Rumors in a Network: Who’s the Culprit?

Tauhid Zaman and Devavrat Shah

September 24, 2009

Abstract

Motivated by applications such as the detection of sources of worms or viruses in computer networks, identification of the origin of infectious diseases, determining the causes of cascading failures in systems such as financial markets, or inferring the leader in a social network, we study the question of inferring the source of a rumor in a network based on the information about rumor infected nodes and the underlying network structure.

We start by proposing a natural, effective model for the spread of the rumor in a network based on the classical SIR model. We obtain an estimator for the rumor source based on the infected nodes and the underlying network structure – it assigns each node a likelihood, which we call the rumor centrality. We show that the node with maximal rumor centrality is indeed the maximum likelihood estimator for regular trees. For general trees, we find the following surprising phase transition: asymptotically in the size of the network, the estimator finds the rumor source with probability 0 if the tree grows like a line and it finds the rumor source with probability strictly greater than 0 if the tree grows at a rate quicker than a line. In a nutshell, our estimator is qualitatively the best possible estimator for these graphs.

Our notion of rumor centrality naturally extends to arbitrary graphs. With extensive simulations, we establish the effectiveness of our notion of rumor centrality. Furthermore, we apply our estimator to identify the most powerful family in the 15th century Florentine elite family marriage network – it indeed finds the correct family (i.e., the Medici) as the power center!

1 Introduction

In the modern world the ubiquity of networks has made us vulnerable to new types of network risks. These network risks arise in many different contexts, but share a common structure: an isolated risk is amplified because it is spread by the network. For example, as we have witnessed in the recent financial crisis, the strong dependencies or ‘network’ between institutions have led to the situation where the failure of one (or few) institution(s) have led to global instabilities. More generally, various forms of social networks allow information and instructions to be disseminated and finding the leader of these networks is of great interest for various purposes – identification of the ‘latent leader’ in a political network, identification of the ‘hidden voice’ in a spy network, or learning the unknown hierarchy of rulers in a historical setup. Finally,
one wishes to identify the source of computer viruses or worms in the Internet and the source of contagious diseases in populations in order to quarantine them.

In essence, all of these situations can be modeled as a rumor spreading through a network. The goal is to find the source of the rumor in these networks in order to control and prevent these network risks based on limited information about the network structure and the ‘rumor infected’ nodes. In this paper, we will provide a systematic study of the question of identifying the rumor source based on the network structure and rumor infected nodes, as well as understand the fundamental limitations on this estimation problem.

1.1 Related Work

Prior work on rumor spreading has primarily focused on viral epidemics in populations. The natural (and somewhat standard) model for viral epidemics is known as the susceptible-infected-recovered or SIR model [1]. In this model, there are three types of nodes: (i) susceptible nodes, capable of being infected; (ii) infected nodes that can spread the virus further; and (iii) recovered nodes that are cured and can no longer become infected. Research in the SIR model has focused on understanding how the structure of the network and rates of infection/cure lead to large epidemics[2],[3]. This motivated various researchers to propose network inference techniques to learn the relevant network parameters [4], [5],[6],[7],[8]. However, there has been little (or no) work done on inferring the source of an epidemic.

The primary reason for the lack of such work is that it is quite challenging. To substantiate this, we briefly describe a closely related (and much simpler) problem of reconstruction on trees [9],[10], or more generally, on graphs [11]. In this problem one node in the graph, call it the root node, starts with a value, say 0 or 1. This information is propagated to its neighbors and their neighbors recursively along a breadth-first-search (BFS) tree of the graph (when the graph is a tree, the BFS tree is the graph). Now each transmission from a node to its neighbor is noisy – a transmitted bit is flipped with a small probability. The question of interest is to estimate or reconstruct the value of the root node, based on the ‘noisy’ information received at nodes that are far away from root. Currently, this problem is well understood only for graphs that are trees or tree-like, after a long history. Now the rumor source identification problem is, in a sense harder, as we wish to identify the location of the source among many nodes based on the infected nodes – clearly a much noisier situation than the reconstruction problem.

1.2 Our Contributions.

In this paper, we provide a systematic study of the question of designing an estimator for the rumor source based on knowledge of the underlying network structure and the rumor infected nodes. To begin, we present a probabilistic model of rumor spreading in a network based on the SIR model. On one hand this is a natural and well studied model for rumor spreading; on the other hand it should be thought of a good starting point to undertake the systematic study of such inference problems.
Following the approach of researchers working on the reconstruction problem and efficient inference algorithm design (i.e. Belief Propagation), we first address the rumor source estimation problem for tree networks. We characterize the maximum likelihood estimator for the rumor source in regular trees. This estimator assigns to each node a likelihood which we call its *rumor centrality*. Rumor centrality strongly depends on the underlying topology of the rumor network as well as the rumor infected nodes. The notion of rumor centrality of a node readily extends to arbitrary tree networks.

For arbitrary trees, we find the following surprising threshold phenomenon about the estimator’s effectiveness. If the number of nodes within a distance $d$ from any node in a tree scales like $d^\alpha$, then for trees with $\alpha = 0$ (i.e. line graphs), the detection probability of our estimator will go to 0 as the network grows in size; but for trees with $\alpha > 0$, the detection probability will always be strictly greater than 0 (uniformly bounded away from 0) irrespective of the network size. In the latter case, we find that estimator error remains finite with probability 1, independent of the network size. In the former case (i.e. $\alpha = 0$), it can be shown that for any estimator the detection probability will go to 0. Thus, our estimator is essentially the optimal for any tree network.

Motivated by these results for trees, we develop a systematic approach to utilize the tree estimator – the rumor centrality – to develop an estimator for general networks. This is possible because in essence, under the SIR model, rumors spread along a (random) sub-tree of the network. We perform extensive simulations to show that this estimator performs extremely well. In addition, we apply our estimator to the 15th century Florentine elite family marriage network and are able to accurately infer the most powerful family in the network – the Medici family.

## 2 Estimator Construction

In this section we start with a description of our rumor spreading model and then we define the maximum likelihood estimator for the rumor source. For regular tree graphs, we equate the maximum likelihood estimator to a novel combinatoric quantity we call *rumor centrality*. We obtain a closed form expression for this quantity. Using rumor centrality, we construct rumor source estimators for general trees and general graphs.

### 2.1 Rumor Spreading Model.

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

We use a variant of the SIR model for the rumor spreading known as the *susceptible-infected* or SI model which does not allow for any nodes to recover, i.e. once a node has the rumor, it keeps it forever. Once a node $i$ has the rumor, it is able to spread it to another node $j$ if and only if there is an edge between them, i.e. if $(i, j) \in E$. The
time for a node $i$ to spread the rumor to node $j$ is modeled by an exponential random variable $\tau_{ij}$ with rate $\lambda$. We assume without loss of generality that $\lambda = 1$. All $\tau_{ij}$’s are independent and identically distributed.

### 2.2 Rumor Source Maximum Likelihood Estimator

We now assume that the rumor has spread in $G(V,E)$ according to our model and that $N$ nodes have the rumor. These nodes are represented by a rumor graph $G_N(V,E)$ which is a subgraph of $G(V,E)$. We will refer to this rumor graph as $G_N$ from here on. The actual rumor source is denoted as $v^*$ and our estimator will be $\hat{v}$. We assume that each node is equally likely to be the source a priori, so the best estimator will be the maximum likelihood estimator. The only data we have available is the final rumor graph $G_N$, so the estimator becomes

$$\hat{v} = \arg \max_{v \in G_N} P(G_N | v^* = v)$$

(1)

In general, $P(G_N | v^* = v)$ will be difficult to evaluate. However, we will show that in regular tree graphs, it can be expressed in a simple closed form.

### 2.3 Rumor Source Estimator for Regular Trees

To simplify our rumor source estimator, we consider the case where the underlying graph is a regular tree where every node has the same degree. In this case, $P(G_N | v^* = v)$ can be exactly evaluated when we observe $G_N$ at the instant when the $N$th node is infected.

First, because of the tree structure of the network, there is a unique sequence of nodes for the rumor to spread to each node in $G_N$. Therefore, to obtain the rumor graph $G_N$, we simply need to construct a permutation of the $N$ nodes subject to the ordering constraints set by the structure of the rumor graph. We will refer to these permutations as permitted permutations. For example, for the network in Figure 1 if node 1 is the source, then $\{1, 2, 4\}$ is a permitted permutation, whereas $\{1, 4, 2\}$ is not because node 2 must have the rumor before node 4.

Second, because of the memoryless property of the rumor spreading time between nodes and the constant degree of all nodes, each permitted permutation resulting in $G_N$ is equally likely. To see this, imagine every node has degree $k$ and we wish to find the probability of a permitted permutation $\sigma$ conditioned on $v^* = v$. A new node can connect to any node with a free edge with equal probability. When it joins, it contributes $k - 2$ new free edges. Therefore, the probability of any $N$ node permitted permutation $\sigma$ for any node $v$ in $G_N$ is

$$P(\sigma | v^* = v) = \frac{1}{k \cdot k + (k - 2) \cdots k + (N - 2)(k - 2)}$$

The probability of obtaining $G_N$ given that $v^* = v$ is obtained by summing the probability of all permitted permutations which result in $G_N$. Because all of the permutations are equally likely, $P(G_N | v^* = v)$ will be proportional to the number of permitted permutations which start with $v$ and result in $G_N$. Because we will find it necessary to count the number of these permutations, we introduce the following definition:
**Definition 1.** Consider a tree $T$. Then $R(v, T)$ is the number of permitted permutations of nodes which start with node $v$ and result in $T$. We refer to $R(v, T)$ as the rumor centrality of node $v$.

With this definition, the likelihood is proportional to $R(v, G_N)$, so we can then rewrite our estimator as

$$\hat{v} = \arg \max_{v \in G_N} P(G_N | v^* = v) = \arg \max_{v \in G_N} R(v, G_N)$$

(2)

Because the maximum likelihood estimator for the rumor source is also the node which maximizes $R(v, G_N)$, we call this term the rumor centrality of the node $v$, and the node which maximizes it the rumor center of the graph.

### 2.4 Rumor Source Estimator for General Trees

To obtain the form of the rumor source estimator in equation (2), we relied on the fact that every permitted permutation was equally likely in a regular tree. However, in a general tree where node degrees may not all be the same, this fact may not hold. This considerably complicates the construction of the maximum likelihood estimator. To avoid this complication, we define the following randomized estimator for general trees. Consider a rumor that has spread on a tree and reached all nodes in the subgraph $G_N$. Then, let the estimate for the rumor source be a random variable $\hat{v}$ with the following distribution.

$$P(\hat{v} = v | G_N) \propto R(v, G_N)$$

(3)

This estimator weighs each node by its rumor centrality. It is not the maximum likelihood estimator as we had for regular trees. However, we will show that this estimator is qualitatively as good as the best possible estimator for general trees.

### 2.5 Rumor Source Estimator for General Graphs

When a rumor spreads in a network, each node receives the rumor from one other node. Therefore, there is a spanning tree corresponding to a rumor graph. If we knew this spanning tree, we could apply the previously developed tree estimators. However, the knowledge of the spanning tree will be unknown in a general graph, complicating the rumor source inference.

To begin constructing a rumor source estimator for a general graph, we first define the set $\mathcal{T}(G_N)$ to be the set of all spanning trees of the rumor graph $G_N$. Then, we can express the likelihood as a sum of likelihoods over all trees in $\mathcal{T}(G_N)$.

$$P(G_N | v = v^*) = \sum_{T \in \mathcal{T}(G_N)} P(T | v^* = v)$$

(4)

We showed that for regular trees every permitted permutation of nodes was equally likely. We now assume this to be true for a general graph. With this assumption, the
Figure 1: Illustration of variables $T^1_2$ and $T^1_7$.

likelihood of any spanning tree $T$ given that the source is $v$ is proportional to its rumor centrality $R(v, T)$. Then the rumor source estimator $\hat{v}$ will be

$$
\hat{v} = \arg\max_{v \in G_N} \left( \sum_{T \in T(G_N)} R(v, T) \right)
$$

We show a practical implementation of this estimator in Section 3.

2.6 Evaluating the Rumor Centrality

The rumor source estimators we have constructed all require us to evaluate the rumor centrality of a tree graph, $R(v, G_N)$. We now show how to evaluate $R(v, G_N)$. To begin, we first define a term which will be of use in our calculations.

**Definition 2.** $T^u_{v_j}$ is the number of nodes in the subtree rooted at node $v_j$, with node $v$ as the source.

To illustrate this definition, a simple example is shown in Figure 1. In this graph, $T^u_2 = 3$ because there are 3 nodes in the subtree with node 2 as the root and node 1 as the source. Similarly, $T^u_7 = 1$ because there is only 1 node in the subtree with node 7 as the root and node 1 as the source.

We now can count the permutations of $G_N$ with $v$ as the source. In the following analysis, we will abuse notation and use $T^u_{v_j}$ to refer to the subtrees and the number of nodes in the subtrees. To begin, we assume $v$ has $k$ neighbors, $v_1, v_2, ..., v_k$. Each of these nodes is the root of a subtree with $T^u_{v_1}, T^u_{v_2}, ..., T^u_{v_k}$ nodes, respectively. Each node in the subtrees can receive the rumor after its respective root has the rumor. We will have $N$ slots in a given permitted permutation, the first of which must be the source node $v$. Then, from the remaining $N - 1$ nodes, we must choose $T^u_{v_1}$ slots for the nodes in the subtree rooted at $v_1$. These nodes can be ordered in $R(v_1, T^u_{v_1})$ different ways. With the remaining $N - 1 - T^u_{v_1}$ nodes, we must choose $T^u_{v_2}$ nodes for the tree rooted at node $v_2$, and these can be ordered $R(v_2, T^u_{v_2})$ ways. We continue this way recursively.
to obtain

\[ R(v, G_N) = \left( \frac{N-1}{T_{v_1}} \right) \left( \frac{N-1 - T_{v_1}}{T_{v_2}} \right) \ldots \]

\[ \left( N - 1 - \sum_{i=1}^{k-1} T_{v_i} \right) \prod_{i=1}^{k} R(v_i, T_{v_i}) \]

\[ = (N-1)! \prod_{i=1}^{k} \frac{R(v_i, T_{v_i})}{T_{v_i}!} \]

Now, to complete the recursion, we expand each of the \( R(v_i, T_{v_i}) \) in terms of the subtrees rooted at the nearest neighbor children of these nodes. To simplify notion, we label the nearest neighbor children of node \( v_i \) with a second subscript, i.e. \( v_{ij} \). We continue this recursion until we reach the leaves of the tree. The leaf subtrees have 1 node and 1 permitted permutation. Therefore, the number of permitted permutations for a given tree \( G_N \) rooted at \( v \) is

\[ R(v, G_N) = (N-1)! \prod_{i=1}^{k} \frac{R(v_i, T_{v_i})}{T_{v_i}!} \]

\[ = (N-1)! \prod_{i=1}^{k} \frac{(T_{v_i} - 1)!}{T_{v_i}!} \prod_{v_{ij} \in T_{v_i}} R(v_{ij}, T_{v_{ij}}) \]

\[ = (N-1)! \prod_{i=1}^{k} \frac{1}{T_{v_i}!} \prod_{v_{ij} \in T_{v_i}} \frac{R(v_{ij}, T_{v_{ij}})}{T_{v_{ij}}!} \]

\[ = N! \prod_{u \in G_N} \frac{1}{T_u} \quad (6) \]

In the last line, we have used the fact that \( T_u = N \). We thus end up with a simple expression for \( R(v, G_N) \) in terms of the size of the subtrees of all nodes in \( G_N \).

### 3 Evaluating the Rumor Source Estimator

In the following sections we present algorithms for evaluating the rumor source estimator for trees and general graphs. For trees, the estimator is the rumor centrality defined earlier. We present a message passing algorithm to evaluate the rumor centrality of all nodes in a tree. Rumor centrality plays an important role in the rumor source estimator for general graphs. We present an algorithm for evaluating the rumor source estimator in a general graph using the rumor centrality algorithm for trees in combination with an algorithm for generating uniformly distributed random spanning trees.

#### 3.1 Trees: A Message Passing Algorithm

In order to find the rumor center of a tree graph of \( N \) nodes \( G_N \), we need to first find the rumor centrality of every node in \( G_N \). To do this we need the size of the subtrees
For all $v$ and $u$ in $G_N$. There are $N^2$ of these subtrees, but we can utilize a local condition of the rumor centrality in order to calculate all the rumor centralities with only $O(N)$ computation. Consider two neighboring nodes $u$ and $v$ in $G_N$. All of their subtrees will be the same size except for those rooted at $u$ and $v$. In fact, there is a special relation between these two subtrees.

\[ T^v_u = N - T^u_v \]  

(7)

For example, in Figure 1, for node 1, $T^1_2$ has 3 nodes, while for node 2, $T^2_1$ has $N - T^1_2$ or 4 nodes. Because of this relation, we can relate the rumor centralities of any two neighboring nodes.

\[ R(u, G_N) = R(v, G_N) \frac{T^v_u}{N - T^u_v} \]  

(8)

This result is the key to our algorithm for calculating the rumor centrality for all nodes in $G_N$. We first select any node $v$ as the source node and calculate the size of all of its subtrees $T^u_v$ and its rumor centrality $R(v, G_N)$. This can be done by having each node $u$ pass two messages up to its parent. The first message is the number of nodes in $u$’s subtree, which we call $t^u_{u\to\text{parent}(u)}$. The second message is the cumulative product of the size of the subtrees of all nodes in $u$’s subtree, which we call $p^u_{u\to\text{parent}(u)}$. The parent node then adds the $t^u_{u\to\text{parent}(u)}$ messages together to obtain the size of their own subtree, and multiply the $p^u_{u\to\text{parent}(u)}$ messages together to obtain their cumulative subtree product. These messages are then passed upward until the source node receives the messages. By multiplying the cumulative subtree products of its children, the source node will obtain its rumor centrality, $R(v, G_N)$. This algorithm will require only $O(N)$ computation.

With the rumor centrality of node $v$, we then evaluate the rumor centrality for the children of $v$ using equation (8). Each node $u$ passes its rumor centrality to its children in a message we define as $r^u_{u\to\text{child}(u)}$. Each node $u$ can calculate its rumor centrality using its parent’s rumor centrality and its own subtree size $T^u_u$. The computational effort of this algorithm is also $O(N)$. Therefore, the overall algorithm obtains the rumor centrality of all $N$ nodes with $O(N)$ computation. The pseudocode for this message passing algorithm is shown for completeness.

### 3.2 General Graphs

For a general graph $G_N$ with $N$ nodes, recall that the rumor source estimator was of the form

\[ \hat{v} = \arg \max_{v \in G_N} \sum_{T \in T(G_N)} R(v, T) \]  

(9)

where $T(G_N)$ was the set of all spanning trees of $G_N$. If we consider the spanning tree $T$ to be a uniformly distributed random variable in the sample space in $T(G_N)$, where each tree has probability $1/|T(G_N)|$, then we can rewrite the sum as an expectation of
Algorithm 1 Rumor Center Message Passing Algorithm

1: Choose a root node $v \in G_N$
2: for $u$ in $G_N$ do
3: if $u$ is a leaf then
4: $t^u_{u-parent(u)} = 1$
5: $p^u_{u-parent(u)} = 1$
6: else
7: if $u$ is source $v$ then
8: $r^v_{down}$
9: else
10: $t^u_{u-parent(u)} = \sum_{j \in \text{children}(u)} t^u_{j-u} + 1$
11: $p^u_{u-parent(u)} = t^u_{u-parent(u)} \prod_{j \in \text{children}(u)} p^u_{j-u}$
12: $r^v_{down}$
13: end if
14: end if
15: end for

the random variable $R(v, T)$ over this uniform distribution.

$$\sum_{T \in \mathcal{T}(G_N)} R(v, T) = |T(G_N)| \sum_{T \in \mathcal{T}(G_N)} \frac{R(v, T)}{|T(G_N)|} = |T(G_N)| \mathbb{E}[R(v, T)]$$

We now need a way to evaluate the above expectation for all nodes in $G_N$. We accomplish this using two algorithms. The first is an algorithm for generating uniformly distributed spanning trees utilizing a random walk on $G_N$ [12]. The second is the previous algorithm for calculating the rumor centrality on a tree.

To generate uniformly distributed random spanning trees, we perform a random walk on $G_N$ in the following manner. The random walk starts at a random node and moves to any of the node’s neighbors with equal probability. This random walk continues this way on $G_N$ until the graph is covered (i.e. until every node is reached).

Once the random walk has covered every node in $G_N$, we obtain a spanning tree with the following construction. We call the first node in the random walk $v_{start}$. For each node $v \in G_N/v_{start}$, we add to the spanning tree the edge $(w, v)$ which corresponds to the first transition into node $v$ in the random walk. For example, consider a random walk on the graph in Figure 2 with the covering random walk node sequence \{1, 2, 4, 2, 1\}. Then the generated tree will consist of edges \{(1, 2), (1, 3), (2, 4)\}, as indicated in the figure. The trees generated by this random walk on $G_N$ will have a uniform distribution and the runtime of this algorithm is given by the cover time of
Figure 2: The random walk \( \{1, 2, 4, 2, 1, 3\} \) generated on the graph \( G_N \) as indicated by the sequence of arrows (left), and the resulting spanning tree (right).

\( G_N \), which for most graphs is \( O(N \log N) \) and for the worst graphs \( O(N^3) \) \cite{12}.

Once we have generated a tree, we use the tree rumor centrality algorithm to calculate the rumor centrality for every node in the tree. We generate many trees and take the average of the rumor centralities for each node. The node with the maximum expected value becomes our estimate of the rumor source. In more detail, if we define the \( i^{th} \) generated tree as \( T_i \), and \( M \) total trees are generated, the our estimator will be

\[
\hat{v} = \arg \max_{v \in G_N} \frac{1}{M} \sum_{i=1}^{M} R(v, T_i)
\]  

(11)

4 Detection Probability: A Threshold Phenomenon

This section examines the behavior of the detection probability of the rumor source estimators for different graph structures. We establish that the asymptotic detection probability has a phase-transition effect: for line graphs it is 0, while for trees with finite growth it is strictly greater than 0.

4.1 Line Graphs: No Detection

We first consider the detection probability for a line graph. This is a regular tree with degree 2, so we use the maximum likelihood estimator for regular trees. We will establish the following result for the performance of the rumor source estimator in a line graph.

**Theorem 1.** Define the event of correct rumor source detection after time \( t \) on a linear graph as \( C_t \). Then the probability of correct detection of the maximum likelihood rumor source estimator, \( P(C_t) \), scales as

\[
P(C_t) = O\left(\frac{1}{\sqrt{t}}\right)
\]
Figure 3: Detection probability for line graphs. The dotted line is a plot of $\sqrt{2/\pi}N^{-1/2}$ and the circles are the empirical detection probability.

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

We generated 1000 rumor graphs per rumor graph size on an underlying linear graph. The detection probability versus the graph size is shown in Figure 3. As can be seen, the detection probability decays as $N^{-1/2}$ as predicted in Theorem 1.

4.2 Proof of Theorem 1

In this section, we present a proof of Theorem 1. The rumor spreading in the line graph is equivalent to 2 independent Poisson processes with rate 1 beginning at the source and spreading in opposite directions. The following theorem, which is proved in the appendix, bounds the number of arrivals in a Poisson process in time $t$.

**Theorem 2.** Consider a Poisson process $N(\cdot)$ with rate 1, and a small positive $\epsilon$. In a time $t$, where $t$ is large, the probability of having less than $t(1-\epsilon)$ arrivals is bounded by

$$P(N(t) \leq t(1-\epsilon)) \leq c(t + \delta)^{1/2}e^{-t\epsilon^2}$$

for some positive $c$ and some small positive $\delta$.

Also, in a time $t$, the probability of having more than $t(1+\epsilon)$ arrivals is bounded by

$$P(N(t) \geq t(1+\epsilon)) \leq e^{-t\epsilon^2}$$

Therefore, with high probability, after a time $t$, for some small $\epsilon$, the number of total nodes $N$, which is the sum of the arrivals in both Poisson processes, will be bounded by

$$2t(1-\epsilon) \leq N \leq 2t(1+\epsilon) \quad (12)$$

If $N$ is fixed, the detection probability can be easily calculated. However, we want the detection probability after a fixed time $t$. Therefore, we define the $C_N$ as the event
of correct detection given \(N\) nodes in the graph. Then we can rewrite \(P(C_t)\) as

\[
P(C_t) = \sum_{N} P(C_N)P(N|t)
\]

\[
\leq \sum_{N=2t(1-\epsilon)} 2t(1+\epsilon) + e^{-t\epsilon^2}(c(t + \delta)^{1/2} + 1)
\]

For large \(t\), we can neglect the exponential term on the right, so the above expression reduces to

\[
P(C_t) \approx 2t(1+\epsilon) \sum_{N=2t(1-\epsilon)} P(C_N)P(N|t)
\]

We now consider \(N\) to be a fixed quantity and evaluate \(P(C_N)\).

Because of the linear structure of the underlying graph, all rumor graphs \(G_N\) with \(N\) nodes are isomorphic (they are all lines on length \(N\)). For any \(G_N\), the estimate for the rumor source \(\hat{v}\) will be the node at the center of the line. The following lemma makes this more precise.

**Lemma 1.** For a linear rumor graph with \(N\) nodes, label nodes a distance \(k\) from one side of the line as \(v_k\). Then, if \(N\) is odd, the rumor source estimator will be node \(v_{(N+1)/2}\). If \(N\) is even, the rumor source estimator is either node \(v_{N/2}\) or node \(v_{N/2+1}\) with equal probability.

To prove this, we first must evaluate the rumor centrality of a node in the line graph. For a node \(v_k\) a distance \(k\) from one end, the rumor centrality is

\[
R(v_k, G_N) = \frac{N!}{k!(N-k)!} \prod_{i=1}^{k-1} \frac{1}{i!(N-i)!}
\]

\[
= \frac{N^k}{k!} \prod_{i=1}^{k-1} \frac{1}{i!} \prod_{j=1}^{N-k} \frac{1}{j!}
\]

\[
= (N-1)! \frac{(N-k)!}{(k-1)!(N-k)!}
\]

\[
= \frac{(N-1)!}{(k-1)!(N-1-(k-1))!}
\]

\[
= \frac{N!}{k!(N'-k')!}
\]

We see that the rumor centrality \(R(v_k, G_N)\) is just the binomial coefficient. It is known that this will be maximized when \(k'\) is \(N'/2\) for even \(N'\), and when \(k'\) is \((N'+1)/2\) or \((N'-1)/2\) for odd \(N'\). In terms of the original labels for the line graph, the rumor centrality is maximized for \(k = (N+1)/2\) for odd \(N\) and for \(k = N/2\) and \(k = N/2 + 1\) for even \(N\). This proves Lemma 1.
Without loss of generality, we now assume that $N$ is odd and that the rumor source estimator $\hat{v}$ is node $v_{(N+1)/2}$. The detection probability $P(C_N)$ will then be equal to the conditional probability that $v^* = v_{(N+1)/2}$ given a graph $G_N$. To evaluate this probability, we express it in terms of the rumor centrality of the nodes.

\[
P(C_N) = \frac{P(v^* = v_{(N+1)/2} | G_N)}{P(G_N)}
\]

\[
= \frac{R(v_{(N+1)/2}, G_N) P(v^* = v_{(N+1)/2})}{\sum_{v \in G_N} R(v, G_N) P(v^* = v)}
\]

\[
= \frac{R(v_{(N+1)/2}, G_N)}{\sum_{v \in G_N} R(v, G_N)}
\]

Now we can evaluate the detection probability.

\[
P(C_N) = \frac{R(v_{(N+1)/2}, G_N)}{\sum_{k=1}^{N} R(v_k, G_N)}
\]

\[
= \left( \frac{(N-1)!}{((N-1)/2)!(N-1)/2)!} \right) \left( \sum_{k=1}^{N} \frac{(N-1)!}{(k-1)!(N-1-(k-1))!} \right)^{-1}
\]

\[
= \left( \frac{N'}{(N'/2)!(N'/2)!} \right) \left( \sum_{k'=0}^{N'} \frac{N'}{k'!(N'-k')!} \right)^{-1}
\]

To simplify the expression above, we use Stirling’s approximation for $N!$,

\[
N! \approx \sqrt{2\pi N} \left( \frac{N}{e} \right)^N
\]

along with the identity

\[
\sum_{k=0}^{N} \frac{N!}{k!(N-k)!} = 2^N
\]
Then, the detection probability becomes

$$P(C_N) \approx 2^{-N'} \frac{\sqrt{2\pi N'} \left( \frac{N'}{e^2} \right)^{N'}}{\left( \sqrt{2\pi N'}/2 \left( \frac{N'}{e^2} \right)^{N'/2} \right)^2}$$

$$\approx 2^{-N'} \frac{\sqrt{2\pi N'} \left( \frac{N'}{e^2} \right)^{N'}}{\pi N' \left( \frac{N'}{e^2} \right)^{N'} 2^{-N'}}$$

$$\approx \sqrt{\frac{2}{\pi N}}$$

$$= O \left( \frac{1}{\sqrt{N}} \right)$$

Now we need to convert this expression from a function of $N$ to a function of $t$. Using equation (13), we obtain

$$P(C_t) \approx \sum_{N=2t(1-\epsilon)}^{2t(1+\epsilon)} P(C_N)P(N|t)$$

$$\approx \sum_{N=2t(1-\epsilon)}^{2t(1+\epsilon)} O \left( \frac{1}{\sqrt{N}} \right) P(N|t)$$

$$\approx O \left( \frac{1}{\sqrt{t}} \right)$$

This completes the proof of Theorem 1.

4.3 Geometric Trees: Non-Trivial Detection

We now consider the detection probability of our estimator in a geometric tree, which is a non-regular tree parameterized by a number $\alpha$. If we let $n(d)$ denote the maximum number of nodes a distance $d$ from any node, then there exist constants $b$ and $c$ such that $b \leq c$ and

$$bd^\alpha \leq n(d) \leq cd^\alpha$$

We use the randomized estimator for geometric trees. For this estimator, we obtain the following result.

Theorem 3. Define the event of correct rumor source detection after time $t$ on a geometric tree with parameter $\alpha > 0$ as $C_t$. Then the probability of correct detection of the randomized rumor source estimator, $P(C_t)$, is strictly greater than 0. That is,

$$\lim_{t} \inf P(C_t) > 0$$

This theorem says that $\alpha = 0$ and $\alpha > 0$ serve as a threshold for non-trivial detection: For $\alpha = 0$, the graph is essentially a linear graph, so we would expect the
detection probability to go to 0 based on Theorem 1. While Theorem 3 only deals with correct detection, one would also be interested in the size of the rumor source estimator error. We obtain the following result for the estimator error.

**Lemma 2.** Define \( d(\hat{v}, v^*) \) as the distance from the rumor source estimator \( \hat{v} \) to the rumor source \( v^* \). Assume a rumor has spread for a time \( t \) on a geometric tree with parameter \( \alpha > 0 \). Then, for any \( \epsilon > 0 \), there exists a \( l \geq 0 \) such that

\[
\liminf_{t} P( d(\hat{v}, v^*) \leq l ) \geq 1 - \epsilon
\]

What this lemma says is that no matter how large the rumor graph becomes, most of the detection probability mass concentrates on a region close to the rumor source \( v^* \).

We generated 1000 instances of rumor graphs per rumor graph size on underlying geometric trees. The \( \alpha \) parameters ranged from 0 to 4. As can be seen in Figure 4, the detection probability remains constant as the tree size grows for strictly positive \( \alpha \) and decays to 0 for \( \alpha = 0 \), as predicted by Theorem 3. Notice that the detection probability for non-zero \( \alpha \) is close to 1. A histogram for the geometric tree with \( \alpha = 1 \) shows that the error is no larger than 4 hops. This indicates that the estimator error remains bounded, in accordance with Lemma 2.

### 4.4 Proof of Theorem 3

In this section we present a proof of Theorem 3. This proof involves 3 steps. First, we show that the rumor graph will have a certain structure with high probability. This allows us to put bounds on \( T_{v^*} \), the sizes of the subtrees with the rumor source as the source node. Then, we express the detection probability in terms of the variables \( T_{v^*} \).

Finally, we show that with this structure for the rumor graphs, the detection probability is bounded away from zero. Throughout we assume that the underlying geometric tree satisfies the property that there exist constants \( b \) and \( c \) such that \( b \leq c \) and the number of nodes a distance \( d \) from any node, \( n(d) \), is bounded by

\[
bd^\alpha \leq n(d) \leq cd^\alpha
\]

(19)
Structure of Rumor Graphs. We wish to understand the structure of a rumor graph on an underlying geometric tree. To do this, we first assume that the rumor has been spreading for a long time $t$. Then, we will formally show that there are two conditions that the rumor graph $G_t$ will satisfy. First, the rumor graph will contain every node within a distance $t(1 - \epsilon)$ of the source node, for some small positive $\epsilon$. Second, there will not be any nodes beyond a distance $t(1 + \epsilon)$ from the source node. Figure 5 shows the basic structure of the rumor graph. It is full up to a distance $t(1 - \epsilon)$ and does not extend beyond $t(1 + \epsilon)$. We now formally state our results for the structure of the rumor graph.

Theorem 4. Consider a geometric tree with parameter $\alpha$ on which a rumor spreads for a long time $t$, and let $\epsilon = t^{-1/2 + \delta}$ for some small $\delta$. Define the resulting rumor graph as $G_t$ and $G_t$ as the set of all rumor graphs which occur after a time $t$ that have the following two properties: every node within a distance $t(1 - \epsilon)$ from the source receives the rumor and there are no nodes with the rumor beyond a distance $t(1 + \epsilon)$ from the source. Then,

$$
\lim_{t \to \infty} P(G_t \in G_t) = 1
$$

(20)

To prove this theorem, we first note that every spreading time is exponentially distributed with an identical parameter, which we assume to be 1 without loss of generality. Then after a time $t$, a node a distance $t(1 - \epsilon)$ from the source having the rumor is equivalent to a Poisson process $N(\cdot)$ with rate 1 having $t(1 - \epsilon)$ arrivals in time $t$. Theorem 2 bounds the number of arrivals in the Poisson process.

Now, we define the following events.

- $E_i =$ Node $i$ which is a distance $t(1 - \epsilon)$ from the source has the rumor
- $F =$ All nodes less than a distance $t(1 - \epsilon)$ from the source have the rumor
- $A_i =$ Node $i$ which is a distance $t(1 + \epsilon)$ from the source has the rumor
- $B =$ All nodes greater than a distance $t(1 + \epsilon)$ from the source do not have the rumor
We begin by proving that all nodes within a distance \( t(1 - \epsilon) \) of the source have the rumor. At a distance \( t(1 - \epsilon) \) there are at most \( c [t(1 - \epsilon)]^\alpha \) nodes for the geometric tree. With this we now apply the union bound to the probability of event \( F \).

\[
P(F) = P \left( \bigcap_{i=1}^{c[t(1-\epsilon)]^\alpha} E_i \right)
\]

\[
= 1 - P \left( \bigcup_{i=1}^{c[t(1-\epsilon)]^\alpha} E_i^c \right)
\]

\[
\geq 1 - \sum_{i=1}^{c[t(1-\epsilon)]^\alpha} P(E_i^c)
\]

\[
\geq 1 - c[t(1-\epsilon)]^\alpha P(E_i^c)
\]

\[
\geq 1 - ct^\alpha P(E_i^c)
\]

Event \( E_i^c \) occurring means a node a distance \( t(1 - \epsilon) \) from the source does not have the rumor. This is equivalent to a Poisson process of rate 1 having less than \( t(1 - \epsilon) \) arrivals in time \( t \). We can use Theorem 2 to lower bound \( P(E_i^c) \).

\[
P(E_i^c) \leq a\sqrt{te^{-te^2}}
\]

Using this bound, we now obtain a lower bound for \( P(F) \).

\[
P(F) \geq 1 - ct^\alpha P(E_i^c)
\]

\[
\geq 1 - act^{\alpha+1/2}e^{-te^2}
\]

We now wish to take the limit as \( t \) approaches infinity. However, the \( \epsilon \) is dependent upon \( t \), so care must be taken. Substituting in the expression for \( \epsilon \) and taking the limit we obtain

\[
\lim_{t \to \infty} P(F) \geq \lim_{t \to \infty} 1 - act^{\alpha+1/2}e^{-te^2}
\]

\[
\geq \lim_{t \to \infty} 1 - act^{\alpha+1/2}e^{-(t^2)}
\]

\[
\geq 1
\]

Now we wish to prove that all nodes beyond a distance \( t(1 + \epsilon) \) from the source do not have the rumor. We will follow a similar procedure as we did for proving the first half of Theorem 4. At a distance \( t(1 + \epsilon) \) there are at most \( c [t(1 + \epsilon)]^\alpha \) nodes for the
geometric tree. With this we now apply the union bound to the probability of event $B$.

$$
P(B) = P \left( \bigcap_{i=1}^{\infty} A_i^c \right)
= 1 - P \left( \bigcup_{i=1}^{\infty} A_i \right)
\geq 1 - \sum_{i=1}^{\infty} P(A_i)
\geq 1 - c \left( t(1+\epsilon) \right)^\alpha P(A_i)
$$

Event $A_i$ occurring means a node a distance $t(1+\epsilon)$ from the source has the rumor. This is equivalent to a Poisson process of rate 1 having more than $t(1+\epsilon)$ arrivals in time $t$. We can use Theorem 2 to lower bound $P(A_i)$.

$$
P(A_i) \leq e^{-t\epsilon^2}
$$

Using this bound, we now obtain an lower bound for $P(B)$.

$$
P(B) \geq 1 - c \left[ t(1+\epsilon) \right]^\alpha P(A_i)
\geq 1 - c \left[ t(1+\epsilon) \right]^\alpha e^{-t\epsilon^2}
$$

We now wish to take the limit as $t$ approaches infinity. Again, we substitute in the expression for $\epsilon$ and take the limit.

$$
\lim_{t \to \infty} P(B) \geq \lim_{t \to \infty} 1 - c \left[ t(1+\epsilon) \right]^\alpha e^{-t\epsilon^2}
\geq \lim_{t \to \infty} 1 - c \left[ t(1+t^{-1/2+\delta}) \right]^\alpha e^{-t^{2\delta}}
\geq 1
$$

This completes the proof of Theorem 4.

**Detection Probability in terms of $T^v_w$.** Our rumor source estimator is a random variable $\hat{v}$ which takes the value $v$ with probability proportional to $R(v,G_t)$. The conditional probability of correct detection given a rumor graph $G_t$ will be the probability of this estimator choosing the source node $v^*$, which is $P(\hat{v} = v^*|G_t)$. We showed that all rumor graphs will belong to the set $G_t$ with probability 1 for large $t$. Therefore,
we lower bound the probability of correct detection \( P(C_t) \) as

\[
\liminf_t P(C_t) = \liminf_t \sum_{G_t} P(\hat{v} = v^*|G_t)P(G_t) \\
\geq \liminf_t \left( \sum_{G_t \in \mathcal{G}_t} P(\hat{v} = v^*|G_t) \right) \\
\liminf_t P(G_t \in \mathcal{G}_t) \\
\geq \liminf_{G_t \in \mathcal{G}_t} P(\hat{v} = v^*|G_t)
\]

We see that the detection probability is lower bounded by the infimum of the conditional detection probability \( P(\hat{v} = v^*|G_t) \) over \( G_t \in \mathcal{G}_t \). Next, we express the detection probability in terms of the size of the subtrees \( T^v_{v_i} \).

\[
\liminf_t P(C_t) \geq \liminf_t \inf_{G_t \in \mathcal{G}_t} P(\hat{v} = v^*|G_t) \\
\geq \liminf_t \inf_{G_t \in \mathcal{G}_t} \frac{R(v^*, G_t)}{R(v, G_t)} \\
\geq \liminf_t \inf_{G_t \in \mathcal{G}_t} \left( \sum_{v \in G_t} \prod_{v_i \in G_t} T^v_{v_i} \right)^{-1} (21)
\]

The structure of rumor graphs in \( \mathcal{G}_t \) will allow us to bound the sizes of subtrees whose source is node \( v^* (T^v_{v_i}) \). Therefore, if we can express \( P(\hat{v} = v^*|G_t) \) in terms of \( T^v_{v_i} \), we will be able to bound the detection probability.

In order to evaluate the detection probability for a general tree, we must relate \( T^v_{v_i} \) to \( T^v_{v^*} \). We have already seen that when node \( v \) is one hop from \( v^* \), all of the subtrees are the same except for those rooted at \( v \) and \( v^* \). In fact, we showed that for a graph with \( N \) total nodes,

\[
T^v_{v} = N - T^v_{v^*} \quad (22)
\]

For a node \( v \) one hop from \( v^* \), the product in equation \( (21) \) becomes

\[
\prod_{v_i \in G_t} \frac{T^v_{v_i}}{T^v_{v_i}} = \frac{T^v_{v^*}}{T^v_{v^*}T^v_{v}} \\
= \frac{T^v_{v^*}}{(N - T^v_{v^*})} \quad (23)
\]
When \( v \) is two hops from \( v^* \), all of the subtrees are the same except for those rooted at \( v \), \( v^* \), and the node in between, which we call node 1. Figure 6 shows an example. In this case, the product in equation 21 becomes

\[
\prod_{v_i \in G_t} T_{v_i}^{v^*} T_{v_i}^{v} T_{v_i}^{v^*}
\]

\[
= T_{v^*}^{v^*} T_{v^*}^{v} T_{v^*}^{v^*} (N - T_{v^*}^{v^*}) (N - T_{v^*}^{v^*})
\]

Continuing this way, we find that in general, for any node \( v \) in \( G_t \),

\[
\prod_{v_i \in G_t} T_{v_i}^{v^*} T_{v_i}^{v} T_{v_i}^{v^*} = \prod_{v_i \in \mathcal{P}(v^*, v)} T_{v_i}^{v^*} (N - T_{v_i}^{v^*})
\]

where \( \mathcal{P}(v^*, v) \) means any node in the path between \( v^* \) and \( v \), not including \( v^* \). The detection probability of the rumor source estimator is then

\[
\liminf_t \mathbb{P}(C_t) \geq \liminf_t \inf_{G_t \in \mathcal{G}_t} \left( 1 + \sum_{v \in G_t \setminus v^*} \prod_{v_i \in \mathcal{P}(v^*, v)} T_{v_i}^{v^*} (N - T_{v_i}^{v^*}) \right)^{-1}
\]

\[
\geq \liminf_t \inf_{G_t \in \mathcal{G}_t} \frac{1}{S}
\]

We call the resulting summation \( S \) and will need to upper bound it in order to get a lower bound on the detection probability.

**Upper Bounding \( S \).** In this section we will show that the sum \( S \) has a finite upper bound. We start with an underlying geometric tree with parameter \( \alpha > 0 \). We then assume we have a rumor graph \( G_t \) with \( N \) nodes which belongs to \( \mathcal{G}_t \). To evaluate the
detection probability, we must upper bound the sum

\[ S = 1 + \sum_{v \in G_t} \prod_{v_i \in \mathcal{P}(v,v)} \frac{T_{v_i}^*}{(N - T_{v_i}^*)} \]  (28)

We know from Theorem 4 that after a time \( t \) the graph will be full up to \( t(1 - \epsilon) \), with \( \epsilon = t^{-1/2 + \delta} \) as before. We will now divide \( G_t \) into two parts as show in Figure 5. The first part is the portion of the graph within a distance \( t(1 - \epsilon) \) from the source and not including the source, and is denoted \( G_0 \). The remaining nodes will form graph \( G_1 \). We can then break the sum \( S \) into two parts.

\[ S = 1 + \sum_{v \in G_0} \prod_{v_i \in \mathcal{P}(v,v)} \frac{T_{v_i}^*}{(N - T_{v_i}^*)} \]

\[ S = 1 + \sum_{v \in G_1} \prod_{v_i \in \mathcal{P}(v,v)} \frac{T_{v_i}^*}{(N - T_{v_i}^*)} + \]

\[ \sum_{v \in G_1} \prod_{v_i \in \mathcal{P}(v,v)} \frac{T_{v_i}^*}{(N - T_{v_i}^*)} \]

\[ S = 1 + S_0 + S_1 \]

First we will upper bound \( S_0 \). To do this, we must first count the number of nodes in \( G_0 \), which we will call \( N_0 \). We know that there are \( d^\alpha \) nodes a distance \( d \) from the source. By summing over \( d \) up to \( t(1 - \epsilon) \) we obtain the following bounds for \( N_0 \).

\[ \sum_{d=1}^{t(1-\epsilon)} bd^\alpha \leq N_0 \leq \sum_{d=1}^{t(1-\epsilon)} cd^\alpha \]

\[ b \frac{[t(1-\epsilon)]^{\alpha+1}}{\alpha+1} \leq N_0 \leq c \frac{[t(1-\epsilon)]^{\alpha+1}}{\alpha+1} \]

\[ N_0^{min} \leq N_0 \leq N_0^{max} \]

We have approximated the sum by an integral, which is valid when \( t \) is large. Now, we must calculate \( N_1 \), the number of nodes in \( G_1 \). To do this, we note that from Theorem 2 there are no nodes beyond a distance \( t(1 + \epsilon) \). Therefore, using the integral approximation again for the sum, we obtain the following bounds for \( N_1 \)

\[ \frac{b \epsilon (\alpha + 1) t^{\alpha+1}}{\alpha+1} \leq N_1 \leq \frac{2 \epsilon (\alpha + 1) t^{\alpha+1}}{\alpha+1} \]

\[ 2b \epsilon t^{\alpha+1} \leq N_1 \leq 2c \epsilon t^{\alpha+1} \]

\[ N_1^{min} \leq N_1 \leq N_1^{max} \]
We used the first order term of the binomial approximation for \((1 \pm \epsilon)^{\alpha + 1}\) above. Now we rewrite \(S_0\) in a more convenient notation.

\[
S_0 = \sum_{v \in G_0} \prod_{v_i \in P(v^*, v)} \frac{T_{v_i}^{v^*}}{(N - T_{v_i}^{v^*})} \tag{29}
\]

\[
= \sum_{v \in G_0} \prod_{v_i \in P(v^*, v)} w_{v_i} \tag{30}
\]

\[
= \sum_{v \in G_0} b_v \tag{31}
\]

Now, to upper bound \(S_0\), we group the \(b_v\) according to the distance of \(v\) from \(v^*\). We denote \(a_d\) as the maximum value of \(b_v\) among the set of nodes a distance \(d\) from the source. Then we can upper bound \(S_0\) as

\[
S_0 \leq \sum_{d=1}^{t(1-\epsilon)} c d^\alpha a_d
\]

Now, to calculate \(a_d\), we first must evaluate the \(w_{v_i}\) term in equation \((30)\). To do this, we consider a node \(v_i \in G_0\) a distance \(i\) from the source. For this node, we upper bound the number of nodes in its subtree by dividing all \(N_0\) nodes in \(G_0\) among the minimum \(b_i^{\alpha}\) nodes a distance \(i\) from the root. Then, to this we add all \(N_1\) nodes in \(G_1\) to get the following upper bound on \(T_{v_i}^{v^*}\)

\[
T_{v_i}^{v^*} \leq \frac{N_0}{b_i^{\alpha}} + N_1
\]

With this, we obtain the following upper bound for \(w_{v_i}\)

\[
w_{v_i} = \frac{T_{v_i}^{v^*}}{N - T_{v_i}^{v^*}} \leq \frac{\frac{N_0}{b_i^{\alpha}} + N_1}{N - \frac{N_0}{b_i^{\alpha}} - N_1} \leq \frac{\frac{N_0}{b_i^{\alpha}} + N_1}{N_0 - \frac{N_0}{b_i^{\alpha}}} \leq \frac{\frac{1}{b_i^{\alpha}} + \frac{N_1}{N_0}}{1 - \frac{1}{b_i^{\alpha}}} \leq \frac{c_i \left( \frac{1}{b_i^{\alpha}} + \frac{2ce(\alpha + 1)}{b(1-\epsilon)^{\alpha+1}} \right)}{1 - \frac{1}{b_i^{\alpha}}}
\]

The constant \(c_i\) is equal to \((1 - 1/b)^{-1}\). Now, we write down an upper bound for \(S_0\),
recalling that \( \epsilon = t^{-1/2+\delta} \).

\[
S_0 \leq \sum_{d=1}^{t(1-\epsilon)} cd^\alpha a_d \\
\leq \sum_{d=1}^{t(1-\epsilon)} cd^\alpha \prod_{i=1}^{d} c_i \left( \frac{1}{c_i^{\alpha}} + \frac{2c\epsilon(\alpha + 1)}{b(1-\epsilon)^{\alpha+1}} \right) \\
\leq \sum_{d=1}^{t(1-t^{-1/2+\delta})} cd^\alpha \prod_{i=1}^{d} c_i \left( \frac{1}{c_i^{\alpha}} + \frac{2c\epsilon^{-1/2+\delta}(\alpha + 1)}{b(1-t^{-1/2+\delta})^{\alpha+1}} \right) \\
\leq \sum_{d=1}^{t(1-t^{-1/2+\delta})} cd^\alpha \prod_{i=1}^{d} c_i \left( \frac{1}{c_i^{\alpha}} + \frac{2c\epsilon^{-1/2+\delta}(\alpha + 1)}{b(1-d^{-1/2+\delta})^{\alpha+1}} \right)
\]

In the last line, we used the fact that \( d \leq t \) to upper bound the product.

We define the terms in the above sum corresponding to a specific value of \( d \) as \( A_d \). Then, we use an infinite sum to upper bound this sum.

\[
S_0 \leq \sum_{d=1}^{t(1-t^{-1/2+\delta})} A_d \\
\leq \sum_{d=1}^{\infty} A_d
\]

If we apply the ratio test to the terms of the infinite sum, we find that

\[
\limsup_d \frac{A_d}{A_{d-1}} = \limsup_d \left( \frac{d}{d-1} \right)^\alpha \\
e_1 \left( \frac{1}{c_i^{\alpha}} + \frac{2c\epsilon^{-1/2+\delta}(\alpha + 1)}{b(1-d^{-1/2+\delta})^{\alpha+1}} \right) = 0
\]

Thus, the infinite sum converges, so \( S_0 \) also converges. Now we only need to show convergence of \( S_1 \).

We upper bound \( S_1 \) in the same way as we did for \( S_0 \). We write the sum as

\[
S_1 = \sum_{v \in G_1} \prod_{v_j \in P(v^*, v)} \frac{T_{v_i}^v}{(N - T_{v_i}^v)} 
\]

\[
= \sum_{v \in G_1} \prod_{v_j \in P(v^*, v)} w_{v_i} \tag{32}
\]

\[
= \sum_{v \in G_1} \prod_{v_j \in P(v^*, v), v_i \in G_0} w_{v_i} \prod_{v_j \in P(v^*, v), v_i \in G_1} w_{v_i} \tag{33}
\]

\[
= \sum_{v \in G_1} \left( \prod_{v_j \in P(v^*, v), v_i \in G_0} w_{v_i} \right) b_v \tag{34}
\]

23
To upper bound \( S_1 \), we group the \( b_v \) according to the distance of \( v \) from the top of \( G_1 \). We denote \( a_d \) as the maximum value of \( b_v \) among the set of nodes a distance \( d \) from the top of \( G_1 \). We also denote the upper bound of the product of \( w_{v_i} \) over nodes in \( P(v^*, v) \) and \( G_0 \) as \( \Gamma \). Then we can upper bound \( S_1 \) as

\[
S_1 \leq \sum_{v \in G_1} \Gamma b_v
\]

\[
S_1 \leq \sum_{d=1}^{2\epsilon t} cd^\alpha \Gamma a_d
\]

Now, to calculate \( a_d \), we upper bound the \( w_{v_i} \) for nodes in \( G_1 \). We assume that every subtree in \( G_1 \) has size \( N_1 \). Then, similar to our procedure for \( S_0 \), we upper bound the weights \( w_{v_i} \) for the nodes in \( G_1 \).

\[
w_{v_i} = \frac{T_{v_i}^*}{N - T_{v_i}^*}
\]

\[
\leq \frac{N_1}{N - N_1}
\]

\[
\leq \frac{N_1}{N_0}
\]

\[
\leq \frac{N_{max}}{N_{min}}
\]

\[
\leq \frac{2\epsilon \alpha + 1}{b(1 - \epsilon)^{\alpha + 1}}
\]

Recalling that \( \epsilon = t^{-1/2+\delta} \), we upper bound \( S_1 \) as

\[
S_1 \leq \sum_{d=1}^{2\epsilon t} cd^\alpha \Gamma a_d
\]

\[
\leq \sum_{d=1}^{2\epsilon t} cd^\alpha \prod_{i=1}^{d} w_{v_i}
\]

\[
\leq \sum_{d=1}^{2\epsilon t} cd^\alpha \Gamma \left( \frac{2\epsilon t^{-1/2+\delta}(\alpha + 1)}{b(1 - t^{-1/2+\delta})^{\alpha+1}} \right)^d
\]

\[
\leq \sum_{d=1}^{2\epsilon t} cd^\alpha \Gamma \left( \frac{2\epsilon d^{-1/2+\delta}(\alpha + 1)}{b(1 - d^{-1/2+\delta})^{\alpha+1}} \right)^d
\]

\[
\leq \sum_{d=1}^{2\epsilon t} B_d
\]

Above we have used the relation that \( d \leq t \). Similar to what was done for \( S_0 \), we
upper bound this sum with an infinite sum.

\[ S_1 \leq \sum_{d=1}^{\infty} B_d \]

\[ \leq \sum_{d=1}^{\infty} B_d \]

If we apply the ratio test to the terms of the infinite sum, we find that

\[ \limsup \frac{B_d}{B_{d-1}} = \limsup \left( \frac{d}{d-1} \right)^\alpha \frac{2cd^{-1/2+\delta}(\alpha + 1)}{b(1 - d^{-1/2+\delta})^{\alpha+1}} = 0 \]

Again, the ratio test proves convergence of the sum \( S_1 \).

We have now shown that the sum \( S = 1 + S_0 + S_1 \) is upper bounded by some finite \( S^\ast \). With this, we can lower bound the detection probability for the geometric tree.

\[ \liminf_t P(C_t) \geq \liminf_t \inf_{G_t \in G} \frac{1}{S} \geq \frac{1}{S^\ast} > 0 \]

This completes the proof of Theorem 3.

4.5 Proof of Lemma 2

We utilize Theorem 3 to prove Lemma 2. First, we rewrite the distribution of the estimator \( \hat{v} \) on a rumor graph \( G_t \) formed after a rumor has spread for a time \( t \).

\[ P(\hat{v} = v) = \frac{R(v, G_t)}{\sum_{v \in G_t} R(v, G_t)} = \frac{R(v, G_t)}{\sum_{v \in G_t} R(v, G_t) / R(v^\ast, G_t)} = \frac{\rho(v, G_t)}{\sum_{v \in G_t} \rho(v, G_t)} \]

where \( \rho(v, G_t) \) is defined as follows using equation 27

\[ \rho(v, G_t) = \prod_{v_i \in P(v^\ast, v)} \frac{T^\ast_{v_i}}{N - T^\ast_{v_i}} \]
We recognize the sum of $\rho(v, G_t)$ over all $v$ in $G_t$ as the sum $S$ which was previously shown to converge to a positive constant $S^*$. Now, let $d(\hat{v}, v^*)$ be the distance between the rumor source estimator and the rumor source. We can write the probability of the estimator error being greater than $l$ hops as

$$P(d(\hat{v}, v^*) > l|G_t) = \frac{\sum_{v : d(v, v^*) > l} \rho(v, G_t)}{\sum_{v \in G_t} \rho(v, G_t)} = \frac{\sum_{v : d(v, v^*) > l} \rho(v, G_t)}{S}$$

We select an $\epsilon > 0$ and define $\epsilon_1 = \epsilon S$. Then, because of the convergence of the sum $S$, there exists an $l \geq 0$ such that

$$\sum_{v : d(v, v^*) > l} \rho(v, G_t) \leq \epsilon_1 \leq \epsilon S$$

Now, using this result along with Theorem 4 we find the limiting behavior of the probability of the error being less than $l$ hops:

$$\lim_{l} \inf_{l} P(d(\hat{v}, v^*) \leq l) = 1 - \lim_{l} \sup_{l} P(d(\hat{v}, v^*) > l)$$

$$= 1 - \lim_{l} \sup_{l} \left( \sum_{G_t \in \mathcal{G}_t} P(d(\hat{v}, v^*) > l|G_t)P(G_t) \right)$$

$$\geq 1 - \lim_{l} \sup_{l} \frac{\sum_{v : d(v, v^*) > l} \rho(v, G_t)}{S} \sup_{l} P(G_t \in \mathcal{G}_t)$$

$$\geq 1 - \lim_{l} \sup_{l} \frac{\epsilon S}{S}$$

$$\geq 1 - \epsilon$$

Thus, for any positive $\epsilon$, there will always be a finite $l$ such that the probability of the estimator being within $l$ hops of the rumor source is greater than $1 - \epsilon$, no matter how large the rumor graph is.

5 Simulation Results for General Graphs

This section provides simulation results for our rumor source estimators on two general graphs: a simulated grid graph and a real network. For the grid graph, several random
Figure 7: Example of a 100 node rumor graph on a grid (top). Detection probability for the grid graph vs. number of nodes $N$ (bottom left). Histogram of the estimator error for a 100 node rumor graph (bottom right).

rumor graph instances were generated on the underlying grid and the statistics of the rumor source estimator were collected. The real network we used is the marriage network of elite families in 15th century Florence. We find that our estimator performs extremely well for both networks.

5.1 Grid Graphs.

Grid graphs are not trees, so we must utilize the general graph rumor source estimator. We generated 100 instances of rumor graphs per rumor graph size on an underlying grid graph. To calculate the expectation value in equation 10, 1000 trees were generated per rumor graph. Figure 7 shows an example of a 100 node rumor graph on a grid. In this case, our estimator was able to find the rumor source exactly. Next is a plot of the detection probability of the estimator versus rumor graph size. We find that for rumor graphs with up to 100 nodes, the detection probability does not go to 0. Finally, we show a histogram of the estimator error for a 100 node rumor graph. As can be seen, we never obtain an error greater than 3 hops. This empirical data indicates that the general graph estimator should have good performance on general graphs.
Figure 8: The 15th century Florentine elite family marriage network. The darkened node is our estimate of the rumor source, which in this case is the Medici family. This family is also the true center of power in this network.

5.2 Florentine Marriage Network: A Future Application

In order to see if our estimator can be applied to situations beyond finding rumor sources, we used it on the marriage network of elite families in 15th century Florence. This is a well known network in the social science literature. The links in this network represent a marriage between families. It is known that the Medici family wielded the most power and so was effectively the center of the network. Even though there was no rumor spreading, our estimator found, rather surprisingly, that the Medici family was the source of this network. This indicates that our estimator may do more than just determine the rumor source. It may also indicate which nodes are important or influential in a network. The Florentine marriage network can be seen in Figure 8.

6 Conclusion and Future Work

We constructed estimators for the rumor source in regular trees, general trees, and general graphs. We defined the maximum likelihood estimator for a regular tree to be a new notion of network centrality which we called rumor centrality. We used rumor centrality as the basis for estimators for general trees and general graphs.

We analyzed the asymptotic behavior of the rumor source estimators for line graphs and geometric trees. For line graphs, it was shown that the detection probability goes to 0 as the network grows in size. However, for geometric trees, it was shown that the estimator detection probability is bounded away from 0 as the graph grows in size. Simulations performed on synthetic graphs agreed with these tree results and also demonstrated that the general graph estimator performed well. The general graph estimator was also able to predict the most powerful family in the 15th century Flo-
rentine elite family marriage network. This indicates that this estimator may be able to find influential nodes in networks in addition to finding rumor sources.

There are several future steps for this work. First, we would like to develop estimators when the spreading times are not identically distributed. Second, we would like to create a message passing algorithm for the general graph estimator in order for it to be applicable to distributed environments. Third, we would like to test our estimators on other real networks to accurately assess their performance.

7 Proof of Theorem 2

To prove the bound for $t(1 - \epsilon)$ arrivals in a Poisson process $N(\cdot)$ of rate 1, we first write down the exact probability of this event

$$P(N(t) \leq t(1 - \epsilon)) = e^{-t} \sum_{i=0}^{t(1-\epsilon)} \frac{t^i}{i!}$$

Next, we upper bound the sum by noting that its terms are monotonically increasing. To see this, we take the ratio of consecutive terms.

$$\frac{t^{i-1}(i)!}{(i-1)!t^{i}} = \frac{i}{t}$$

This ratio is less than 1 if $i < t$, which is true for the sum. Therefore, we upper bound the sum by taking all terms equal to the largest term.

$$P(N(t) \leq t(1 - \epsilon)) \leq e^{-t} \sum_{i=0}^{t(1-\epsilon)} \frac{t^{i(1-\epsilon)}}{(t(1-\epsilon))!} \leq (t(1 - \epsilon) + 1)e^{-t} \frac{t^{t(1-\epsilon)}}{(t(1-\epsilon))!}$$

We apply Stirling’s approximation to the factorial in the denominator to obtain

$$P(N(t) \leq t(1 - \epsilon)) \leq \frac{(t(1 - \epsilon) + 1)e^{-t}t^{t(1-\epsilon)}}{\sqrt{2\pi t(1 - \epsilon)t(1 - \epsilon)t(1-\epsilon)}}$$

$$\leq \sqrt{\frac{1 - \epsilon}{2\pi}} \left(t + \frac{1}{t(1 - \epsilon)^2}\right)e^{-t(\epsilon + (1 - \epsilon)\log(1-\epsilon)}$$

$$\leq a\sqrt{t} + \delta e^{-t(\epsilon + (1 - \epsilon)\log(1-\epsilon)}$$

(36)

where we have defined $a$ and $\delta$ as

$$a = \sqrt{\frac{1 - \epsilon}{2\pi}}$$

$$\delta = \frac{1}{t(1 - \epsilon)^2}$$

29
Now, in order to simplify the exponent, we approximate \( \log(1 - \epsilon) \) as \(-\epsilon\) for small \( \epsilon \).

Inserting this into equation (36) we obtain the first part of Theorem 2:

\[
P(N(t) \leq t(1 - \epsilon)) \leq a\sqrt{t} + \delta e^{-t(1-\epsilon)\epsilon} \leq a\sqrt{t} + \delta e^{-te^2}
\]

To prove the bound on \( t(1 + \epsilon) \), we use the Chernoff bound. For a \( \theta > 0 \), we have

\[
P(N(t) \geq t(1 + \epsilon)) \leq e^{-\theta t(1+\epsilon)} E\left[e^{\theta N(t)}\right]
\]

For a Poisson process, the above expectation is

\[
E\left[e^{\theta N(t)}\right] = e^{t(e^{\theta} - 1)}
\]

We insert this into the Chernoff bound to obtain

\[
P(N(t) \geq t(1 + \epsilon)) \leq e^{-t[\theta(1+\epsilon)+(e^{\theta} - 1)]}
\]

To obtain the tightest possible bound, we maximize the expression inside the brackets in the exponent. The maximum is achieved for \( \theta = \log(1 + \epsilon) \). Using \( \epsilon \) as an approximation for \( \log(1 + \epsilon) \), we obtain the second result of Theorem 2.

\[
P(N(t) \geq t(1 + \epsilon)) \leq e^{-te^2}
\]

References

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

[2] M. E. J. Newman, “The spread of epidemic disease on networks”, *Phys. Rev. E*, 66:016128, (2002).

[3] A. Ganesh, L. Massoulie, and D. Towsley, “The effect of network topology on the spread of epidemics,” *Proc. 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, vol. 2, pp. 1455-1466, (2005).

[4] N. Demiris and P. D. O’Neill, “Bayesian inference for epidemics with two levels of mixing”, *Scandinavian J. of Statistics*, vol. 32, pp. 265 - 280 (2005).

[5] G. Streftaris and G. J. Gibson, “Statistical inference for stochastic epidemic models,” *Proc. 17th international Workshop on Statistical Modeling*, pp. 609-616. (2002).

[6] P. D. O’Neill, “A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods,” *Mathematical Biosciences* vol. 180, pp. 103-114. (2002).
[7] N. Demiris and P. D. O’Neill, “Bayesian inference for stochastic multitype epi-
demics in structured populations via random graphs,” J. Roy. Statist. Soc. B, vol. 67, pp. 731-745. (2005).

[8] H. Okamura, K. Tateishi, and T. Doshi, “Statistical inference of computer virus propagation using non-homogeneous Poisson processes,” Proc. 18th IEEE International Symposium on Software Reliability, vol. 5-9, pp. 149 - 158. (2007).

[9] W. Evans, C. Kenyon, Y. Peres, and L. Schulman, “Broadcasting on trees and the Ising model”, Ann. Appl. Prob., vol. 10, pp. 410-433. (2000).

[10] E. Mossel, “Reconstruction on trees: beating the second eigenvalue”, Ann. Appl. Prob., vol. 11, pp. 285-300. (2001).

[11] A. Gerschenfeld and A. Montanari, “Reconstruction for models on random graphs,” Proc. 48’th IEEE Symp. Found. Comp. Sci. pp. 194-204 (2007).

[12] A. Broder, “Generating random spanning trees”, Proc. 30’th IEEE Symp. Found. Comp. Sci..pp. 442-447. (1989).

[13] J. F. Padgett and C. K. Ansell, “Robust action and the rise of the Medici,1400-1436”, American J. of Sociology, vol 98, pp 1259-1319.