Rooting out the Rumor Culprit from Suspects

Wenxiang Dong*,Wenyi Zhang* and Chee Wei Tan†
University of Science and Technology of China*, City University of Hong Kong†
javin@mail.ustc.edu.cn, wenyizha@ustc.edu.cn and cheewtan@cityu.edu.hk

***Abstract*—Suppose that a rumor originating from a single source among a set of suspects spreads in a network, how to root out this rumor source? With the *a priori* knowledge of suspect nodes and a snapshot observation of infected nodes, we construct a maximum *a posteriori* (MAP) estimator to identify the rumor source using the susceptible-infected (SI) model. We propose to use a notion of *local rumor center* to characterize $\mathbf{P_c}(n)$, the correct detection probability of the source estimator upon observing n infected nodes, in both the finite and asymptotic regimes, for regular trees of node degree δ. First, when all nodes are suspects, $\lim_{n\to\infty} \mathbf{P_c}(n)$ grows from 0.25 to 0.307 as δ increases from three to infinity, a result first established in Shah and Zaman (2011, 2012) via a different approach; furthermore, $\mathbf{P_c}(n)$ monotonically decreases with n and increases with δ even in the finite-n regime. Second, when the suspect nodes form a connected subgraph of the network, $\lim_{n\to\infty} \mathbf{P_c}(n)$ significantly exceeds the *a priori* probability if $\delta \geq 3$, and reliable detection is achieved as δ becomes sufficiently large; furthermore, $\mathbf{P_c}(n)$ monotonically decreases with n and increases with δ. Third, when there are only two suspect nodes, $\lim_{n\to\infty} \mathbf{P_c}(n)$ is at least 0.75 if $\delta \geq 3$; and $\mathbf{P_c}(n)$ increases with the distance between the two suspects. Fourth, when there are multiple suspect nodes, among all possible connection patterns, that all the suspects form a single connected subgraph yields the smallest $\mathbf{P_c}(n)$. Our analysis leverages ideas from the Pólya's urn model in probability theory and sheds insight into the behavior of the rumor spreading process not only in the asymptotic regime but also for the general finite-n regime.**

I. Introduction

Spreading of epidemics and information cascades through social networks is ubiquitous in the modern world [1]. Examples include the propagation of infectious diseases, information diffusion in the Internet, tweeting and retweeting of popular Twitter topics. In general, any of these situations can be modeled as an epidemic-like rumor spreading in a network [2]. A key challenge in network inference is to identify the rumor source by leveraging the network topology, suspect characteristics and the observation of infected nodes. Finding this rumor source has practical applications and also allows us to better understand the amplification role of the network in information cascades.

In the seminal work [2], [3], the authors tackled the single rumor source identification problem using the SI model. For regular tree-type networks, they constructed a maximum likelihood (ML) estimator and derived its asymptotic performance. When the node degree is $\delta = 2$, i.e., a linear network, the asymptotic correct detection probability is zero; when $\delta = 3$, it is 0.25; when $\delta > 3$, it is a positive constant value $\phi_1(\delta)$, which approaches 0.307 as δ grows large. Later, other models were also considered, such as multiple rumor sources [4] and the susceptible-infected-recovered (SIR) model [5]. However, these works assume that every node in the network is equally likely to be a suspect for spreading the rumor.

In practice, we may have *a priori* knowledge that only a portion of nodes, called *suspect nodes*, can initiate the rumor spreading. For example, when an infectious disease is spread, the frequent travellers from earlier infected cities have a higher suspicion of causing an epidemic outbreak in a new city. When the cyberspace is hit by rumors or computer viruses, most of the victims are in fact innocent and thus not all victims should be suspected as the culprit. Only the suspicious nodes, which may have nontrivial connection patterns in the network or scaling behavior (e.g., the number of suspects and infected nodes can scale differently), need to be examined first. The suspect characteristics thus affect detectability, and add an interesting dimension to identifying the source reliably.

We study the problem of reliably identifying a single rumor source from suspect nodes with a uniform *a priori* distribution of being the rumor source. In particular, for regular tree-type networks of node degree δ, we characterize $\mathbf{P_c}(n)$, the correct detection probability using MAP estimation upon observing n infected nodes. Our main contributions are as follows:

- If every infected node is a suspect, we characterize $\mathbf{P_c}(n)$ in both the finite and asymptotic regimes (in the number n), thus generalizing the results in [2], [3] (which only have the asymptotic results). This is achieved by exploiting the Pólya's urn model [6] in probability theory. Furthermore, $\mathbf{P_c}(n)$ monotonically decreases with n and increases with δ even in the finite-n regime.
- If k suspect nodes form a connected subgraph of the network, we characterize $\mathbf{P_c}(n)$ in both the asymptotic and finite regimes, and find that $\lim_{n\to\infty} \mathbf{P_c}(n)$ is a constant value $\phi_2(\delta, k)$ that significantly exceeds the *a priori* probability $1/k$ if $\delta > 2$. Furthermore, the MAP estimator achieves reliable detection as the node degree becomes sufficiently large. Similarly, $\mathbf{P_c}(n)$ monotonically decreases with n and increases with δ.
- If there are only two suspect nodes with their shortest path distance d (measured by the number of hops), $\mathbf{P_c}(n)$ is at least 0.75 if $\delta > 2$, and increases with d, thereby achieving reliable detection if the node degree becomes sufficiently large.

The work of W. Dong and W. Zhang has been supported by National Basic Research Program of China (973 Program) through grant 2012CB316004, and by the 100 Talents Program of Chinese Academy of Sciences. The work of C. W. Tan has been supported by the Research Grants Council of Hong Kong under Project No. RGC CityU 125212.

- When there are multiple suspect nodes, among all possible connection patterns, the MAP estimator has the smallest correct detection probability when all suspects form a connected subgraph of the network.

The rest of the paper is organized as follows. Section II describes the SI spreading model and the MAP rumor source estimator. We establish analytical results on the MAP detection probability for regular trees in Section III, and we illustrate its performance numerically in Section IV.

II. Rumor Spreading Model and Rumor Source Estimator

In this section, we describe the SI rumor spreading model, give the MAP estimator for the rumor source in regular trees, and introduce the notion of *local rumor center*.

A. Rumor Spreading Model

In general, an undirected network $G = (V, E)$ consists of a set of nodes V and a set of edges E. V is assumed countably infinite so as to avoid any boundary effect, and any pair of nodes may infect each other if and only if they are connected by an edge in E. The rumor spreading process is modeled by the SI model, in which once a node gets infected it keeps the rumor forever.

Throughout this paper, we focus on the case where G is a regular tree. We consider the scenario that a single node in an *a priori* specified suspect set S can be the rumor source, where $S \subseteq V$ has cardinality k. The *a priori* distribution over S is assumed to be uniform; namely, $\mathbf{P_s}(s)$, the probability that s is the rumor source, is equal to $1/k$ for any $s \in S$.

The rumor spreading process unfolds as follows. Initially, only a single node $s^* \in S$ possesses a rumor. An infected node may infect its neighbors, independent of all other nodes. Let τ_{ij} be the time it takes for node j to receive the rumor from its neighbor i after i has possessed the rumor, where $(i, j) \in E$. We assume that $\{\tau_{ij}, (i, j) \in E\}$ are mutually independent and all exponentially distributed with unit mean.

B. MAP Rumor Source Estimator

(1) MAP estimator for regular trees

Suppose that a rumor originates from a node $s^* \in S$, and we get to observe G at some point, finding a snapshot of n infected nodes with the rumor, which are collectively denoted by G_n. Due to the SI model, G_n must form a connected subgraph of G and contain at least a node in S. Our goal is to construct an estimator to identify a node $\hat{s}$ as the rumor source.

Now, conditioned on G_n, the source node has a uniform distribution over $S \bigcap G_n$. Utilizing the Bayes rule, the maximum *a posteriori* (MAP) estimator of s^* maximizes the *average* correct detection probability and is given by

$$\hat{s} \in \underset{s \in \{S \bigcap G_n\}}{\arg\max}\, \mathbf{P_G}\left(G_n | s\right), \tag{1}$$

where $\mathbf{P_G}(G_n|s)$ is the probability of observing G_n with s being the rumor source. Note that a key difference from the model in [2] is that in our work the rumor source resides in the *a priori* suspect set $S \subseteq V$.

Since the evaluation of $\mathbf{P_G}\left(G_n|s\right)$ may be computationally prohibitive, we leverage the concept of rumor centrality, first developed in [2], to design our estimator. In particular, the MAP estimator for a regular tree is given by

$$\hat{s} \in \underset{s \in \{S \bigcap G_n\}}{\arg\max}\, \mathbf{P_G}\left(G_n | s\right) = \underset{s \in \{S \bigcap G_n\}}{\arg\max}\, R\left(s, G_n\right), \tag{2}$$

where $R(s, G_n)$ is the rumor centrality of node s in G_n and can be computed by

$$R(s, G_n) = n! \prod_{u \in G_n} \left|T_u^s\right|^{-1}, \tag{3}$$

where T_u^s is the subtree rooted at u with s as the source in G_n and $\left|T_u^s\right|$ is the number of nodes in T_u^s; e.g., see Fig. 1.

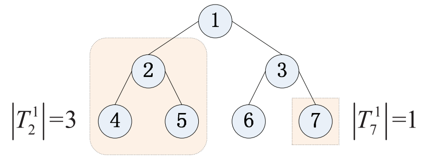

Fig. 1. Illustration of a subtree T_u^s.

(2) Local rumor center

In the following, we develop a notion of local rumor center, which enables efficient implementation of the estimator (2) and will be instrumental for our subsequent analysis. Consider a node ω with a neighbor set $N(\omega)$ and a sub-neighborhood $N_l(\omega) \subseteq N(\omega)$. If $R(\omega, G_n) \geq R(u, G_n)$ for all $u \in N_l(\omega)$, then ω is called the *local rumor center* with respect to (w.r.t.) a sub-neighborhood $N_l(\omega)$ of G_n. We have the following proposition; for its proof see [7, Proposition 1].

Proposition 1. *i) Given a tree G_n of n nodes, if node ω is the local rumor center w.r.t. a sub-neighborhood $N_l(\omega) \subset G_n$, then for any $u \in N_l(\omega)$, we have $\left|T_u^\omega\right| \leq n/2$; and for any $u' \in T_u^\omega \setminus \{u\}$, we have $R(u', G_n) < R(\omega, G_n)$.*
ii) If there is a node ω such that $\left|T_u^\omega\right| \leq n/2$ for all $u \in N_l(\omega)$, then ω is a local rumor center w.r.t. the sub-neighborhood $N_l(\omega) \subset G_n$.
iii) Furthermore, if node ω is the local rumor center w.r.t. a sub-neighborhood $N_l(\omega) \subset G_n$, then there is at most a node $u \in N_l(\omega)$ such that $R(u, G_n) = R(\omega, G_n)$, which holds if and only if $\left|T_u^\omega\right| = n/2$.

Remark 1: In fact, the local rumor center is a generalization of the rumor center in [2], which is the node with the maximal rumor centrality in G_n. However, the rumor center may belong to $G_n \setminus S$ and thus is not the solution of the estimator (2). The notion of local rumor center not only generalizes the concept of rumor center, but also will prove to be a key to tackle the MAP estimator with suspects.

III. Detection Probability Analysis

In this section, we analyze the performance of the rumor source estimator for regular trees under four scenarios using the Pólya's urn model. Let us first introduce our main results.

A. Main Results

We focus on the correct detection probability $\mathbf{P_c}(n)$, the probability of the estimator (2) to correctly identify the rumor source from the suspect nodes in S, upon observing n infected nodes. For a regular tree with $G = (V, E)$ node degree δ, we have the following characterizations for $\mathbf{P_c}(n)$.

Theorem 2. *Suppose $S = V$, i.e., every infected node is a suspect node, then:*
i) When $\delta = 2$ (linear network),

$$\mathbf{P_c}(n) = \frac{1}{2^{n-1}}\binom{n-1}{\lfloor (n-1)/2 \rfloor}, \tag{4}$$

and $\mathbf{P_c}(n) = O(1/\sqrt{n})$ with sufficiently large n.
ii) When $\delta = 3$,

$$\mathbf{P_c}(n) = \frac{1}{4} + \frac{3}{4}\frac{1}{2\lfloor n/2 \rfloor + 1}, \tag{5}$$

and $\mathbf{P_c}(n) = 0.25 + O(1/n)$ with sufficiently large n.
iii) When $\delta > 3$, we use Algorithm 1 in [7, Section 4] to compute the exact value of $\mathbf{P_c}(n)$ with finite n. Besides,

$$\lim_{n\to\infty} \mathbf{P_c}(n) = \phi_1(\delta) := 1 - \delta\left(1 - I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right)\right), \tag{6}$$

where $I_x(\alpha, \beta)$ is the incomplete Beta function with parameters α and β; and $\phi_1(\delta) \to 1 - \ln 2 \approx 0.307$ as $\delta \to \infty$.

Remark 2: When $S = V$, the MAP estimator (2) in effect reduces to the ML estimator established in [2], [3].[1] The asymptotic parts of Theorem 2 have been established in [2], [3] using a different approach from ours that explicitly relies on the exponential distribution of the infection time. Therefore, without any *a priori* knowledge, the estimator achieves a strictly positive correct detection probability if the network is not linear, but it is asymptotically upper bounded by 0.307. In addition, we have the following corollary as complement to Theorem 2.

Corollary 1 *Suppose $S = V$, i.e., every infected node is a suspect node, then:*
i) $\mathbf{P_c}(n)$ monotonically decreases with n;
ii) $\mathbf{P_c}(n)$ monotonically increases with δ.

Theorem 3. *Suppose that S forms a connected subgraph of G, then:*
i) When $\delta = 2$ (linear network),

$$\mathbf{P_c}(n) = \frac{1}{k}\left(1 + \frac{k-1}{2^{n-1}}\binom{n-1}{\lfloor (n-1)/2 \rfloor}\right), \tag{7}$$

and $\mathbf{P_c}(n) = 1/k + O(1/\sqrt{n})$ with sufficiently large n.
ii) When $\delta = 3$,

$$\mathbf{P_c}(n) = \frac{k+1}{2k} + \frac{k-1}{k}\frac{1}{4\lfloor n/2 \rfloor + 2}, \tag{8}$$

and $\mathbf{P_c}(n) = (k+1)/(2k) + O(1/n)$ with sufficiently large n.
iii) When $\delta > 3$, we use Algorithm 2 in [7, Section 4] to compute the exact value of $\mathbf{P_c}(n)$ with finite n. Besides,

$$\lim_{n\to\infty} \mathbf{P_c}(n) = \phi_2(\delta, k) := 1 - \frac{2k-2}{k}\left(1 - I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right)\right), \tag{9}$$

where $\phi_2(\delta, k) \to 1$ as $\delta \to \infty$, and $\phi_2(\delta, k) \to 2I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right) - 1$ as $k \to \infty$.

Remark 3: For linear networks, $\mathbf{P_c}(n)$ can barely exceed the *a priori* probability $1/k$. When $\delta \geq 3$, $\mathbf{P_c}(n)$ is at least $\max\{1/k, 1/2\}$, which is in sharp contrast to that $\mathbf{P_c}(n)$ is at most 1/2 with no *a priori* knowledge as in Theorem 2. Furthermore, the MAP estimator achieves reliable detection as δ grows sufficiently large. Therefore, the performance of the MAP estimator is significantly improved and reliable detection can be achieved when the *a priori* knowledge on the suspect nodes is given. In addition, we have the following corollary as complement to Theorem 3.

Corollary 2 *Suppose that S forms a connected subgraph of G, then:*
i) $\mathbf{P_c}(n)$ monotonically decreases with n;
ii) $\mathbf{P_c}(n)$ monotonically increases with δ.

Theorem 4. *Suppose S only contains two suspect nodes, and denote by d their shortest path distance ($d < n$), then:*
i) When $\delta = 2$ (linear network),

$$\mathbf{P_c}(n) = \begin{cases} \dfrac{1}{2} - \dfrac{1}{2^n}\displaystyle\sum_{z_1=(n-d-1)/2}^{(n+d+1)/2}\binom{n-1}{z_1}, & (n-d)\ \textit{is odd}; \\ \dfrac{1}{2} - \dfrac{1}{2^n}\displaystyle\sum_{z_1=(n-d)/2}^{(n+d-2)/2}\binom{n-1}{z_1}, & (n-d)\ \textit{is even}. \end{cases} \tag{10}$$

ii) When $\delta = 3$, we use Algorithm 3 in [7, Section 4] to compute the exact value of $\mathbf{P_c}(n)$ with finite n. Besides, $\lim_{n\to\infty} \mathbf{P_c}(n) = 0.75$ when $d = 1$ and ≈ 0.886 when $d = 2$.
iii) When $\delta > 3$, we use Algorithm 3 in [7, Section 4] to compute the exact value of $\mathbf{P_c}(n)$ with finite n. Besides,

$$\lim_{n\to\infty} \mathbf{P_c}(n) = \phi_3(\delta) := I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right), d = 1, \tag{11}$$

and $\phi_3(\delta) \to 1$ as $\delta \to \infty$.
iv) In general, $\mathbf{P_c}(n)$ monotonically increases with d.

Remark 4: Theorem 4 formalizes our intuition that it is more difficult to correctly identify the rumor source if the two suspect nodes are closer. We see that when $\delta \geq 3$, the *a priori* probability 1/2 can be significantly exceeded.

Theorem 5. *Suppose S contains k suspect nodes, then:*
i) When $\delta = 2$ (linear network),

$$\mathbf{P_c}(n) \geq \frac{1}{k}\left(1 + \frac{k-1}{2^{n-1}}\binom{n-1}{\lfloor (n-1)/2 \rfloor}\right), \tag{12}$$

and $\mathbf{P_c}(n) \geq 1/k + O(1/\sqrt{n})$ with sufficiently large n.
ii) When $\delta = 3$,

$$\mathbf{P_c}(n) \geq \frac{k+1}{2k} + \frac{k-1}{k}\frac{1}{4\lfloor n/2 \rfloor + 2}, \tag{13}$$

[1] Strictly speaking, our setup of a uniform *a priori* distribution of the rumor source over S is not well defined when $S = V$, since V is a countably infinite set. Our remedy is that we consider ML estimation for Theorem 2 thus returning to the setup in [2], [3], and consider MAP estimation for the other two cases.

and $\mathbf{P_c}(n) \geq (k+1)/(2k) + O(1/n)$ *with sufficiently large n.*
iii) When $\delta > 3$,

$$\lim_{n\to\infty} \mathbf{P_c}(n) \geq \phi_2(\delta, k) := 1 - \frac{2k-2}{k}\left(1 - I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right)\right). \quad (14)$$

iv) In general, $\mathbf{P_c}(n)$ *is minimized, among all possible* S*'s with* $|S| = k$*, when the k suspect nodes constitute a connected subgraph in G as in Theorem 3.*

Remark 5: Theorem 5 can be established as a consequence of Theorems 3 and 4. When there are k suspect nodes, the MAP estimator achieves the smallest correct detection probability in the scenario of Theorem 3. This result formalizes our intuition that the more clustered the suspects are the more difficult it is to identify the rumor source.

Due to space limit, we will only present the proofs of Theorem 2-ii, Theorem 3-iii, and Theorem 4-iv; the other parts can be proved following similar techniques (see [7]).

B. Proof of Theorem 2: Suspecting all Nodes

In the case of $S = V$, we only need to consider an arbitrary node $s^* \in G$ as the rumor source by symmetry. For a source s^* with m ($m \leq \delta$) neighboring nodes $N_l(s^*) = \{v_1, \ldots, v_m\} \subset S$, let a random variable X_j be the number of nodes in subtree $T_{v_j}^{s^*}$ ($1 \leq j \leq m$) of G_n. Then, we have the following lemma (for Theorem 2, $m = \delta$); for its proof see [7, Lemma 6].

Lemma 6. *To correctly identify source* s^* *with m neighboring suspect nodes as the estimate* $\hat{s}$*, we have*

$$\begin{cases} p_1 := \mathbf{P_c}\left(\hat{s} = s^* \middle| \max\{X_j, 1 \leq j \leq m\} < n/2\right) = 1, \\ p_{1/2} := \mathbf{P_c}\left(\hat{s} = s^* \middle| \max\{X_j, 1 \leq j \leq m\} = n/2\right) = \dfrac{1}{2}, \\ p_0 := \mathbf{P_c}\left(\hat{s} = s^* \middle| \max\{X_j, 1 \leq j \leq m\} > n/2\right) = 0. \end{cases} \quad (15)$$

Remark 6: Lemma 6 is deduced from Proposition 1. In order to prove Theorem 2-ii, we should find the conditions, under which s^* is the local rumor center w.r.t. $N_l(s^*)$ of G_n, such that the estimator (2) can correctly identify s^* as the source.

We outline a proof of Theorem 2-ii when $\delta = 3$; the other situations are proved in [7].

Proof of Theorem 2-ii. When $\delta = 3$, the joint distribution of $\{X_j, 1 \leq j \leq 3\}$ is given by

$$\mathbf{P_G}\left[\bigcap_{j=1}^{3}(X_j = x_j)\right] = \frac{2}{n(n+1)}. \quad (16)$$

Since $S = V$, the source s^* has $m = 3$ neighboring suspect nodes. Using Lemma 6, the correct detection probability is

$$\mathbf{P_c}(n) = \frac{1}{4} + \frac{3}{4}\frac{1}{2\lfloor n/2 \rfloor + 1}. \quad (17)$$

□

C. Proof of Theorem 3: Connected Suspects

Now, consider the case where S with cardinality k forms a connected subgraph of G. By the Bayes' rule and the *a priori* knowledge that $\mathbf{P_s}(s^*) = 1/k$ for any $s^* \in S$, we have

$$\mathbf{P_c}(n) = \sum_{i=1}^{k} \mathbf{P_s}(s_i)\mathbf{P_c}(n|s_i) = \frac{1}{k}\sum_{s^* \in S} \mathbf{P_c}(n|s^*). \quad (18)$$

We first find $\mathbf{P_c}(n|s^*)$ for each $s^* \in S$. For a source s^* with m ($m \leq \delta$) neighboring nodes $N_l(s^*) = \{v_1, \ldots, v_m\} \subset S$, let X_j be the number of nodes in subtree $T_{v_j}^{s^*}$ ($1 \leq j \leq m$) of G_n. Then, we have the following lemma; for its proof see [7, Lemma 7].

Lemma 7. *Define* $E_j = \{X_j < n/2\}$ *and* $F_j = \{X_j \leq n/2\}$, $1 \leq j \leq m$*. To correctly identify source* s^* *with m neighboring suspect nodes, we have*

$$1 - m\mathbf{P_G}(E_1^c) \leq \mathbf{P_c}(n|s^*) \leq 1 - m\mathbf{P_G}(F_1^c), \quad (19)$$

where $\mathbf{P_G}(E_1^c)$ *and* $\mathbf{P_G}(F_1^c)$ *are the probabilities that the complements of events* E_1 *and* F_1 *occur, respectively.*

Remark 7: Lemma 7 is deduced from Proposition 1, and is a generalization of the statement claimed in [3, Section 4.1.2]. In order to prove and Theorem 3-iii, we should show that the lower and upper bounds asymptotically coincide.

We outline a proof of Theorem 3-iii when $\delta > 3$; the other situations are proved in [7].

Proof of Theorem 3-iii. Since $E_1 = \{X_1 < n/2\}$ and $F_1 = \{X_1 \leq n/2\}$, i.e. $E_1 = \{X_1/n < 1/2\}$ and $F_1 = \{X_1/n \leq 1/2\}$, we have

$$\mathbf{P_G}(E_1) = \mathbf{P_G}(F_1) = I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right) + \xi(n, \delta), \quad (20)$$

where $I_x(\alpha, \beta)$ is the incomplete Beta function with parameters $\alpha = 1/\delta$ and $\beta = (\delta-1)/(\delta-2)$, and $\lim_{n\to\infty} \xi(n, \delta) = 0$.

Using Lemma 7, for a source s^* with m neighbors in S, we have

$$\mathbf{P_c}(n|s^*) = 1 - m\left(1 - I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right) - \xi(n, \delta)\right). \quad (21)$$

Note that $\mathbf{P_c}(n|s^*)$ is one subtracted by a common factor m times, each of which accounting for one neighboring suspect node of s^* connected by an edge. Since there are $k-1$ edges connecting the k suspect nodes of S in G, each edge will account for a reduction of the factor twice. Therefore, from (18) and (21), we have

$$\begin{aligned} \mathbf{P_c}(n) &= \frac{1}{k}\left(k - 2(k-1)\cdot\left(1 - I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right) - \xi(n, \delta)\right)\right) \\ &= 1 - \frac{2k-2}{k}\left(1 - I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right) - \xi(n, \delta)\right), \quad (22) \end{aligned}$$

where $\lim_{n\to\infty} \xi(n, \delta) = 0$.

Besides, $I_{1/2}(0, 1) = 1$ (by setting $\delta \to \infty$), and thus $\mathbf{P_c}(n) = 1$ for $n \to \infty$ and $\delta \to \infty$. Importantly, note that the growth of δ does not need to depend on the growth of n. □

D. Proof of Theorem 4: Two Suspects

In the case where $S = \{s_1, s_2\}$ contains only two suspect nodes, let d be the shortest path distance between s_1 and s_2 on G. We assume s_1 to be the rumor source s^* by symmetry, let $\mathcal{P} = \{v_0 = s_1, v_1, \cdots, v_d = s_2\}$ be the shortest path from s_1 to s_2, and define a random variable Z_h to be the number of nodes in the subtree $T_{v_h}^{s^*}$ $(1 \le h \le d)$. It is clear that $Z_h \ge Z_{h+1} + 1$ if $Z_h > 0$ for all $1 \le h \le d-1$.

We outline a proof of Theorem 4-iv; the other situations are proved in [7]. A useful recursive relationship is that, for any two neighboring nodes u and v in a tree G_n [2]:

$$R(u, G_n) = R(v, G_n)\frac{\left|T_u^v\right|}{n - |T_u^v|}. \tag{23}$$

Proof of Theorem 4-iv. Equivalently, we prove that $\mathbf{P_e}(n)$ decreases with d. Without loss of generality, we assume $Z_h > 0$ for all $1 \le h \le d$. The proof will be completed by showing that if $R(v_d, G_n) \ge R(s^*, G_n)$ then $R(v_{d-1}, G_n) > R(s^*, G_n)$ for all $d \ge 2$, which is to be verified by contradiction to the assumption $R(v_d, G_n) \ge R(s^*, G_n)$.

Suppose $R(v_{d-1}, G_n) \le R(s^*, G_n)$, then $Z_{d-1} \le n/2$. Otherwise, $Z_{d-1} > n/2$ and thus $Z_h > n/2$ for all $1 \le h \le d-1$; namely, $Z_h/(n - Z_h) > 1$ for all $1 \le h \le d-1$. Repeatedly using (23), we have

$$R(v_{d-1}, G_n) = R(s^*, G_n)\frac{Z_1}{n - Z_1}\frac{Z_2}{n - Z_2}\cdots\frac{Z_{d-1}}{n - Z_{d-1}}, \tag{24}$$

which leads to the contradiction that $R(v_{d-1}, G_n) > R(s^*, G_n)$. As a result, we have $Z_{d-1} \le n/2$.

As $Z_{d-1} \le n/2$, then $Z_d < n/2$ and thus $Z_d/(n - Z_d) < 1$. As a result, we have

$$R(v_d, G_n) = R(v_{d-1}, G_n)\frac{Z_d}{n - Z_d} < R(v_{d-1}, G_n) \le R(s^*, G_n), \tag{25}$$

which is in contradiction to the assumption of $R(v_d, G_n) \ge R(s^*, G_n)$. □

IV. Simulation Results

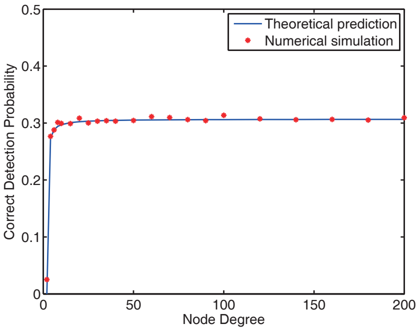

Fig. 2. Detection probability when $S = V$.

In this section, we carry out simulation experiments to corroborate and illustrate our analysis. For verifying the asymptotic results, in each experiment run we let $n = 1000$ nodes be eventually infected by a rumor source node uniformly randomly chosen from the suspect set, following the SI model, and use the estimator (2) to identify this source.

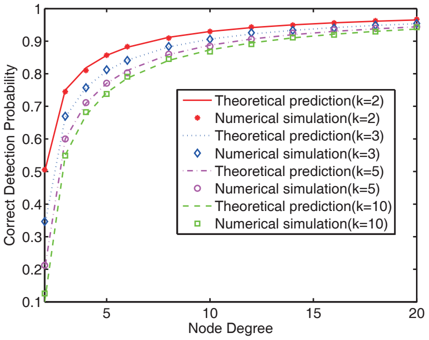

Fig. 3. Detection probability when S forms a connected subgraph of G.

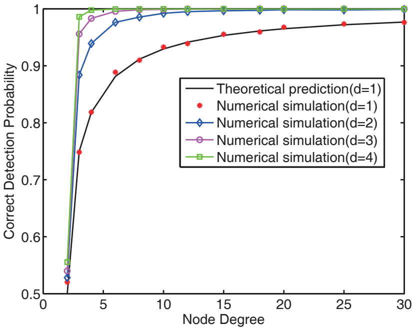

Fig. 4. Detection probability when S contains two suspect nodes.

For the first scenario of $S = V$, it is shown in Fig. 2 that the correct detection probability is increasing with the node degree δ, from virtually zero when $\delta = 2$ to 0.307 as δ exceeds 50. In Fig. 3, we consider the scenario where there are k connected suspect nodes. We observe that the correct detection probability significantly exceeds $1/k$ when $\delta > 2$, and that reliable detection is achieved as δ grows large. In Fig. 4, we consider the scenario where there are two suspect nodes with a shortest path distance d. We observe that the correct detection probability significantly exceeds the prior 1/2 when $\delta > 2$. Furthermore, the correct detection probability increases as d increases, and reliable detection is achieved when either δ or d is sufficiently large.

References

[1] D. Kempe, J. Kleinberg, and Éva Tardos, “Maximizing the spread of influence through a social network,” in *Proc. ACM SIGKDD*, 2003.

[2] D. Shah and T. Zaman, “Rumors in a network: who’s the culprit?” *IEEE Trans. Inform. Theory*, vol. 57, no. 8, pp. 5163–5181, 2011.

[3] ——, “Rumor centrality: a universal source detector,” in *Proc. ACM SIGMETRICS*, 2012.

[4] W. Luo, W. P. Tay, and M. Leng, “Identifying infection sources and regions in large networks,” *IEEE Trans. Signal Processing*, 2013.

[5] K. Zhu and L. Ying, “Information source detection in the SIR model: a sample path based approach,” in *Proc. ITA Workshop*, 2013.

[6] N. Johnson and S. Kotz, *Urn Models and Their Application: An Approach to Modern Discrete Probability Theory*. John Wiley & Sons, 1977.

[7] W. Dong, W. Zhang, and C. W. Tan, “Rooting out the rumor culprit from suspects,” [Online] http://arxiv.org/abs/1301.6312.