Detecting Sources of Computer Viruses in Networks: Theory and Experiment

Devavrat Shah
Massachusetts Institute of Technology
Cambridge, Massachusetts
USA
devavrat@mit.edu

Tauhid Zaman
Massachusetts Institute of Technology
Cambridge, Massachusetts
USA
zlisto@mit.edu

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 an 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.

Categories and Subject Descriptors

G.2.2 [**Graph Theory**]: Network problems, Graph algorithms; G.2.1 [**Combinatorics**]: Combinatorial algorithms, Counting problems, Permutations and combinations

General Terms

Theory, Performance, Security, Algorithms

Keywords

Epidemics, Estimation

SIGMETRICS'10, June 14–18, 2010, New York, New York, USA.

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 prob-

lem 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 it being able 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.

[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.

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 τ_{ij} with rate λ. We assume without loss of generality that $\lambda = 1$. All τ_{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 $\widehat{v}$. 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

$$\widehat{v} = \arg \max_{v \in G_N} \mathbf{P}(G_N | v^* = v) \tag{1}$$

In general, $\mathbf{P}(G_N|v^* = v)$ 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, $\mathbf{P}(G_N|v^* = v)$ 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.

In general, these permutations have different probabilities. However, we find that on a regular tree, they are all equally likely. This is because of the memoryless property of the virus spreading time between nodes and the constant degree

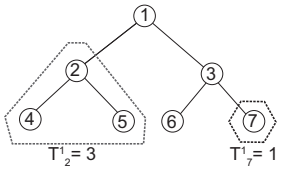

Figure 1: Illustration of variables T_2^1 and T_7^1.

of all nodes. To see this, imagine every node in a regular tree has degree k and we wish to find the probability of a permitted permutation σ 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 σ for any node v in G_N is

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

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. All of the permutations are equally likely, so $\mathbf{P}(G_N|v^* = v)$ will be proportional to the number of permitted permutations which start with v and result in G_N. Thus, we need to count the number of these permutations, which we now define:

DEFINITION 1. *R(v,T) is the number of permitted permutations of nodes that result in a tree T and begin with node $v \in T$.*

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

$$\widehat{v} = \arg\max_{v \in G_N} \mathbf{P}(G_N|v^* = v) = \arg\max_{v \in G_N} R(v, G_N) \qquad (2)$$

$R(v, G_N)$ counts the number of ways a virus, or more generally a rumor or any information, can spread. Therefore, we call $R(v, G_N)$ the **rumor centrality** of the node v, and refer to the node which maximizes it as the **rumor center** of the graph.

2.4 Evaluating Rumor Centrality

We now show how to evaluate rumor centrality. To begin, we first define a term which will be of use in our calculations.

DEFINITION 2. *$T_{v_j}^v$ 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, $T_2^1 = 3$ because there are 3 nodes in the subtree with node 2 as the root and node 1 as the source. Similarly, $T_7^1 = 1$ 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 $T_{v_j}^v$ 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 $T_{v_1}^v, T_{v_2}^v, ..., T_{v_k}^v$ 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

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

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

$$\begin{aligned} R(v, G_N) &= (N-1)!\prod_{i=1}^{k}\frac{R(v_i, T_{v_i}^v)}{T_{v_i}^v!} \\ &= (N-1)!\prod_{i=1}^{k}\frac{(T_{v_i}^v-1)!}{T_{v_i}^v!}\prod_{v_{ij}\in T_{v_i}^v}\frac{R(v_{ij}, T_{v_{ij}}^v)}{T_{v_{ij}}^v!} \\ &= (N-1)!\prod_{i=1}^{k}\frac{1}{T_{v_i}^v}\prod_{v_{ij}\in T_{v_i}^v}\frac{R(v_{ij}, T_{v_{ij}}^v)}{T_{v_{ij}}^v!} \\ &= N!\prod_{u\in G_N}\frac{1}{T_u^v} \qquad (3) \end{aligned}$$

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 subtrees 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.

$$T_u^v = N - T_v^u \qquad (4)$$

For example, in Figure 1, for node 1, T_2^1 has 3 nodes, while for node 2, T_1^2 has $N - T_2^1$ or 4 nodes. Because of this relation, we can relate the rumor centralities of any two neighboring nodes.

$$R(u, G_N) = R(v, G_N)\frac{T_u^v}{N - T_u^v} \qquad (5)$$

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 $t^{up}_{u\to parent(u)}$. The second message is the cumulative product of the size of the subtrees of all nodes in u's subtree, which we call $p^{up}_{u\to parent(u)}$. The parent node then adds the $t^{up}_{u\to parent(u)}$ messages together to obtain the size of its own subtree, and multiplies the $p^{up}_{u\to parent(u)}$ 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 to its children in a message we define as $r^{down}_{u\to child(u)}$. Each node u can calculate its rumor centrality using its parent's rumor centrality and its own subtree size T_u^v. 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

Choose a root node $v \in G_N$
for u in G_N **do**
  if u is a leaf **then**
    $t^{up}_{u\to parent(u)} = 1$
    $p^{up}_{u\to parent(u)} = 1$
  else
    if u is source v **then**
      $r^{down}_{v\to child(v)} = \frac{N!}{N \prod_{j\in children(v)} p^{up}_{j\to v}}$
    else
      $t^{up}_{u\to parent(u)} = \sum_{j\in children(u)} t^{up}_{j\to u} + 1$
      $p^{up}_{u\to parent(u)} = t^{up}_{u\to parent(u)} \prod_{j\in children(u)} p^{up}_{j\to u}$
      $r^{down}_{u\to child(u)} = r^{down}_{parent(u)\to u} \frac{t^{up}_{u\to parent(u)}}{N - t^{up}_{u\to parent(u)}}$
    end if
  end if
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 $\widehat{v}$ with the following distribution.

$$\mathbf{P}(\widehat{v} = v|G_N) \propto R(v, G_N) \tag{6}$$

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.

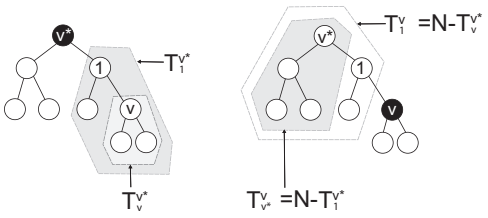

Figure 2: T_i^j variables for source nodes 2 hops apart.

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, $T_{bfs}(v)$. 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.

$$\widehat{v} = \arg\max_{v\in G_N} R(v, T_{bfs}(v)) \tag{7}$$

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.*

$$T_v^{v^*} \le \frac{N}{2} \tag{8}$$

If there is a node u such that for all $v \neq u$

$$T_v^u \le \frac{N}{2} \tag{9}$$

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,

$$T_u^v = N - T_v^u \tag{10}$$

For a node v one hop from v^*, we find

$$\frac{R(v,T)}{R(v^*,T)} = \frac{T_{v^*}^{v^*} T_v^{v^*}}{T_{v^*}^{v} T_v^{v}} = \frac{T_v^{v^*}}{(N - T_v^{v^*})}$$

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

$$\frac{R(v,T)}{R(v^*,T)} = \frac{T_v^{v^*} T_1^{v^*}}{(N - T_1^{v^*})(N - T_v^{v^*})}$$

Continuing this way, we find that in general, for any node v in T,

$$\frac{R(v,T)}{R(v^*,T)} = \prod_{i \in \mathcal{P}(v^*,v)} \frac{T_i^{v^*}}{(N - T_i^{v^*})} \tag{11}$$

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

$$\frac{R(v,T)}{R(v^*,T)} = \prod_{i \in \mathcal{P}(v^*,v)} \frac{T_i^{v^*}}{(N - T_i^{v^*})} \leq 1 \tag{12}$$

For a node v one hop from v^*, this gives us that

$$T_v^{v^*} \leq \frac{N}{2} \tag{13}$$

For any node u in subtree $T_v^{v^*}$, we will have $T_u^{v^*} \leq T_v^{v^*} - 1$. 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 find that

$$\frac{R(v,T)}{R(v^*,T)} = \prod_{i \in \mathcal{P}(v^*,v)} \frac{T_i^{v^*}}{(N - T_i^{v^*})} \leq 1 \tag{14}$$

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 $T_v^{v^*} < N/2$. 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 $T_v^{v^*} = N/2$. Then, $T_{v^*}^{v} = N/2$ also, and all other subtrees $T_u^v < N/2$, 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

$$D(v,G) = \sum_{j \in G} d(v,j) \tag{15}$$

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 $v \neq v_D$*

$$T_v^{v_D} \leq \frac{N}{2} \tag{16}$$

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_ℓ which is ℓ hops from v_D, and label a node on the path between v_ℓ 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 $T_{v_h}^{v_D}$ but not in subtree $T_{v_{h+1}}^{v_D}$, we have $d(v_\ell, j) = d(v_D, j) + d - 2h$. Using this, we find

$$\begin{aligned}
D(v_D,T) \leq & D(v_\ell,T) \\
\sum_{j \in T} d(v_D,j) \leq & \sum_{v \in T} d(v_\ell,j) \\
& \leq \sum_{j \in T} d(v_D,j) + \ell(N - T_{v_1}^{v_D}) + \\
& \quad (\ell-2)(T_{v_1}^{v_D} - T_{v_2}^{v_D}) + ... + (\ell - 2\ell)(T_{v_\ell}^{v_D}) \\
\sum_{h=1}^{\ell} T_{v_h}^{v_D} \leq & \sum_{h=1}^{\ell} (N - T_{v_h}^{v_D})
\end{aligned} \tag{17}$$

If we consider a node v_1 adjacent to v_D, we find the same condition we had for the rumor center. That is,

$$T_{v_1}^{v_D} \leq \frac{N}{2} \tag{18}$$

For any node u in subtree $T_{v_1}^{v_D}$, we will have $T_u^{v_D} \leq T_{v_1}^{v_D} - 1$. 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. □

For a general graph which is not a tree, the rumor center will be the node chosen by the general graph estimator which utilizes BFS trees. In a general graph, as can be seen in Figure 3, this general graph rumor center is not always equivalent to the distance center as it was for trees. We will see later that the general graph rumor center will be a better estimator of the virus source than the distance center. The intuition for this is that the distance center is evaluated using only the shortest paths in the graph, whereas the general graph rumor centrality utilizes more of the network structure for estimation of the source.

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.

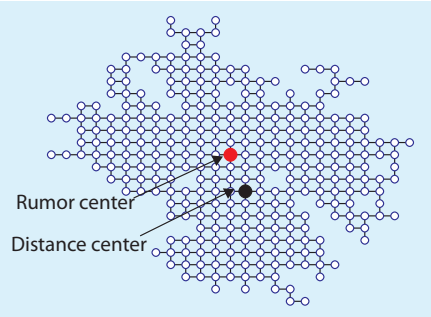

Figure 3: A network where the distance center does not equal the general graph rumor center.

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 $\mathcal{C}_t$. Then the probability of correct detection of the ML virus source estimator, $\mathbf{P}(\mathcal{C}_t)$, scales as*

$$\mathbf{P}(\mathcal{C}_t) = O\left(\frac{1}{\sqrt{t}}\right)$$

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. The proof of this theorem is omitted due to space constraints.

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 $\mathcal{C}_t$. Then the probability of correct detection of the ML virus source estimator, $\mathbf{P}(\mathcal{C}_t)$ is bounded uniformly away from 0. That is,*

$$\liminf_t \mathbf{P}(\mathcal{C}_t) > 0$$

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 α. 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

$$bd^\alpha \leq n(d) \leq cd^\alpha \tag{19}$$

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 $\mathcal{C}_t$. Then the probability of correct detection of the randomized virus source estimator, $\mathbf{P}(\mathcal{C}_t)$, is bounded uniformly away from 0. That is,*

$$\liminf_t \mathbf{P}(\mathcal{C}_t) > 0$$

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 $d(\widehat{v}, v^*)$ as the distance from the virus source estimator $\widehat{v}$ 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*

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

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.

4.1 Tree Networks

The detection probability of the virus source estimator versus the graph size for different trees is show 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

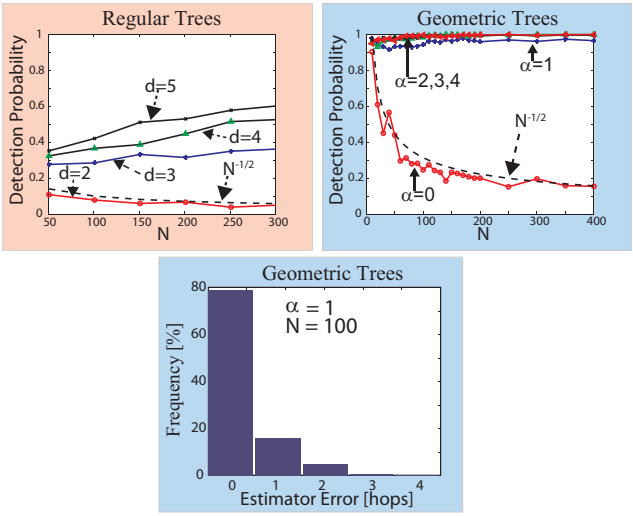

Figure 4: Virus source estimator detection probability for regular trees (left) and geometric trees (right) vs. number of nodes N, and a histogram of the error for a 100 node geometric tree with $\alpha = 1$ (bottom). The dotted lines are plots of $N^{-1/2}$.

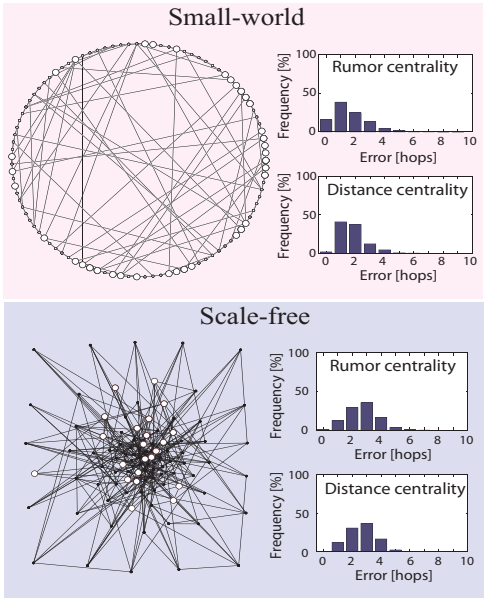

Figure 5: An example of a virus graph (infected nodes in white) and a histogram of the rumor centrality and distance centrality estimator error for a 400 node virus network on small-world (top) and scale-free networks (bottom).

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 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

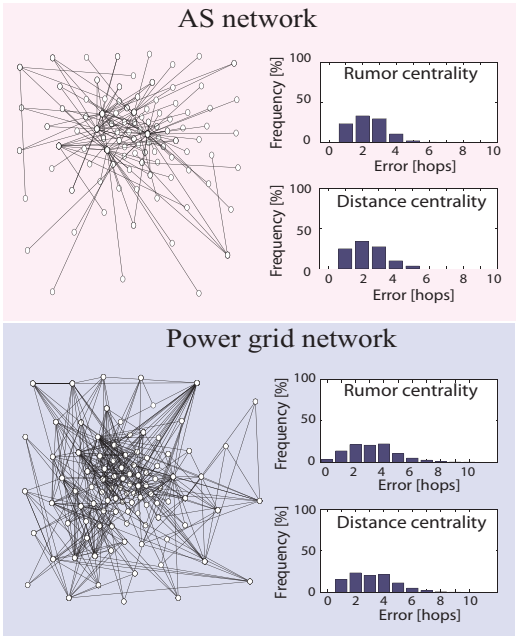

Figure 6: An example of a virus graph and a histogram of the rumor centrality and distance centrality estimator error for a 400 node virus network on an AS network (top) and the U.S. electric power grid network (bottom).

the event of correct detection $\mathcal{C}_t$ as

$$\mathcal{C}_t = \left\{ \omega \Big| \max_{i \in 1,2,\ldots,d} N_i(t,\omega) \leq \frac{1}{2} \sum_{j=1}^{d} N_i(t,\omega) \right\} \tag{20}$$

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 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

$$S_n = \sum_{i=1}^{n} T_i \tag{21}$$

Because of the complexity of event $\mathcal{C}_t$, we define the following sequence of events. Let $\mathcal{D}_n(t)$ 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,

$$\mathcal{D}_n(t) = \left\{ \omega \Big| \bigcap_{i=1}^{d} n \leq N_i(\omega,t) \leq (d-1)n \right\} \tag{22}$$

With $\mathcal{D}_n(t)$ we now can lower bound $\mathbf{P}(\mathcal{C}_t)$.

$$\mathbf{P}(\mathcal{C}_t) \geq \mathbf{P}\left(\bigcup_{i=1}^{\infty} \mathcal{D}_i(t) \right) \geq \mathbf{P}(\mathcal{D}_n(t)) \quad \forall n \in \{1,2,\ldots\} \tag{23}$$

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

$$\begin{aligned}
\mathbf{P}(\mathcal{D}_n(t)) =& \mathbf{P}\left(\bigcap_{i=1}^{d} \left(S_n \leq t \bigcap S_{(d-1)n} \geq t \right) \right) \\
=& \left(1 - \mathbf{P}\left(S_n \geq t\right) - \mathbf{P}\left(S_{(d-1)n} \leq t\right)\right)^d \\
=& \left(\mathbf{P}\left(S_n \leq t\right) - \mathbf{P}\left(S_{(d-1)n} \leq t\right)\right)^d \\
=& \left(\int_0^t \left(f_{S_n}(\tau) - f_{S_{(d-1)n}}(\tau) \right) d\tau \right)^d
\end{aligned} \tag{24}$$

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*

$$f_{S_n}(t) = \prod_{i=1}^{n-1} \left(1 + \frac{1}{ia}\right) e^{-t} (1 - e^{-at})^{n-1} \tag{25}$$

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. *$df_{S_n}(t)/dt > 0 \quad \forall t \in (0, t_n)$*
2. *$\limsup_n f_{S_n}(t_n) \leq C_a$ and $\liminf_n f_{S_n}(\tau_n) \geq B_a$ for some finite $C_a, B_a > 0$.*
3. *$\exists\, \gamma \in (0,1)$ such that $\limsup_n f_{S_{(d-1)n}}(t)/f_{S_n}(t) \leq (1-\gamma) \quad \forall t \in (0, t_n)$*

PROOF. We derive the density by induction. For $n = 1$, we have

$$f_{S_1}(t) = e^{-t} \tag{26}$$

Now, we do the induction step.

$$\begin{aligned}
f_{S_{n+1}}(t) &= f_{S_n}(t) * f_{T_{n+1}}(t) \\
f_{S_{n+1}}(t) &= \prod_{i=1}^{n-1} \left(1 + \frac{1}{ia}\right)(1 + an) \\
&\quad \int_0^t e^{-\tau} (1 - e^{-a\tau})^{n-1} e^{-(1+an)(t-\tau)} d\tau \\
&= C(n) e^{-(1+an)t} \int_0^t e^{an\tau} (1 - e^{-a\tau})^{n-1} d\tau
\end{aligned}$$

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

$$\int_0^t e^{an\tau} (1 - e^{-a\tau})^{n-1} d\tau = \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \frac{e^{at(n-i)} - 1}{a(n-i)}$$

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

$$\begin{aligned}
f_{S_{n+1}}(t) =& C(n)e^{-(1+an)t}\sum_{i=0}^{n-1}\binom{n-1}{i}(-1)^i\frac{e^{at(n-i)}-1}{a(n-i)}\\
=& \frac{C(n)e^{-t}}{an}\sum_{i=0}^{n-1}\binom{n}{i}(-1)^i(e^{-ati}-e^{-ant})\\
=& \frac{C(n)e^{-t}}{an}((1-e^{-at})^n-(-e^{at})^n+(-e^{-at})^n)\\
=& \prod_{i=1}^{n-1}\left(1+\frac{1}{ai}\right)\frac{1+an}{an}e^{-t}(1-e^{-at})^n\\
=& \prod_{i=1}^{n}\left(1+\frac{1}{ai}\right)e^{-t}(1-e^{-at})^n.
\end{aligned}$$

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

$$\begin{aligned}
df_{S_n}(t)/dt &> 0\\
(e^{-at}(an-a+1)-1) &> 0\\
1/a\log(an-a+1) &> t\\
t_n &> t.
\end{aligned}$$

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

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

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

where $\zeta(2)$ is the Riemann zeta function. Then, we can bound $f_{S_n}(\tau_n)$ as

$$\begin{aligned}
f_{S_n}(\tau_n) \geq& (n-1)^{1/a}e^{-\zeta(2)/a}e^{-1/a\log(n)-1/a\log(3/4a)}\\
&(1-e^{-log(3/4an)})^{n-1}\\
\geq& \left(\frac{n-1}{n}\right)^{1/a}e^{-\zeta(2)/a-1/a\log(3/4a)}\left(1-\frac{3}{4an}\right)^{n-1}
\end{aligned}$$

Therefore,

$$\liminf_n f_{S_n}(\tau_n)\geq B_a>0. \tag{27}$$

We also find that

$$\begin{aligned}
f_{S_n}(t_n) \leq& n^{1/a}e^{-1/a\log(na-a+1)+1}(1-e^{-log(na-a+1)})^{n-1}\\
\leq& e\left(\frac{1}{a+(a-1)/n}\right)^{1/a}\left(1-\frac{1}{na-a+1}\right)^{n-1}.
\end{aligned}$$

Therefore,

$$\limsup_n f_{S_n}(t_n)\leq C_a<\infty. \tag{28}$$

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

$$\begin{aligned}
\log\left(\frac{f_{S_{(d-1)n}}(t)}{f_{S_n}(t)}\right) =& \sum_{i=n}^{(a+1)n-1}\log\left(1+\frac{1}{ai}\right)+na\log(1-e^{-at})\\
\leq& \frac{1}{a}\sum_{i=n}^{(a+1)n-1}\frac{1}{i}+na\log(1-e^{-at_n})\\
\leq& \frac{1}{a}\log\left(\frac{(a+1)n-1}{n}\right)+na\log(1-\frac{1}{an-a+1})\\
\leq& \frac{1}{a}\log\left(\frac{(a+1)n}{n}\right)+na\log(1-\frac{1}{an})\\
\leq& \frac{1}{a}\log(a+1)-1+\frac{1}{2na}\leq\log(1-\gamma)<0
\end{aligned}$$

Therefore,

$$\limsup_n\frac{f_{S_{(d-1)n}}(t)}{f_{S_n}(t)}\leq(1-\gamma)<1\quad\forall t\in(0,t_n) \tag{29}$$

□

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

$$\begin{aligned}
\int_0^t(f_{S_n}(\tau)-f_{S_{(d-1)n}}(\tau))d\tau \geq& \int_0^t(f_{S_n}(\tau)-(1-\gamma)f_{S_n}(\tau))\,d\tau\\
\geq& \gamma\int_{\tau_n}^{t_n}f_{S_n}(\tau)d\tau-\gamma\int_{t_{n-1}}^{t_n}f_{S_n}(\tau)d\tau\\
\geq& \gamma f_{S_n}(\tau_n)(t_n-\tau_n)-\gamma f_{S_n}(t_n)(t_n-t_{n-1})\\
\geq& \gamma B_a(1/a\log(4/3))-\gamma C_a\frac{1}{a}\log\left(\frac{n}{n-1}\right).
\end{aligned}$$

For large n, the second term will go to 0, and therefore we have

$$\begin{aligned}
\liminf_n\mathbf{P}(\mathcal{C}_t) &\geq \liminf_n\mathbf{P}(\mathcal{D}_n(t))\\
&\geq(\gamma B_a(1/a\log(4/3)))^d>0.
\end{aligned}$$

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 $T_v^{v^*}$, the sizes of the subtrees with the virus source as the source node. Then, we express the detection probability in terms of the variables $T_v^{v^*}$. 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

$$bd^\alpha\leq n(d)\leq cd^\alpha \tag{30}$$

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 ϵ. Second, there will not be any nodes beyond a distance $t(1+\epsilon)$

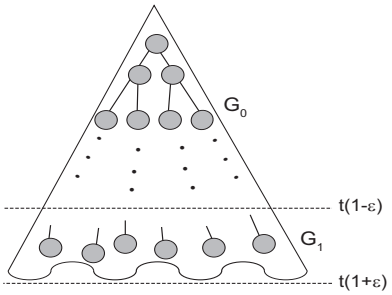

Figure 7: Partitioning of geometric tree for evaluating S.

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 now formally state our results for the structure of the virus graph (the proof is omitted due to space constraints).

THEOREM 4. *Consider a geometric tree with parameter α on which a virus spreads for a long time t, and let $\epsilon = t^{-1/2+\delta}$ for some small δ. Define the resulting virus graph as G_t. Also define $\mathcal{G}_t$ 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,*

$$\lim_{t\to\infty} \mathbf{P}(G_t \in \mathcal{G}_t) = 1$$

Detection Probability in terms of $T_v^{v^*}$. Our virus source estimator is a random variable $\widehat{v}$ 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 $\mathbf{P}(\widehat{v} = v^*|G_t)$. We showed that all virus graphs will belong to the set $\mathcal{G}_t$ with probability 1 for large t. Therefore, we lower bound the probability of correct detection $\mathbf{P}(\mathcal{C}_t)$ as

$$\begin{aligned}\liminf_t \mathbf{P}(\mathcal{C}_t) =& \liminf_t \sum_{G_t} \mathbf{P}(\widehat{v} = v^*|G_t)\mathbf{P}(G_t)\\ \geq& \liminf_t \inf_{G_t\in\mathcal{G}_t} \mathbf{P}(\widehat{v} = v^*|G_t)\end{aligned}$$

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

$$\begin{aligned}\mathbf{P}(\mathcal{C}_t) \geq& \inf_{G_t\in\mathcal{G}_t} \mathbf{P}(\widehat{v} = v^*|G_t)\\ \geq& \inf_{G_t\in\mathcal{G}_t} \left(\sum_{v\in G_t}\prod_{v_i\in G_t} \frac{T_{v_i}^{v^*}}{T_{v_i}^{v}}\right)^{-1}\\ \geq& \inf_{G_t\in\mathcal{G}_t} \left(\sum_{v\in G_t}\prod_{v_i\in\mathcal{P}(v^*,v)} \frac{T_{v_i}^{v^*}}{N - T_{v_i}^{v^*}}\right)^{-1} \geq \inf_{G_t\in\mathcal{G}_t} \frac{1}{S}\end{aligned}$$

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 $\mathcal{G}_t$ (Theorem 4) will allow us to bound the sizes of the subtrees $T_{v_i}^{v^*}$, and thus bound S.

Upper Bounding S. To evaluate the detection probability, we must upper bound the sum

$$S = \sum_{v\in G_t}\prod_{v_i\in\mathcal{P}(v^*,v)} \frac{T_{v_i}^{v^*}}{\left(N - T_{v_i}^{v^*}\right)} \tag{31}$$

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.

$$\begin{aligned}S =& \sum_{v\in G_0}\prod_{v_i\in\mathcal{P}(v^*,v)} \frac{T_{v_i}^{v^*}}{\left(N - T_{v_i}^{v^*}\right)} + \sum_{v\in G_1}\prod_{v_i\in P(v^*,v)} \frac{T_{v_i}^{v^*}}{\left(N - T_{v_i}^{v^*}\right)}\\ S =& S_0 + S_1\end{aligned}$$

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

$$\begin{aligned}\sum_{d=1}^{t(1-\epsilon)} bd^\alpha \leq& N_0 \leq \sum_{d=1}^{t(1-\epsilon)} cd^\alpha\\ b\frac{[t(1-\epsilon)]^{\alpha+1}}{\alpha+1} \leq& N_0 \leq c\frac{[t(1-\epsilon)]^{\alpha+1}}{\alpha+1}\end{aligned}$$

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 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

$$\begin{aligned}b\frac{t^{\alpha+1}}{\alpha+1}\left((1+\epsilon)^{\alpha+1} - (1-\epsilon)^{\alpha+1}\right) \leq N_1 \leq\\ c\frac{t^{\alpha+1}}{\alpha+1}\left((1+\epsilon)^{\alpha+1} - (1-\epsilon)^{\alpha+1}\right)\\ 2b\epsilon t^{\alpha+1} \leq N_1 \leq 2c\epsilon t^{\alpha+1}\end{aligned}$$

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.

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

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

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

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 bi^α 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 $T_{v_i}^{v^*}$

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

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

$$\begin{aligned} w_{v_i} &= \frac{T_{v_i}^{v^*}}{N - T_{v_i}^{v^*}} \\ &\leq \frac{\frac{N_0}{bi^\alpha} + N_1}{N - \frac{N_0}{bi^\alpha} - N_1} \\ &\leq \frac{\frac{N_0}{bi^\alpha} + N_1}{N_0 - \frac{N_0}{bi^\alpha}} \\ &\leq c_1 \left(\frac{1}{bi^\alpha} + \frac{2c\epsilon(\alpha+1)}{b(1-\epsilon)^{\alpha+1}} \right) \end{aligned}$$

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

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

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.

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

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

$$\begin{aligned} \limsup_d \frac{A_d}{A_{d-1}} &= \limsup_d \left(\frac{d}{d-1} \right)^\alpha \\ & \quad c_1 \left(\frac{1}{cd^\alpha} + \frac{2cd^{-1/2+\delta}(\alpha+1)}{b(1-d^{-1/2+\delta})^{\alpha+1}} \right) = 0 \end{aligned}$$

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

$$\begin{aligned} S_1 &= \sum_{v \in G_1} \prod_{v_i \in P(v^*,v), v_i \in G_0} w_{v_i} \prod_{v_i \in P(v^*,v), v_i \in G_1} w_{v_i} \\ &= \sum_{v \in G_1} \left(\prod_{v_i \in P(v^*,v), v_i \in G_0} w_{v_i} \right) b_v \end{aligned}$$

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 Γ. Then we can upper bound S_1 as

$$S_1 \leq \sum_{v \in G_1} \Gamma b_v \leq \sum_{d=1}^{2t\epsilon} cd^\alpha \Gamma a_d$$

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.

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

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

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

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.

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

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

$$\begin{aligned} \limsup_d \frac{B_d}{B_{d-1}} &= \limsup_d \left(\frac{d}{d-1} \right)^\alpha \frac{2cd^{-1/2+\delta}(\alpha+1)}{b(1-d^{-1/2+\delta})^{\alpha+1}} \\ &= 0 \end{aligned}$$

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.

$$\liminf_t \mathbf{P}(\mathcal{C}_t) \geq \liminf_t \inf_{G_t \in \mathcal{G}_t} \frac{1}{S} \geq \frac{1}{S^*} > 0$$

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 $\widehat{v}$ on a virus graph G_t formed after a virus has spread for a time t.

$$\mathbf{P}(\widehat{v} = v) = \frac{R(v, G_t)/R(v^*, G_t)}{\sum_{v \in G_t} R(v, G_t)/R(v^*, G_t)} = \frac{\rho(v, G_t)}{\sum_{v \in G_t} \rho(v, G_t)}$$

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

$$\rho(v, G_t) = \prod_{v_i \in \mathcal{P}(v^*, v)} \frac{T_{v_i}^{v^*}}{N - T_{v_i}^{v^*}}$$

We recognize the sum of $\rho(v, G_t)$ over all v in G_t as the sum S which was previously shown to be bounded by a positive constant S^*. Now, let $d(\widehat{v}, v^*)$ 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

$$\begin{aligned} \mathbf{P}(d(\widehat{v}, v^*) > l | G_t) &= \left(\sum_{v \in G_t} \rho(v, G_t) \right)^{-1} \sum_{v : d(v, v^*) > l} \rho(v, G_t) \\ &= S^{-1} \sum_{v : d(v, v^*) > l} \rho(v, G_t) \end{aligned}$$

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:

$$\begin{aligned}\liminf_t \mathbf{P}(d(\widehat{v}, v^*) \leq l) =& 1 - \limsup_t \mathbf{P}(d(\widehat{v}, v^*) > l) \\ =& 1 - \limsup_t \sum_{G_t \in \mathcal{G}_t} \mathbf{P}(d(\widehat{v}, v^*) > l | G_t) \mathbf{P}(G_t) \\ \geq& 1 - \limsup_t S^{-1} \sum_{v: d(\widehat{v}, v^*) > l} \rho(v, G_t) \\ \geq& 1 - \limsup_t \frac{\epsilon S}{S} \geq 1 - \epsilon\end{aligned}$$

Thus, for any positive ϵ, 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. ACKNOWLEDGMENTS

This work was supported by AFOSR Complex Networks Program SubAward 00006517

8. REFERENCES

[1] The CAIDA AS relationships dataset. *http://www.caida.org/data/active/as-relationships/*, August 30th, 2009.

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

[3] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[4] P. Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92:1170–1182, 1987.

[5] L. C. Freeman. A set of measure of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.

[6] 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)*, 2:1455–1466, 2005.

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

[8] M. E. J. Newman. The spread of epidemic disease on networks. *Phys. Rev. E*, 66:016128, 2002.

[9] 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*, 5:149 – 158, 2007.

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

[11] G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31:581–603, 1966.

[12] G. Streftaris and G. J. Gibson. Statistical inference for stochastic epidemic models. *Proc. 17th international Workshop on Statistical Modeling*, pages 609–616, 2002.

[13] D. J. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.