Rumor Source Detection with Multiple Observations: Fundamental Limits and Algorithms

Zhaoxu Wang
Dept. of EEIS
University of Science and Technology of China
Hefei, China
wzxboa@mail.ustc.edu.cn

Wenxiang Dong
Dept. of EEIS
University of Science and Technology of China
Hefei, China
javin@mail.ustc.edu.cn

Wenyi Zhang
Dept. of EEIS
University of Science and Technology of China
Hefei, China
wenyizha@ustc.edu.cn

Chee Wei Tan
College of Science and Engineering
City University of Hong Kong
Hong Kong SAR, China
cheewtan@cityu.edu.hk

ABSTRACT

This paper addresses the problem of a single rumor source detection with multiple observations, from a statistical point of view of a spreading over a network, based on the susceptible-infectious model. For tree networks, multiple sequential observations for one single instance of rumor spreading cannot improve over the initial snapshot observation. The situation dramatically improves for multiple independent observations. We propose a unified inference framework based on the union rumor centrality, and provide explicit detection performance for degree-regular tree networks. Surprisingly, even with merely two observations, the detection probability at least doubles that of a single observation, and further approaches one, i.e., reliable detection, with increasing degree. This indicates that a richer diversity enhances detectability. For general graphs, a detection algorithm using a breadth-first search strategy is also proposed and evaluated. Besides rumor source detection, our results can be used in network forensics to combat recurring epidemic-like information spreading such as online anomaly and fraudulent email spams.

Categories and Subject Descriptors

G.2.2 [**Graph Theory**]: Graph algorithms, Network problems

General Terms

Analytical modeling, Anomaly detection, Social networks

SIGMETRICS'14, June 16-20, 2014, Austin, Texas, USA.

ACM 978-1-4503-2789-3/14/06 ...$15.00.
http://dx.doi.org/10.1145/2591971.2591993.

Keywords

Graph networks, inference algorithms, maximum likelihood detection, multiple observations, rumor spreading

1. INTRODUCTION

Identification of malicious information sources in a network, be it in the case of an online spam spreading in the Internet or a misinformation or rumor propagating in an online social network, allows timely quarantine of the epidemic-like spreading to limit the damage caused. For example, law enforcement agencies may be interested in identifying the perpetrators of false information used to manipulate the market prices of certain stocks. It is challenging to understand how the source of the spreading can be identified in a reliable manner from a number of snapshot observations of the spread conditions, e.g., infected nodes in virus propagation or users possessing the rumor, and the underlying connectivity of the network. This can lead to a fundamental understanding of the role of network in aiding or constraining epidemic-like spreading [1, 2].

In this paper, we consider the issue of reliably detecting a single rumor source with multiple observations from a statistical point of view of a spreading. Our goal is to find the rumor source in order to control and prevent network risks based on limited information about the network structure and multiple snapshot observations.

1.1 Related Works

This network inference problem is a combinatorial problem that is generally hard to solve, and has remained surprisingly unexplored until late. In the seminal work [3, 4], Shah and Zaman studied the rumor source detection problem, and conducted asymptotic performance analysis of the Maximum Likelihood (ML) detector for a single rumor source using the susceptible-infectious (SI) model. Then, this work was extended in [5] for random graphs. The authors in [6] studied the Maximum A Posteriori (MAP) detection based on various suspect set characteristics. In [7], the authors investigated the detection of multiple rumor sources in the SI model. In

[8], the authors considered the case where nodes randomly report their infection state in SI spreading, which means we only get an incomplete picture of the infection state. Besides, other types of models have also been considered in [9]-[11]. In [9], the authors analyzed the detection under susceptible-infected-recovered (SIR) spreading, and proposed a probabilistic method based on sample path estimation. In [10], the authors proposed an approach based on message passing algorithms to detect a single rumor source in the SIR model. The authors in [11] studied the detection in susceptible-infected-susceptible (SIS) spreading of regular tree networks.

1.2 Our Contributions

One key question in rumor source detection is to determine the right number of observations or adaptive measurements such that the source can be correctly detected. The existing works, however, rely critically on the assumption that only a single snapshot observation is made in the detection. In practice, there can possibly be multiple observations of a *recurrent* online anomaly, e.g., fraudulent email spams and recurring malcode, which are usually originated from a common culprit. *Dependent* and *independent* multiple observations, respectively, refer to snapshot observations made on different spreading instances (from the same single source) that are dependent and independent from one another. Obviously, the interdependency between observations and network structure affect detectability. Single observation detection does not leverage all degrees of freedom and can perform poorly. Indeed, the multiple observations add an interesting dimension to detecting the source reliably. Can we do more with less by having as few observations as possible and yet enough to provide detectability guarantees? This paper answers this question. To the best of our knowledge, there has been no work done on rumor source detection with multiple observations.

In summary, the contributions of the paper are as follows:

(1) We show that, in SI spreading over a tree network, multiple *dependent* observations, i.e., sequential observations for a single instance of rumor spreading, cannot improve detectability over the initial snapshot observation. On the other hand, we show that multiple *independent* observations lead to dramatic improvement in detectability.

(2) For multiple *independent* observations in a tree network, we propose a unified detection framework based on a graph topological quantity called union rumor centrality and design an efficient algorithm to identify the union rumor center. For degree-regular trees, we characterize and evaluate the exact and asymptotic detection probability.

(3) For general graphs, we extend our detection framework by applying a breadth-first search (BFS) strategy to obtain induced trees from multiple observations. Our numerical results show that the algorithm performs more effectively than that for a single observation in small-world and scale-free networks.

1.3 Organization

The structure of this paper is as follows. We introduce the system model in Section 2. In Section 3, we study the problem with multiple sequential observations. In Section 4, we study the problem with multiple independent observations, and propose a detection framework for tree networks. We analytically characterize the detection performance for regular tree networks. In Section 5, for tree-type networks, we propose algorithms to calculate the exact correct detection probability and identify the union rumor center under multiple independent observations. We extend our framework to the general network case in Section 6. In Section 7, we numerically evaluate the performance of our algorithms. Section 8 provides proofs of the analytical results. We conclude the paper in Section 9.

2. MODEL AND PRELIMINARIES

In this section, we introduce the SI rumor spreading model, and describe the ML detector for the rumor source in regular trees, general trees and graphs, respectively. We list the key parameters used in the paper in Table 1.

Table 1: Key Terms and Symbols

Symbol	Definition
δ	node degree
s^*	original rumor source
$\widehat{s}$	detected rumor source
$T_u^{s^*}$	subtree rooted at node u, with node s^* as source
G_{n_j} $(1\leq j\leq k)$	a snapshot observation of n_j infected nodes for the jth rumor spreading
$\{\cap G_{1\to k}\}$	the intersection of observations, i.e., $\{G_{n_1}\cap\ldots\cap G_{n_k}\}$
$\{\cup G_{1\to k}\}$	the union of observations, i.e., $\{G_{n_1}\cup\ldots\cup G_{n_k}\}$
$R(u,G_n)$	rumor centrality of node u in G_n
$R_k(u,G_{n_1}\ldots G_{n_k})$	union rumor centrality of node u in $\{\cap G_{1\to k}\}$
$\mathbf{P}_G(\cdot)$	probability distribution of an infection sample in a rumor spreading process, or of an infection sample equivalently constructed using the Pólya's urn model
$P_c(\cdot)$	correct detection probability
$P_e(\cdot)$	error detection probability

2.1 Rumor Spreading Model

We consider a network of nodes modeled as an undirected graph $G=(V,E)$, where the set of vertices V represents the nodes in the network, and the set of edges E represents the links between the network nodes. We assume V is countably infinite in order to avoid boundary effects and consider the case where initially only one node $s^*\in V$ is the rumor source.

We use a variant of the SIR model for the rumor spreading known as the SI model, where a node that is infected with the rumor retains it forever. A rumor is spread from node i to node j if and only if there is an edge between them, i.e., if $(i,j)\in E$. Let τ_{ij} be the spreading time from i to j for all $(i,j)\in E$, which are mutually independent and have exponential distribution with parameter λ. Without loss of generality, take λ=1.

2.2 ML Rumor Source Detector

(1) ML Detector

Suppose that a rumor originates from a node $s^*\in V$. We observe the network G at some time and find n infected nodes, which are collectively denoted by G_n. Our goal is to construct a detector to detect a node $\hat{s}$ as the rumor source s^*. The ML detector of s^* given G_n maximizes the correct detection

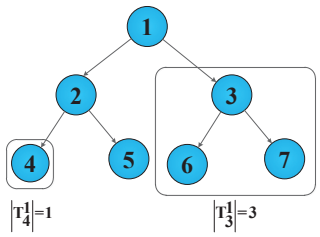


Figure 1: Illustration of subtree $T_u^{s^*}$

probability and is given by

$$\widehat{s} \in \arg\max_{s^* \in G_n} \mathbf{P}_G\left(G_n \mid s^*\right), \tag{1}$$

where $\mathbf{P}_G\left(G_n|s^*\right)$ is the probability of observing G_n assuming s^* to be the rumor source.

(2) Exact ML detector for regular trees

In general, the evaluation of $\mathbf{P}_G\left(G_n|s^*\right)$ is computationally prohibitive since it is related to counting the number of linear extensions of a partially ordered set. By leveraging the concept of rumor centrality, first introduced in [4], the exact ML detector for a regular tree is given by

$$\widehat{s} \in \arg\max_{s^* \in G_n} \mathbf{P}_G\left(G_n \mid s^*\right) = \arg\max_{s^* \in G_n} R(s^*, G_n), \tag{2}$$

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

$$R(s^*, G_n) = (n-1)! \prod_{u \in \text{child}(s^*)} \frac{R(s^*, T_u^{s^*})}{|T_u^{s^*}|!}, \tag{3}$$

where $T_u^{s^*}$ denotes the subtree rooted at node u, with node s^* as the source, and $|T_u^{s^*}|$ denotes the number of nodes in the subtree $T_u^{s^*}$; e.g., see Fig. 1.

In the following analysis, we will abuse notation and use $T_u^{s^*}$ to refer to both a subtree and the number of nodes in the subtree.

(3) Approximate detector

For general cases, intuitively, a rumor tends to travel from the source to each infected node along a minimum-distance path [4], and this serves, along with the rumor centrality, as a reasonable and useful heuristic for constructing approximate detectors. For general trees, an approximate detector is thus given by [4]

$$\widehat{s} \in \arg\max_{s^* \in G_n} P(\sigma_s^* \mid s^*) \cdot R(s^*, G_n), \tag{4}$$

where σ_s^* represents the breadth-first search (BFS) ordering of nodes in the tree.

For general graphs, we approximate the diffusion tree by a BFS tree; that is, we assume that if node s^* was the source, then the rumor spreads along a BFS tree rooted at s^*, denoted by $T_b(s^*)$. So an approximate detector is given by

$$\widehat{s} \in \arg\max_{s^* \in G_n} P(\sigma_s^* \mid s^*) \cdot R\big(s^*, T_b(s^*)\big), \tag{5}$$

where σ_s^* represents the BFS ordering of nodes in the BFS tree $T_b(s^*)$.

Besides, a message-passing algorithm has been proposed in [4] to compute the rumor centralities for all the nodes in a general tree with n nodes using only $O(n)$ computation steps.

3. SEQUENTIAL OBSERVATIONS

In this section, we first define the ML rumor source detector under multiple sequential observations for one single instance of rumor spreading for general trees. Then, we analyze the performance of the ML detector and show that sequential observations cannot help improve the correct detection probability for any general tree.

3.1 ML Detection

Suppose that a rumor originates from a node $s^* \in V$. We observe the network at k different times. In the jth $(1 \le j \le k)$ observation, we find n_j infected nodes carrying the rumor, which are collectively denoted by G_{n_j}. Without loss of generality, assume observation is in turn according to the time order. Due to the SI spreading, G_{n_j} $(1 \le j \le k)$ must form a connected subgraph and contain at least a node, which is the rumor source, i.e., $s^* \in G_{n_1} \subseteq G_{n_2} \ldots \subseteq G_{n_k}$. Note that we assume a uniform prior probability of the source node among all the nodes in the network. Our goal is to construct a detector under k sequential observations to detect a node $\hat{s}$ as the rumor source s^*.

For clarity, in the following analysis, we denote the sets $\{G_{n_1} \cap \ldots \cap G_{n_k}\}$ and $\{G_{n_1} \cup \ldots \cup G_{n_k}\}$ by $\{\cap G_{1\to k}\}$ and $\{\cup G_{1\to k}\}$, respectively. Then, the ML detector under multiple sequential observations maximizes the correct detection probability and is given by

$$\begin{aligned}\widehat{s} \in & \arg\max_{s^* \in \{\cap G_{1\to k}\}} \mathbf{P}_G(G_{n_1}, G_{n_2}, \ldots, G_{n_k} \mid s^*) \\ = & \arg\max_{s^* \in \{\cap G_{1\to k}\}} \mathbf{P}_G(G_{n_1}|s^*) \cdot \mathbf{P}_G(G_{n_2}|s^*, G_{n_1}) \cdot \\ & \mathbf{P}_G(G_{n_3}|s^*, G_{n_1}, G_{n_2}) \cdots \mathbf{P}_G\big(G_{n_k}|s^*, G_{n_1}, \ldots, G_{n_{k-1}}\big). \end{aligned} \tag{6}$$

3.2 Impossibility Result

For ML detector under multiple sequential observations for general trees, it is intuitively apparent that only the first observation is useful in detecting the rumor source. This intuition is made formal in the following proposition.

PROPOSITION 1. *Suppose the rumor spreads on a general tree-type graph as per the SI model, multiple sequential observations for one single instance of rumor spreading cannot improve detectability over the first snapshot observation.*

Remark: The impossibility of performance improvement demonstrates the need for early detection.

The argument of Proposition 1 basically follows from the fact that any $\mathbf{P}_G\big(G_{n_j} | s^*, G_{n_1}, G_{n_2} \ldots G_{n_{j-1}}\big)$ $(2 \le j \le k)$ is conditionally independent of the source s^*. Hence, from (6), the ML detector of s^* under arbitrary k sequential observations becomes

$$\begin{aligned}\widehat{s} \in & \arg\max_{s^* \in \{G_{n_1} \cap G_{n_2} \ldots \cap G_{n_k}\}} \mathbf{P}_G(G_{n_1}, G_{n_2}, \ldots, G_{n_k} \mid s^*) \\ = & \arg\max_{s^* \in \{G_{n_1} \cap G_{n_2} \ldots \cap G_{n_k}\}} \mathbf{P}_G(G_{n_1} \mid s^*). \end{aligned} \tag{7}$$

Therefore, for general tree graphs, the ML detector under sequential observations $(G_{n_1}, \ldots, G_{n_k})$ is nothing but the M-L detector based on the first observation G_{n_1}. This means

that multiple sequential observations cannot help improve the correct detection probability.

4. INDEPENDENT OBSERVATIONS

In this section, we describe the ML rumor source detector under multiple independent observations, which will be shown to be equivalent to a topological quantity that we call the union rumor centrality, for regular tree graphs. Then we leverage the concept of union rumor centrality to develop analytical performance results of the ML detector for regular tree-type networks.

4.1 ML Detection

Suppose k different rumors originate from a common node in the network, regarded as k times independent rumor spreading with the same rumor source. For the jth $(1 \leq j \leq k)$ rumor spreading, at some time, we observe a snapshot of n_j infected nodes carrying the rumor, which are collectively denoted by G_{n_j}. For arbitrary k independent observations, due to the SI model, each G_{n_j} $(1 \leq j \leq k)$ must form a connected subgraph and the rumor source must belong to $\{G_{n_1} \cap \ldots G_{n_k}\}$. We assume a uniform prior probability of the source node among all the nodes in the network. The ML detector under multiple independent observations $G_{n_1},\ldots,G_{n_k}$ maximizes the correct detection probability, as given by

$$\begin{aligned}\widehat{s} \in & \underset{s^* \in \{G_{n_1} \cap G_{n_2} \ldots \cap G_{n_k}\}}{\arg\max} \mathbf{P}_G(G_{n_1}, G_{n_2}, \ldots, G_{n_k} \mid s^*) \\ = & \underset{s^* \in \{G_{n_1} \cap G_{n_2} \ldots \cap G_{n_k}\}}{\arg\max} \mathbf{P}_G(G_{n_1} | s^*) \cdots \mathbf{P}_G(G_{n_k} | s^*). \end{aligned} \quad (8)$$

For a δ-regular tree, since all nodes have the same degree, every permitted permutation has the same probability that is independent of the source [4]. Hence, we have

$$\mathbf{P}_G(G_n \mid s^*) = \sum_{\sigma \in \Omega(s^*, G_n)} P(\sigma | s^*) = R(s^*, G_n) P(\sigma | s^*), \quad (9)$$

where $\Omega(s^*, G_n)$ is the set of all permitted permutations starting with s^* and resulting in G_n, and $P(\sigma|s^*) = \prod_{i=1}^{n-1} \frac{1}{di-2(i-1)} \equiv P(d,n)$.

Then, by substituting (9) into (8), the ML detector under arbitrary k observations for a regular tree becomes

$$\begin{aligned}\widehat{s} = & \underset{s^* \in \{G_{n_1} \cap \ldots G_{n_k}\}}{\arg\max} \Big(R(s^*, G_{n_1}) \cdots R(s^*, G_{n_k})\Big) \Big(P(d, n_1) \cdots P(d, n_k)\Big) \\ = & \underset{s^* \in \{G_{n_1} \cap \ldots G_{n_k}\}}{\arg\max} R(s^*, G_{n_1}) \cdots R(s^*, G_{n_k}). \end{aligned} \quad (10)$$

Note that in the model we do not require that all the rumors originate simultaneously from the original source, or that all the snapshots are observed simultaneously. In fact these k rumor spreadings can be completely asynchronous, as long as they are independent.

4.2 Union Rumor Centrality for Regular Trees

We need to evaluate $R(s^*,G_{n_1}) \cdots R(s^*,G_{n_k})$ in (10) for multiple independent observations in a regular tree network. We call this quantity the union rumor centrality, denoted by $R_k(s^*,G_{n_1} \ldots G_{n_k})$, which enables efficient implementation of the ML detector. Note that the rumor source $s^* \in \{\cap G_{1\to k}\}$. If $R_k(s^*,G_{n_1} \ldots G_{n_k}) \geq R_k(u,G_{n_1} \ldots G_{n_k})$ for all $u \in \{\cap G_{1\to k}\}$, then s^* is called a union rumor center. For the union rumor center, we have the following proposition.

PROPOSITION 2. *For a rumor source s^* with m $(0 \leq m \leq \delta)$ neighbors in $\{\cap G_{1\to k}\}$, given multiple independent observations $G_{n_1}, \ldots, G_{n_k}$ of $n_1, \ldots, n_k$ nodes respectively, then:*
(1) When $m = 0$, there is only one node in $\{\cap G_{1\to k}\}$, which is the rumor source s^*; when $1 \leq m \leq \delta$, node s^* is a union rumor center, if and only if for any $u \in \{\cap G_{1\to k} \setminus s^*\}$, we have

$$\frac{T^{s^*}_{u,G_{n_1}} \cdots T^{s^*}_{u,G_{n_k}}}{(n_1 - T^{s^*}_{u,G_{n_1}}) \cdots (n_k - T^{s^*}_{u,G_{n_k}})} \leq 1, \quad (11)$$

where $T^{s^*}_{u,G_{n_i}}$ denotes the number of nodes in $T^{s^*}_{u,G_{n_i}}$ of the observation G_{n_i} $(1 \leq i \leq k)$.
(2) If there is a node s^* such that s^* has the maximum union rumor centrality among all its neighbors in $\{\cap G_{1\to k} \setminus s^*\}$, then s^* is a union rumor center.
(3) Furthermore, the intersection $\{\cap G_{1\to k}\}$ under multiple independent observations has two union rumor centers, say, s^* and s', if and only if

$$T^{s^*}_{s',G_{n_1}} \cdots T^{s^*}_{s',G_{n_k}} = (n_1 - T^{s^*}_{s',G_{n_1}}) \cdots (n_k - T^{s^*}_{s',G_{n_k}}), \quad (12)$$

and these two nodes are neighbors. In addition, there can be at most two union rumor centers in $\{\cap G_{1\to k}\}$.

Remark: The union rumor center generalizes the rumor center introduced in [4]. Notably, the union rumor center is a graph topological quantity that has a more restricted range $\{\cap G_{1\to k}\}$ than each observation G_{n_j} $(1 \leq j \leq k)$. Due to space limit, we only present a part of the proof of Proposition 2(3).

Proof of Proposition 2(3)

Here, rewriting the assumptions, we will prove that neighboring nodes s^* and s' with equal union rumor centrality are two union rumor centers. For this purpose we need the following lemma, which indicates that the error detection events are all disjoint.

LEMMA 1. *For any node s^* with m $(0 \leq m \leq \delta)$ neighbors in $\{\cap G_{1\to k}\}$, i.e., $s^*_1, \ldots, s^*_m$, let random variable X_{i,n_j} be the number of nodes in subtree $T^{s^*}_{s^*_i,G_{n_j}}$ of a snapshot observation G_{n_j} $(1 \leq i \leq m, 1 \leq j \leq k)$. Then, there is at most a neighbor of the node s^* satisfying*

$$\prod_{j=1}^{k} \Big(\frac{x_{i,n_j}}{n_j}\Big) \geq \prod_{j=1}^{k} \Big(1 - \frac{x_{i,n_j}}{n_j}\Big), \ i \in [1, m]. \quad (13)$$

Based on Lemma 1, since s' and s^* are neighbors that have the same union rumor centrality, while s' is the only neighbor of s^* with a larger union rumor centrality than all other neighbors, s^* also has a larger union rumor centrality than its neighbors, thus leading to Proposition 2(3).

4.3 Detection Performance for Regular Trees

This section examines the performance of the ML detector leveraging the union rumor center. We characterize analytically the performance for finite and asymptotically large number of nodes, for regular trees with an arbitrary degree δ. We present the proofs in Section 8.

The following two theorems study the special cases of node degree $\delta = 2$ and 3.

THEOREM 1. *Suppose the rumor spreads on a regular tree with degree $\delta=2$, i.e., a linear network, given two independent observations G_{n_1} and G_{n_2}, then:*
(1) When $n_2=n_1=n\geq 1$, the correct detection probability $P_{\rm c}$ is given by

$$P_{\rm c}=\binom{2n}{n}2^{-2n+1},\qquad n\geq 1. \tag{14}$$

As $n_1=n_2=n\to\infty$, $P_{\rm c}$ asymptotically scales like $\frac{2}{\sqrt{\pi n}}$.
(2) When $n_1=2$, $n_2>n_1$, $P_{\rm c}$ is given by

$$P_{\rm c}=\left\{\sum_{m=0}^{\lfloor\frac{n_2}{n_1}\rfloor}\binom{n_2-1}{m}-\mathbb{I}\Big(\frac{n_2}{n_1}\Big)\binom{n_2-1}{\lfloor\frac{n_2}{n_1}\rfloor}\right\}\cdot 2^{-n_2+1},$$

where

$$\mathbb{I}(\frac{n_2}{n_1})=\begin{cases}1 & \frac{\rm n_2}{\rm n_1}\in \rm Z\\ 0 & \textit{others.}\end{cases} \tag{15}$$

(3) When $n_2>n_1\geq 3$, $P_{\rm c}$ is given by

$$P_{\rm c}=\left\{\sum_{m=0}^{n_1-1}\Big[\binom{n_2-1}{\lfloor m\frac{n_2}{n_1}\rfloor}\binom{n_1}{m}+S_{n_2-1}(m)\binom{n_1-1}{m}\Big]-2\binom{n_2-1}{\lfloor\frac{n_2}{n_1}\rfloor}-\mathbb{I}\Big(\frac{n_2}{n_1}\Big)\binom{n_2-1}{\lfloor\frac{n_2}{n_1}\rfloor}\right\}\cdot 2^{-(n_1+n_2)+2},$$

where $S_{n_2-1}(m)=\sum_{i=\lfloor m\frac{n_2}{n_1}\rfloor+1}^{\lfloor(m+1)\frac{n_2}{n_1}\rfloor-1}\binom{n_2-1}{i}$.

Remark: For linear networks, the above results provide exact correct detection probability under two independent observations. The performance does not improve much compared with that of a single observation in [4] and [6]; also see Fig. 4(a).

In the following, we see that when the degree is greater than two, multiple independent observations dramatically boost the performance.

THEOREM 2. *Suppose the rumor spreads on a regular tree with degree $\delta=3$, given two independent observations G_{n_1} and G_{n_2}, then:*
(1) When $n_1=n, n_2=qn$ ($q\in\mathrm{Z}^+$), $P_{\rm c}$ is given by

$$P_{\rm c}=\frac{qn+q+2}{2(qn+1)}. \tag{16}$$

(2) When $n_1=n, n_2=qn+1$ ($q\in\mathrm{Z}^+$), (16) still holds.
(3) When $n_1=n, n_2=qn+t$ ($q\in\mathrm{Z}^+$), $t<n$, we have

$$P_{\rm c}=\frac{qn+q+2}{2(qn+1)}+\Delta P_c, \tag{17}$$

with $\Delta P_c<\frac{1}{2(qn+1)}$. This demonstrates that as $n\to\infty$, the asymptotic correct detection probability $\lim_{n\to\infty}P_{\rm c}=1/2$.

Remark: Note that in [4], as $\delta\to\infty$, $P_{\rm c}$ asymptotically approaches 0.307 under a single observation, while here even with only two independent observations, we achieve $\lim_{n\to\infty}P_{\rm c}=1/2$ for $\delta=3$.

For general k, the following theorem describes certain basic properties of the correct detection probability pertaining to its monotonicity.

THEOREM 3. *Suppose the rumor spreads on a δ-regular tree, for k independent observations $G_{n_1},\ldots,G_{n_k}$, then:*
(1) The correct detection probability $P_{\rm c}$ is increasing with δ. As δ grows sufficiently large, $P_{\rm c}$ approaches 1.
(2) When fixing $n_1,\ldots,n_{k-1}$, $P_{\rm c}$ is non-increasing with the number of nodes of the kth observation G_{n_k}, i.e., n_k.

Remark: Theorem 3 reveals that, detection under multiple observations exhibits definite monotonicity behaviors with node degree and snapshot sizes. Such properties shed insights into understanding how the network structure and observations affect the ML detector. In addition, the following corollary is a consequence of Theorem 3.

COROLLARY 1. *Suppose the rumor spreads on a δ-regular tree. For two independent observations G_{n_1} and G_{n_2}, given the number of nodes of G_{n_1}, i.e., n_1, and considering $n_2=qn_1$ and $n_2=qn_1+1$, $q\in Z^+$, these two cases have the same correct detection probability, i.e.,*

$$P_{\rm c}(n_2=qn_1)=P_{\rm c}(n_2=qn_1+1).$$

For any finite $n_1,\ldots,n_k$, we can use Algorithm 1 in Section 5.1 to calculate the exact value of the correct detection probability $P_{\rm c}$. Furthermore, the following theorem precisely quantifies the asymptotic behaviors of $P_{\rm c}$, for large snapshot sizes, degree δ, and number of observations k.

THEOREM 4. *Suppose the rumor spreads on a δ-regular tree, given k independent observations $G_{n_1},\ldots,G_{n_k}$, then:*
(1) When $\delta\geq 3$, we have

$$\lim_{n_1,\ldots,n_k\to\infty}P_{\rm c}=\phi_k(\delta):=1-\delta\left(1-\varphi_k\left(\frac{1}{\delta-2},\frac{\delta-1}{\delta-2}\right)\right), \tag{18}$$

where

$$\varphi_k(\alpha,\beta)=\int\cdots\int_{\prod_{j=1}^k\frac{x_j}{1-x_j}\leq 1}\frac{\Gamma(\alpha+\beta)^k}{\Gamma(\alpha)^k\Gamma(\beta)^k}\prod_{j=1}^k\Big(x_j^{\alpha-1}(1-x_j)^{\beta-1}\Big)\,\mathrm{d}x_1\ldots\mathrm{d}x_k,$$

$\alpha=\frac{1}{\delta-2}$, $\beta=\frac{\delta-1}{\delta-2}$.

As δ grows sufficiently large, $\phi_k(\delta)\to 1$; also, as k grows sufficiently large, $\phi_k(\delta)\to 1$.
(2) In particular, when $\delta=3$, $\phi_k(3)$ can be written as

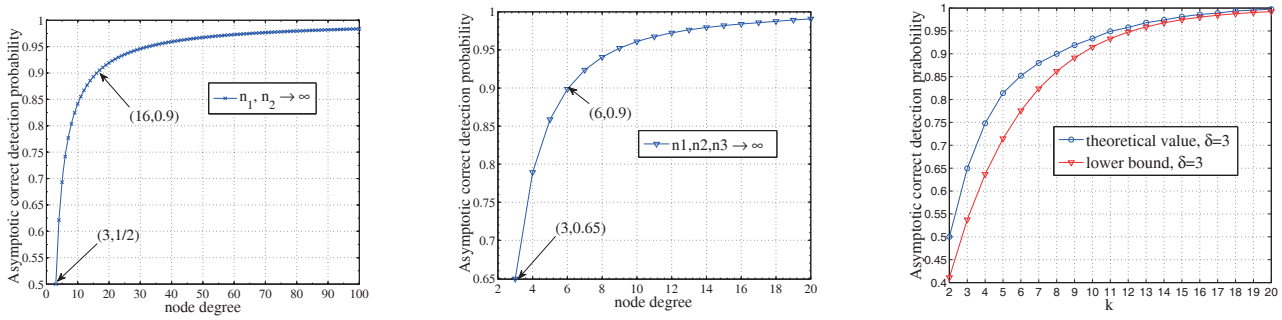
$$\phi_k(3)=1-3\cdot 2^{k-2}\int_0^1\cdots\int_0^1\frac{\prod_{j=1}^{k-1}\big(x_j(1-x_j)\big)}{\prod_{j=1}^{k-1}x_j+\prod_{j=1}^{k-1}(1-x_j)}\mathrm{d}x_1\ldots\mathrm{d}x_{k-1}, \tag{19}$$

which is increasing with k and is bounded as follows:

$$1-\frac{3}{4}(\frac{\pi}{4})^{k-1}<\lim_{n_1,\ldots,n_k\to\infty}P_{\rm c}<1, k\in\mathrm{Z}^+; \tag{20}$$

that is, as k grows sufficiently large, $\lim_{n_1,\ldots,n_k\to\infty}P_{\rm c}\to 1$ exponentially.

Remark: For general k observations, we mainly provide the results in the asymptotic regime as $n\to\infty$. The lower bound in case of $\delta=3$ provides an intuitive insight that $\phi_k(3)$ increases with the number of observations k, and further indicates the convergence to be exponentially fast; see Fig. 2(c). According to Theorem 3 that $P_{\rm c}$ increases with δ and decreases with n, we have as $\delta>2$, $k>1$, $P_{\rm c}\geq P_{\rm c}(\delta=3)>\phi_k(3)\geq\phi_2(3)=1/2$. Indeed $P_{\rm c}$ may by far exceed $1/2$, if $\delta>2$, $k>1$; and as any of k and δ grows sufficiently large, $P_{\rm c}$ converges to 1. Therefore, Theorem 4 reveals that multiple independent observations significantly improve the detection performance.



(a) Two independent observations (b) Three independent observations (c) $\phi_k(\delta)$ vs k, $\delta = 3$

Figure 2: Asymptotic correct detection probability under multiple independent observations

COROLLARY 2. *Suppose the rumor spreads on a δ-regular tree, given two independent observations G_{n_1} and G_{n_2}, then: (1) When $\delta \geq 3$, we have*

$$\lim_{n_1,n_2\to\infty} P_{\rm c} = \phi_2(\delta) = 1-\delta\left(1-\varphi_2\left(\frac{1}{\delta-2},\frac{\delta-1}{\delta-2}\right)\right),$$

where
$\varphi_2(\alpha,\beta) = \iint_{x_1+x_2\leq 1} \frac{\Gamma(\alpha+\beta)^2}{\Gamma(\alpha)^2\Gamma(\beta)^2} x_1^{\alpha-1}x_2^{\alpha-1}(1-x_1)^{\beta-1}(1-x_2)^{\beta-1}\mathrm{d}x_1\mathrm{d}x_2$,
$\alpha = \frac{1}{\delta-2}$, $\beta = \frac{\delta-1}{\delta-2}$. *As δ grows sufficiently large, $\phi_2(\delta)\to 1$. (2) In particular, when $\delta = 3$, the asymptotic correct detection probability equals to $1/2$, i.e.,*

$$\lim_{n_1\to\infty,n_2\to\infty} P_{\rm c} = \frac{1}{2}; \tag{21}$$

when $\delta = 4$, the asymptotic correct detection probability is

$$\lim_{n_1\to\infty,n_2\to\infty} P_{\rm c} = \frac{16}{\pi^2} - 1 \approx 0.621. \tag{22}$$

Remark: Corollary 2 is a special case of Theorem 4 for $k = 2$, illustrating that the asymptotic correct detection probability under merely two independent observations already significantly exceeds 1/2; see Fig. 2(a). Besides, Fig. 2(b) displays an example of Theorem 4 in the case of three independent observations. As observed, when $\delta = 3$ or 6, the asymptotic correct detection probability equals to 0.65 or 0.9, in sharp contrast to that of 0.25 or 0.307 for $\delta = 3$ or $\delta\to\infty$ under a single observation in [4] and [6].

5. ALGORITHMS FOR TREES

5.1 Detection Probability for Regular Trees

In this section, we present Algorithm 1 to calculate the exact correct detection probability under multiple independent observations $G_{n_1},\ldots,G_{n_k}$ in the finite regime for regular trees, based on the theoretical results established in Section 8.

5.2 Union Rumor Centrality Calculation

To find the union rumor center from k different independent observations $G_{n_1},\ldots,G_{n_k}$, we need to evaluate the maximum union rumor centrality in $\{\cap G_{1\to k}\}$. In [4], a message-passing algorithm is proposed to compute the rumor centralities for one time observation in a general tree G_n with n nodes, using $\mathcal{O}(n)$ computations. In this section, we propose an algorithm for calculating the union rumor centralities among all infected nodes based on message-passing, using $\mathcal{O}(|G_{n_1}\cup\ldots G_{n_k}|)$ computation steps. A baseline algorithm for comparison is to naively execute k instances of the single observation message-passing algorithm in [4] to find the union rumor center, using $\mathcal{O}(\sum_{j=1}^k n_j)$ computations. Since $|G_{n_1}\cup\ldots G_{n_k}| < \sum_{j=1}^k n_j$, our algorithm is more computationally efficient.

Algorithm 1 Correct Detection Probability Calculation

Input: δ, k, G_{n_j} $(j=1,\ldots,k)$
Initialize $P_e = 0$
for $x_{1,n_k} = 1 \to n_k - 1$ **do**
  .....
  for $x_{1,n_1} = 1 \to n_1 - 1$ **do**
    if $\prod_{j=1}^k x_{1,n_j} > \prod_{j=1}^k (n_j - x_{1,n_j})$ **then**
      $P_e = P_e + \prod_{j=1}^k \mathbf{P}_{G_{n_j}}\left(X_{1,n_j} = x_{1,n_j}\right)$
    end if
    if $\prod_{j=1}^k x_{1,n_j} = \prod_{j=1}^k (n_j - x_{1,n_j})$ **then**
      $P_e = P_e + \frac{1}{2}\cdot\prod_{j=1}^k \mathbf{P}_{G_{n_j}}\left(X_{1,n_j} = x_{1,n_j}\right)$
    end if
  end for
end for
Output: $P_c = 1 - \delta\cdot P_e$

According to (3) and (10), the union rumor centrality of node s ($s\in\{\cap G_{1\to k}\}$) can be written as

$$R_k(s,G_{n_1}\ldots G_{n_k}) = R_k(s^*,G_{n_1}\cdots G_{n_k})\prod_{i\in\mathbb{P}(s^*,s)}\frac{T^{s^*}_{i,G_{n_1}}\cdots T^{s^*}_{i,G_{n_k}}}{(n_1-T^{s^*}_{i,G_{n_1}})\cdots(n_k-T^{s^*}_{i,G_{n_k}})}$$
$$= \frac{(n_1-1)!\cdots(n_k-1)!}{\prod_{u\in\{G_{n_1}\setminus s\}} T^s_{u,G_{n_1}}\cdots\prod_{u\in\{G_{n_k}\setminus s\}} T^s_{u,G_{n_k}}}, \tag{23}$$

where $\mathbb{P}(s^*,s)$ is the set of nodes in the path between the source s^* and s, not including s^*. This is the key to our algorithm for calculating the union rumor centralities for all the nodes in $\{\cap G_{1\to k}\}$. Let us briefly describe the idea of our algorithm. We consider the extended infected tree-type graph formed by the k observations, $\{\cup G_{1\to k}\}$. We first select an arbitrary node $v\in\{\cap G_{1\to k}\}$ as the rumor source and calculate the size of all of its subtrees and its union rumor centrality. This is accomplished by having each node u pass three mes-

sages up to its parent in $\cup G_{1\to k}$. The first message, denoted by $t^{\text{up}}_{u\to\text{parent}(u,\cup G_{1\to k})}$, contains up to k variables, each of which is denoted by $t^{\text{up}}_{u,G_{n_j}\to\text{parent}(u,G_{n_j})}$, representing the number of nodes in u's subtree of the observation G_{n_j}, as u and its parent both belong to G_{n_j}. The second message contains two variables, one of which is the cumulative product of the variables in the first message, which we call $t^{\text{up}*}_{u\to\text{parent}(u,\cup G_{1\to k})}$, and the other of which is an optional variable, denoted by $t^{\text{up}'}_{u\to\text{parent}(u,\cup G_{1\to k})}$, when both $u,v\in\cap G_{1\to k}$, thus assisting in calculating the union rumor centralities of neighboring nodes. The third message is the cumulative product of the sizes of the subtrees of all nodes in u's subtree of each G_{n_j}, which we call $p^{\text{up}}_{u\to\text{parent}(u,\cup G_{1\to k})}$. These messages are passed upward until the source node receives its messages. By multiplying the cumulative subtree products of the children of the source v, we obtain the union rumor centrality $R_k(v,G_{n_1}\dots G_{n_k})$.

Algorithm 2 Union Rumor Centrality Calculation

Input: Graphs $G_{n_j}, j\in[1,k]$

1: Choose an arbitrary node $v\in\cap G_{1\to k}$ as a root.
2: **for** each $u\in\cup G_{1\to k}$ **do**
3: **if** u is a leaf **then**
4: $t^{\text{up}}_{u\to\text{parent}(u,\cup G_{1\to k})}$: $t^{\text{up}}_{u,G_{n_j}\to\text{parent}(u,G_{n_j})}=1$, $j\in\{j: u,\text{parent}(u)\in G_{n_j}, j\in[1,k]\}$
5: $t^{\text{up}*}_{u\to\text{parent}(u,\cup G_{1\to k})}=1$, $p^{\text{up}}_{u\to\text{parent}(u,\cup G_{1\to k})}=1$
6: **else if** u is root v **then**
7: $v',v\in\cap G_{1\to k},\forall v'\in\text{child}(v,\cap G_{1\to k}), r^{\text{down}}_{v\to v'}=\frac{\prod_{j=1}^{k}(n_j-1)!}{\prod_{i\in\text{child}(v,\cup G_{1\to k})}p^{\text{up}}_{i\to v}}$
8: **else**
9: $t^{\text{up}}_{u\to\text{parent}(u,\cup G_{1\to k})}$: $t^{\text{up}}_{u\to\text{parent}(u,G_{n_j})}=\sum_{i\in\text{child}(u,G_{n_j})}t^{\text{up}}_{i\to u}+1$, $j\in\{j: u,\text{parent}(u)\in G_{n_j}, j\in[1,k]\}$
10: $t^{\text{up}*}_{u\to\text{parent}(u,\cup G_{1\to k})}=\prod_{\{j:u,\text{parent}(u)\in G_{n_j},j\in[1,k]\}}t^{\text{up}}_{u\to\text{parent}(u,G_{n_j})}$
11: **if** $u\in\cap G_{1\to k}$ **then**
12: $t^{\text{up}'}_{u\to\text{parent}(u,\cup G_{1\to k})}=\prod_{\{j:u,\text{parent}(u)\in G_{n_j},j\in[1,k]\}}\left(n_j-t^{\text{up}}_{u\to\text{parent}(u,G_{n_j})}\right)$
13: $\forall u'\in\text{child}(u,\cap G_{1\to k})$, $r^{\text{down}}_{u\to u'}=r^{\text{down}}_{\text{parent}(u,\cup G_{1\to k})\to u}\cdot\frac{t^{\text{up}*}_{u\to\text{parent}(u,\cup G_{1\to k})}}{t^{\text{up}'}_{u\to\text{parent}(u,\cup G_{1\to k})}}$
14: **end if**
15: $p^{\text{up}}_{u\to\text{parent}(u,\cup G_{1\to k})}=t^{\text{up}*}_{u\to\text{parent}(u,\cup G_{1\to k})}\cdot\prod_{i\in\text{child}(u,\cup G_{1\to k})}p^{\text{up}}_{i\to u}$
16: **end if**
17: **end for**

Output: Union rumor centralities of all the nodes in $\{\cap G_{1\to k}\}$.

Based on the union rumor centrality of node v, we evaluate the union rumor centralities of the children of v using (23). Each node in $\cap G_{1\to k}$ passes its union rumor centrality to its children in a message which we denote by $r^{\text{down}}_{u\to u'}$, for $\forall u'\in\text{child}(u,\cap G_{1\to k})$. Each node calculates its union rumor centrality using its parent's union rumor centrality and its messages $t^{\text{up}*}_{u\to\text{parent}(u,\cup G_{1\to k})}$ and $t^{\text{up}'}_{u\to\text{parent}(u,\cup G_{1\to k})}$. Therefore, our union rumor centrality calculation algorithm calculates the union rumor centralities of all infected nodes by message-passing, using $\mathcal{O}(|G_{n_1}\cup\dots G_{n_k}|)$ computation steps.

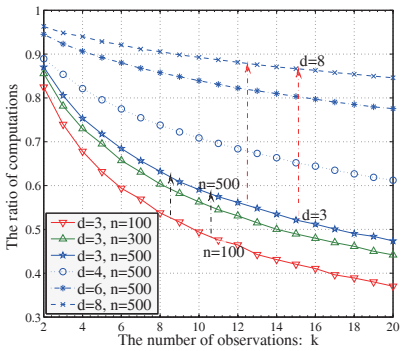


Figure 3: Ratio of computations between Algorithm 2 and the direct implementation of multiple instances of single observation message-passing algorithm.

We perform simulations with different node degrees and infected set sizes to empirically evaluate the computational efficiency. We define r_a as the ratio of computations between Algorithm 2 and the direct execution of multiple instances of single observation message-passing algorithm in [4]. For each simulation experiment we conduct the Monte-Carlo simulation 10000 times. As shown in Fig. 3, r_a decreases with the number of observations, and increases with the infected set size and node degree. These imply that Algorithm 2 is advantageous in cases with small node degrees and large number of observations. For example, when the node degree is 3, the number of observations is 20, and 500 infected nodes are observed in each observation, Algorithm 2 almost reduces half of the computations as compared to running multiple single observation message-passing algorithms.

6. DETECTOR FOR GENERAL GRAPHS

For general graphs, owing to the lack of knowledge of the spanning tree, we use the BFS heuristic to obtain an induced tree network from each observation. We assume that if node $s\in G_{n_j}$ $(1\le j\le k)$ is the source, then the rumor spreads along a BFS tree rooted at s, denoted by $T_b(s,G_{n_j})$. Therefore, based on Section 2.2(3), we obtain the following rumor source detector:

$$\widehat{s}=\arg\max_{s^*\in\{\cap G_{1\to k}\}}\prod_{j=1}^{k}P(\sigma_s^*\mid s,G_{n_j})\cdot R_k(s^*,G_{n_1}\dots G_{n_k}),\qquad(24)$$

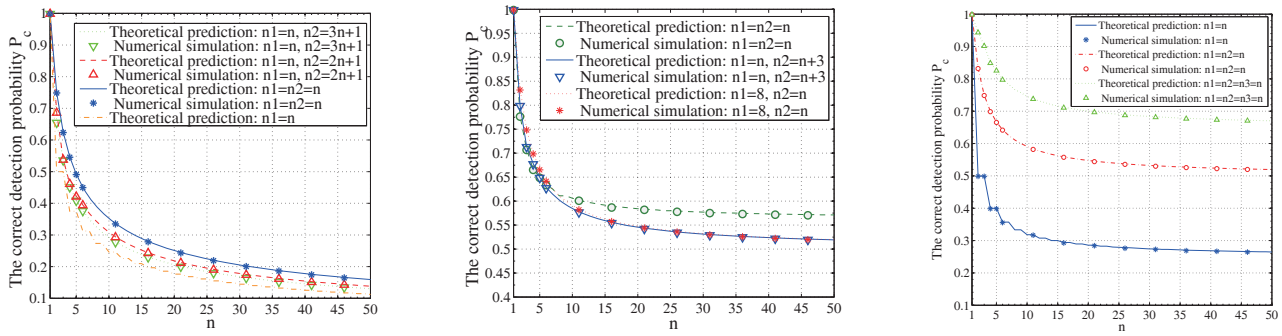
where σ_s^* represents the BFS ordering of nodes in $T_b(s,G_{n_j})$. In the next section, we will show with simulations the performance of this detector for different networks.

7. NUMERICAL SIMULATIONS

In this section, we evaluate the performance of the proposed rumor source detector on different networks.

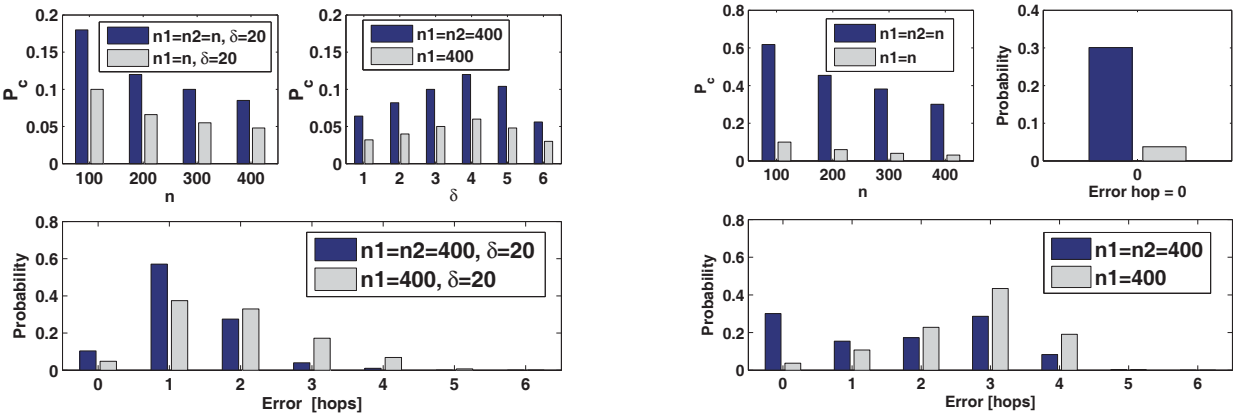
7.1 Tree Networks

Here we provide simulation results for regular trees in order to corroborate the analysis in Section 4. For each configuration, we conduct the Monte-Carlo simulation 10000 times to assess the correct detection probability.



(a) $\delta = 2$ (b) $\delta = 3$ (c) P_c vs n,$\delta = 3$

Figure 4: Simulation results for regular trees



(a) Small-world network (b) Scale-free network

Figure 5: Simulation results for general graph networks

In Fig. 4(a), for $\delta = 2$, the simulation results under two independent observations coincide with the theoretical values as predicted by Theorem 1. Similarly, in Fig. 4(b), for $\delta = 3$, we observe the consistency between the simulation results and the theoretical prediction in Theorem 2. When n becomes large, P_c is close to 1/2, as predicted by Theorem 2.

In Fig. 4(c), we show that when $\delta = 3$, P_c increases with the number of observations and decreases with the number of infected nodes. As shown, multiple independent observations significantly improve the performance, compared with a single observation.

7.2 General Graph Networks

We perform simulations on small-world networks [12] and scale-free networks [13]. For both topologies, each underlying graph contains 5000 nodes and we let the rumor spread to infect up to 400 nodes in each observation.

For both Fig. 5(a) and Fig. 5(b), in the top left subplots, we study the correct detection probability with the number of infected nodes. For both networks, we have a larger correct detection probability with two independent observations compared with a single observation, and the correct detection probability decreases as the number of infected nodes increases. Then for small-world networks, we change the average degree δ to investigate its effect as displayed in the top right subplot of Fig. 5(a). We observe that even when δ is large, two observations still enhance the detectability considerably.

In the bottom subplots of Fig. 5(a) and Fig. 5(b), we investigate the error (i.e., the number of hops between the detected source and the actual source) on the two networks, respectively. Comparing two independent observations with a single observation, we have a better performance in both networks, especially the scale-free network (see the top right subplot of Fig. 5(b)). The reason for this may be that scale-free networks have a highly heterogeneous degree distribution and thus contain many high-degree hubs.

8. PROOFS

8.1 Lemmas

We need the following technical lemmas, whose proofs (along with that of Lemma 1) are in the appendix. For a rumor source s^* with m $(1 \leq m \leq \delta)$ neighbors in $\{\cap G_{1\to k}\}$, i.e., $s_1^*, \ldots, s_m^*$,

let random variable X_{i,n_j} be the number of nodes in subtree $T^{s^*}_{s^*_i,G_{n_j}}$ of observation G_{n_j} $(1 \le i \le m, 1 \le j \le k)$.

LEMMA 2. *For the source s^* and the inferred $\hat{s}$, introduce*

$$\begin{cases} P_1 := P\left(\hat{s}=s^* \middle| \forall i, 1\le i\le m, \prod_{j=1}^k x_{i,n_j} < \prod_{j=1}^k (n_j - x_{i,n_j})\right) = 1, \\ P_{\frac{1}{2}} := P\left(\hat{s}=s^* \middle| \exists i, 1\le i\le m, \prod_{j=1}^k x_{i,n_j} = \prod_{j=1}^k (n_j - x_{i,n_j})\right) = \frac{1}{2}, \\ P_0 := P\left(\hat{s}=s^* \middle| \exists i, 1\le i\le m, \prod_{j=1}^k x_{i,n_j} > \prod_{j=1}^k (n_j - x_{i,n_j})\right) = 0. \end{cases}$$

The correct detection probability $P_{\rm c}$ is given by

$$P_{\rm c} = P_{\frac{1}{2}} \cdot \sum_{\prod_{j=1}^k x_{i,n_j} = \prod_{j=1}^k (n_j - x_{i,n_j})} \prod_{j=1}^k \mathbf{P}_{G_{n_j}}\left(\bigcap_{i=1}^{\delta} \{X_{i,n_j} = x_{i,n_j}\}\right) + P_1 \cdot \sum_{\prod_{j=1}^k x_{i,n_j} < \prod_{j=1}^k (n_j - x_{i,n_j})} \prod_{j=1}^k \mathbf{P}_{G_{n_j}}\left(\bigcap_{i=1}^{\delta} \{X_{i,n_j} = x_{i,n_j}\}\right). \quad (25)$$

LEMMA 3. *The correct detection probability $P_{\rm c}$ in Lemma 2 can be rewritten as*

$$P_{\rm c} = 1 - \delta \left\{ \frac{1}{2} \sum_{\prod_{j=1}^k x_{i,n_j} = \prod_{j=1}^k (n_j - x_{i,n_j})} \prod_{j=1}^k \mathbf{P}_{G_{n_j}}\left(\bigcap_{i=1}^{\delta} \{X_{i,n_j} = x_{i,n_j}\}\right) + \sum_{\prod_{j=1}^k x_{i,n_j} > \prod_{j=1}^k (n_j - x_{i,n_j})} \prod_{j=1}^k \mathbf{P}_{G_{n_j}}\left(\bigcap_{i=1}^{\delta} \{X_{i,n_j} = x_{i,n_j}\}\right) \right\}. \quad (26)$$

Remark: Algorithm 1 in Section 5.1 is based on Lemma 3.

LEMMA 4. *Suppose a rumor source s^* spreads on a δ-regular tree, resulting in an observation G_n. As $n \to \infty$, the probability that a subtree $T^{s^*}_{s^*_i,G_n}$ $(1 \le i \le \delta)$ is empty asymptotically vanishes:*

$$\lim_{n\to\infty} \mathbf{P}_{G_n}(X_i = 0) = 0, \qquad 1 \le i \le \delta\ . \quad (27)$$

LEMMA 5. *Suppose a rumor source s^* spreads on a δ-regular tree, with multiple independent observations $G_{n_1}, \ldots, G_{n_k}$. For $1 \le i \le \delta$, define*

$$E_i = \left\{ \frac{x_{i,n_1}}{n_1} \cdots \frac{x_{i,n_k}}{n_k} < (1 - \frac{x_{i,n_1}}{n_1}) \cdots (1 - \frac{x_{i,n_k}}{n_k}) \right\},$$
$$D_i = \left\{ \frac{x_{i,n_1}}{n_1} \cdots \frac{x_{i,n_k}}{n_k} = (1 - \frac{x_{i,n_1}}{n_1}) \cdots (1 - \frac{x_{i,n_k}}{n_k}) \right\},$$
$$F_i = \left\{ \frac{x_{i,n_1}}{n_1} \cdots \frac{x_{i,n_k}}{n_k} \le (1 - \frac{x_{i,n_1}}{n_1}) \cdots (1 - \frac{x_{i,n_k}}{n_k}) \right\}.$$

As $n_1, \ldots, n_k \to \infty$, we have

$$\lim_{n_1,\ldots,n_k\to\infty} P_{\rm c} = 1 - \delta \cdot \lim_{n_1,\ldots,n_k\to\infty} \mathbf{P}_G(E_1^{\rm c}) = 1 - \delta \cdot \lim_{n_1,\ldots,n_k\to\infty} \mathbf{P}_G(F_1^{\rm c}). \quad (28)$$

8.2 Theorem 1

Here we only present the proof of Theorem 1(1) for $n_1 = n_2 = n$; the other cases can be deduced similarly.

Since the rumor source s^* has m neighbors $(0 \le m \le \delta)$, we have $\delta+1$ cases depending on m. For clarity, we abbreviate x_{i,n_1} and x_{i,n_2} by x_i and y_i $(1 \le i \le \delta)$, respectively. Our proof proceeds in two steps. First, for each of the $\delta+1$ cases, we calculate $P_{\rm c}$ using Lemma 2. Second, we sum up all the cases.

When $\delta=2$, the joint distributions of $\{X_i, 1 \le i \le 2\}$ and $\{Y_i, 1 \le i \le 2\}$ are respectively (see [6])

$$\mathbf{P}_{G_{n_1}}\left(\bigcap_{i=1}^{2}\{X_i = x_i\}\right) = \frac{(n-1)!}{x_1! x_2!} \frac{1}{2^{n-1}}, \mathbf{P}_{G_{n_2}}\left(\bigcap_{i=1}^{2}\{Y_i = y_i\}\right) = \frac{(n-1)!}{y_1! y_2!} \frac{1}{2^{n-1}}. \quad (29)$$

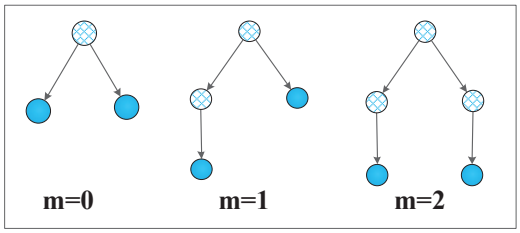


Figure 6: Illustration of a rumor source s^* with m neighbors $(0 \le m \le \delta)$ in $\{\cap G_{1\to k}\}$, $\delta = 2$. Latticed circles represent infected nodes in $\{\cap G_{1\to k}\}$.

Consider the case of $m=2$; see Fig. 6. According to Proposition 2, (x_1, x_2) and (y_1, y_2) should satisfy

$$\begin{cases} x_1 + y_1 \le n, \\ x_2 + y_2 \le n, \\ x_1 + x_2 = n-1, \quad x_1 \ge 1, x_2 \ge 1, \\ y_1 + y_2 = n-1, \quad y_1 \ge 1, y_2 \ge 1. \end{cases} \quad (30)$$

Then when multiplied by the joint distributions $P_G(X)$ and $P_G(Y)$, from Lemma 2, the contribution to $P_{\rm c}$ is given by

$$\begin{aligned} P_{\rm c,2} = &\left\{ \binom{n-1}{1}\left(\binom{n-1}{1} + \frac{1}{2}\binom{n-1}{2}\right) + \sum_{x_1=2}^{n-3} \binom{n-1}{x_1}\left(\frac{1}{2}\binom{n-1}{x_1-1} + \binom{n-1}{x_1}\left(\frac{1}{2}\binom{n-1}{x_1+1}\right) + \binom{n-1}{x_1}\left(\binom{n-1}{x_1} + \frac{1}{2}\binom{n-1}{x_1+1}\right) + \binom{n-1}{n-2}\left(\binom{n-1}{n-2} + \frac{1}{2}\binom{n-1}{n-3}\right)\right\} \cdot 2^{-2(n-1)} \\ = &\left(\frac{1}{2}\binom{2n}{n} - 2n\right) \cdot 2^{-2(n-1)}. \end{aligned} \quad (31)$$

Likewise, the contributions to $P_{\rm c}$ in the case of $m=1$ and $m=0$ are $P_{\rm c,1} = 2(n-1) \cdot 2^{-2(n-1)}$ and $P_{\rm c,0} = 2 \cdot 2^{-2(n-1)}$, respectively. Summing up all the cases from $m=0$ to 2, Theorem 1(1) is proved.

As $n \to \infty$, by Stirling's formula, from (14), we have

$$P_{\rm c} = \frac{1}{2^{2n-1}} \cdot \frac{(2n)!}{n!n!} \approx \frac{1}{2^{2n-1}} \frac{\sqrt{2\pi n}(\frac{2n}{e})^{2n}}{(\sqrt{2\pi n}(\frac{2n}{e})^n)^2} = \frac{2}{\sqrt{\pi n}}. \quad (32)$$

8.3 Theorem 2

Here we only present the proof of Theorem 2(1); the other cases can be deduced similarly.

Our proof proceeds in two steps. First, we identify error detection events from Proposition 2 and Lemma 1. Second, we calculate $P_{\rm c}$ by Lemma 3.

Since $\delta=3$, the joint distributions of $\{X_i, 1 \le i \le 3\}$ and $\{Y_i, 1 \le i \le 3\}$ are $\mathbf{P}_{G_{n_1}}\left(\bigcap_{i=1}^{3}\{X_i = x_i\}\right) = \frac{2}{n(n+1)}$, $\mathbf{P}_{G_{n_2}}\left(\bigcap_{i=1}^{3}\{Y_i = y_i\}\right) = \frac{2}{qn(qn+1)}$, respectively. Hence, the marginal distributions of X_1

and Y_1 are

$$\mathbf{P}_{G_{n_1}}(X_1{=}x_1){=}\frac{2(n{-}x_1)}{n(n{+}1)},\quad \mathbf{P}_{G_{n_2}}(Y_1{=}y_1){=}\frac{2(n{-}y_1)}{qn(qn{+}1)},\quad (33)$$

respectively.

From Proposition 2 and Lemma 1, when X_1 and Y_1 satisfy

$$\frac{x_1y_1}{(n{-}x_1)(qn{-}y_1)}\geq 1, \quad (34)$$

the error detection events happen.

Applying Lemma 3, when enumerating the error detection events, we have

$$P_{\mathrm{e}}{=}\sum_{x_1=1}^{n-1}\frac{2(n{-}x_1)}{n(n{+}1)}\Bigg(\sum_{y_1=qn-qx_1+1}^{qn-1}\frac{2(n{-}y_1)}{qn(qn{+}1)} + \frac{1}{2}\sum_{y_1=qn-qx_1}\frac{2(n{-}y_1)}{qn(qn{+}1)}\Bigg)=\frac{qn-q}{6(qn+1)}. \quad (35)$$

So P_{c} is

$$P_{\mathrm{c}}{=}1-3{\cdot}\frac{qn-q}{6(qn{+}1)}{=}\frac{qn{+}q{+}2}{2(qn{+}1)}. \quad (36)$$

Theorem 2(1) is thus proved.

8.4 Theorem 3

a. Theorem 3(1)

From Lemma 3, in order to prove that P_{c} is increasing with δ, we just need to demonstrate that for any $(x_{1,n_1},\ldots,x_{1,n_k})$ satisfying $\prod_{j=1}^{k}x_{1,n_j}{\geq}\prod_{j=1}^{k}(n_j{-}x_{1,n_j})$, $\delta\prod_{j=1}^{k}\mathbf{P}_{G_{n_j}}(X_{1,n_j}{=}x_{1,n_j})$ is decreasing with δ. Note that the marginal distribution of X_1 is given by

$$\begin{aligned}&\mathbf{P}_G(X_1{=}x_1)\\ &{=}\binom{n-1}{x_1}\frac{\prod_{i=1}^{2}b_i(b_i{+}\delta{-}2)\cdots(b_i{+}(x_i{-}1)(\delta{-}2))}{\delta(\delta{+}\delta{-}2)\cdots(\delta{+}(n{-}2)(\delta{-}2))},\end{aligned} \quad (37)$$

where $x_2 = n-1-x_1$, $b_1 = 1$ and $b_2 = \delta-1$. By defining $f(a,b) = \frac{\frac{1}{\delta-2}+a}{\frac{\delta}{\delta-2}+b}$ and $F(\delta) = \delta\prod_{j=1}^{k}\mathbf{P}_{G_{n_j}}(X_{1,n_j}=x_{1,n_j})$, we have

$$\begin{aligned}F(\delta){=}&\delta\prod_{j=1}^{k}\binom{n_j{-}1}{x_{1,n_j}}\frac{\prod_{z_1=0}^{x_{1,n_j}-1}(1{+}z_1(\delta{-}2))\prod_{z_2=0}^{n_j-x_{1,n_j}-1}(\delta{-}1{+}z_2(\delta{-}2))}{\delta(\delta{+}\delta{-}2)\cdots(\delta{+}(n{-}2)(\delta{-}2))}\\ {=}&\delta^{-(k-1)}\prod_{j=1}^{k}\Big(f(1,1)f(2,2)\cdots f(x_{1,n_j}{-}1,x_{1,n_j}{-}1)\\ &\cdot f(1,x_{1,n_j})\cdots f(n_j-2-x_{1,n_j},n_j-2)\Big).\end{aligned} \quad (38)$$

Rearranging $f(a,b)$, we have

$$f(a,b)=\frac{a\delta-a+1}{(1+b)\delta-b}=\frac{b+1}{a}+\frac{\frac{b+1-a}{a}}{(1+b)\delta-b}.$$

Hence if $b\geq a$, $f(a,b)$ is decreasing with δ. So all factors of (38) are decreasing with δ, and so is $F(\delta)$. As δ grows large, $\lim_{\delta\to\infty}f(a,b)=\frac{b+1}{a}$ is a finite number. By applying (38), $\lim_{\delta\to\infty}F(\delta)=\lim_{\delta\to\infty}\delta^{-(k-1)}\cdot C=0$, where C is a bounded number. Due to the fact that the total number of the error detection events is finite, as δ grows without bound, the correct detection probability P_{c} approaches 1 asymptotically.

b. Theorem 3(2)

Here, we will prove that P_{c} is non-increasing as n_k increases to $n_k{+}1$, with given $G_{n_1},\ldots,G_{n_{k-1}}$ (i.e., fixing $n_1,\ldots,n_{k-1}$). We define events $D(\cdot)$ and $E(\cdot)$ as follows:

$$D(n_k){=}\left\{\frac{x_{1,n_k}}{n_k}=\frac{\prod_{j=1}^{k-1}(n_j{-}x_{1,n_j})}{\prod_{j=1}^{k-1}(n_j{-}x_{1,n_j})+\prod_{j=1}^{k-1}x_{1,n_j}}\right\}, \quad (39)$$

$$E(n_k){=}\left\{\frac{x_{1,n_k}}{n_k}>\frac{\prod_{j=1}^{k-1}(n_j{-}x_{1,n_j})}{\prod_{j=1}^{k-1}(n_j{-}x_{1,n_j})+\prod_{j=1}^{k-1}x_{1,n_j}}\right\}. \quad (40)$$

Thus $D(n_k){\cup}E(n_k)$ is

$$D(n_k){\cup}E(n_k)=\left\{\frac{\prod_{j=1}^{k}x_{1,n_j}}{\prod_{j=1}^{k}(n_j{-}x_{1,n_j})}\geq 1\right\}. \quad (41)$$

It can then be shown that the correct detection probability in Lemma 3 is equivalent to

$$\begin{aligned}P_{\mathrm{c}}{=}1{-}\delta\Bigg\{&\frac{1}{2}\sum_{x_{1,n_k}=1}^{n_k-1}\mathbf{P}_{G_{n_k}}(X_{1,n_k}{=}x_{1,n_k})\sum_{D(n_k)}\prod_{j=1}^{k-1}\mathbf{P}_{G_{n_j}}(X_{1,n_j}{=}x_{1,n_j})\\ &+\sum_{x_{1,n_k}=1}^{n_k-1}\mathbf{P}_{G_{n_k}}(X_{1,n_k}{=}x_{1,n_k})\sum_{E(n_k)}\prod_{j=1}^{k-1}\mathbf{P}_{G_{n_j}}(X_{1,n_j}{=}x_{1,n_j})\Bigg\}.\end{aligned} \quad (42)$$

We assume $|G_{n_j}|{=}n_j$ $(1{\leq}j{\leq}k)$ and $|G_{n_k}|{=}n_k^{'}{=}n_k{+}1$, and evaluate the change of P_{c} as $|G_{n_k}|$ changes from n_k to $n_k^{'}$.

Here we outline a sketch of the subsequent proof procedure. For any fixed $(x_{1,n_1},\ldots,x_{1,n_{k-1}})$, we can determine the solution set of x_{1,n_k} as well as that of $x_{1,n_k^{'}}$, through (41). For the fixed $(x_{1,n_1},\ldots,x_{1,n_{k-1}})$, we can then evaluate the change in the correct detection probability. Therein, several possible cases need to be considered. Finally, based on (42), we conclude the proof by observing that for each fixed $(x_{1,n_1},\ldots,x_{1,n_{k-1}})$, the change of P_{c} from n_k to $n_k^{'}$ is bounded by zero from above.

In the following, we use X_j $(1\leq j\leq k-1)$, Y_k and $Y_k^{'}$ to denote X_{1,n_j} $(1\leq j\leq k-1)$, X_{1,n_k} and $X_{1,n_k^{'}}$, respectively; furthermore, define

$$\bar{y}_k=\min\Big\{y_k:(x_1,\ldots,x_{k-1},y_k)\in\{D(n_k){\cup}E(n_k)\}\Big\}, \quad (43)$$

$$\bar{y}_k^{'}=\min\Big\{y_k^{'}:(x_1,\ldots,x_{k-1},y_k^{'})\in\{D(n_k^{'}){\cup}E(n_k^{'})\}\Big\}. \quad (44)$$

For arbitrary $(x_1^*,x_2^*,\ldots,x_{k-1}^*)$, take $x_k^*=\bar{y}_k$. If $x_k^*\in E(n_k)$, due to the fact that $\frac{x_k^*+1}{n_k+1}>\frac{x_k^*}{n_k}>\frac{x_k^*}{n_k+1}$, we have $x_k^*+1\in E(n_k^{'})$; but x_k^* may not belong to either $E(n_k^{'})$ or $D(n_k^{'})$. Besides, note that x_k^*-1 cannot belong to $E(n_k^{'})$ or $D(n_k^{'})$, because $x_k^*{-}1$ does not belong to $E(n_k)$ and $\frac{x_k^*}{n_k}>\frac{x_{k-1}^*}{n_k}>\frac{x_k^*-1}{n_k+1}$. Hence, for given $(x_1^*,x_2^*,\ldots,x_{k-1}^*)$, when $x_k^*=\bar{y}_k\in E(n_k)$, we have $\bar{y}_k^{'}=x_k^*{+}1\in E(n_k^{'})$ or $\bar{y}_k^{'}=x_k^*\in E(n_k^{'}){\cup}D(n_k^{'})$. Similarly, if $x_k^*=\bar{y}_k\in D(n_k)$, according to the definition of $D(n_k)$, we have $\bar{y}_k^{'}=x_k^*+1\in E(n_k^{'})$. In summary, for arbitrary $(x_1^*,x_2^*,\ldots,x_{k-1}^*)$, we have four possible cases:

(1) $\bar{y}_k = x_k^* \in E(n_k)$, $\bar{y}_k^{'} = x_k^* + 1 \in E(n_k^{'})$ and $\bar{y}_k^{'} = \bar{y}_k + 1$;
(2) $\bar{y}_k = x_k^* \in E(n_k)$, $\bar{y}_k^{'} = x_k^* \in E(n_k^{'})$ and $\bar{y}_k^{'} = \bar{y}_k$;
(3) $\bar{y}_k = x_k^* \in E(n_k)$, $\bar{y}_k^{'} = x_k^* \in D(n_k^{'})$ and $\bar{y}_k^{'} = \bar{y}_k$;
(4) $\bar{y}_k = x_k^* \in D(n_k)$, $\bar{y}_k^{'} = x_k^* + 1 \in E(n_k^{'})$ and $\bar{y}_k^{'} = \bar{y}_k + 1$.

To proceed, we need the following fact from [6]:

$$\sum_{y_k^{'}=x_k^*+1}^{n_k} \mathbf{P}_{G_{n_k+1}}\big(Y_k^{'} = y_k^{'}\big) = \frac{1+x_k^*(\delta-2)}{2+n_k(\delta-2)}\mathbf{P}_{G_{n_k}}\left(Y_k = x_k^*\right) + \sum_{y_k=x_k^*+1}^{n_k-1} \mathbf{P}_{G_{n_k}}\left(Y_k = y_k\right). \tag{45}$$

We also notice a complementary relationship that, if some $\overrightarrow{x^*} = (x_1^*, \ldots, x_{k-1}^*)$ satisfies Case 1, then $\overleftarrow{x^*} = \big((n_1 - x_1^*), (n_2 - x_2^*), \ldots, (n_{k-1} - x_{k-1}^*)\big)$ has to satisfy Case 2. Hence, in the following we combine Case 1 and Case 2 to evaluate the change of P_{c}.

In Case 1, for $\overrightarrow{x^*}$, we have $\bar{y}_k = x_k^* \in E(n_k)$, $\bar{y}_k^{'} = x_k^* + 1 \in E(n_k^{'})$ and $\bar{y}_k^{'} = \bar{y}_k + 1$. Let $y_0 = n_k + 1 - x_k^*$. From (45), by letting $P_d(\overrightarrow{x^*}) = \big(\sum_{y_k^{'}=y_0}^{n_k} \mathbf{P}_{G_{n_k+1}}(Y_k^{'} = y_k^{'}) - \sum_{y_k=y_0}^{n_k-1} \mathbf{P}_{G_{n_k}}(Y_k = y_k)\big) \cdot \prod_{j=1}^{k-1} \mathbf{P}_{G_{n_j}}(X_j = x_j^*)$, we have

$$P_d(\overrightarrow{x^*}) = \Big(\frac{1+x_k^*(\delta-2)}{2+n_k(\delta-2)} - 1\Big)\mathbf{P}_{G_{n_k}}(Y_k = x_k^*)\prod_{j=1}^{k-1}\mathbf{P}_{G_{n_j}}(X_j = x_j^*). \tag{46}$$

Correspondingly, in Case 2, by letting $P_d(\overleftarrow{x^*}) = \big(\sum_{y_k^{'}=y_0}^{n_k} \mathbf{P}_{G_{n_k+1}}(Y_k^{'} = y_k^{'}) - \sum_{y_k=y_0}^{n_k-1} \mathbf{P}_{G_{n_k}}(Y_k = y_k)\big) \cdot \prod_{j=1}^{k-1} \mathbf{P}_{G_{n_j}}(X_j = x_j^*)$, we have

$$P_d(\overleftarrow{x^*}) = \frac{1+(n_k - x_k^*)(\delta-2)}{2+n_k(\delta-2)}\mathbf{P}_{G_{n_k}}(Y_k = n_k - x_k^*)\prod_{j=1}^{k-1}\mathbf{P}_{G_{n_j}}(X_j = n_j - x_j^*). \tag{47}$$

By adding up (46) and (47), we can show that the total change in Case 1 and Case 2 as n_k increases by one is positive, i.e.,

$$P_d(\overrightarrow{x^*}) + P_d(\overleftarrow{x^*}) > 0,$$

in which we use the fact that $\frac{\mathbf{P}_{G_n}(X=n-x)}{\mathbf{P}_{G_n}(X=x)} = \frac{x}{n-x}$, which can be verified by (37).

Following similar way, for all the four cases we have that for any fixed $n_1, \ldots, n_{k-1}$, the change in the error detection probability, as n_k changes to $n_k + 1$, is lower bounded by zero, i.e., that the correct detection probability is monotonically non-increasing with n_k.

8.5 Theorem 4

a. Theorem 4(1)

For the rumor source s^* with m ($1 \le m \le \delta$) neighbors, i.e., s_1^*, s_2^*, ..., s_m^*, let random variable X_{i,n_j} be the number of nodes in subtree $T_{s_i^*,G_{n_j}}^{s^*}$ of observation G_{n_j} ($1 \le i \le m, 1 \le j \le k$). Note that as $n_j \to \infty$, the limiting marginal distribution of the ratio $X_{1,n_j}/n_j$ converges to a Beta distribution with a density function given by

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{48}$$

where $\alpha = \frac{1}{\delta-2}$, $\beta = \frac{\delta-1}{\delta-2}$. So for k independent observations, $\mathbf{P}_G(E_1)$ is regarded as the cumulative distribution function of the ratios $\frac{X_{1,n_1}}{n_1}, \ldots, \frac{X_{1,n_k}}{n_k}$, given by

$$\lim_{n_1,\ldots n_k \to \infty} \mathbf{P}_G(E_1) = \int\!\!\cdot\!\!\int_{\prod_{j=1}^k \frac{x_j}{1-x_j} \le 1} \frac{\Gamma(\alpha+\beta)^k}{\Gamma(\alpha)^k\Gamma(\beta)^k} \prod_{j=1}^k \Big(x_j^{\alpha-1}(1-x_j)^{\beta-1}\Big)\overline{d_k}, \tag{49}$$

where $\overline{d_k} = \mathrm{d}x_1 \cdots \mathrm{d}x_k$. From Lemma 5, we obtain (18). From Theorem 3, $\phi_k(3) \le \phi_k(\delta) < 1$ ($\delta \ge 3$). Combining with the property that as $k \to \infty$, $\phi_k(3)$ goes to 1, which we will prove next, we have that as k grows without bound, $\phi_k(\delta)$ approaches one for any $\delta \ge 3$.

b. Theorem 4(2)

According to (18), when $\delta = 3$, as $n_1, \ldots n_k \to \infty$,

$$\lim_{n_1,\ldots n_k \to \infty} P_{\mathrm{c}} = 1 - 3 \cdot (1 - \varphi_k(1,2)), \tag{50}$$

where

$$\begin{aligned}
&1 - \varphi_k(1,2) \\
&= 1 - \int\!\!\cdot\!\!\int_{\prod_{j=1}^k \frac{x_j}{1-x_j} \le 1} 2^k \cdot \prod_{j=1}^k (1-x_j)\ \overline{d_k} \\
&= 2^k \int_0^1\!\!\cdot\!\int_0^1 \Big(\int_{\frac{\prod_{j=1}^{k-1}(1-x_j)}{\prod_{j=1}^{k-1}x_j + \prod_{j=1}^{k-1}(1-x_j)}}^1 \prod_{j=1}^k (1-x_j)\mathrm{d}x_k\Big)\overline{d_{k-1}} \\
&= 2^k \int_0^1\!\!\cdot\!\int_0^1 \frac{\prod_{j=1}^{k-1} x_j^2 \prod_{j=1}^{k-1}(1-x_j)}{2\Big(\prod_{j=1}^{k-1}(1-x_j) + \prod_{j=1}^{k-1} x_j\Big)^2}\ \overline{d_{k-1}} \\
&\overset{(a)}{=} 2^{k-2} \int_0^1\!\!\cdot\!\int_0^1 \frac{\prod_{j=1}^{k-1} x_j^2(1-x_j) + \prod_{j=1}^{k-1}(1-x_j)^2 x_j}{\big(\prod_{j=1}^{k-1}(1-x_j) + \prod_{j=1}^{k-1} x_j\big)^2}\ \overline{d_{k-1}} \\
&= 2^{k-2} \int_0^1\!\!\cdot\!\int_0^1 \frac{\prod_{j=1}^{k-1} x_j(1-x_j)}{\prod_{j=1}^{k-1}(1-x_j) + \prod_{j=1}^{k-1} x_j}\ \overline{d_{k-1}},
\end{aligned} \tag{51}$$

where (a) follows from substituting $x_j = 1 - x_j^{'}$, $1 \le j \le k-1$. By applying (51) into (50), we get (19). By rearranging (19),

$$\begin{aligned}
&\lim_{n_1,\ldots n_k \to \infty} P_{\mathrm{c}} \\
&= 1 - 3 \cdot 2^{k-2} \int_0^1\!\!\cdot\!\int_0^1 \frac{\prod_{j=1}^{k-1}\left(x_j \cdot (1-x_j)\right)}{\prod_{j=1}^{k-1} x_j + \prod_{j=1}^{k-1}(1-x_j)}\ \overline{d_{k-1}} \\
&\overset{(b)}{\ge} 1 - 3 \cdot 2^{k-2} \int_0^1\!\!\cdot\!\int_0^1 \frac{\prod_{j=1}^{k-1}\left(x_j \cdot (1-x_j)\right)}{2\sqrt{\prod_{j=1}^{k-1}\left(x_j \cdot (1-x_j)\right)}}\ \overline{d_{k-1}} \\
&\overset{(c)}{=} 1 - 3 \cdot 2^{k-3} \cdot (\frac{\pi}{8})^{k-1} = 1 - \frac{3}{4} \cdot (\frac{\pi}{4})^{k-1},
\end{aligned} \tag{52}$$

where (b) follows from the inequality of arithmetic and geometric means, and (c) follows from $\int_0^1 \sqrt{x(1-x)}\mathrm{d}x = \pi/8$. This leads to the desired result.

9. CONCLUSION

We studied the rumor source detection problem for the SI model with multiple observations. By providing characterization through the interdependency of observations and network topology, we established a number of explicit analytical results for regular tree-type networks. We showed that having multiple independent observations dramatically enhances detectability: even two observations can more than double the detection probability of a single observation. We may thus find the right number of observations needed to provide detectability guarantees. We also showed that for degrees greater than two, the detection probability increases with the number of observations and decreases with the number of infected nodes. The detection probability increases with the degree as well as the number of observations, i.e., richer connectivity and diversity both enhance detection. We provided a unified inference framework based on message passing for tree networks and leveraged it as an effective heuristic for general graphs. In the next step of work, it would be interesting to explore exact detectability results on general graphs. Other future topics also include developing efficient network forensics protocols, and addressing practical issues, e.g., non-unique initial sources, complicated spreading models, and outliers.

10. ACKNOWLEDGMENTS

The research has been supported by the Doctoral Program of Higher Education (SRFDP) and Research Grants Council Earmarked Research Grants (RGC ERG) Joint Research Scheme through Specialized Research Fund 20133402140001, National Natural Science Foundation of China through Grant 61379003, and the Research Grants Council of Hong Kong under Project M-CityU107/13.

11. REFERENCES

[1] A. Ganesh, L. Massoulie, and D. Towsley, "The effect of network topology on the spread of epidemics," in *Proc. of IEEE INFOCOM*, 2005, pp. 1455–1466.

[2] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[3] D. Shah and T. Zaman, "Detecting sources of computer viruses in networks: theory and experiment," in *Proc. of ACM SIGMETRICS*, vol. 38, no. 1, 2010, pp. 203–214.

[4] ——, "Rumors in a network: Who's the culprit?" *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5163–5181, 2011.

[5] ——, "Rumor centrality: a universal source detector," in *Proc. of ACM SIGMETRICS*, vol. 40, no. 1, 2012, pp. 199–210.

[6] W. Dong, W. Zhang, and C. W. Tan, "Rooting out the rumor culprit from suspects," in *Proc. of IEEE ISIT*, 2013, pp. 2671–2675.

[7] W. Luo, W. P. Tay, and M. Leng, "Identifying infection sources and regions in large networks," *IEEE Transactions on Signal Processing*, vol. 61, no. 11, pp. 2850–2865, 2013.

[8] N. Karamchandani and M. Franceschetti, "Rumor source detection under probabilistic sample," in *Proc. of IEEE ISIT*, 2013, pp. 2184–2188.

[9] K. Zhu and L. Ying, "Information source detection in the SIR model: A sample path based approach," in *Proc. of IEEE ITA*, 2013, pp. 1–9.

[10] A. Y. Lokhov, M. Mezard, H. Ohta, and L. Zdeborova, "Inferring the origin of an epidemy with dynamics message-passing algorithm," *arXiv-v1303.5315*, 2013.

[11] W. Luo and W. P. Tay, "Finding an infection source under SIS model," in *Proc. of IEEE ICASSP*, 2013, pp. 2930–2934.

[12] D. J. Watts and S. H. Strogatz, "Collective dynamics of ąősmall-worldąŕnetworks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[13] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

APPENDIX

Proof of Lemma 1

We illustrate the proof for ruling out the hypothetic case where node v has two neighbors satisfying the condition in Lemma 1. Denote the two neighbors by v_1 and v_2, which satisfy $x_{1,n_1}x_{1,n_2}\cdots x_{1,n_k} \geq (n_1-x_{1,n_1})(n_2-x_{1,n_2})\cdots(n_k-x_{1,n_k})$ and $x_{2,n_1}x_{2,n_2}\cdots x_{2,n_k} \geq (n_1-x_{2,n_1})(n_2-x_{2,n_2})\cdots(n_k-x_{2,n_k})$, respectively. Summing up the two inequalities, we have

$$\prod_{j=1}^{k} x_{1,n_j} + \prod_{j=1}^{k} x_{2,n_j} \geq \prod_{j=1}^{k}(n_j - x_{1,n_j}) + \prod_{j=1}^{k}(n_j - x_{2,n_j}). \quad (53)$$

Noting that v_1 and v_2 are neighbors of node v, we have $x_{1,n_j} + x_{2,n_j} \leq n_j - 1$, $\forall j \in [1,k]$. It follows that $x_{1,n_j} < n_j - x_{2,n_j}$ and $x_{2,n_j} < n_j - x_{1,n_j}$, $\forall j \in [1,k]$. Adopting into the left hand side of (53), we obtain

$$\prod_{j=1}^{k} x_{1,n_j} + \prod_{j=1}^{k} x_{2,n_j} < \prod_{j=1}^{k}(n_j - x_{1,n_j}) + \prod_{j=1}^{k}(n_j - x_{2,n_j}),$$

which contradicts (53). Hence it is impossible to have two neighbors of node v satisfying the condition in Lemma 1. Similarly, the possibility of having more than two neighbors can be ruled out, and hence there can be at most a neighbor of node v that satisfies the condition in Lemma 1.

Proof of Lemma 2

If for all $i \in [1,m]$, $\prod_{j=1}^{k} x_{i,n_j} < \prod_{j=1}^{k}(n_j - x_{i,n_j})$, according to Proposition 2, s^* is the unique union rumor center. Therefore, we can make sure to correctly detect s^* as the rumor source. If for some $1 \leq i \leq m$, $\prod_{j=1}^{k} x_{i,n_j} = \prod_{j=1}^{k}(n_j - x_{i,n_j})$, according to Proposition 2 and Lemma 1, there are two union rumor centers. Therefore, the probability to correctly detect s^* as the rumor source is $1/2$. If for some $1 \leq i \leq m$, $\prod_{j=1}^{k} x_{i,n_j} > \prod_{j=1}^{k}(n_j - x_{i,n_j})$, then s^* does not have the maximum union rumor centrality. Therefore s^* is not a union rumor center, and we cannot detect it as the rumor source. Then Lemma 2 follows from summarizing all the correct detection events.

Proof of Lemma 3

Denote $P_e = 1 - P_c$ as the error detection probability. Define $D_j = \left\{ \prod_{i=1}^{k} x_{j,n_i} = \prod_{i=1}^{k}(n_i - x_{j,n_i}) \right\}$, and $F_j = \left\{ \prod_{i=1}^{k} x_{j,n_i} > \prod_{i=1}^{k}(n_i - x_{j,n_i}) \right\}$, $(1 \leq j \leq \delta)$. According to Lemma 1, $D_j, F_j, 1 \leq j \leq \delta$ are all disjoint. Hence, based on

Proposition 2, $P_{\rm e}$ can be evaluated by

$$P_{\rm e}=\frac{1}{2}\mathbf{P}_G\Big(\bigcap_{j=1}^{\delta}D_j\Big)+\mathbf{P}_G\Big(\bigcap_{j=1}^{\delta}F_j\Big)=\delta\left(\frac{1}{2}\mathbf{P}_G\big(D_1\big)+\mathbf{P}_G\big(F_1\big)\right),$$

where the second equality follows from symmetry.

Proof of Lemma 4

Note that the marginal distribution of X_1 is given by

$$\begin{aligned}&\mathbf{P}_G(X_1=x_1)\\&=\frac{(n-1)!}{x_1!x_2!}\cdot\frac{\prod_{i=1}^{2}b_i(b_i+\delta-2)\cdots(b_i+(x_i-1)(\delta-2))}{\delta(\delta+\delta-2)\cdots(\delta+(n-2)(\delta-2))},\end{aligned}\tag{54}$$

where $x_2=n-1-x_1$, $b_1=1$, $b_2=\delta-1$. As $n\to\infty$, we have

$$\begin{aligned}&\mathbf{P}_G(X_1=0)\\&=\frac{(n-1)!}{(n-1)!}\frac{(\delta-1)\cdot(\delta-1+(\delta-2))\cdots(\delta-1+(n-2)(\delta-2))}{\delta\cdot(\delta+(\delta-2))\cdot(\delta+2(\delta-2))\cdots(\delta+(n-2)(\delta-2))}\\&=\frac{\Gamma(\frac{\delta}{\delta-2})}{\Gamma(\frac{\delta-1}{\delta-2})}\cdot\frac{\Gamma(n+\frac{1}{\delta-2})}{\Gamma(n+\frac{2}{\delta-2})}\overset{(a)}{\sim}\frac{\Gamma(\frac{\delta}{\delta-2})}{\Gamma(\frac{\delta-1}{\delta-2})}\cdot\frac{\Gamma(n)n^{\frac{1}{\delta-2}}}{\Gamma(n)n^{\frac{2}{\delta-2}}}\to 0,\end{aligned}$$

where (a) follows from $\lim_{n\to\infty}\frac{\Gamma(n+\alpha)}{\Gamma(n)n^{\alpha}}=1$, $\alpha\in\mathbb{R}$.

Proof of Lemma 5

First, according to Lemma 4, as $n_1,\ldots,n_k\to\infty$, the probability for the rumor source to have less than δ infected neighbors (i.e., empty subtrees) asymptotically vanishes. So we only need to consider the case where the rumor source has $m=\delta$ infected neighbors, and denote by $P_{\rm c,m=\delta}$ the correct detection probability in that case.

According to Proposition 2 and Lemma 2,

$$\begin{aligned}P_{\rm c,m=\delta}\geq&\mathbf{P}_G\Big(\bigcap_{j=1}^{\delta}E_j\Big)=1-\mathbf{P}_G\Big(\bigcup_{j=1}^{\delta}E_j^c\Big)\\&\overset{(a)}{\geq}1-\sum_{j=1}^{\delta}\mathbf{P}_G\big(E_j^c\big)\overset{(b)}{=}1-\delta\cdot\mathbf{P}_G\big(E_1^c\big),\end{aligned}\tag{55}$$

where (a) is the union bound, and (b) follows from symmetry.

Again using Proposition 2 and Lemma 2, we have

$$\begin{aligned}P_{\rm c,m=\delta}\leq&\mathbf{P}_G\Big(\bigcap_{j=1}^{\delta}F_j\Big)=1-\mathbf{P}_G\Big(\bigcup_{j=1}^{\delta}F_j^c\Big)\\&\overset{(c)}{=}1-\sum_{j=1}^{\delta}\mathbf{P}_G\big(F_j^c\big)\overset{(d)}{=}1-\delta\cdot\mathbf{P}_G\big(F_1^c\big),\end{aligned}\tag{56}$$

where (c) follows from that $F_1^c,\ldots,F_\delta^c$ are disjoint (Lemma 1), and (d) follows from symmetry.

Therefore, we have

$$\begin{aligned}&\lim_{n_1,\ldots,n_k\to\infty}P_{\rm c}=\lim_{n_1,\ldots,n_k\to\infty}P_{\rm c,m=\delta}\\&=1-\delta\cdot\lim_{n_1,\ldots,n_k\to\infty}\mathbf{P}_G(E_1^c)=1-\delta\cdot\lim_{n_1,\ldots,n_k\to\infty}\mathbf{P}_G(F_1^c).\end{aligned}$$