1/(d-2)$. In summary, the fraction of balls of type 1 equals the ratio $Z_1^k(t)/(\sum_{j=1}^k Z_j^k(t))$ which is equal to the ratio of $T_1^k(t)/(\sum_{j=1}^d T_j^k(t))$ as $t \to \infty$, since these quantities go to infinity as $t \to \infty$ and $Z_i^k(t)$ is linearly related to $T_i^k(t)$ for $1 \le i \le d$. Since the limit law of the fraction of balls of type 1 has density on $[0,1]$, we conclude that as $t \to \infty$

$$\mathbf{P}\Big(E_1^c\Big) \;=\; \mathbf{P}\Big(F_1^c\Big) \;=\; 1 - I_{1/2}\Big(k-1+\frac{1}{d-2},\; 1+\frac{1}{d-2}\Big). \tag{4.8}$$

To evaluate E_i^c (and F_i^c) for $2 \le i \le d$, in the corresponding Polya's urn model, we start with 1 ball of type 1 and $k(d-2)+1$ balls of type 2. Therefore, using an identical sequence of arguments, we obtain that for $2 \le i \le d$ as $t \to \infty$,

$$\mathbf{P}\Big(E_i^c\Big) \;=\; \mathbf{P}\Big(F_i^c\Big) \;=\; 1 - I_{1/2}\Big(\frac{1}{d-2},\; k+\frac{1}{d-2}\Big). \tag{4.9}$$

From (4.6)-(4.9), it follows that

$$\begin{aligned}\lim_{t\to\infty} \mathbf{P}\Big(C_{n(t)}^k\Big) =& I_{1/2}\Big(k-1+\frac{1}{d-2},\; 1+\frac{1}{d-2}\Big) \\ &+ (d-1)\Big(I_{1/2}\Big(\frac{1}{d-2},\; k+\frac{1}{d-2}\Big) - 1\Big).\end{aligned} \tag{4.10}$$

This establishes (3.2) for all k and hence completes the proof of Theorem 3.1.

4.2. *Proof of Corollary 1.* Simple analysis yields Corollary 1. We start by defining the asymptotic probability for a d-regular tree as $\lim_{t\to\infty} \mathbf{P}\left(C_{n(t)}^1\right) = \alpha_d$. This quantity then becomes

$$\begin{aligned}\alpha_d =& d\left(I_{1/2}\left(\frac{1}{d-2}, 1+\frac{1}{d-2}\right)\right) - d + 1 \\ =& 1 - d\frac{\Gamma(1+\frac{2}{d-2})}{\Gamma(\frac{1}{d-2})\Gamma(1+\frac{1}{d-2})}\int_{\frac{1}{2}}^1 t^{\frac{1}{d-2}-1}(1-t)^{\frac{1}{d-2}}dt\end{aligned}$$

We then take the limit as d approaches infinity.

$$\begin{aligned}\lim_{d\to\infty}\alpha_d &= \lim_{d\to\infty} 1 - d\frac{\Gamma(1+\frac{2}{d-2})}{\Gamma(\frac{1}{d-2})\Gamma(1+\frac{1}{d-2})}\int_{\frac{1}{2}}^{1} t^{\frac{1}{d-2}-1}(1-t)^{\frac{1}{d-2}}dt \\ &= 1 - \lim_{d\to\infty}\frac{d\Gamma(1+\frac{2}{d-2})}{(d-2-\gamma+O\left(d^{-1}\right))\Gamma(1+\frac{1}{d-2})}\int_{\frac{1}{2}}^{1} t^{\frac{1}{d-2}}\left(1-t\right)^{\frac{1}{d-2}}dt \\ &= 1 - \int_{\frac{1}{2}}^{1} t^{-1}dt \\ &= 1 - \ln\left(2\right)\end{aligned}$$

In above, γ is the Euler-Mascheroni constant and we have used the following approximation for $\Gamma(x)$ for small x:

$$\Gamma\left(x\right) = \frac{1}{x} - \gamma + O\left(x\right)$$

4.3. *Proof of Corollary 2.* Corollary 2 follows from (3.2) and monotonicity of the Γ function over $[1,\infty)$ as follows: for $k \geq 2$,

$$\begin{aligned}\lim_{t\to\infty}\mathbf{P}\Big(C^k_{n(t)}\Big) &= I_{1/2}\left(k-1+\frac{1}{d-2}, 1+\frac{1}{d-2}\right) \\ &\qquad + (d-1)\left(I_{1/2}\left(\frac{1}{d-2}, k+\frac{1}{d-2}\right)-1\right) \\ &\leq I_{1/2}\left(k-1+\frac{1}{d-2}, 1+\frac{1}{d-2}\right) \\ &= \frac{\Gamma(k+\frac{2}{d-2})}{\Gamma(k-1+\frac{1}{d-2})\Gamma(1+\frac{1}{d-2})}\int_0^{\frac{1}{2}} t^{k+\frac{1}{d-2}-2}(1-t)^{\frac{1}{d-2}}dt \\ &\overset{(a)}{\leq} \frac{\Gamma(k+\frac{2}{d-2})}{\Gamma(k-1+\frac{1}{d-2})\Gamma(1+\frac{1}{d-2})}\int_0^{\frac{1}{2}} t^{k-2}dt \\ &\overset{(b)}{\leq} \frac{4e^2\Gamma(k+2)}{\Gamma(k-1)\Gamma(1)}\int_0^{\frac{1}{2}} t^{k-2}dt \\ &\leq \frac{4e^2k(k+1)}{k-1}\left(\frac{1}{2}\right)^{k-1} \\ &= \exp\Big(-k\ln 2+\varepsilon(k)\Big) \;=\; \exp\big(-\Theta(k)\big),\end{aligned}$$

where $\varepsilon(k) = \Theta(\log k)$. In above (a) follows from the fact that $t < 1$ and hence $t^{k-2+1/(d-2)} \leq t^{k-2}$. For (b), we need to use property of the Γ function carefully. It is well known that Γ achieves minimum value in $[1,2]$, $\Gamma(1) = \Gamma(2) = 1$, and

$\Gamma(x)$ increases for $x \geq 2$. Now the minimum value of Γ function is at least $1/(2e)$. Since $d \geq 3$, $1 + 1/(d-2) \in [1,2]$. Therefore, $\Gamma(1+1/(d-2)) \leq \Gamma(1)/(2e)$. Similarly, $\Gamma(k-1+1/(d-2)) \geq \Gamma(k-1)/(2e)$ for all $k \geq 2$ and $d \geq 3$. Therefore, (b) follows.

4.4. *Proof of Theorem 3.2: correct detection for random trees.* The goal is to establish that there is a strictly positive probability of detecting the source correctly as the rumor center when the rumor starts at the root of a generic randomly generated tree: this is with respect to the joint probability distribution induced by the tree construction and the SI rumor spreading model with independent spreading times with distribution $F(t)$. We establish this result along the lines of the proof for Theorem 3.1. However, it requires additional details due to the irregularity and randomness in the node degrees in the tree and the arbitrary spreading time distribution $F(t)$.

4.4.1. *Notations.* We quickly recall some notations. To start with, as before let v_1 be the root node of the tree. It has η_0 children distributed as per $\mathcal{D}_0$. By assumption of Theorem 3.2, $\mathbf{P}(\eta_0 \geq 3) > 0$. We shall condition on this positive probability event that $\eta_0 \geq 3$ and let $d = \eta_0$ for the remainder of the proof. Let $u_1, \ldots, u_d$ be the d children of v_1. The random tree $\mathcal{G}$ is constructed by adding a random number of children to $u_1, \ldots, u_d$ recursively as per distribution $\mathcal{D}$ as explained in Section 3.2. The rumor starts at v_1 at time 0 and spreads as per SI model on $\mathcal{G}$ with spreading times whose cumulative density function $F : \mathbb{R} \to [0,1]$ is such that $F(0) = 0$, $F(0^+) = 0$ and $\lim_{t\to\infty} F(t) = 1$.

Let G be the sub-tree of $\mathcal{G}$ that is infected at time t, let $n(t)$ be the number of nodes in G at time t, and let $T_i(t)$ denote the subtree of G rooted at node u_i at time t, for $1 \leq i \leq d$. We shall abuse notation as before by using $T_i(t)$ as the subtree size as well. By definition $T_i(0) = 0$ for $1 \leq i \leq d$. Let $Z_i(t)$ denote the size of the rumor boundary of $T_i(t)$; initially $Z_i(0) = 1$.

Since we are interested in evaluating the probability of detection with respect to the joint distribution over the tree generation and SI spreading model, we model the evolution of $Z_i(\cdot)$ and $T_i(\cdot)$ as follows. Each node in the rumor boundary has its own independent clock with distribution $F(t)$. When the clock of a particular node ticks, it dies and it is replaced by η new nodes where η is an independent random variable distributed as per $\mathcal{D}$. If the node that died belonged to $Z_i(\cdot)$ (i.e. tree $T_i(\cdot)$), then the new nodes are added to $Z_i(\cdot)$. Therefore, each $Z_i(\cdot)$ is a general time branching process with $Z_i(0) = 1$ for all $1 \leq i \leq d$. Now we recall some facts about branching processes that will be useful for establishing the non-triviality of the probability of correct detection for such randomly generated trees.

First we define what is known as the Malthusian parameter for the branching process.

DEFINITION 4. *[1, pp. 146] The Malthusian parameter for a constant γ and a distribution F is the root, provided it exists, of the equation*

$$\gamma \int_0^\infty e^{-\alpha y} dF(y) = 1. \tag{4.11}$$

We denote it by $\alpha = \alpha(\gamma, F)$.

For any $\gamma > 1$ the Malthusian parameter always exists. The Malthusian parameter characterizes the growth rate of a general time branching process. Consider for example a Markov branching process with exponential spreading times of rate λ and let $\mathbf{E}[\eta] = m > 1$. The spreading time distribution is $F(t) = 1 - e^{-\lambda t}$ Then the Malthusian parameter for this process $\alpha(m, F)$ is given by

$$\begin{aligned} m \int_0^\infty e^{-\alpha y} \lambda e^{-\lambda y} dy &= 1 \\ m \frac{\lambda}{\alpha + \lambda} &= 1 \\ \lambda(m-1) &= \alpha. \end{aligned}$$

This is exactly the growth rate for the mean of a Markov branching process [1]. The Malthusian parameter also describes the growth of the mean of general time branching processes, as the following theorem shows.

THEOREM 4.1. *[1, Theorem 3A, pp. 152] Let $Z(\cdot)$ be a continuous time branching process as described above: $Z(0) = 1$; each node in $Z(t)$ has an independent clock with distribution whose CDF is F as described above, and it dies upon the tick of the clock; upon death of a node, it is replaced by η new nodes chosen independently for each node, and so on. If $\mathbf{E}[\eta] = m > 1$ then as $t \to \infty$,*

$$\lim_{t \to \infty} \frac{\mathbf{E}\left[Z(t)\right]}{c' e^{\alpha t}} = 1,$$

where α is the Malthusian parameter for (m, F) and

$$c' = \frac{m-1}{\alpha m^2 \int_0^\infty y e^{-\alpha y} dF(y)}.$$

This theorem says that the mean of the branching process $Z(t)$ has exponential growth with rate given by the Malthusian parameter $\alpha(m, F)$. As an example of

this general theorem, consider again the Markov branching process with exponential spreading times with rate λ. Then we have already seen that the Malthusian parameter $\alpha = \lambda(\mathbf{E}[\eta] - 1)$. The constant c' evaluates to

$$\begin{aligned} c' &= \frac{m-1}{\alpha m^2 \int_0^\infty y e^{-\alpha y} \lambda e^{-\lambda y} dy} \\ &= \frac{(m-1)(\alpha+\lambda)^2}{\lambda \alpha m^2} \\ &= \frac{(m-1)(\lambda(m-1)+\lambda)^2}{\lambda^2 (m-1) m^2} \\ &= \frac{(m-1)\lambda^2 m^2}{\lambda^2 (m-1) m^2} \\ &= 1 \end{aligned}$$

Therefore, we have that for the Markov branching process, $\mathbf{E}[Z(t)] = e^{\lambda(\mathbf{E}[\eta]-1)}$, a well known result [1]. For the general time branching process, we have the following result.

THEOREM 4.2. *[1, Theorem 2, pp. 172] Let $Z(\cdot)$ be a continuous time branching process as described above: $Z(0) = 1$; each node in $Z(t)$ has an independent clock with distribution whose CDF is given by F and it dies upon the tick of the clock; upon death of a node, it is replaced by η new nodes chosen independently for each node, and so on. Let α be the Malthusian parameter for $(\mathbf{E}[\eta], F)$ and c' be defined as in Theorem 4.1. If $\mathbf{E}[\eta] > 1$ and $\mathbf{E}[\eta \log \eta] < \infty$, then*

$$Z(t)/c' e^{\alpha t} \to W, \qquad \textit{in distribution},$$

where W is such that

$$\mathbf{P}\big(W = 0\big) = q, \tag{4.12}$$

$$\mathbf{P}\big(W \in (x_1, x_2)\big) = \int_{x_1}^{x_2} w(y) dy, \quad \textit{for} \quad 0 < x_1 < x_2 < \infty, \tag{4.13}$$

where $q \in (0,1)$ is the smallest root of the equation $f_\eta(s) = s$ in $[0,1]$ and $w(\cdot)$ is absolutely continuous with respect to the Lebesgue measure so that $\int_0^\infty w(y)dy = 1 - q$. Here $f_\eta(s) = \sum_{k=0}^\infty s^k \mathbf{P}(\eta = k)$.

As we will see, this theorem will be key to proving Theorem 3.2.

4.4.2. *Correct detection.* Recall from the proof for regular trees, the probability of the event of correct detection at time t, $C^1_{n(t)}$, is lower bounded as

$$\mathbf{P}\Big(C^1_{n(t)}\Big) \geq \mathbf{P}\Big(\cap_{i=1}^d \Big\{2T_i(t) < \sum_{j=1}^d T_j(t)\Big\}\Big). \tag{4.14}$$

We shall establish a non-trivial lower bound for the right hand side of (4.14) using Theorem 4.2. This will be done in the two steps: (i) identify an event $\mathcal{E} \subset \cap_{i=1}^d \Big\{2Z_i(t) < \sum_{j=1}^d Z_j(t)\Big\}$ and then establish a non-trivial lower bound on $\mathcal{E}$ using Theorem 4.2; (ii) establish that as $t \to \infty$, $\mathcal{E} \subset \cap_{i=1}^d \Big\{2T_i(t) < \sum_{j=1}^d T_j(t)\Big\}$. This will immediately imply that $\mathbf{P}(\mathcal{E})$ is a non-trivial lower bound on $\mathbf{P}\Big(C^1_{n(t)}\Big)$.

4.4.3. *A non-trivial event.* The event $2Z_i(t) < \sum_{j=1}^d Z_j(t)$ is equivalent to $\frac{Z_i(t)c'^{-1}e^{-\alpha t}}{\sum_{j=1}^d Z_j(t)c'^{-1}e^{-\alpha t}} < 1/2$, with the Malthusian parameter α and c' defined as in Theorem 4.1. For any $x > 0$, define event $\mathcal{E}(x)$ as

$$\mathcal{E}(x) = \cap_{i=1}^d \Big\{ Z_i(t)c'^{-1}e^{-\alpha t} \in (x, (d-1)x)\Big\}. \tag{4.15}$$

Under event $\mathcal{E}(x)$, since each $Z_i(t)c'^{-1}e^{-\alpha t}$ is at least x and at most $(d-1)x$, it follows immediately that

$$\mathcal{E}(x) \subset \Big\{2 \max_{i=1}^d Z_i(t) < \sum_{j=1}^d Z_j(t)\Big\} \;=\; \cap_{i=1}^d \Big\{2Z_i(t) < \sum_{j=1}^d Z_j(t)\Big\}. \tag{4.16}$$

Now $Z_i(\cdot)$ are independent across $1 \leq i \leq d$. By Theorem 4.2 it follows that $Z_i(t)c'^{-1}e^{-\alpha t}$ converges to W_i (because $\mathbf{E}[\eta^2] < \infty$ and hence $\mathbf{E}[\eta \log \eta] < \infty$) which are independent across i and W_i are distributed as per (4.12)-(4.13). From this it follows that as $t \to \infty$,

$$\mathbf{P}\Big(\mathcal{E}(x)\Big) = p(x)^d, \qquad \text{where} \quad p(x) \stackrel{\triangle}{=} \int_x^{(d-1)x} w(y)dy \;>\; 0. \tag{4.17}$$

From the above discussion, it follows that as $t \to \infty$

$$\mathbf{P}\Big(\cap_{i=1}^d \Big\{2Z_i(t) < \sum_{j=1}^d Z_j(t)\Big\}\Big) \geq \Big(\sup_{x>0} p(x)\Big)^d \;>\; 0. \tag{4.18}$$

4.4.4. *Equivalence of boundary and tree processes.* For regular trees, the fraction $Z_i(t)/(\sum_{j=1}^d Z_j(t))$ and $T_i(t)/(\sum_{j=1}^d T_j(t))$ converge to the same limit primarily because each of them converge to ∞ and $Z_i(t)$ and $T_i(t)$ are related linearly. Such a relation does not exist for the random tree. However, with an additional, careful argument we shall establish the same fact for random trees as well under the event $\mathcal{E}(x)$ for $x > 0$.

To that end, for $1 \leq i \leq d$ and $n \geq 0$, define

$$M_{n+1}^i = M_n^i + Q_{n+1}^i(\eta_{n+1} - \mu), \tag{4.19}$$

where Q_{n+1}^i is the probability that the $n+1$ rumor infected node is added to the tree $T_i(\cdot)$, η_{n+1} is the number of children of this infected node, and $\mu = \mathbf{E}[\eta]$. We define $M_0^i = 0$. It can be checked that M_n^i is a martingale with respect to the filtration $\mathcal{F}_n$, where $\mathcal{F}_n$ contains all the information about the part of the graph $\mathcal{G}$ to which the rumor has spread including the rumor boundary. Now M_n^i is a martingale with respect to $\mathcal{F}_n$ because, (a) Q_{n+1}^i is a binary random variable with its probability of being 1 determined by $\mathcal{F}_n$, and (b) η_{n+1} is independent of $\mathcal{F}_n$ and distributed as per $\mathcal{D}$. Now $|M_{n+1}^i - M_n^i| \leq \eta_{n+1}$ and η_{n+1} has a well defined mean and finite second moment. By the property of the martingale, therefore it follows that

$$\mathbf{E}[(M_n^i)^2] \leq n\mathbf{E}[\eta^2].$$

Therefore, a straightforward application of Chebychev's inequality will lead to the following 'weak law of large numbers':

$$\frac{1}{n} M_n^i \to \mathbf{E}[M_0^i] \;=\; 0, \qquad \text{in probability.} \tag{4.20}$$

Now under event $\mathcal{E}(x)$ (for any $x > 0$), $Z_i(t)$ scales like $e^{\alpha t}$ for all $1 \leq i \leq d$. Therefore, it can be checked that $T_i(t)$ also scales like $e^{\alpha t}$ (since $Z_i(t)$ represents the 'rate' at which $T_i(t)$ is growing). Precisely, we have that

$$\begin{aligned}\mathbf{E}\left[Z_i(t)\right] &= 1 + \mathbf{E}\left[\sum_{l \in T_i(t)} \eta_l\right] \\ &= 1 + \mu \mathbf{E}\left[T_i(t)\right]\end{aligned}$$

Now using Theorem 4.1, we have that

$$\begin{aligned}\lim_{t\to\infty} \frac{\mathbf{E}\left[T_i(t)\right]}{c' e^{\alpha t}} &= \frac{1}{\mu} \lim_{t\to\infty} \frac{\mathbf{E}\left[Z_i(t)\right]}{c' e^{\alpha t}} - \frac{1}{c' e^{\alpha t}} \\ &= \frac{1}{\mu}\end{aligned}$$

Therefore, we see that the means of both $Z_i(t)$ and $T_i(t)$ grow as $e^{\alpha t}$.

Since d is finite, it further follows that $n(t)$, the number of rumor infected nodes till time t, or equivalently $\sum_{i=1}^{d} T_i(t)$ scales like $e^{\alpha t}$. Using these facts under event $\mathcal{E}(x)$ and an application of (4.20) with n replaced by $n(t)$ (and taking $t \to \infty$ or equivalently $n(t) \to \infty$), we obtain that for $1 \leq i \leq d$ as $t \to \infty$,

$$\frac{T_i(t)}{n(t)}\Big(\frac{1}{T_i(t)} M^i_{n(t)}\Big) \to 0. \tag{4.21}$$

Since under event $\mathcal{E}(x)$, $T_i(t)/n(t)$ remains bounded away from 0 as $t \to \infty$, from (4.21), it follows that

$$\frac{1}{T_i(t)} M^i_{n(t)} \to 0. \tag{4.22}$$

But

$$\begin{aligned} M^i_{n(t)} &= \sum_{j \in T_i(t)} (\eta_j - \mu) \\ &= T_i(t)\Big(\frac{1}{T_i(t)} \sum_{j \in T_i(t)} \eta_j - \mu\Big). \end{aligned} \tag{4.23}$$

From (4.22) and (4.23), it follows that under event $\mathcal{E}(x),\ x > 0$ for $1 \leq i \leq d$, as $t \to \infty$,

$$\frac{1}{T_i(t)} \sum_{j \in T_i(t)} \eta_j \to \mu. \tag{4.24}$$

Now we are ready to conclude the proof of Theorem 3.2 by establishing that under the event $\mathcal{E}(x),\ x > 0$, $Z_i(t)/(\sum_{j=1}^{d} Z_j(t))$ and $T_i(t)/(\sum_{j=1}^{d} T_j(t))$ converge to the same quantity. To that end, observe that for $1 \leq i \leq d$,

$$\begin{aligned} \frac{Z_i(t)}{\sum_{j=1}^{d} Z_j(t)} &= \frac{1 + \sum_{\ell \in T_i(t)} \eta_\ell}{d + \sum_{\ell'=1}^{n(t)} \eta_{\ell'}} \\ &= \frac{T_i(t)}{n(t)} \frac{\frac{1}{T_i(t)} + \frac{1}{T_i(t)}\Big(\sum_{\ell \in T_i(t)} \eta_\ell\Big)}{\frac{d}{n(t)} + \frac{1}{n(t)}\Big(\sum_{\ell'=1}^{n(t)} \eta_{\ell'}\Big)}. \end{aligned} \tag{4.25}$$

Now as $t \to \infty$, under the event $\mathcal{E}(x),\ x > 0$, the right most term in (4.25) goes to 1 since the numerator and denominator both go to μ due to $T_i(t), n(t) \to \infty$ and (4.24) (with its application to all the subtrees). This concludes that under event $\mathcal{E}(x)$, the ratio $Z_i(t)/(\sum_{j=1}^{d} Z_j(t))$ and $T_i(t)/(\sum_{j=1}^{d} T_j(t))$ are asymptotically

equal as $t \to \infty$. Therefore, from (4.18) and the fact that the initial conditioned event $\mathbf{P}(\eta_0 \geq 3)$ has strictly positive probability, it follows that

$$\liminf_{t\to\infty} \mathbf{P}\left(C^1_{n(t)}\right) > 0.$$

This concludes the proof of Theorem 3.2.

4.5. *Proof of Theorem 3.3.* To obtain the upper bound in Theorem 3.3 for $\lim_{t\to\infty} \mathbf{P}\left(C^k_{n(t)}\right)$ we assume that after time t at least k nodes have been infected ($n(t) \geq k$), with the k^{th} infected node being defined as v_k which has degree d. The number of children of v_k is $\eta_k = d - 1 \geq 2$. $C^k_{n(t)}$ is the event that v_k is the rumor center after time t. To upper bound the probability of this event, we will use the memoryless property of the exponential spreading times crucially. There are d subtrees neighboring v_k. We define the time when v_k is infected as $t_k < t$. We define the size of the rumor boundary at t_k as $Z(t_k)$ which consists of all uninfected nodes neighboring infected nodes. We have that $Z(t_k) = 1 + \sum_{i=1}^{k} (\eta_i - 1)$. Because of the memoryless property of the exponential spreading times and the way in which the random tree is constructed, at time t_k each node in the rumor boundary is an independent copy of identically distributed subtree random processes which we will refer to as $X_j(t)$, for $1 \leq j \leq Z(t_k)$ in the rumor boundary.

We now use the notation from Section 4.1.4 for the subtree processes. Specifically, let us imagine the node v_k as the global root and with respect to it, let $T^k_1(t)$ denoted the size of the subtree rooted at v_{k-1} at time t. Let $T^k_i(t)$ for $i = 2, ..., d$ be the other subtrees rooted at the children of v_k (which were not infected at time t_k but were only part of the rumor boundary). Then, we have that $T^k_1(t_k) = k - 1$ and $T^k_i(t_k) = 0$ for $2 \leq i \leq d$. In $T^k_1(t_k)$, there are $Z(t_k) - \eta_k$ nodes on the rumor boundary, each of which will be the source for i.i.d. subtree process $X_j(t)$, $1 \leq j \leq Z(t_k) - \eta_k$ starting at $t = t_k$. Therefore,

$$T^k_1(t) = k - 1 + \sum_{j=1}^{Z(t_k)-\eta_k} X_j(t) \tag{4.26}$$

Now, we will upper bound $\mathbf{P}\left(C^k_{n(t)}\right)$ as follows.

$$\begin{aligned}
\mathbf{P}\left(C^k_{n(t)}\right) &\leq \mathbf{P}\left(\cap_{i=1}^{d}\left\{2T^k_i(t) \leq \sum_{j=1}^{d} T^k_j(t)\right\}\right) \\
&\leq \mathbf{P}\left(T^k_1(t) \leq \sum_{j=2}^{d-1} T^k_j(t)\right) \\
&\leq \mathbf{P}\left(k-1+\sum_{i=1}^{Z(t_k)-\eta_k} X_i(t) \leq \sum_{j=Z(t_k)-\eta_k+1}^{Z(t_k)} X_j(t)\right) \\
&\leq \mathbf{P}\left(\sum_{i=1}^{Z(t_k)-\eta_k} X_i(t) \leq \sum_{j=Z(t_k)-\eta_k+1}^{Z(t_k)} X_j(t)\right)
\end{aligned}$$

By the condition of Theorem 3.3, we have that $Z(t_k) \geq (ck+1)\eta_k$ for some $c > 1$. Define $X'_i(t) = \sum_{j=(i-1)\eta_k+1}^{i\eta_k} X_j(t)$. Then

$$\begin{aligned}
\mathbf{P}\left(C^k_{n(t)}\right) &\leq \mathbf{P}\left(\sum_{i=1}^{\lfloor (ck+1)\rfloor \eta_k} X_i(t) \leq \sum_{j=Z(t_k)-\eta_k+1}^{Z(t_k)} X_j(t)\right) \\
&\leq \mathbf{P}\left(\sum_{i=1}^{\lfloor (ck+1)\rfloor} X'_i(t) \leq X'_{\lfloor (ck+1)\rfloor+1}(t)\right) \\
&\leq \mathbf{P}\left(\bigcap_{i=1}^{\lfloor (ck+1)\rfloor} \left\{X'_i(t) \leq X'_{\lfloor (ck+1)\rfloor+1}(t)\right\}\right) \qquad (4.27) \\
&\leq \frac{1}{\lfloor (ck+1)\rfloor + 1} \\
&\leq \frac{1}{k} \qquad (4.28)
\end{aligned}$$

Above we have used the fact that for any fixed $t > t_k$, there are in total $\lfloor (ck+1)\rfloor + 1$ independent copies of the random variables $X'_i(t)$ and the probability in equation (4.27) is the probability that one of these random variables is larger than rest, which is $\frac{1}{\lfloor (ck+1)\rfloor+1}$ by symmetry. The above bound is independent of t, so this establishes Theorem 3.3.

4.6. *Proof of Theorem 3.4: geometric trees.* The proof of Theorem 3.4 uses the characterization of the rumor center provided by Proposition 1. That is, we

wish to show that for all n large enough, the probability of the event that the size of the d^* rumor infected sub-trees of the source v^* are essentially 'balanced' with high enough probability. To establish this, we shall use coarse estimations on the size of each of these sub-trees using the standard concentration property of renewal processes 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 sub-trees 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 structures 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^* sub-trees of v^* for any given time t. We are assuming here that the spreading times have distribution with CDF F with mean $\mu > 0$ and exponential tail (precisely, if X is random variable with F as its CDF, then $\mathbf{E}[\exp(\theta X)] < \infty$ for $\theta \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$). Initially, at time 0 the source node v^* has the rumor. It starts spreading along its d^* children (neighbors). Let $T_i(t)$ denote the size of the rumor infected subtree, denoted by $G_i(t)$, rooted at the ith child (or neighbor) of node v^*. Initially, $T_i(0) = 0$. The $T_i(\cdot)$ is a renewal 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 3.4, the following will be satisfied: for small enough $\epsilon > 0$ (a) every node within a distance $\frac{t}{\mu}(1-\epsilon)$ of v^* is in one of the $G_i(t)$, and (b) no node beyond distance $\frac{t}{\mu}(1+\epsilon)$ 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 $T_i(t)$ that will result in establishing Theorem 3.4. This result is summarized below with its proof in Section 4.6.1.

PROPOSITION 1. *Consider a geometric tree with parameters $\alpha > 0$ and $0 < b \leq c$ as assumed in Theorem 3.4 and let the rumor spread from source v^* starting at time 0 as per the SI model with spreading time distribution whose cumulative density function is F such that the mean is μ and $\mathbf{E}[\exp(\theta X)] < \infty$ for $\theta \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$ where X is distributed as per F. Define $\epsilon = t^{-1/2+\delta}$ for any small $0 < \delta < 1/2$. 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 sub-trees rooted at v^* (rumor graphs) such that all nodes within distance $\frac{t}{\mu}(1-\epsilon)$ from v^* are in the tree and no node beyond distance $\frac{t}{\mu}(1+\epsilon)$ from v^* is in the tree. Then*

$$\mathbf{P}(G_t \in \mathcal{G}_t) = 1 - O\big(e^{-t^\delta}\big) \stackrel{t\to\infty}{\longrightarrow} 1.$$

Define $\mathcal{E}_t$ as the event that $G_t \in \mathcal{G}_t$. Under event $\mathcal{E}_t$, consider the sizes of the

sub-trees $T_i(t)$ for $1 \leq i \leq d_{v^*}$. Due to the polynomial growth condition and $\mathcal{E}_t$, we obtain the following bounds on each $T_i(t)$ for all $1 \leq i \leq d_{v^*}$:

$$\sum_{r=1}^{\frac{t}{\mu}(1-\epsilon)-1} br^{\alpha} \leq T_i(t) \leq \sum_{r=1}^{\frac{t}{\mu}(1+\epsilon)-1} cr^{\alpha}.$$

Now bounding the summations by Riemann'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 $1 \leq i \leq d_{v^*}$

$$\frac{b}{1+\alpha}\left(\frac{t}{\mu}(1-\epsilon)-2\right)^{\alpha+1} \leq T_i(t) \leq \frac{c}{1+\alpha}\left(\frac{t}{\mu}(1+\epsilon)\right)^{\alpha+1}.$$

In the most 'unbalanced' situation, $d_{v^*}-1$ of these sub-trees have minimal size $T_{\min}(t)$ and the remaining one sub-tree has size $T_{\max}(t)$ where

$$T_{\min}(t) = \frac{b}{1+\alpha}\left(\frac{t}{\mu}(1-\epsilon)-2\right)^{\alpha+1},$$
$$T_{\max}(t) = \frac{c}{1+\alpha}\left(\frac{t}{\mu}(1+\epsilon)\right)^{\alpha+1}.$$

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 $\delta \in (0, 1/2)$, we have

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

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

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

Therefore, by Proposition 1 it follows that the rumor center is unique and equals v^*. We also have that $\mathcal{E}_t \subset C^1_{n(t)}$. Thus, from above and Theorem 1 we obtain

$$\begin{aligned}\lim_t \mathbf{P}\big(C^1_{n(t)}\big) &\ge \lim_t \mathbf{P}\big(\mathcal{E}_t\big) \\ &= 1.\end{aligned}$$

This completes the proof of Theorem 3.4.

4.6.1. *Proof of Proposition 1.* We recall that Theorem 1 stated that for a rumor spreading for time t as per the SI model with a general distribution with mean spreading time μ the rumor graph on a geometric tree is full up to a distance $\frac{t}{\mu}(1-\epsilon)$ and does not extend beyond $\frac{t}{\mu}(1+\epsilon)$, for $\epsilon = t^{-1/2+\delta}$ for some positive $\delta \in (0, 1/2)$. To establish this, we shall use the following well known concentration property of renewal processes. We provide its proof later for completeness.

PROPOSITION 2. *Consider a renewal process $P(\cdot)$ with holding times with mean μ and finite moment generating function in interval $(-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$. Then for any $t > 0$ and any $\gamma \in (0, \varepsilon')$ for a small enough $\varepsilon' > 0$, there exists a positive constant c such that*

$$\mathbf{P}\left(\left|P(t) - \frac{t}{\mu}\right| \geq \frac{t\gamma}{\mu}\right) \leq 2e^{-\frac{\gamma^2\mu}{8c}t}$$

Now we use Proposition 2 to establish Proposition 1. Recall that the spreading time along each edge is an independent and identically distributed random variable with mean μ. 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 renewal process, say $P(t)$, and node v is infected by time t if and only if $P(t) \geq r$. Therefore, from Proposition 2 it follows that for any node v that is at distance $\frac{t}{\mu}(1-\epsilon)$ for $\epsilon = t^{-\frac{1}{2}+\delta}$ for some $\delta \in (0, 1/2)$ (for all t large enough),

$$\begin{aligned}\mathbf{P}\big(v \text{ is not rumor infected}\big) &\leq 2e^{-\frac{\epsilon^2\mu t}{8c}} \\ &= 2e^{-\frac{\mu}{8c}t^{2\delta}}.\end{aligned}$$

Now the number of such nodes at distance $\frac{t}{\mu}(1-\epsilon)$ from v^* is at most $O\left(\left(\frac{t}{\mu}\right)^{\alpha+1}\right)$ (which follows from arguments similar to those in the proof of Theorem 3.4). Therefore, by an application of the union bound it follows that

$$\begin{aligned}&\mathbf{P}\left(\text{a node at distance } \frac{t}{\mu}(1-\epsilon) \text{ from } v^* \text{ isn't infected}\right) \\ &= O\left(2\left(\frac{t}{\mu}\right)^{\alpha+1} e^{-\frac{\mu}{8c}t^{2\delta}}\right) \\ &= O\left(e^{-\frac{\mu}{8c}t^{\delta}}\right).\end{aligned}$$

Using similar argument and another application of Theorem 2, it can be argued that

$$\begin{aligned}&\mathbf{P}\left(\text{a node at distance } t(1+\epsilon) \text{ from } v^* \text{ is infected}\right) \\ &= O\left(e^{-\frac{\mu}{8c}t^{\delta}}\right).\end{aligned}$$

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 $1 - O\left(e^{-\frac{\mu}{8c}t^{\delta}}\right)$, all nodes at distance up to $\frac{t}{\mu}(1-\epsilon)$ from v^* are infected and all nodes beyond distance $\frac{t}{\mu}(1+\epsilon)$ from v^* are not infected. This completes the proof of Proposition 1.

4.6.2. *Proof of Proposition 2.* We wish to provide bounds on the probability of $P(t) \leq \mu t(1-\gamma)$ and $P(t) \geq \mu t(1+\gamma)$ for a renewal process $P(\cdot)$ with holding times with mean μ and finite moment generating function. Define the n^{th} arrival time S_n as

$$S_n = \sum_{i=1}^{n} X_i$$

where X_i are non-negative i.i.d. random variables with a well defined moment generating function $M_X(\theta) = \mathbf{E}[\exp(\theta X)] < \infty$ for $\theta \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$ and mean $\mathbf{E}[X_i] = \mu > 0$. We can relate the arrival times to the renewal process by the following relations:

$$\mathbf{P}(P(t) \leq n) = \mathbf{P}(S_n \geq t)$$

and

$$\mathbf{P}(P(t) \geq n) = \mathbf{P}(S_n \leq t)$$

The first relation says that the probability of less than n arrivals in time t is equal to the probability that the nth arrival happens after time t. The second relation says that the probability of more than n arrivals in time t is equal to the probability that the nth arrival happens before time t.

We now bound $\mathbf{P}(S_n \geq t)$. To that end, for $\theta \in (0, \varepsilon)$ it follows from the Chernoff bound that

$$\begin{aligned}\mathbf{P}(S_n \geq t) &= \mathbf{P}\left(e^{\theta S_n} \geq e^{\theta t}\right) \\ &\leq M_X(\theta)^n e^{-\theta t}\end{aligned}$$

We can use the following approximation for $M_X(\theta)$ which is valid for small θ, say $\theta \in (0, \varepsilon^+)$ for $0 < \varepsilon^+ \leq \varepsilon$.

$$\begin{aligned}M_X(\theta) &= 1 + \theta\mu + \theta^2 \frac{\mathbf{E}\left[X^2\right]}{2} + \theta^3 \sum_{i=3}^{\infty} \theta^{i-3} \frac{\mathbf{E}\left[X^i\right]}{i!} \\ &\leq 1 + \theta\mu + c_1\theta^2\end{aligned}$$

for some finite positive constant c_1. Using this along with the inequality $\log(1+x) \leq x$ for $-1 < x$, we obtain

$$\begin{aligned}\log(\mathbf{P}(S_n \geq t)) &\leq n \log(M_X(\theta)) - \theta t \\ &\leq n \log\left(1 + \theta\mu + c_1\theta^2\right) - \theta t \\ &\leq \theta(\mu n - t) + n c_1 \theta^2\end{aligned}$$

To minimize this probability, we find the θ that minimizes $\theta(\mu n - t) + nc_1\theta^2$. This happens for $\theta = \frac{1}{2c_1}\left(\frac{t}{n} - \mu\right)$. We set $n = \frac{t}{\mu}(1-\gamma)$, so the minimum value is achieved for $\theta^* = \frac{\gamma\mu}{2c_1(1-\gamma)}$. Therefore, there exists $\varepsilon_1 > 0$ so that for $\gamma \in (0, \varepsilon_1)$, the corresponding $\theta^* = \frac{\gamma\mu}{2c_1(1-\gamma)} < \varepsilon^+$ so that the quadratic approximation of $M_X(\theta)$ is valid. Given this, we obtain

$$\begin{aligned}
\log\left(\mathbf{P}\left(S_{\frac{t}{\mu}(1-\gamma)} \geq t\right)\right) &\leq -\frac{\gamma\mu}{2c_1(1-\gamma)}(\gamma t) + \frac{tc_1}{\mu}(1-\gamma)\frac{\gamma^2\mu^2}{4c_1^2(1-\gamma)^2} \\
&\leq -\frac{\gamma^2\mu t}{2c_1(1-\gamma)} + \frac{\gamma^2\mu t}{4c_1(1-\gamma)} \\
&\leq -\frac{\gamma^2\mu t}{4c_1(1-\gamma)} \\
&\leq -\frac{\gamma^2\mu t}{8c_1}
\end{aligned}$$

With this result, we obtain

$$\mathbf{P}\left(P(t) \leq \frac{t}{\mu}(1-\gamma)\right) \leq e^{-\frac{\gamma^2\mu t}{8c_1}},$$

for any t and $\gamma \in (0, \varepsilon_1)$. For the upper bound, we have for $\theta > 0$

$$\begin{aligned}
\mathbf{P}(S_n \leq t) &= \mathbf{P}\left(e^{-\theta S_n} \geq e^{-\theta t}\right) \\
&\leq M_X(-\theta)^n e^{\theta t}
\end{aligned}$$

We can use the following approximation for $M_X(-\theta)$ which is valid for small enough $\theta \in (0, \varepsilon^-)$ with $0 < \varepsilon^- \leq \varepsilon$.

$$\begin{aligned}
M_X(-\theta) &= 1 - \theta\mu + \theta^2\frac{\mathbf{E}\left[X^2\right]}{2} - \theta^3\sum_{i=3}^{\infty}\theta^{i-3}(-1)^{i-3}\frac{\mathbf{E}\left[X^i\right]}{i!} \\
&\leq 1 - \theta\mu + c_2\theta^2
\end{aligned}$$

for some finite positive constant c_2. Using this we obtain

$$\begin{aligned}
\log(\mathbf{P}(S_n \leq t)) &\leq n\log(M_X(-\theta)) + \theta t \\
&\leq n\log\left(1 - \theta\mu + c_2\theta^2\right) + \theta t \\
&\leq \theta(t - \mu n) + nc_2\theta^2
\end{aligned}$$

To minimize this probability, we find the θ that minimizes $\theta(t - \mu n) + nc_2\theta^2$. This happens for $\theta = \frac{1}{2c_2}\left(\mu - \frac{t}{n}\right)$. We set $n = \frac{t}{\mu}(1+\gamma)$, so the minimum value

is achieved for $\theta^* = \frac{\gamma\mu}{2c_2(1+\gamma)}$. There exists, $\varepsilon_2 > 0$ so that for all $\gamma \in (0, \varepsilon_2)$, $\theta^* = \frac{\gamma\mu}{2c_2(1+\gamma)} \leq \varepsilon^-$ and thus guaranteeing the validity of quadratic approximation of $M_X(-\theta)$ that we have assumed. Subsequently, we obtain

$$\begin{aligned}
\log\left(\mathbf{P}\left(S_{\frac{t}{\mu}(1+\gamma)} \leq t\right)\right) &\leq -\frac{\gamma\mu}{2c_2(1+\gamma)}(\gamma t) + \frac{tc_2}{\mu}(1+\gamma)\frac{\gamma^2\mu^2}{4c_2^2(1+\gamma)^2} \\
&\leq -\frac{\gamma^2\mu t}{2c_2(1+\gamma)} + \frac{\gamma^2\mu t}{4c_2(1+\gamma)} \\
&\leq -\frac{\gamma^2\mu t}{4c_2(1+\gamma)} \\
&\leq -\frac{\gamma^2\mu t}{8c_2}
\end{aligned}$$

With this result, we obtain

$$\mathbf{P}\left(P(t) \geq \frac{t}{\mu}(1+\gamma)\right) \leq e^{-\frac{\gamma^2\mu t}{8c_2}},$$

for any t and $\gamma \in (0, \varepsilon_2)$.

If we set $c = \max(c_1, c_2)$ and $\varepsilon' = \min(\varepsilon_1, \varepsilon_2)$ and combine the upper and lower bounds then we obtain

$$\mathbf{P}\left(\left|P(t) - \frac{t}{\mu}\right| \geq \frac{t\gamma}{\mu}\right) \leq 2e^{-\frac{\gamma^2\mu}{8c}t},$$

for any t and $\gamma \in (0, \varepsilon')$ with $\varepsilon' > 0$. This completes the proof of Proposition 2.

5. Conclusion. Finding the source of a rumor in a network is an important and challenging problem. Here we characterized the performance of the rumor source estimator known as rumor centrality for generic tree graphs. Our analysis was based upon multi-type continuous time branching processes/generalized Polya's urn models. As an implication of this novel analysis method, we recovered all the previous results for regular trees [10] as a special case. We also showed that for rumor spreading on a random regular graph, the probability that the estimated source is more than k hops away from the true source decays exponentially in k. Additionally, we showed that for general random trees and hence for sparse random graphs like Erdos-Renyi graph, there is a strictly positive probability of correct rumor source detection. In summary, we have established the universality of rumor centrality as a source estimator across a variety of tree structured graphs and across variety of SI spreading time distributions.

Acknowledgments. D. Shah would like to acknowledge conversations with David Gamarnik and Andrea Montanari at the Banff Research Institute that seeded this line of works. Both authors would like to acknowledge support of AFOSR Complex Networks Project, MURI on Tomography of Social Networks and the MIT-Shell Graduate Student Fellowship.

REFERENCES

[1] K. B. Athreya and P. E. Ney. *Branching Processes*. Springer-Verlag, 1972.

[2] G. Brightwell and P. Winkler. Counting linear extensions is# p-complete. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pages 175–181. ACM, 1991.

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

[4] A. Ganesh, L. Massoulie, and D. Towsley. The effect of network topology on the spread of epidemics. In *Proc. 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, volume 2, pages 1455–1466, 2005.

[5] A. Gerschenfeld and A. Montanari. Reconstruction for models on random graphs. In *Proc. 48th IEEE Symp. Found. Comp. Sci. (FOCS)*, pages 194–204, 2007.

[6] A. Karzanov and L. Khachiyan. On the conductance of order markov chains. *Order*, 8(1):7–15, 1991.

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

[8] E. Mossel. Reconstruction on trees: Beating the second eigenvalue. *Ann. Appl. Prob.*, 11:285–300, 2001.

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

[10] D. Shah and T. Zaman. Finding sources of computer viruses in networks: Theory and experiment. In *Proc. ACM Sigmetrics*, volume 15, pages 5249–5262, 2010.

DEVAVRAT SHAH
77 MASSACHUSETTS AVE.
CAMBRIDGE, MA 02139
E-MAIL: devavrat@mit.edu

TAUHID ZAMAN
77 MASSACHUSETTS AVE.
CAMBRIDGE, MA 02139
E-MAIL: zlisto@mit.edu