ABSTRACT
We provide a systematic study of the problem of finding the source of a computer virus in a network. We model virus spreading in a network with a variant of the popular SIR model and then construct an estimator for the virus source. This estimator is based upon a novel combinatorial quantity which we term rumor centrality. We establish that this is an ML estimator for a class of graphs. We find the following surprising threshold phenomenon: on trees which grow faster than a line, the estimator always has non-trivial detection probability, whereas on trees that grow like a line, the detection probability will go to 0 as the network grows. Simulations performed on synthetic networks such as the popular small-world and scale-free networks, and on real networks such as the internet AS network and the U.S. electric power grid network, show that the estimator either finds the source exactly or within a few hops in different network topologies. We compare rumor centrality to another common network centrality notion known as distance centrality. We prove that on trees, the rumor center and distance center are equivalent, but on general networks, they may differ. Indeed, simulations show that rumor centrality outperforms distance centrality in finding virus sources in networks which are not tree-like.

1. INTRODUCTION
Imagine a computer virus has spread through a large network, and all that is known about the outbreak is which computers are infected and which computers have been communicating with each other. In this case, is it possible to reliably find the source of the computer virus? At first glance, this problem seems extremely challenging and even impossible. In a large and complex network, how can one find this elusive virus source? However, in this work, we will present a simple algorithm for reliably detecting this source, prove that it works well, and demonstrate its performance with simulations on synthetic and real network topologies.

1.1 Related Work
Prior work on computer virus spreading has utilized models for viral epidemics in populations. The natural (and somewhat standard) model for viral epidemics is known as the susceptible-infected-recovered or SIR model [2]. This model, while initially developed for human viruses, is also commonly used to model the spread of computer viruses [6], [8]. 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 [8], [6], [7], [10]. This motivated various researchers to propose network inference techniques to learn the relevant network parameters [12], [9]. However, there has been little (or no) rigorous work done on inferring the source of a viral epidemic.

The primary reason for the lack of such work in finding epidemic sources is that the problem is quite challenging. It is not clear how to construct the proper estimator given the complexity of the network and knowledge only of which nodes are infected, but not when they were infected. Despite the complexity of inferring the virus source in a network, a simple heuristic is to say that the source is the center of the network. There are many notions of network centrality [5],[4],[11] but a very common one is known as distance centrality. The graph theoretic properties of distance centrality have been extensively studied [11]. However, there has been no rigorous work done to justify distance centrality or any other network centrality as the proper estimator for a virus source in a network.

1.2 Our Contributions
In this paper, we provide a systematic study of the problem of finding the virus source in a network. We construct the virus source estimator in Section 2. We use a simple virus spreading model based upon the SIR model (c.f. Ganesh, et. al., [6]) and then cast finding the virus source as a maximum likelihood (ML) estimation problem. For a general network this seems to be a daunting task, so we begin by addressing the virus source estimation problem for trees. For regular trees, we are able to reduce the ML estimator to a novel combinatorial quantity we call rumor centrality. In principle, rumor centrality involves the sum of an exponential number of terms. However, for trees we find a structural property that allows for a linear time message-passing algorithm for evaluating rumor centrality.\(^1\) We extend this notion of rumor centrality to construct estimators for general trees. For a general graph, we note that there is an underlying tree which corresponds to the first time each node becomes infected. Using this intuition, we develop estimators for general graphs which utilize rumor centrality and breadth-first search (BFS) trees (the idea being that the virus would spread fastest along a tree that is close to the BFS tree).

To understand the estimator performance in terms of its ability to correctly find the virus source, we study its performance on general trees in Section 3. Somewhat surprisingly, we find the following threshold phenomenon about the estimator’s effectiveness. If a tree grows like a line, then the detection probability of the ML virus source estimator will go to 0 as the network grows in size; but for trees growing faster than a line, the detection probability of our estimator will always be strictly greater than 0 (uniformly bounded away from 0) irrespective of the network size. In the latter case, we find that when the estimator makes an error, the wrong prediction is within a few hops of the actual source. Thus, our estimator is essentially the optimal for any tree network. The proofs of these results (found in Section 5) are non-trivial and require novel analytic techniques which may be of general interest in the context of graphical inference and percolation.

We study the performance of the general graph virus source estimator through extensive simulations in Section 4. As representative results, we test the estimator’s performance on the popular small-world and scale-free networks, and also on a real internet autonomous system (AS) network and the U.S. electrical power grid network. Virus spreading on the AS network corresponds to the spread of a computer virus, while virus spreading on the power grid network could instead represent a cascading failure or a blackout. We find that the estimator performs well on all of these different networks.

We compare the new notion of rumor centrality with the more common distance centrality. We show that on trees, the rumor center is equivalent to the distance center. This indicates that distance centrality is the correct estimator for trees and tree-like networks. However, on general networks the rumor center and the distance center can be different. This is because distance centrality only considers the shortest paths in the network, whereas rumor centrality utilizes a richer structure. Through simulations, we find that rumor centrality is a better estimator for the virus source than distance centrality on networks which are not tree-like, such as the small-world and power grid networks.

2. VIRUS SOURCE ESTIMATOR

2.1 Virus 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 commonly used SIR model for the virus 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 virus, it keeps it forever. Once a node \(i\) has the virus, 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 virus 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 Virus Source Maximum Likelihood Estimator

We now assume that the virus has spread in \(G(V, E)\) according to our model and that \(N\) nodes have the virus. These nodes are represented by a virus graph \(G_N(V, E)\) which is a subgraph of \(G(V, E)\). We will refer to this virus graph as \(G_N\) from here on. The actual virus 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 ML estimator. The only data we have available is the final virus graph \(G_N\), so the estimator becomes

\[
\hat{v} = \arg \max_{v \in G_N} \Pr(G_N | v^* = v)
\]

In general, \(\Pr(G_N | v^* = v)\) will be difficult to evaluate. However, we will show that in regular tree graphs, ML estimation is equivalent to a combinatorial problem.

2.3 Virus Source Estimator for Regular Trees

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

Consider for example that all nodes in the network in Figure 1 are infected. If node 1 was the source, then \(\{1, 2, 4\}\) is a permitted infection sequence or permutation, whereas \(\{1, 4, 2\}\) is not because node 2 must have the virus before node 4. In general, to obtain the virus 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 virus graph. We will refer to these permutations as permitted permutations. The likelihood of the virus graph given a source can then be calculated by adding up the probabilities of all permitted permutations which begin with the source.

\(^1\)We note that this message-passing algorithm has no relation to standard Belief Propagation or its variants, other than that it is an iterative algorithm.
We now show how to evaluate rumor centrality. 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 $T_{v_1}^v, T_{v_2}^v, ..., T_{v_k}^v$ nodes, respectively. Each node in the subtrees can receive the virus after its respective root has the virus. 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 $T_{v_1}^v$ slots for the nodes in the subtree rooted at $v_1$. These nodes can be ordered in $R(v_1, T_{v_1}^v)$ different ways. With the remaining $N - 1 - T_{v_1}^v$ nodes, we must choose $T_{v_2}^v$ nodes for the tree rooted at node $v_2$, and these can be ordered $R(v_2, T_{v_2}^v)$ ways. We continue this way recursively to obtain

$$R(v, G_N) = \frac{(N - 1)!}{T_{v_1}^v} \frac{(N - 1 - T_{v_1}^v)!}{T_{v_2}^v} \cdots \frac{(N - 1 - \sum_{i=1}^{k-1} T_{v_i}^v)!}{T_{v_k}^v} \prod_{i=1}^{k} R(v_i, T_{v_i}^v)$$

Now, to complete the recursion, we expand each of the $R(v_i, T_{v_i}^v)$ 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

$$R(v, G_N) = (N - 1)! \prod_{i=1}^{k} \frac{R(v_i, T_{v_i}^v)}{T_{v_i}^v} \prod_{v_{ij} \in T_{v_i}^v} \frac{R(v_{ij}, T_{v_{ij}}^v)}{T_{v_{ij}}^v} = (N - 1)! \prod_{i=1}^{k} \frac{T_{v_i}^v}{T_{v_i}^v} \prod_{v_{ij} \in T_{v_i}^v} \frac{T_{v_{ij}}^v}{T_{v_{ij}}^v} = N! \prod_{v \in G_N} \frac{1}{T_v^v}$$

In the last line, we have used the fact that $T_v^v = N$. We thus end up with a simple expression for rumor centrality in terms of the size of the subtrees of all nodes in $G_N$.

### 2.5 Calculating Rumor Centrality: A Message-Passing Algorithm

In order to find the rumor center of an $N$ node tree $G_N$, we need to first find the rumor centrality of every node in $G_N$. To do this we need the size of the subtree $T_u^v$ 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:

$$T_u^v = N - T_v^u$$

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 $T_{v_1}^v, T_{v_2}^v, ..., T_{v_k}^v$ nodes, respectively. Each node in the subtrees can receive the virus after its respective root has the virus. 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 $T_{v_1}^v$ slots for the nodes in the subtree rooted at $v_1$. These nodes can be ordered in $R(v_1, T_{v_1}^v)$ different ways. With the remaining $N - 1 - T_{v_1}^v$ nodes, we must choose $T_{v_2}^v$ nodes for the tree rooted at node $v_2$, and these can be ordered $R(v_2, T_{v_2}^v)$ ways. We continue this way recursively to obtain

$$R(v, G_N) = \frac{(N - 1)!}{T_{v_1}^v} \frac{(N - 1 - T_{v_1}^v)!}{T_{v_2}^v} \cdots \frac{(N - 1 - \sum_{i=1}^{k-1} T_{v_i}^v)!}{T_{v_k}^v} \prod_{i=1}^{k} R(v_i, T_{v_i}^v)$$

Now, to complete the recursion, we expand each of the $R(v_i, T_{v_i}^v)$ 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

$$R(v, G_N) = (N - 1)! \prod_{i=1}^{k} \frac{R(v_i, T_{v_i}^v)}{T_{v_i}^v} \prod_{v_{ij} \in T_{v_i}^v} \frac{R(v_{ij}, T_{v_{ij}}^v)}{T_{v_{ij}}^v} = (N - 1)! \prod_{i=1}^{k} \frac{T_{v_i}^v}{T_{v_i}^v} \prod_{v_{ij} \in T_{v_i}^v} \frac{T_{v_{ij}}^v}{T_{v_{ij}}^v} = N! \prod_{v \in G_N} \frac{1}{T_v^v}$$

In the last line, we have used the fact that $T_v^v = N$. We thus end up with a simple expression for rumor centrality in terms of the size of the subtrees of all nodes in $G_N$.
For example, in Figure 1, for node 1, $T^*_2$ has 3 nodes, while for node 2, $T^*_2$ has $N - T^*_2$ or 4 nodes. Because of this relation, we can relate the rumor centralities of any two neighboring nodes.

$$R(u,G_N) = R(v,G_N) \frac{T^*_u}{N - T^*_u}$$

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 $T^*_u$ 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 $t_{up-parent(u)}$. The second message is the cumulative product of the size of the subtrees of all nodes in $u$’s subtree, which we call $p_{up-parent(u)}$. The parent node then adds the $t_{up-parent(u)}$ and $p_{up-parent(u)}$ messages together to obtain the size of its own subtree, and multiplies the $p_{up-parent(u)}$ messages together to obtain its 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, $R(v,G_N)$.

With the rumor centrality of node $v$, we then evaluate the rumor centrality for the children of $v$ using equation (5). Each node passes its rumor centrality up to its children in a message we define as $t_{up-child(u)}$. Each node can calculate its rumor centrality using its parent’s rumor centrality and its own subtree size $T^*_u$. We recall that the rumor centrality of a node is the number of permitted permutations that result in $G_N$. Thus, this message-passing algorithm is able to count the (exponential) number of permitted permutations for every node in $G_N$ using only $O(N)$ computations. The pseudocode for this message-passing algorithm is included for completeness.

**Algorithm 1** Rumor Centrality 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: $t_{up-parent(u)} = 1$
5: $p_{up-parent(u)} = 1$
6: else
7: if $u$ is source $v$ then
8: $t_{down-v-child(u)} = \frac{N}{\prod_{j \in \text{children}(u)} p_{j-\text{parent}}}$
9: else
10: $t_{up-parent(u)} = \sum_{j \in \text{children}(u)} t_{up-parent(u)} + 1$
11: $p_{up-parent(u)} = t_{up-parent(u)} \prod_{j \in \text{children}(u)} p_{j-\text{parent}}$
12: $t_{down-v-child(u)} = t_{down-parent(u)-u} \frac{p_{up-parent(u)}}{T^*_u}$
13: end if
14: end if
15: end for

2.6 Virus Source Estimator for General Trees

Rumor centrality is an exact ML virus source estimator for regular trees. In general trees where node degrees may not all be the same, this is no longer the case, as all permitted permutations may not be equally likely. This considerably complicates the construction of the ML estimator. To avoid this complication, we define the following randomized estimator for general trees. Consider a virus that has spread on a tree and reached all nodes in the subgraph $G_N$. Then, let the estimate for the virus source be a random variable $\hat{v}$ with the following distribution.

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

This estimator weighs each node by its rumor centrality. It is not the ML 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.

2.7 Virus Source Estimator for General Graphs

For a general graph, there is an underlying tree which corresponds to the first time each node becomes infected. Therefore, there is a spanning tree corresponding to a virus 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 virus source estimation.

We circumvent the issue of not knowing the underlying spanning tree with the following heuristic. We assume that if node $v \in G_N$ was the source, then it spread the virus along a breadth first search (BFS) tree rooted at $v$, $T_{bf}(v)$. The intuition is that if $v$ was the source, then the BFS tree would correspond to all the closest neighbors of $v$ being infected as soon as possible. With this heuristic, we define the following virus source estimator for a general virus graph $G_N$.

$$\hat{v} = \arg \max \limits_{v \in G_N} R(v,T_{bf}(v))$$

We will show with simulations that this estimator performs well on different network topologies.

2.8 Properties of Rumor Centrality

We now look at some properties of rumor centrality in order to gain an intuition about it.

**Proposition 1.** On an $N$ node tree, if node $v'$ is the rumor center, then any subtree with $v'$ as the source has the following property.

$$T^*_v \leq \frac{N}{2}$$
If there is a node $u$ such that for all $v \neq u$
\[ T^u_v \leq \frac{N}{2} \]  
then $u$ is a rumor center. Furthermore, a tree can have at most 2 rumor centers.

Proof. We showed that for a tree with $N$ total nodes, for any neighboring nodes $u$ and $v$,
\[ T^u_v = N - T^u_u \]  
then the ratios in (11) will all be less than or equal to 1.

For a node $v$ one hop from $v^*$, we find
\[ \frac{R(v, T)}{R(v^*, T)} = \frac{T^v_v T^{v^*}_{v^*}}{(N - T^{v^*}_{v^*})} \]  
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 2 shows an example. In this case, we find
\[ \frac{R(v, T)}{R(v^*, T)} = \frac{T^v_v T^{v^*}_{v^*}}{(N - T^{v^*}_{v^*})} \]  
Continuing this way, we find that in general, for any node $v$ in $T$,
\[ \frac{R(v, T)}{R(v^*, T)} = \frac{\prod_{i \in \mathcal{P}(v^*, v)} T^v_i}{(N - T^v_u)} \]  
where $\mathcal{P}(v^*, v)$ means any node in the path between $v^*$ and $v$, not including $v^*$.

Now imagine that $v^*$ is the rumor center. Then we have
\[ \frac{R(v, T)}{R(v^*, T)} = \frac{\prod_{i \in \mathcal{P}(v^*, v)} T^v_i}{(N - T^v_u)} \leq 1 \]  
For a node $v$ one hop from $v^*$, this gives us that
\[ T^v_u \leq \frac{N}{2} \]  
For any node $u$ in subtree $T^v_u$, we will have $T^v_u \leq T^v_u - 1$. Therefore, (13) will hold for any node $u \in T$. This proves the first part of Proposition 1.

Now assume that the node $v^*$ satisfies (13) for all $v \neq v^*$. Then the ratios in (11) will all be less than or equal to 1. Thus, we have
\[ \frac{R(v, T)}{R(v^*, T)} = \frac{\prod_{i \in \mathcal{P}(v^*, v)} T^v_i}{(N - T^v_u)} \leq 1 \]  
Thus, $v^*$ is the rumor center, as claimed in the second part of Proposition 1.

Finally, assume that $v^*$ is a rumor center and that all of its subtrees satisfy $T^v_\ell < N/2$. Then, any other node $v$ will have at least one subtree that is larger than $N/2$, so $v^*$ is will be the unique rumor center. Now assume that $v^*$ has a neighbor $v$ such that $T^v_u = N/2$. Then, $T^v_\ell = N/2$ also, and all other subtrees $T^v_\ell < N/2$, so $v$ is also a rumor center. There can be at most 2 nodes in a tree with subtrees of size $N/2$, so a tree can have at most 2 rumor centers. □

2.9 Rumor Centrality vs. Distance Centrality
We now wish to compare rumor centrality to another popular type of network centrality known as distance centrality. For a graph $G$, the distance centrality of node $v \in G$, $D(v, G)$, is defined as
\[ D(v, G) = \sum_{j \in v} d(v, j) \]  
where $d(v, j)$ is the shortest path distance from node $v$ to node $j$. The distance center of a graph is the node with the smallest distance centrality. Intuitively, it is the node closest to all other nodes. We will present two important results in this section. First, on a tree, we will show the distance center is equivalent to the rumor center. Thus, we now have the proper justification for distance centrality to be the correct estimator for a virus source on a tree. Second, we will see that in a general network which is not a tree, the rumor center and distance center need not be equivalent.

We will prove the following proposition for the distance center of a tree.

**PROPOSITION 2.** On an $N$ node tree, if $v_D$ is the distance center, then, for all $v \neq v_D$
\[ T^v_{v_D} \leq \frac{N}{2} \]  
Furthermore, if there is a unique rumor center on the tree, then it is equivalent to the distance center.

Proof. Assume that node $v_D$ is the distance center of a tree $T$ which has $N$ nodes. The distance centrality of $v_D$ is less than any other node. We consider a node $v_j$ which is $\ell$ hops from $v_D$, and label a node on the path between $v_j$ and $v_D$ which is $h$ hops from $v_D$ by $v_h$. Now, because we are dealing with a tree, we have the following important property. For a node $j$ which is in subtree $T^v_D$ but not in subtree $T^v_{v_h}$, we have $d(v_j) = d(v_D, j) + d_D - 2h$. Using this, we find
\[ D(v_D, T) \leq D(v_D, T) \leq \sum_{j \in T^v_D} d(v_D, j) \leq \sum_{j \in T} d(v_D, j) \leq \sum_{j \in T} (\ell - 2)(T^v_D - T^v_{v_h}) + ... + (\ell - 2\ell)(T^v_D) \]  
\[ \sum_{h=0}^{\ell} T^v_{v_h} \leq (N - T^v_D) \]  
If we consider a node $v_1$ adjacent to $v_D$, we find the same condition we had for the rumor center. That is,
\[ T^v_D \leq \frac{N}{2} \]  
For any node $u$ in subtree $T^v_D$, we will have $T^v_D \leq T^v_D - 1$. Therefore, (18) will hold for any node $u \in T$. This proves the first half of Proposition 2.

If $v_D$ is a rumor center, then, it also satisfies (18) as previously shown. Thus, when unique, the rumor center is equivalent to the distance center on a tree. This proves the second half of Proposition 2. □
3. MAIN RESULTS: THEORY

This section examines the behavior of the detection probability of the virus 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 which grow faster than a line it is strictly greater than 0.

3.1 Line Graphs: No Detection

We first consider the detection probability for a line graph. We will establish the following result.

**Theorem 1.** Define the event of correct virus source detection after time $t$ on a line graph as $C_t$. Then the probability of correct detection of the ML virus source estimator, $P(C_t)$, scales as

$$P(C_t) = O \left( \frac{1}{\sqrt{t}} \right)$$

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 estimator provides very little information because of the line graph’s trivial structure. This theorem is proved in the appendix.

3.2 Regular Expander Trees: Non-Trivial Detection

We next consider detection on a regular degree expander tree. We assume each node has degree $d > 2$. For $d = 2$, the tree is a line, and we have seen that the detection probability goes to 0 as the network grows in size. For a regular tree with $d > 2$ we obtain the following result.

**Theorem 2.** Define the event of correct virus source detection after time $t$ on a regular expander tree with degree $d > 2$ as $C_t$. Then the probability of correct detection of the ML virus source estimator, $P(C_t)$, is bounded uniformly away from 0. That is,

$$\liminf_t P(C_t) > 0$$

The intuition here is that when $d > 2$, there is enough complexity in the network that allows us to perform non-trivial detection of the virus source. This theorem is proved in Section 5.

3.3 Geometric Trees: Non-Trivial Detection

The previous results cover trees which grow linearly and exponentially. We now consider the detection probability of our estimator in trees which grow polynomially, known as geometric trees. These are non-regular trees parameterized by a number $\alpha$. If we let $n(d)$ denote the maximum number of nodes a distance $d$ from any node in the tree, 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 virus 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 virus source estimator, $P(C_t)$, is bounded uniformly away from 0. That is,

$$\liminf_t P(C_t) > 0$$

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 line graph, so we would expect the detection probability to go to 0 based on Theorem 1, but for $\alpha > 0$, we always have a positive probability of detection. While Theorem 3 only deals with correct detection, one would also be interested in the size of the virus source estimator error. We obtain the following result for the estimator error.

**Corollary 1.** Define $d(\hat{v}, v^*)$ as the distance from the virus source estimator $\hat{v}$ to the virus source $v^*$. Assume a virus has spread for a time $t$ on a geometric tree with parameter $\alpha > 0$. Then, for any $\epsilon > 0$, there exists an $l \geq 0$ such that

$$\liminf_t P(d(\hat{v}, v^*) \leq l) \geq 1 - \epsilon$$

What this corollary says is that no matter how large the virus graph becomes, most of the detection probability mass concentrates on a region close to the virus source $v^*$. Both of these results are proved in Section 5.

4. MAIN RESULTS: EXPERIMENT

We simulated virus propagation on several different network topologies using our simple SI model. For all networks, 1000 virus graphs were generated per virus graph size. The virus source estimator performance is evaluated for these different networks in this section.
Figure 4: Virus source estimator detection probability for regular trees (left) and geometric trees (right) vs. number of nodes $N$, and a histogram of the error for a 100 node geometric tree with $\alpha = 1$ (bottom). The dotted lines are plots of $N^{-1/2}$.

4.1 Tree Networks
The detection probability of the virus source estimator versus the graph size for different trees is shown in Figure 4. As can be seen, the detection probability decays as $N^{-1/2}$ as predicted in Theorem 1 for the graphs which grow like lines ($d = 2$ and $\alpha = 0$). For regular degree trees with $d > 2$ and for geometric trees with $\alpha > 0$, we see that the detection probability does not decay to 0, as predicted by Theorems 2 and 3, and is very close to 1 for the geometric trees.

A histogram for a 100 node virus graph on a geometric tree with $\alpha = 1$ shows that most of the estimator error is less than 1 hop, whereas the average virus graph diameter was 9 hops. This indicates that the estimator error remains bounded, as predicted by Corollary 1.

4.2 General Networks
We performed simulations on synthetic small-world [13] and scale-free [3] networks. These are two very popular models for networks and so we would like our virus source estimator to perform well on these topologies. For both topologies, the underlying graph contained 5000 nodes. Figure 5 shows an example of virus spreading in a small-world and a scale-free network. The graphs show the virus infected nodes in white. Also shown is the histogram of the virus source estimator error and distance centrality estimator error for 400 node virus graphs in each network. Most of the error in these simulations was below 4 hops, while the average virus graph diameter was 22 hops for the small-world and 12 hops for scale-free networks. Thus, we are seeing good performance of the general graph estimator for both small-world and scale-free networks.

The distance centrality estimator performs very similarly to the rumor centrality estimator. However, we see that on the small-world network, rumor centrality is better able to correctly find the source (0 error) than distance centrality (16% correct detection versus 2%). For the scale-free network used here, the average ratio of edges to nodes in the 400 node virus graphs is 1.5 and for the small-world network used here, the average ratio is 2.5. For a tree, the ratio would be 1, so the small-world virus graphs are less tree-like. This may explain why rumor centrality does better than distance centrality at correctly identifying the source on the small-world network. Also, we note that neither estimator had correct detection for the scale-free network. This may be due to this network having many high degree nodes. Our estimator assumes all permutations are equally likely, but this assumption breaks down if some nodes have very high degree. In essence, the estimator is being fooled to always select higher degree nodes. However, by assigning appropriate prior probabilities to each node based upon its degree, we can compensate for the tendency of the estimator to favor higher degree nodes.

4.3 Real Networks
We performed simulations on an internet autonomous system (AS) network [1] and the U.S electric power grid network [13]. These are two important real networks so we would like our virus source estimator to perform well on these topologies. The AS network contained 32,434 nodes and the power grid network contained 4941 nodes. Figure 6 shows an example of virus spreading in both of these net-
networks. Also shown is the histogram of the rumor centrality and distance centrality estimator error for 400 node virus graphs in each network. Most of the error in these simulations was below 4 hops, while the average virus graph diameter was 8 and 17 hops for the AS and the power grid networks, respectively. Thus, we are seeing good performance of the general graph estimator for both of these real networks.

We see that rumor centrality and distance centrality have similar performance, but we see that for the power grid network, rumor centrality is better able to correctly find the source than distance centrality (3% correct detection versus 0%). For the power grid network, the average ratio of edges to nodes in the 400 node virus graphs is 4.2, and for the AS network the average ratio is 1.3. Thus, the virus graphs on the power grid network are less tree-like. Similar to the small-world networks, this may explain why rumor centrality outperforms distance centrality on the power grid network.

5. MAIN RESULTS: PROOFS

5.1 Proof of Theorem 2

In this section we prove the result on detection probability for regular expander trees (Theorem 2). First, we need to know under what conditions we have correct detection. We saw earlier that the rumor center has the property that all other subtrees have less than half of the total nodes. For a degree $d$ regular tree, there are $d$ subtrees connected to the source node. We define the number of nodes in each of these $d$ subtrees at time $t$ as $N_i(t)$. With this definition, we define the event of correct detection $C_t$ as

$$C_t = \left\{ \omega \mid \max_{i, 1 \leq i \leq d} N_i(t, \omega) \leq \frac{1}{2} \sum_{j=1}^{d} N_j(t, \omega) \right\} \quad (20)$$

Now, we must find the distribution for $N_i(t)$. However, this distribution has no closed form, so we instead work with another related process. We define the time between the $(n-1)^{th}$ and $n^{th}$ infections in one of the subtrees as the random variable $T_n$. The root of the subtree is connected to the source by 1 edge, and so its infection time $T_1$ is just an exponential of rate 1. The regularity of the tree means that if there are $n$ nodes in the subtree, then there are $1 + (d - 2)(n - 1)$ edges going out from these nodes which can infect new nodes. Thus, the inter-infection time of the $n^{th}$ node ($T_n$) is the minimum of $1 + (d - 2)(n - 1)$ exponential random variables of rate 1, which is an exponential random variable with rate $1 + (d - 2)(n - 1)$. We define the total time for the $n^{th}$ infection as $S_n$ which is given by

$$S_n = \sum_{i=1}^{n} T_i \quad (21)$$

Because of the complexity of event $C_t$, we define the following sequence of events. Let $D_n(t)$ occur when all the $d$ subtrees have between $n$ and $(d - 1)n$ infected nodes at time $t$. This way no subtree can have more than half of the total nodes. More precisely,

$$D_n(t) = \left\{ \omega \mid \bigcap_{i=1}^{d} n \leq N_i(\omega, t) \leq (d - 1)n \right\} \quad (22)$$

With $D_n(t)$ we now can lower bound $P(C_t)$.

$$P(C_t) \geq P \left( \bigcup_{t=1}^{\infty} D_t(t) \right) \geq P(D_n(t)) \quad \forall n \in \{1, 2, \ldots\} \quad (23)$$

We now show how to bound $P(D_n(t))$. For $D_n(t)$ to occur, it must be that $S_n \leq t$ and that $S_{(d-1)n} \geq t$. Therefore, we can write

$$P(D_n(t)) = P \left( \bigcap_{i=1}^{d} (S_{(d-1)n} \geq t) \right)$$

$$= (1 - P(S_n \geq t) - P(S_{(d-1)n} \leq t))^d$$

$$= \left( \int_{0}^{\infty} f_{S_n}(\tau) - f_{S_{(d-1)n}}(\tau) \ d\tau \right)^d \quad (24)$$

If we can show the above integral to be strictly positive, we will prove Theorem 2. To begin, we first show some properties of the random variable $S_n$ in the following lemma.

**Lemma 1.** The density of $S_n$ for a degree $d$ regular tree, $f_{S_n}(t)$ is given by

$$f_{S_n}(t) = \prod_{i=1}^{n-1} \left( 1 + \frac{1}{it} \right) e^{-t} (1 - e^{-at})^{n-1} \quad (25)$$
where \( a = d - 2 \). Furthermore, let \( t_n = 1/a \log(na - a + 1) \) and \( \tau_n = 1/a \log(n) + 1/a \log(3/4a) \). Then we have that

1. \( df_n(t)/dt > 0 \quad \forall t \in (0, t_n) \)

2. \( \limsup_n f_n(t_n) \leq C_a \) and \( \liminf_n f_n(\tau_n) \geq B_a \) for some finite \( C_a, B_a > 0 \).

3. \( \exists \gamma \in (0, 1) \) such that \( \limsup_n f_n(\tau_n) \leq (1 - \gamma) \quad \forall t \in (0, t_n) \)

Proof. We derive the density by induction. For \( n = 1 \), we have

\[
 f_s(1) = e^{-t} \tag{26}
\]

Now, we do the induction step.

\[
 f_{n+1}(t) = f_n(t) \ast f_{n+1}(t)
\]

where we put all terms not involving \( t \) or \( \tau \) into \( C(n) \). To do the integral, we must expand \((1 - e^{-at})^{n-1}\) and then integrate term by term. The resulting integral is then

\[
 \int_0^t e^{anr}(1 - e^{-at})^{n-1} d\tau = \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i e^{it(n-i) - 1} / a(n-i)
\]

When we combine this with \( C(n) \) we obtain

\[
 f_{n+1}(t) = C(n) e^{-(1+a)n} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i e^{it(n-i) - 1} / a(n-i)
\]

This completes the induction. Next, we show that the density is strictly increasing on \((0, t_n)\). If we take the derivative of the density and set it to be positive, we obtain

\[
 df_n(t)/dt > 0 \quad (e^{-at}(an - a + 1) - 1) > 0 \quad 1/a \log(an - a + 1) > t \quad t_n > t.
\]

Now, we show item 2 of Lemma 1. First, we bound the constant term in front of the distribution.

\[
 \frac{1}{a} \sum_{i=1}^{n-1} \frac{1}{i} \geq \log \left( \prod_{i=1}^{n-1} \left( 1 + \frac{1}{ai} \right) \right) \geq \frac{1}{a} \sum_{i=1}^{n-1} \left( \frac{1}{i} - \frac{1}{n} \right)
\]

\[
 1 + \frac{1}{a} \log(n) \geq \log \left( \prod_{i=1}^{n-1} \left( 1 + \frac{1}{ai} \right) \right) \geq \frac{1}{a} \log(n - 1) - \frac{(2)}{a}
\]

where \( (2) \) is the Riemann zeta function. Then, we can bound \( f_n(\tau_n) \) as

\[
 f_n(\tau_n) \geq (n-1)^{-1/a} e^{-\zeta(2)/a} e^{-1/a \log(n-1) - 1/a \log(3/4a)}
\]

\[
 \left( 1 - e^{-\log(3/4a)} \right)^{n-1}
\]

Therefore,

\[
 \liminf_n f_n(\tau_n) \geq B_a > 0. \tag{27}
\]

We also find that

\[
 f_n(t_n) \leq (n-1)^{-1/a} e^{-\zeta(2)/a} e^{-1/a \log(n-1) - 1/a \log(3/4a)}
\]

\[
 \left( 1 - e^{-\log(3/4a)} \right)^{n-1}
\]

Therefore,

\[
 \limsup_n f_n(t_n) \leq C_a < \infty. \tag{28}
\]

Finally, we establish item 3 of Lemma 1 We take the logarithm of the ratio of the distributions.

\[
 \log \left( \frac{f_n(\tau_n)}{f_n(t_n)} \right) = \sum_{i=0}^{(a-1)n-1} \log \left( 1 + \frac{1}{ai} \right) + na \log(1 - e^{-at})
\]

\[
 \leq \frac{1}{a} \sum_{i=n}^{(a+1)n-1} \frac{1}{i} + na \log(1 - e^{-an})
\]

\[
 \leq \frac{1}{a} \log \left( \frac{(a+1)n-1}{n} \right) + na \log(1 - \frac{1}{an})
\]

\[
 \leq \frac{1}{a} \log (a + 1) - 1 + \frac{1}{2an} \leq \log(1 - \gamma) < 0
\]

Therefore,

\[
 \limsup_n \frac{f_n(\tau_n)}{f_n(t_n)} \leq (1 - \gamma) < 1 \quad \forall t \in (0, t_n) \tag{29}
\]

Now we choose an \( n \) such that \( t_{n-1} \leq t \leq t_n \) and we lower bound the integral in (24).

\[
 \int_0^t (f_n(\tau) - f_n(\tau_n)) d\tau \geq \int_0^t (f_n(\tau) - (1 - \gamma) f_n(\tau)) d\tau
\]

\[
 \geq \gamma \int_{t_n}^{\infty} f_n(\tau) d\tau - \gamma \int_{t_n}^{t_n} f_n(\tau) d\tau
\]

\[
 \geq \gamma f_n(\tau_n) (t_n - \tau_n) - \gamma f_n(\tau_n) (t_n - t_{n-1})
\]

\[
 \geq B_n (1/a \log(4/3)) - \gamma C_n \frac{1}{a} \log \left( \frac{n}{n-1} \right).
\]
For large \( n \), the second term will go to 0, and therefore we have

\[
\liminf_{n} P(C_i) \geq \liminf_{n} P(D_n(t)) \\
\geq (\gamma B_d(1/a \log(4/3)))^d > 0.
\]

This completes the proof of Theorem 2.

### 5.2 Proof of Theorem 3

In this section we present a proof of Theorem 3. This proof involves 3 steps. First, we show that the virus 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 virus 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 virus 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

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

**Structure of Virus Graphs.** We wish to understand the structure of a virus graph on an underlying geometric tree. To do this, we first assume that the virus has been spreading for a long time \( t \). Then, we will formally show that there are two conditions that the virus graph \( G_t \) will satisfy. First, the virus 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 with the virus. Figure 7 shows the basic structure of the virus graph. It is full up to a distance \( t(1-\epsilon) \) and does not extend beyond \( t(1+\epsilon) \). We formally state our results for the structure of the virus graph, which is proved in the appendix.

**Theorem 4.** Consider a geometric tree with parameter \( \alpha \) on which a virus spreads for a long time \( t \), and let \( \epsilon = t^{-1/2+\delta} \) for some small \( \delta \). Define the resulting virus graph as \( G_t \). Also define \( G_\epsilon \) as the set of all virus 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 virus and there are no nodes with the virus beyond a distance \( t(1+\epsilon) \) from the source. Then,

\[
\lim_{t \to \infty} P(G_t \in G_\epsilon) = 1
\]

**Detection Probability in terms of \( T_{v}^{*} \).** Our virus 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 virus 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 virus graphs will belong to the set \( G_\epsilon \), with probability 1 for large \( t \). Therefore, we lower bound the probability of correct detection \( P(C_t) \) as

\[
\liminf_{t} P(C_t) = \liminf_{t} \sum_{G_t} P(\hat{v} = v^* | G_t) P(G_t) \\
\geq \liminf_{t} \inf_{G_t \in G_\epsilon} P(\hat{v} = v^* | G_t)
\]

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

\[
P(C_t) \geq \inf_{G_t \in G_\epsilon} P(\hat{v} = v^* | G_t) \\
\geq \inf_{G_t \in G_\epsilon} \left( \sum_{v \in G_t, v \in P(v^*, v)} \frac{T_{v}^{*}}{N - T_{v}^{*}} \right)^{-1} \\
\geq \inf_{G_t \in G_\epsilon} \frac{1}{S}
\]

Above we used the result from Section 2.8. We call the resulting summation \( S \) and will need to upper bound it in order to get a lower bound on the detection probability. The structure of virus graphs in \( G_\epsilon \) (Theorem 4) will allow us to bound the sizes of the subtrees \( T_{v} \), and thus bound \( S \).

**Upper Bounding \( S \).** To evaluate the detection probability, we must upper bound the sum

\[
S = \sum_{v \in G_\epsilon} \prod_{v \in P(v^*, v)} \frac{T_{v}^{*}}{N - T_{v}^{*}}
\]

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 7. The first part is the portion of the graph within a distance \( t(1-\epsilon) \) from 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.

\[
S = \sum_{v \in G_0} \prod_{v \in P(v^*, v)} \frac{T_{v}^{*}}{N - T_{v}^{*}} + \sum_{v \in G_1} \prod_{v \in P(v^*, v)} \frac{T_{v}^{*}}{N - T_{v}^{*}}
\]

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 \( d^{d} \) nodes a distance \( d \) from the source. By summing over \( d \) up to \( t(1-\epsilon) \) we obtain the following bounds for \( N_0 \).

\[
\frac{t(1-\epsilon)}{d} \leq N_0 \leq \frac{t(1-\epsilon)}{d}
\]

\[
\frac{t(1-\epsilon)}{d+1} \leq N_0 \leq \frac{t(1-\epsilon)}{d+1}
\]

\[
\frac{t(1-\epsilon)}{d+1} \leq N_0 \leq \frac{t(1-\epsilon)}{d+1}
\]
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 5, 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 $N_1$:

$$b_t^{\alpha+1} \left( (1 + \epsilon)^{\alpha+1} - (1 - \epsilon)^{\alpha+1} \right) \leq N_1 \leq 2b_t^{\alpha+1} \left( (1 + \epsilon)^{\alpha+1} - (1 - \epsilon)^{\alpha+1} \right)$$

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.

$$S_0 = \sum_{v \in G_0} \prod_{i \in P(v^*, v)} w_{v_i} = \sum_{v \in G_0} b_v \quad (32)$$

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

$$S_0 \leq \sum_{d=1}^{t(1-\epsilon)} c_d a_d$$

Now, to calculate $a_d$, we first must evaluate the $w_{v_i}$ term in equation (32). 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 $b^\alpha$ 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 $T_{v_i}^*$:

$$T_{v_i}^* \leq \frac{N_0}{b^\alpha} + N_1$$

With this, we obtain the following upper bound for $w_{v_i}$

$$w_{v_i} = \frac{T_{v_i}^*}{N - T_{v_i}^*} \leq \frac{\frac{N_0}{b^\alpha} + N_1}{N - \frac{N_0}{b^\alpha}} \leq \frac{\frac{N_0}{b^\alpha} + N_1}{\frac{N_0}{b^\alpha} - N_1} \leq c_1 \left( \frac{1}{b^\alpha} + \frac{2\epsilon(\alpha + 1)}{b(1 - \epsilon)^{\alpha+1}} \right)$$

The constant $c_1$ is equal to $(1 - 1/b)^{-1}$. Now, we write down an upper bound for $S_0$, recalling that $\epsilon = t^{-1/2+\delta}$.

$$S_0 \leq \sum_{d=1}^{t(1-\epsilon^{-1/2+\delta})} c_d d^{\alpha} \prod_{i=1}^{d} c_1 \left( \frac{1}{b^\alpha} + \frac{2\epsilon(\alpha + 1)}{b(1 - \epsilon)^{\alpha+1}} \right)$$

We have 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.

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

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

$$\limsup_{d} A_d \geq 1 \quad \limsup_{d} (d^{\alpha} \prod_{i=1}^{d} c_1 \left( \frac{1}{b^\alpha} + \frac{2\epsilon(\alpha + 1)}{b(1 - \epsilon)^{\alpha+1}} \right)) = 0$$

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

$$S_1 = \sum_{v \in G_1} \prod_{i \in P(v^*, v)} w_{v_i} = \sum_{v \in G_1} \left( \prod_{i \in P(v^*, v)} w_{v_i} \right) b_v$$

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

$$S_1 \leq \sum_{v \in G_1} \Gamma b_v \leq \sum_{d=1}^{2^{1/2+\delta}} c_d \Gamma a_d$$

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 $G_1$.

$$w_{v_i} = \frac{T_{v_i}^*}{N - T_{v_i}^*} \leq \frac{N_1}{N_0} \leq \frac{2\epsilon(\alpha + 1)}{b(1 - \epsilon)^{\alpha+1}}$$

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

$$S_1 \leq \sum_{d=1}^{2^{1/2+\delta}} c_d \Gamma \left( \frac{2\epsilon(\alpha + 1)}{b(1 - \epsilon)^{\alpha+1}} \right)^d$$

Above we have used the relation that $d \leq t$. We define the terms in the final sum as $B_d$ and as was done for $S_0$, we upper bound this sum with an infinite sum.

$$S_1 \leq \sum_{d=1}^{2^{1/2+\delta}} B_d \leq \sum_{d=1}^{\infty} B_d$$

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

$$\limsup_{d} B_d \geq 1 \quad \limsup_{d} \left( \frac{2\epsilon(\alpha + 1)}{b(1 - \epsilon)^{\alpha+1}} \right) = 0$$

Again, the ratio test proves convergence of the sum $S_1$. We have now shown that the sum $S = S_0 + S_1$ is upper bounded by some finite $S^*$. With this, we can lower bound the detection probability for the geometric tree.

$$\liminf_{t} P(G_t) \geq \liminf_{t} \inf_{G_t \in G_t} \frac{1}{S^*} \geq \frac{1}{S^*} > 0$$

This completes the proof of Theorem 3.
5.3 Proof of Corollary 1
We utilize Theorem 3 to prove Corollary 1. First, we rewrite the distribution of the estimator \( \hat{v} \) on a virus graph \( G_t \) formed after a virus has spread for a time \( t \):

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

where \( \rho(v, G_t) \) is defined as follows using equation 11

\[
\rho(v, G_t) = \prod_{v_i \in P(v^*, v)} \frac{T^*_{v_i}}{N - T^*_{v_i}}
\]

We recognize the sum of \( \rho(v, G_t) \) over all \( v \) in \( G_t \) as the sum \( S \). Now, let \( d(\hat{v}, v^*) \) be the distance between the virus source estimator and the virus source. We can write the probability of the estimator error being greater than \( l \) hops as

\[
P(d(\hat{v}, v^*) > l | G_t) = \left( \sum_{v \in G_t} \rho(v, G_t) \right)^{-1} \sum_{v \in G_t} \rho(v, G_t)
\]

\[
= S^{-1} \sum_{v \in G_t} \rho(v, G_t)
\]

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

\[
\sum_{v \in G_t} \rho(v, G_t) \leq \epsilon_1 \leq \epsilon S
\]

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

\[
\liminf_l P(d(\hat{v}, v^*) \leq l) = 1 - \limsup_l P(d(\hat{v}, v^*) > l)
\]

\[
= 1 - \limsup_l \sum_{G_t \in \mathcal{G}_t} P(d(\hat{v}, v^*) > l | G_t) P(G_t)
\]

\[
\geq 1 - \limsup_l S^{-1} \sum_{v \in G_t} \rho(v, G_t)
\]

\[
\geq 1 - \limsup_l \frac{cS}{S} \geq 1 - \epsilon
\]

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 virus source is greater than \( 1 - \epsilon \), no matter how large the virus graph is.

6. CONCLUSION AND FUTURE WORK
This paper has provided, to the best of the authors’ knowledge, the first systematic study of the problem of finding virus sources in networks. Using the well known SIR model, we constructed an estimator for the virus source in regular trees, general trees, and general graphs. We defined the ML estimator for a regular tree to be a new notion of network centrality which we called rumor centrality and used this as the basis for estimators for general trees and general graphs.

We analyzed the asymptotic behavior of the virus source estimator for regular trees and geometric trees. For line graphs, it was shown that the detection probability goes to 0 as the network grows in size. However, for trees which grew faster than lines, it was shown that there was always non-trivial detection probability and that for geometric trees the estimator error was bounded. Simulations performed on synthetic graphs agreed with these tree results and also demonstrated that the general graph estimator performed well in different network topologies, both synthetic (small-world, scale-free) and real (AS, power grid).

On trees, we showed that the rumor center is equivalent to the distance center. However, these were not equivalent in a general network. Also, it was seen that in networks which are not tree-like, rumor centrality is a better virus source estimator than distance centrality.

The next step of this work would be to refine the general graph estimator by choosing appropriate prior probabilities for the nodes in order to compensate for the fact that the node permutations have different probabilities. This would improve the performance of the estimator on networks which have nodes with very high degree, such as scale-free networks.

7. REFERENCES
APPENDIX

A. PROOF OF THEOREM 1

In this section, we present a proof of Theorem 1. The virus spreading in the line graph is equivalent to 2 independent Poisson processes with rate 1 beginning at the source and spreading in opposite directions. We refer to these processes as $N_1(t)$ and $N_2(t)$.

The rumor center of a line will be the center of the line. This can be seen from the results of Section 2.8. Thus, we will correctly detect the source if the two Poisson processes on each side of the source have exactly the same number of arrivals. If we define the event of correct detection at time $t$ as $C_t$, then we can write $P(C_t)$ as

$$P(C_t) = P(N_1(t) = N_2(t))$$

$$= \sum_{k=0}^{\infty} \left( \frac{e^{-t} t^k}{k!} \right)^2 = \sum_{k=0}^{\infty} a_k$$

We first look at the ratio of consecutive terms in this sum.

$$\frac{a_k}{a_{k-1}} = \left( \frac{t}{k} \right)^2$$

This ratio will be greater than 1 until $t = k$, after which it is less than 1. Therefore, we see that the terms increase up to $k = t$ and then decrease. Next we look at the value of this term $a_t$. We take the logarithm of $a_t$ and apply Stirling’s approximation to obtain

$$\log(a_t) = -2t + 2t \log(t) - 2 \log(t!)$$

$$\approx -2t + 2t \log(t) - 2t \log(t) + 2t - \log(t)$$

$$\approx -\log(t)$$

Therefore, $a_t \approx t^{-1}$. Now, to approximate the sum, we divide it into groups of length $\sqrt{t}$. We will calculate the sum of terms within each block using a rectangular approximation. If we pick a block which starts at $t + \sqrt{t}$ and extends to $t + (l+1)\sqrt{t}$ for some integer $l$, we obtain

$$\sum_{k=t+\sqrt{t}}^{t+(l+1)\sqrt{t}} a_k \leq \sqrt{t} a_{t+1+\sqrt{t}}$$

We must now find a simple expression for $a_{t+1+\sqrt{t}}$. These are terms which come after $a_t$. We relate these terms to $a_t$.

$$\frac{a_{t+1+\sqrt{t}}}{a_t} = \left( \frac{\sqrt{t}}{\prod_{j=1}^{\sqrt{t}}(t+j)} \right)^2$$

$$= \frac{\sqrt{t}}{\prod_{j=1}^{l} (1+j/t)}$$

$$\approx 1 + \frac{\sqrt{t}}{t 2^{l/4}} \approx e^{-l^2}$$

For terms that occur before $a_t$, a similar analysis yields

$$\frac{a_{t-1-l\sqrt{t}}}{a_t} \approx e^{-l^2}$$

Now, using these approximations, we can calculate our desired sum.

$$\sum_{k=1}^{\infty} a_k \leq \sqrt{t} a_t \sum_{l=1}^{\sqrt{t}} e^{-l^2} + \sqrt{t} a_t \sum_{l=1}^{\infty} e^{-l^2} \leq \frac{A}{\sqrt{t}}$$

for some constant $A$. A similar technique can be used to lower bound the sum, where each block’s height is chosen to be the minimum value of $a_k$ in the block instead of the maximum value. In this case, one finds that $P(C_t) \geq B/\sqrt{t}$ for some constant $B < A$. Thus, we see that $P(C_t) = O(1/\sqrt{t})$. This complete the proof of Theorem 1.

B. PROOF OF THEOREM 4

We recall that Theorem 4 stated that the virus graph on a geometric tree is full up to a distance $t(1-\epsilon)$ and does not extend beyond $t(1+\epsilon)$, for $\epsilon = t^{-1/2+\delta}$ for some positive $\delta$. The following useful theorem, which is proved in the appendix, bounds the number of arrivals in a Poisson process in time $t$.

**Theorem 5.** Consider a Poisson process $N(t)$ 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

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

for some positive $c$ and some small positive $\delta t$. Also, in a time $t$, the probability of having more than $t(1+\epsilon)$ arrivals is bounded by

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

To prove Theorem 4, 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 5 bounds the number of arrivals in the Poisson process.

Now, we define the following events.

- $E_1 = \text{Node } i \text{ which is a distance } t(1-\epsilon) \text{ from the source has the rumor}$
- $F = \text{All nodes less than a distance } t(1-\epsilon) \text{ from the source have the rumor}$
- $A_1 = \text{Node } i \text{ which is a distance } t(1+\epsilon) \text{ from the source has the rumor}$
- $B = \text{All nodes greater than a distance } t(1+\epsilon) \text{ from the source do not have the rumor}$

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 $F$.

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

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

$$\geq 1 - c\alpha P(E_i^c)$$
Event $E_i$ 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 5 to lower bound $P(E_i)$.

$$P(E_i) \leq a\sqrt{t}e^{-t^2/2}$$

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

$$P(F) \geq 1 - act^{\alpha+1/2}e^{-t^2/2}$$

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

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

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 $B$.

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

$$\geq 1 - \sum_{i=1}^{c(t+1)\alpha} P(A_i)$$

$$\geq 1 - c[t(1 + \epsilon)]^{\alpha} P(A_i)$$

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 5 to lower bound $P(A_i)$.

$$P(A_i) \leq e^{-t^{1/2}}$$

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

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

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

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

This completes the proof of Theorem 4.

**C. PROOF OF THEOREM 5**

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

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

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

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

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.

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

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

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

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

$$a = \sqrt{1 - \epsilon}$$

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

Now, in order to simplify the exponent, we approximate $\log(1 - \epsilon)$ as $-\epsilon$ for small $\epsilon$. Inserting this into equation (33) we obtain the first part of Theorem 5.

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

$$\leq a\sqrt{1 + \delta}e^{-t(1 - \epsilon)^{2}}$$

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

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

For a Poisson process, the above expectation is

$$E[e^{\theta N(t)}] = e^{\theta e^{(1 + \epsilon)}}$$

We insert this into the Chernoff bound to obtain

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

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 5.

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