Rumor Source Detection in Finite Graphs with Boundary Effects by Message-passing Algorithms

Abstract—**Detecting information source in viral spreading has important applications such as to root out the culprit of a rumor spreading in online social networks. In particular, given a snapshot observation of the network topology of nodes having rumor, how to accurately identify the initial source of the spreading? In the seminal work by Shah and Zaman in 2011, this problem was solved as a maximum likelihood estimation problem using a rumour centrality approach for *infinite graphs* that are degree-regular. This however is optimal only if the underlying potential number of susceptible nodes is countably infinite, i.e., no boundary effects. In general, all practical real-world networks are finite, and therefore these boundary effects can not be ignored. In this paper, we solve the constrained maximum likelihood estimation problem and characterize the rumor centrality solution to spreading in finite graphs with boundary effects. We derive message-passing algorithms that are optimal for degree-regular graph with a single end vertex and near-optimal for general tree graph, and evaluate their performance.**

I. Introduction

Networks represent a fundamental medium for the spreading and diffusion of information. Network nodes are said to be infected when they possess this information, and network topologies govern how the spreading processes increase the susceptibility of other nodes to be infected leading to the successive spread of information from a few initial nodes to a much larger set. An example of such a viral spreading phenomenon is rumor spreading in online social networks. From cybersecurity enforcement viewpoint, this begs the question of detecting and rooting out malicious information sources in a reliable manner [1]. In particular, given a snapshot observation of the infected nodes, who is the culprit source of the spreading?

In a recent seminal work in [9], [10], Shah and Zaman formulated this as a maximum likelihood estimation problem, and proposed *rumor centrality*, a form of network centrality, to solve this problem exactly for degree-regular tree graphs assuming that the underlying number of susceptible nodes is countably infinite. This rumor centrality approach counts the number of ways each infected node can spread to the rest and to find the node with the maximum value called the *rumor center* that coincides with the maximum likelihood estimate. The rumor centrality was subsequently extended to various problem settings, e.g., extension in [13] to random trees, extension in [16] to suspect sets, extension in [17] to multiple source detection and extension in [18] to detection with multiple snapshot observations. It is shown in [8] that this rumor center is equivalent to the graph centroid.

There are however several open issues in the rumor centrality approach. A key limitation in all the aforementioned work is that the underlying graph is infinite. This is not true in general of practical real-world networks where the number of underlying susceptible nodes is countably finite. This introduces boundary effects that can not be ignored and makes this constrained maximum likelihood estimation problem a much harder combinatorial problem. When the boundary effects of the dynamical spreading process in a finite underlying graph are taken into account, spreading cannot continue further at the boundary end vertices, i.e., the susceptible nodes with only a single neighbour, and this in fact increases the likelihood of the nodes near the boundary to be the culprit. This means that the number of these end vertices and their location in the underlying graph can significantly shape spreading and thus the optimal estimation performance. To be exact, existing algorithms in the literature, e.g., [9], [10], [13], [16]–[18], are no longer optimal *even with the presence of a single infected end vertex* in degree-regular graphs with boundary effects.

Rumor source detection over a finite underlying graph is clearly more challenging, but is more realistic and also significantly generalizes all previous work [9], [10], [13], [16]–[18]. In passing, we mention another open issue which is the presence of cycles in network topology whether the underlying graph is finite or infinite. In the case of infinite underlying graphs, suboptimal heuristics based on the rumor centrality (e.g., see the breadth-first search heuristic in [9], [10]) have performed reasonably well. As we shall see in this paper, the finite graph generalization analysis offers a surprisingly unique perspective to tackle some network topologies with cycles that are in fact the optimal maximum-likelihood estimate. A focus in this paper is thus to extend the rumor centrality approach and propose efficient algorithm design. In particular, adding to the combinatorial barrier with boundary effect is the computational barrier due to the network size when the number of infected nodes can be very large or asymptotically approach infinity. We investigate the use of message passing algorithms (see [12] for an introduction) to design detection algorithms with practical computational complexity.

A. Our Contributions

The main contributions of this paper are summarized as follows.

- We propose an extended rumor centrality to solve the maximum likelihood estimation problem when the underlying susceptible graph is finite and with boundary effects of spreading. For a finite *degree-regular tree*

TABLE I
TABLE OF NOTATION

Notation	Remark
$v^\star$	Actual rumor source
$\hat{v}$	Maximum likelihood estimator for $v^\star$
v_c	*Centroid* of G_n (also the *rumor center* in [13])
v_e	End vertex of G_n
G	Underlying graph(network)
G_n	Infected subgraph of G of n nodes
$P(v\|G_n)$	Probability that $v = v^\star$, when G_n is observed
$P(\hat{v}\|G_n)$	Probability that $\hat{v} = v^\star$, i.e., correct detection probability

graph with a single end vertex, we propose an optimal message-passing algorithm to detect the single source, and characterize its performance.

- We extend our analysis to degree-regular tree graphs with multiple end vertices and propose suboptimal algorithms that we evaluate to be near-optimal in performance.
- For the finite general graph, we propose a breadth-first-search heuristic in combination with this extended rumor centrality and also show how this analysis offers a unique perspective to tackle some network topologies with cycles.

II. PRELIMINARIES OF RUMOR CENTRALITY

We model an online social network of nodes by an undirected graph $G = (V, E)$, where the set of vertices V represents the nodes in the underlying network, and the set of edges E represents the links between the nodes. We shall assume that V is *countably finite* (this is the crucial departing point from the previous assumption of infinite graph in the literature [9], [10], [13], [16]–[18]). Following [9], [10], we use the Susceptible-Infectious (SI) model in [11] to model rumor spreading. Nodes that possess the rumor are called *infected nodes* and otherwise they are *susceptible nodes*. The spreading is initiated by a single node $v^\star \in V$ that we call the rumor source. Once a node is infected (i.e., possesses the rumor), it stays infected and can in turn infect its susceptible neighbours. A rumor is spread from node i to node j if and only if there is an edge between them (i.e., $(i, j) \in E$). Let τ_{ij} be the spreading time from i to j, which are independent and have exponential distribution with parameter λ. Without loss of generality, let $\lambda = 1$. Hence, we have a discrete-time dynamical spreading model over an underlying *finite graph*. Let G_n be a subgraph of order n of G, that modles a snapshot observation of the spreading when there are n infected nodes, i.e., $|G_n| = n$. Clearly, G_1 is the *actual rumor source*, i.e., $v^\star$. The rumor source detection problem is thus to find $v^\star$ given this observation of G_n.

First, we review the maximum likelihood (ML) estimation problem of the rumor source in a tree network. The ML estimator for the rumor source is the vertex v with the maximum probability $P(G_n|v)$ [9], [10]. We focus on characterizing $P(G_n|v)$ for degree-regular tree networks.

Definition II.1. For a given G_n over the underlying graph G, $\hat{v}$ is an ML-estimator for the source in G_n, i.e., $P(\hat{v}|G_n) = \max\limits_{v_i \in G_n} P(v_i|G_n)$.

By Bayes theorem, $P(G_n|v)$ is the probability that v is the *real rumor source culprit* that leads to G_n. Now, let σ_i be the possible ordered infection sequence starting from v, and $S(v, G_n)$ be the collection of all σ_i when v is the source in G_n. Then, we have

$$P(G_n|v) = \sum_{\sigma_i \in S(v,G_n)} P(\sigma_i|v). \tag{1}$$

In particular, for a d-regular tree, we have [9], [10]:

$$P(\sigma_i|v_1) = \prod_{k=1}^{n-1} \frac{1}{dk - 2(k-1)}. \tag{2}$$

Now, if the spreading has not reached the end vertices, then $P(\sigma_i|v_1) = P(\sigma_j|v_1)$ for all $\sigma_i, \sigma_j \in S(v, G_n)$. By combining (1) and (2), we have

$$\begin{aligned} P(G_n|v) &= \sum_{\sigma_i \in S(v,G_n)} P(\sigma_i|v) \\ &= |S(v, G_n)| \cdot P(\sigma|v) \quad \forall \sigma_i \in S(v, G_n) \\ &= |S(v, G_n)| \cdot \prod_{k=1}^{n-1} \frac{1}{dk - 2(k-1)} \end{aligned}$$

which means that $P(G_n|v)$ is proportional to $|S(v, G_n)|$. This quantity $|S(v, G_n)|$ denoted by $R(v, G_n)$ is the *rumor centrality* in [9], [10] that is crucial to solving the maximum likelihood estimation for degree-regular trees. In particular, the vertex having the maximum rumor centrality called the *rumor center*.

Definition II.2. Let G be a tree, for any $v \in G$, define $weight(v) = \max\limits_{c \in child(v)} |T_c^v|$, where T_c^v is the subtree rooted at c by removing the edge (v, c) from G.

Definition II.3. Let G_n be an infected subgraph of the underlying graph G. The *centroid* of G_n is the vertex v with minimum $weight(v)$.

The *centroid* of G_n is equal to the *rumor center* in the sense of [13] when G_n is a tree and $|V| \to \infty$ [8]. Since we are studying an *extended rumor centrality* for finite $|V|$ (and thus an extended rumor center) in this paper, we will call the *rumor center* in the sense of [13] as the *centroid*.

In the following, we first motivate studying the single end vertex case by comparing with the prediction result in the special case of *infinite graph*. This will demonstrate that ignoring the boundary effect can lead to wrong estimate and thus a need to characterize the boundary effect on spreading and extend the rumor centrality for the general case of *finite graph*.

III. TREES WITH A SINGLE END VERTEX

Let us consider the case when G is a regular tree that is finite, i.e., there are leaf vertices each with degree one. In addition, consider $G_n \subseteq G$, where only one vertex is a leaf called the *end vertex* that can only receive the rumor but can no longer spread it further. In this section, we will study how the influence of this *end vertex* in G_n affects the performance of the maximum likelihood estimation.

A. Impact of Boundary Effects On $P(G_n|v)$

Example 1. Let us use the example in Section II with G now being a finite order $3-regular$ tree and $G_5 \subseteq G$ as shown in Figure 1. Consider $P(G_5|v_1)$, for the spreading order $\sigma : v_1 \to v_2 \to v_5 \to v_3 \to v_4$, we have $P(\sigma|v_1) = \frac{1}{3}\cdot\frac{1}{4}\cdot\frac{1}{3}\cdot\frac{1}{4}$. Now, if v_5 were not the end vertex, then $P(\sigma|v_1) = \frac{1}{3}\cdot\frac{1}{4}\cdot\frac{1}{5}\cdot\frac{1}{6}$. This demonstrates that the time when the rumor spreads to v_5 has impact on the probability $P(\sigma|v_1)$. Let us list all the spreading orders sorted according to the position of v_5 in each σ_i on Table II. So, $P(G_5|v_1) = \frac{34}{720}$. We also have $P(G_5|v_4) = P(G_5|v_3) = \frac{7}{720}$ by symmetry, and $P(G_5|v_2) = \frac{40}{720}$. In particular, even although v_1 is the *centroid*, but $P(G_5|v_1) < P(G_5|v_2)$, and thus v_1 is not the ML-estimator.

Example 1 reveals some interesting properties of boundary effects due to end vertices:

- The earlier the end vertex appears in σ_i (ordered from left to right of σ_i) in comparison to σ_j, $j \neq i$, the larger $P(\sigma_i|v)$ will be.
- When there is at least one end vertex in G_n, then $P(G_n|v)$ is no longer proportional to $|S(v, G_n)|$.

Since $P(\sigma_i|v)$ is no longer a constant for each i, and is dependent on the position of the *end vertex*, we need to compute the spreading order separately taking into account the *end vertex*. For brevity of notation, let v_e be the end vertex and define

$$M_v^{v_e}(G_n, k) = \{\sigma | v_e \ is \ on \ k_{th} \ position \ of \ \sigma\}$$
$$P_v^{v_e}(G_n, k) = P(\sigma|v), \ for \ \sigma \in M_v^{v_e}(G_n, k),$$

$M_v^{v_e}(G_n, k)$ collects all the spreading orders starting from v and with v_e at the k_{th} position. Denote $|M_v^{v_e}(G_n, k)| = m_v^{v_e}(G_n, k)$ and let $P_v^{v_e}(G_n, k)$ be the corresponding probability for the same k. Let D be the distance from v to v_e.

Lemma 1. *$m_v^{v_e}(G_n, k)$ is a non-strictly increasing sequence from $k = D + 1$ to n.*

Proof. Given $D + 1 \leq k \leq n$. Define a function

$$f_k(\sigma) : M_v^{v_e}(G_n, k) \longmapsto M_v^{v_e}(G_n, k + 1),$$

i.e., $f_k(\sigma)$ swaps the kth and the $(k + 1)$th elements in σ. Now, $f_k(\sigma)$ is an one-to-one function from $M_{v_c}^{v_e}(G_n, k)$ to $M_{v_c}^{v_e}(G_n, k + 1)$, so we have $|M_{v_c}^{v_e}(G_n, k + 1)| \geq |M_{v_c}^{v_e}(G_n, k)|$. This implies that $m_v^{v_e}(G_n, k)$ is a non-strictly increasing sequence over k. □

Now, we can rewrite $P(G_n|v)$ for the finite tree:

$$P(G_n|v) = \sum_{k=D+1}^{n} m_v^{v_e}(G_n, k) \cdot P_v^{v_e}(G_n, k). \quad (3)$$

Again, our goal is to find out the vertex v having $P(G_n|v) = \max_{v_i \in G_n} P(v_i|G_n)$, but now $P(G_n|v)$ is not proportional to $S(v, G_n)$.

B. Computing $P(G_n|v)$

We now describe how to compute $P(G_n|v)$ in G_n. First, consider $P_v^{v_e}(G_n, k)$ and let $z_d(i) = (i - 1)(d - 2)$, then

$$P_v^{v_e}(G_n, k) = \prod_{i=1}^{k} \frac{1}{d + z_d(i)} \cdot \prod_{i=k}^{n-2} \frac{1}{d + z_d(i) - 1}. \quad (4)$$

The first factor of $P_v^{v_e}(G_n, k)$ in (4) is the probability before the end vertex receives the rumor, and the second factor is the probability after the rumor reaches the end vertex. To compute $m_v^{v_e}(G_n, k)$, similarly, we decompose $m_v^{v_e}(G_n, k)$ into two parts with the details given in Algorithm 1.

Definition III.1. The *contraction of a pair of vertices v_i and v_j of a graph produces a graph in which the two nodes v_i and v_j are replaced by a single vertex v such that v is adjacent to the union of the nodes to which v_i and v_j were originally adjacent. Conceptually, v_i and v_j are together into a single node.*

In Algorithm 1, after we go through all $(k-1)$-subtrees that contain v_e's parent, we finish the computation of $m_v^{v_e}(G_n, k)$ for a given k. When G_n is a arbitrary tree on G, then it's almost impossible to go through all $(k - 1)$-subtrees on G_n, especially when k getting larger. Fortunately, when G_n is a line, then we have a closed form expression for $m_v^{v_e}(G_n, k)$, so we can now compute this special case to see how end point affect the probability $P(v|G_n)$.

Algorithm 1 Computing $m_v^{v_e}(G_n, k)$ for a tree graph with a single end vertex

Input: G_n,k,v,v_e
Step 1: Starting from v, find a $(k - 1)$-subtree of G_n containing v_e's parent.
Step 2: Apply the rumor centrality algorithm given in [10] on the $(k-1)$-subtree with root=v and set the output to P_1.
Step 3: Contract all vertices in $(k - 1)$-subtree and v_e, get a new graph G_{n-k+1} with a new root v'.
Step 4: Apply the rumor centrality algorithm again on G_{n-k+1} with root v', and set the output to P_2.
Output: $(P_1 \cdot P_2)$

C. Degree-Regular Tree ($d \geq 3$) Special Case: G_n is Line Graph

Suppose G is a degree-regular tree with finite order and G_n is a subtree of G with a single end vertex. Without loss of generality, suppose n is odd(to ensure that there will be only one centroid) and $n = 2t + 1$ for some t. Let $z_d(i)$ as defined above. Label all the vertices on G_n as shown in Figure 2. To compute $P(G_{2t+1}|v_i)$ for $v_i \in G_n$, from (3) and (4), we have $P_{v_i}^{v_e}(G_n, k)$ already, so now we need $m_{v_i}^{v_e}(G_n, k)$.

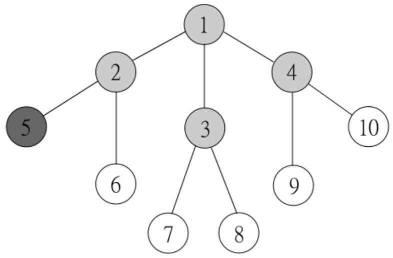
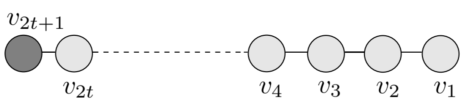


Fig. 1. In this example, G is a finite 3-regular tree, and G_n is a subtree with a single end vertex $v_e = v_5$. The maximum likelihood estimate is v_2, while the naive application of the rumor centrality in [13] yields v_1.

TABLE II
EXAMPLE OF COMPUTING $P(\sigma_i|G_5)$ USING THE EXAMPLE SHOWN IN FIGURE 1

σ_i	Spreading Order	$P(\sigma_i\|G_5)$	σ_i	Spreading Order	$P(\sigma_i\|G_5)$
σ_1	v_1, v_2, v_5, v_3, v_4	$\frac{1}{144}$	σ_7	v_1, v_2, v_3, v_4, v_5	$\frac{1}{360}$
σ_2	v_1, v_2, v_5, v_4, v_3	$\frac{1}{144}$	σ_8	v_1, v_2, v_4, v_3, v_5	$\frac{1}{360}$
σ_3	v_1, v_3, v_2, v_5, v_4	$\frac{1}{240}$	σ_9	v_1, v_3, v_2, v_4, v_5	$\frac{1}{360}$
σ_4	v_1, v_4, v_2, v_5, v_3	$\frac{1}{240}$	σ_{10}	v_1, v_3, v_4, v_2, v_5	$\frac{1}{360}$
σ_5	v_1, v_2, v_3, v_5, v_4	$\frac{1}{240}$	σ_{11}	v_1, v_4, v_2, v_3, v_5	$\frac{1}{360}$
σ_6	v_1, v_2, v_4, v_5, v_3	$\frac{1}{240}$	σ_{12}	v_1, v_4, v_3, v_2, v_5	$\frac{1}{360}$

Fig. 2. G_n as a line graph with a single end vertex v_{2t+1}



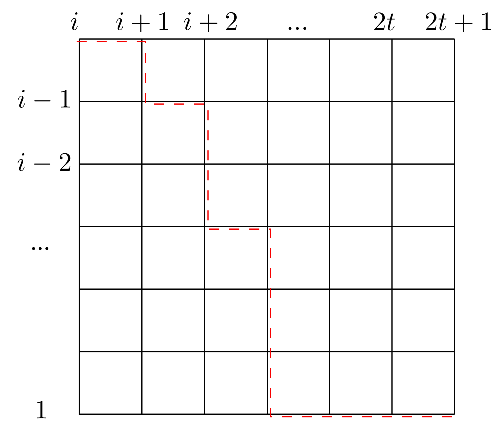
Observe that the number of spreading orders started from v_i is equivalent to finding all possible ways from the coordinate i to the lower right-hand corner coordinate in Figure 3. This can be accomplished with a path-counting message-passing algorithm (see, e.g., Chapter 16 in [12]) with polynomial-time complexity. Or we can just write a close formula for $m_{v_i}^{v_e}(G_n, k)$. We deduce $m_{v_i}^{v_e}(G_n, k) = \binom{k-2}{k-2t+i-2} or \binom{k'+(2t-i)}{k'}$, where $k' = k - 2t + i - 2$. Finally, we have

$$P(G_{2t+1}|v_i) = \sum_{k=2t+2-i}^{2t+1} m_{v_i}^{v_e}(G_n, k) P_{v_i}^{v_e}(G_n, k) \quad (5)$$

for $i = 1, 2, ..., 2t$, where $m_{v_i}^{v_e}(G_n, k) = \binom{k-2}{k-2t+i-2}$ and $P_{v_i}^{v_e}(G_n, k)$ is as described in (4). Also, we have $P(G_{2t+1}|v_{2t+1}) = \prod_{l=1}^{2t} \frac{1}{z_d(l)+1}$, when $i = 2t+1$.

$k' = 0$ means that in Figure 3, i goes horizontally right to the end of the grid and then goes down vertically down to the end, i.e., v_i spreads the rumor straight to the end vertex before spreading to the other vertices. As such, k' can be viewed as a parameter related to the time period when the rumor reaches the end vertex. Using (5), we can numerically compute $P(G_{2t+1}|v_i)$ for all v_i. We illustrate one result in Figure 4, where G is a 4-regular tree and G_n is a line with a single end vertex for $n = 7, 8, 9, 10$. The x-axis is the node v_i where $i = 1, 2, ..., n$, and the y-axis plots the probability $P(v_i = v^\star|G_n)$. It can be observed in Figure 4, the influence from the end vertex on $P(v_i = v^\star|G_n)$ is dominate the influence from centroid when $n = 7, 8, 9$ but when $n = 10$ the influence of centroid on $P(v_i|G_n)$ is greater then that from end vertex. The following theorem state that for any d-regular underlying graph, when G_n is a line with a single end vertex,

Fig. 3. The dash line on the grid corresponds to a spreading order started from i and ended at $2t+1$, and the order is $\sigma = (i, i+1, i-1, i+2, i-2, ..., 1, 2t-1, 2t, 2t+1)$



the influence of end vertex on $P(v_i|G_n)$ decreases as n grows. Asymptotically, this in fact reduces to the result in [13], when $n \to \infty$, i.e., the centroid is the maximum likelihood estimator of the source.

Theorem 1. *G is a d-regular graph ($d > 2$) with finite order, if G_n is a line-graph with a single end vertex, then $\exists j$ such that $P(v_c|G_n) > P(v_e|G_n)$ when $n > j$.*

From the above theorem, we can found that in Figure 4, in our example we have $j = 9$. The proof of Theorem 1 is shown in appendix.

D. Location of the ML-Estimator

From (3), we know that in addition to the spreading order as illustrated above, the distance (number of hops) from the end vertex also affects the likelihood probability $P(v|G_n)$. Suppose there are two adjacent vertices v_c and v_p, with $R(v_c, G_n) > R(v_c, G_n)$. If G_n has no end vertex, then we can deduce that $P(G_n|v_c) > P(G_n|v_p)$. Now, G_n has an end vertex v_e, from the properties we observed in the end of Section III Part A, intuitively, we may assume one more thing that v_c is more closer to the end vertex than v_p does,

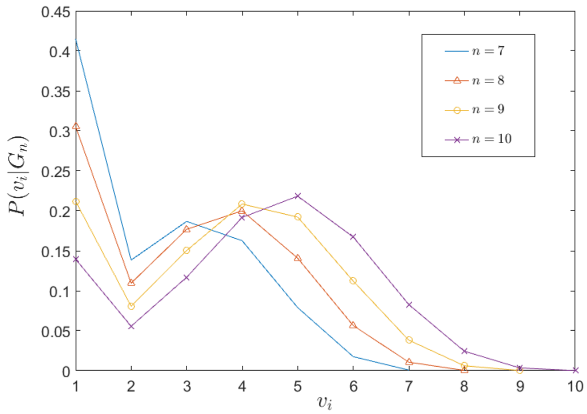
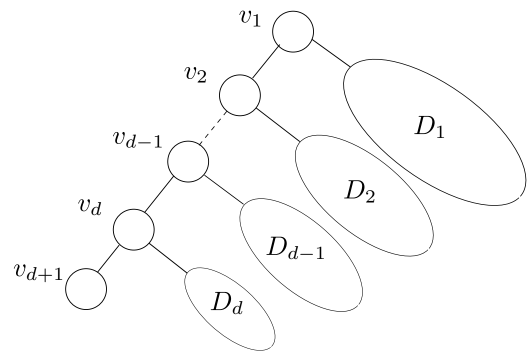


Fig. 4. $P(v|G_n)$, where G_n is a line graph with a 4-regular underlying finite graph

Fig. 5. G_n with an end vertex v_{d+1}



so these two assumptions may let us reach to the result $P(G_n|v_c) > P(G_n|v_p)$. In the following, base on the concept above, we give several results and a theorem to describe the relation between any two adjacent vertices and location of the ML-estimator for the rumor source on G_n, when G_n has a single end vertex.

Let $G_n \subseteq G$ be an infected subtree with a single end vertex v_e and a centroid v_c. Let $T^{v_1}_{v_2}$ be as defined in Definition II.2. v_p is a neighbor of v_c with $|T^{v_p}_{v_c}| - |T^{v_c}_{v_p}| = 1$. And if $G_n \backslash \{v_e\}$, then $|T^{v_p}_{v_c}| = |T^{v_c}_{v_p}|$. Let $distance(v_c, v_e) = D$ and denote the path from v_c to v_e as $P = (p_0, p_1, p_2, ..., p_D)$ where $v_c = p_0$ and $v_e = p_D$. $M^{v_e}_{v_c}(G_n, k)$ and $m^{v_e}_{v_c}(G_n, k)$ are as defined above.

Lemma 2. Let G_n,v_p as defined above, and $distance(v_c, v_e)$ denoted as D. Then we have,

$$m^{v_e}_{v_c}(G_n, i) \geq m^{v_e}_{v_p}(G_n, i),$$

for $D + 1 \leq i \leq n$.

Theorem 2. *Let G be a tree with finite order and $G_n \subseteq G$ is a subtree of G with an end vertex $v_e \in G_n$, then the vertex $v^\star$ with maximum probability $P(v|G_n)$ is located on the path from the v_c to v_e. In particular, Algorithm 1 computes the $P(v_i|G_n)$ for all v_i on this path.*

Although computing $P(G_n|v)$ is nearly impossible when n is large, but with Theorem 2, we only need to care about those vertices on the path from v_c to v_e.

E. G_n's Topology and $P(v|G_n)$

From [9], [10], we knew that for any two infected subgraph G_n and G'_n under G, if G_n is not isomorphic to G'_n, then the probability $P(v|G_n)$ for each v of them may not be the same. Topology of G_n do have great effect on the probability $P(v|G_n)$. For G_n and G'_n, if there exists an axes of symmetry of G_n such that G'_n can be reached by rotating G_n alone this axis, then for any $v_i \in G_n$ and its corresponding vertex $v'_i \in G_n$, we have $P(v_i|G_n) = P(v'_i|G_n)$.

Now, consider two graphs G_{n_1} and G_{n_2} under G where $n_1 = n_2$. G_{n_1} is a line graph with an end vertex v_e, and G_{n_2} is a "complete" tree with an end vertex v_e.(Here, the meaning of "complete" is the same as "complete" binary tree.) We have $\frac{P(v_c|G_{n_2})}{P(v_e|G_{n_2})} > \frac{P(v_c|G_{n_1})}{P(v_e|G_{n_1})}$, i.e. the effect of the end vertex on $P(v|G_{n_1})$ is less than that on $P(v|G_{n_2})$. Since when n is fixed, $\frac{R(v_c,G_n)}{R(v_e,G_n)}$ reach to minimum when G_n is a line. The following is an example of the effect from G_n's topology on the boundary effect.

Example 2. Let G be a d-regular graph, and G_n is an infected subgraph with a single end vertex. As shown in Figure **??**, G_n is a star graph with one vertex in the middle(the grey vertex which is also the centroid) and the rest vertices(white vertices) are connected to the centroid. One of G_n's leaves is an end vertex denoted as v_e. The following theorem is an extreme case compared with the result in Theorem 1.

Theorem 3. *Assume that d is large enough(d is always larger than n). v_e is always the ML-estimator no matter how many new vertices added to the centroid.*

Note that in Theorem 3, we assume that d is extremely large($d > n$), so we can keep adding vertices to the centroid and reach to the result that v_e is the ML-estimator. But when d is fixed, then as we adding vertices continuously to the centroid(if $degree(v_c) = d$, then add new vertex to other existing vertex rather than v_c), the ML-estimator on G_n will finally be v_c, just the same result in Theorem 1.

IV. Trees with Multiple End Vertices

In the previous section, we describe the location of ML-estimator when G_n has only one end vertex. But in reality, there will always be more than one end vertex(if G is not a line). Since once the rumor reach to one of the leaves in G, the rumor is also close to other leaves which are near the first end vertex. We start with a special case: The entire finite network is infected (all end vertices being infected), i.e., $G_n =$

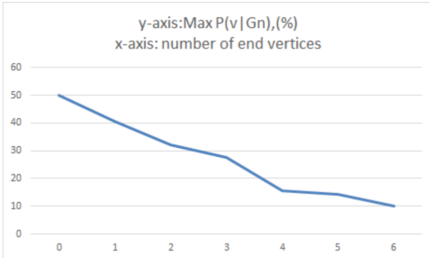


Fig. 6. Detection Probability and Number of end vertices

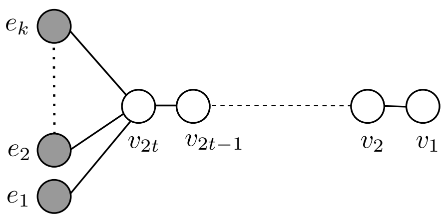


Fig. 7. G_n as a line graph with k star-like end vertices e_1 to e_k

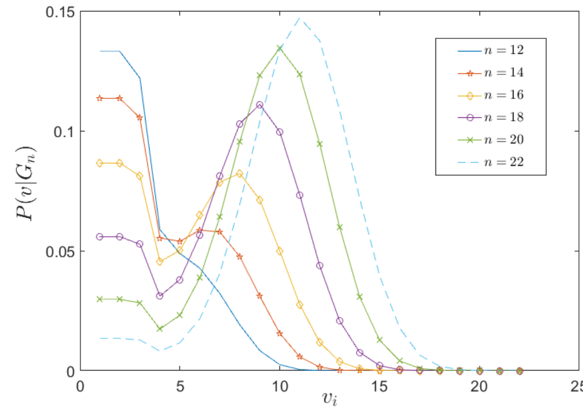


Fig. 8. Probability distribution of each node on G_n with 2 end verices where G is 3-regular, y-axis is the probability $P(v_i|G_n)$ and x-axis is vertex v_i's number i, v_1, v_2 are end vertices

G, then the probability $P(G_n|v)$ of each vertex in G_n is $\frac{1}{n}$, since any vertex is able to spread the rumor to all the vertices in G in $n-1$ steps when $G_n = G$. In this case, $P(\hat{v}|G_n)$ is exactly the minimum detection probability of this source finding problem. For example, consider a 3-regular underlying graph G with 10 vertices. In Figure 6, it shows the detection probability of the infected subgraph of G as the number of end vertices increasing. Note that when there is no end vertex, G_n is composed of 4 vertices of the inner part of G. We can see that as the number of end vertices increases, $P(\hat{v}|G_n)$ decreases to $1/10$. Unlike those previous results, when G is infinitely large, then as G_n growing, the detection probability will converge to some specific number rather than decreasing to $1/n$. So the bound of $P(\hat{v}|G_n)$ given in previous study are not suitable for the case with end vertices. Therefore, when simulating the rumor spreading in a network, we will set a upper bound n/k of the number of end vertices where k is some integer greater than 1, once the number of end vertices in G_n reach to n/k, then we will stop the spreading process.

A. Degree-Regular Tree ($d \geq 3$) Special Case: G_n is Broom-Shaped

In Section III-B, we have compute the $P(v|G_n)$ for each vertex on a line graph with an end vertex on one end. It shows that when G_n is large enough, then the effect of one end vertex on $P(v|G_n)$ is less than the effect of centroid. In the following, we will show the effect of multiple end vertices and centroid on a class of graph. As we consider the line graph in section 3, now we add more end vertices to v_{2t}, so when G is d-regular, then there will be at most $d-1$ end vertices in G_n. For simplicity, we call this class of graph *broom*. We can compute $P(v|G_n)$ by extending the result in section 3 part B. Define $P_{v_i}^{\{e_1,e_2,\ldots,e_k\}}(\{h_1, h_2, \ldots h_k\}, G_n)$ be the probability of the spreading order starting from v_i with the end vertex set $\{e_1, e_2, ..., e_k\}$ and their position set $\{h_1, h_2, ..., h_k\}$ in this spreading order. So as $m_{v_i}^{\{e_1,e_2,\ldots e_k\}}(\{h_1, h_2, \ldots h_k\}, G_n)$. Here we don't assume h_i is the position of e_i, it can be the position of any end vertex in G_n. The detail of the $P_{v_i}^{\{e_1,e_2,\ldots,e_k\}}(\{h_1, h_2, ..., h_k\}, G_n)$ is skipped here, it's just the same as (4) but with d parts multiplied together. To compute $m_{v_c}^{\{e_1,e_2,\ldots,e_k\}}(\{h_1, h_2, ..., h_k\}, G_n)$, first we consider the line-shaped part of G_n i.e. the part $\{v_1, v_2, ..., v_{2t}\}$ say G'_n. From the previous discussion, we have $m_{v_i}^{v_{2t}}(G'_n, j) = \binom{j+(2t+i-1)}{j}$, and for each spreading order that v_{2t} lies on the $j-th$ position, the end vertices $e_1, e_2, ..., e_k$ can be place to any position after the $j-th$ position. So for each spreading order in $m_{v_i}^{v_{2t}}(G'_n, j)$, there are $k! \cdot \binom{n-k-j+1}{k}$ corresponding spreading orders on G_n. We deduce that

$$m_{v_i}^{\{e_1..e_k\}}(G_n, \{h_1..h_k\}) = k! \sum_{j=2t-i+1}^{h_1-1} \binom{j-2}{2t-i-1} \quad (6)$$

With $P_{v_i}^{\{e_1,e_2,\ldots e_k\}}$ and $m_{v_i}^{\{e_1,e_2,\ldots e_k\}}$, now we can compute the probability $P(v_i|G_n)$ by go through all possible $\{h_1, h_2..h_k\}$. Figure 9 shows that even though there are five end vertices, the effect of the centroid on $P(v|G_n)$ finally dominate that of end vertices as n getting from 37 to 39.(In Figure 8, this happens when n grows from 16 to 18). These results imply that: When there are more end vertices in G_n, n needs to get larger to offset the effect of end vertices or to enhance the effect of centroid. For other d and n in *broom* graph, as the proof of Theorem 1, we can prove this in the same way, so we conclude that, if we fix the number of end vertex, the probability $P(v_c|G_n)$ will greater than $P(v_e|G_n)$ when n is large enough.

B. Heuristic Algorithm for ML-Estimator in Tree with Multiple End Vertices

In this section, we are going to provide a heuristic algorithm to find where is the ML-estimator for the source on the finite

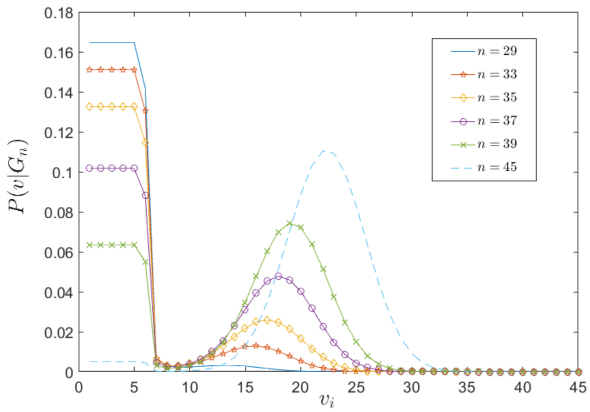


Fig. 9. Probability distribution of each node on G_n with 5 end verices where G is 6-regular, y-axis is the probability $P(v_i|G_n)$ and x-axis is vertex v_i's number i, $v_1 ... v_5$ are end vertices

regular tree G, this algorithm is based on all the observation and result given in previous sections. The following are the summary:

1) If there is only one end vertex v_e in G_n, then the ML-estimator is on the path from centroid v_c to v_e
2) If $G_n = G$, then for all $v_i \in G_n$, $P(v_i|G_n) = 1/n$
3) The greater $\frac{R(v_c,G_n)}{R(v_e,G_n)}$, the less influence of the end vertex.
4) Suppose G_n has q end vertices, there exists an n' such that if $n > n'$ then the $P(v_c|G_n) > \max_{1 \le i \le k} \{P(v_{e_i}|G_n)\}$, moreover n' will increase as q increases
5) If two vertices v_1 and v_2 are on the symmetric position of G_n, then $P(v_1|G_n) = P(v_2|G_n)$.

Property 1), 5) help us to find the location of the ML-estimator in the whole graph which is corresponding to the main part of the algorithm below. 3),4) shows the location of the ML-estimator on a subtree t_{ML} of G_n where t_{ML} is the subtree contains ML-estimator.

Algorithm 2 Message Passing Algorithm Computing ML-Estimator for a tree graph with Multiple End Vertices G_n

Input: G_n, $MLSET = \{\}$

Step 1: Find the Centroid v_c, by using the properties of centroid given in [9].

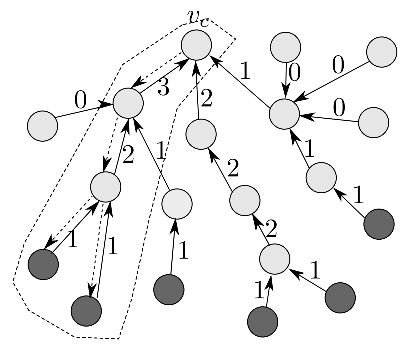
Step 2: Choose v_c as root, use the message passing algorithm to count the number of end vertices on each branch of v_c.

Step 3: Start a walk from v_c to end vertices, each time choose the child with maximum number of end vertex(if there were two children with the same number of end vertex, then choose both.) The route construct a subtree t_{ML} with root v_c.

Output: $MLSET = \{leaves\ of\ t_{ML}, v_c\}$

In Algorithm 2, the first step is to find the centroid of G_n, which can be done in $O(n)$ time. The detail of topological

Fig. 10. This figure illustrate how the algorithm works on a tree. t_{ML} is the subtree with five vertices in the dotted line.



properties of centroid in a graph can be found in [9]. The second step in Algorithm 2 is to compute the number of end vertices of each branch of the centroid. This can also be done in $O(n)$ time. The third step is to collect vertices on the subtree, where the ML-estimator is, which construct a subtree of G_n denoted as t_{ML}. Note that, t_{ML} in a graph with multiple end vertices is like the *path* from centroid to end vertex in a single end vertex graph. Finally, we will get a set $MLSET$.

Note that in the multiple end vertices problem: 1. The end vertex may be the ML-estimator 2. There may be some pairs of vertices (v_i, v_j) in the "symmetric" position of G_n, so by summary 5) we deduce that there maybe more than one vertices in $MLSET$, also based on the observation in previous section, we put centroid in to the set. (That's why the detection probability is very low when the underlying graph is finite.)

C. Simulation Results for Finite d-regular Tree Networks

We simulate the rumor spreading in the degree regular tree network G for d=3,4,5,6 with $|G| = 1000$ and $|G_n| = 100$. For each degree, we simulate 1,000 times rumor spreading on G, and compare the average performance of Algorithm 2 and the algorithm given in [2], [14]. Algorithm 2 output a set with k vetices. So base on the result given by [2], [14], we use the algorithm to find a set which contains k vertices with top-k rumor centrality in all vertices of G_n. The number $k = |MLSET|$ depends on the topology of G_n, and it's not a constant. Now define the error of a set S,

$$error(S) = min\{distance(v, source)|v \in S\}.$$

The simulation results are illustrated in the following. In Figure 11 and 12 shows the distribution of error hops in detail for both algorithm when underlying graph G is 3-regular and 4-regular respectively. We can deduce that algorithm 2 can make a good guess($P(error \le 1\ hop) > 75\%$) in most case, but there are still some situations that the error may be very large. For example, the case when the source lies in one end vertex but we guess other end vertices(and there distance is almost the diameter of the graph) as an ML-estimator. Table III shows that the average.

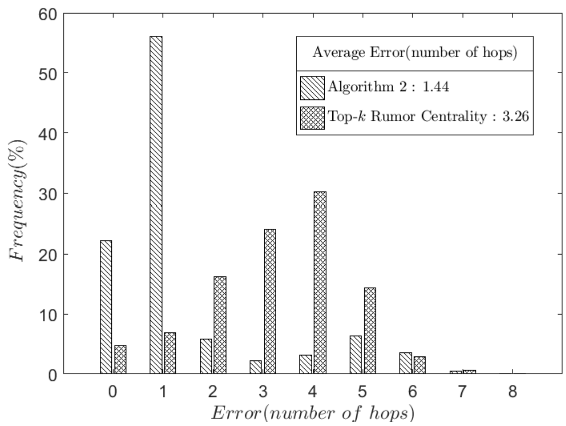


Fig. 11. Comparison of error distribution of algorithm 2 and the algorithm of top-k rumor centrality where G is a 3-regular finite graph

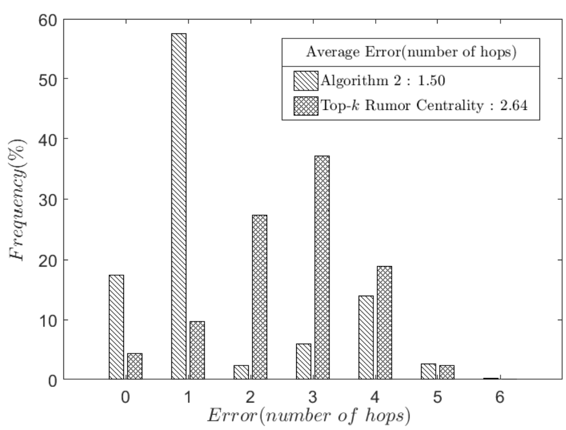


Fig. 12. Comparison of error distribution of algorithm 2 and the algorithm of top-k rumor centrality where G is a 4-regular finite graph

TABLE III
AVERAGE ERROR(HOPS) OF TWO ALGORITHMS

	Algorithm 2	Top-k Rumor Centrality
$d = 3$	1.44	3.26
$d = 4$	1.50	2.64
$d = 5$	1.48	2.36
$d = 6$	1.40	2.32

V. CONCLUSION

In this research, we study the rumor source detection problem for a finite degree-regular tree graph. First, we described the influence on $P(v|G_n)$ as "end vertex" appears in G_n. Then we consider the case with one single vertex, we demonstrated the unique property that the "most probable source" lies on the path from the centroid to the end vertex with an example when G_n is a line graph which is computable. Furthermore, we consider the case with multiple end vertices, starting with a computable example "line graph with star end vertices". Finally, we gave a heuristic massage passing algorithm based on the results in previous discussion to find the set of ML-estimator. This new algorithm turns out to outperform then the original one.

We know that both end vertices and centroid has influence on $P(v_c|G_n)$ but we don't know the exact relation between them. In fact, from Figure 9 and other numeric results, we believe that the ratio n'/q(when $n > n'$, the influence of centroid will exceed end vertices and q is the number of end vertices) can be computed for some specific graph. To find this "ration" can help us improve the performance of the algorithm.

REFERENCES

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

[2] S. Bubeck, L. Devroye and G. Lugosi, "Finding Adam in random growing trees," *arXiv.*, 2015.

[3] G. Fanti , P. Kairouz and S. Oh, "Rumor Source Obfuscation on Irregular Trees," *Proc. of ACM SIGMETRICS*, 2016.

[4] G. Fanti , P. Kairouz, S. Oh, K. Ramchandran and P. Viswanath "Hiding the Rumor Source," *arXiv.*, 2015.

[5] L. Vassio , F. Fagnani, P. Frasca and A. Ozdaglar, "Message Passing Optimization of Harmonic Influence Centrality," *IEEE Trans. On Control*, no. 1, vol. 1, pp. 109-120, 2014.

[6] N. Karamchandani, M. Franceschetti, "Rumor Source Detection under Probabilistic Sampling," *Proc. of IEEE ISIT*, 2013.

[7] B. Zelinka, "Medians and peripherians of trees," *Arch. Math.*, vol. 4, no. 2, pp. 87–95, 1968.

[8] C. W. Tan, P. D. Yu, C. K. Lai, H. L Fu "Optimal detection of influential spreaders in online social networks," *CISS*, 145-150, 2016.

[9] D. Shah and T. Zaman, "Detecting sources of computer viruses in networks: theory and experiment," *Proc. of ACM SIGMETRICS*, 2010.

[10] D. Shah and T. Zaman, "Rumors in a network: Whos's the culprit?," *IEEE Trans. on Information Theory*, no. 8, vol. 57, pp. 5163–5181, 2011.

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

[12] D. J. C. Mackay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, first ed., 2003.

[13] D. Shah and T. Zaman, "Rumor centrality: a universal source detector," *Proc. of ACM SIGMETRICS*, 2012.

[14] J. Khim and P. L. Loh, "Confidence Sets for the Source of a Diffusion in Regular Trees," *arXiv*, 2015.

[15] V. Jog and P. L. Loh, "Persistence of centrality in random growing trees," *arXive*, 2015.

[16] W. Dong, W. Zhang, and C. W. Tan, "Rooting out the rumor culprit from suspects," *Proc. of IEEE ISIT*, 2013.

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

[18] Z. Wang, W. Dong, W. Zhang, and C. W. Tan, "Rumor source detection with multiple observations: Fundamental limits and algorithms," *Proc. of ACM SIGMETRICS*, 2014.

[19] D. B. West, "Introduction to Graph Theory 2 ed," Prentice Hall, Upper Saddle River, New Jersy, 2011.

APPENDIX

A. Proof of Theorem1

Proof. G and G_n are as defined in Theorem 1, let $n = 2t+1$ and v_c represents the centroid of G_n, v_e represents the end vertex. For v_e,

$$\begin{aligned} P(v_e|G_n) &= m_{v_e}^{v_e}(G_n, 1) \cdot P_{v_e}^{v_e}(G_n, 1) \\ &= 1 \cdot P_{v_e}^{v_e}(G_n, 1) \\ &= \prod_{i=0}^{2t-1} \frac{1}{1 + i(d-2)}. \end{aligned}$$

For v_c, it's more simple to consider the last term of the Equation (3) only, that is,$P_{v_c}^{v_e}(G_n, n) \cdot m_{v_c}^{v_e}(G_n, n)$. Note that $m_{v_c}^{v_e}(G_n, n) = R(v_e, G'_n)$ where $G'_n = G_n \setminus \{v_e\}$. We have,

$$\begin{aligned} P_{v_c}^{v_e}(G_n, n) \cdot m_{v_c}^{v_e}(G_n, n) &= P_{v_c}^{v_e}(G_n, n) \cdot \frac{(2t)!}{2t(t-1)!t!} \\ &= \frac{(2t)!}{2t(t-1)!t!} \cdot \prod_{i=0}^{2t-1} \frac{1}{d + i(d-2)} \end{aligned}$$

Now we consider the ratio $\frac{P_{v_c}^{v_e}(G_n,n) \cdot m_{v_c}^{v_e}(G_n,n)}{P(v_e|G_n)}$, we have

$$\begin{aligned} &\frac{P_{v_c}^{v_e}(G_n, n) \cdot m_{v_c}^{v_e}(G_n, n)}{P(v_e|G_n)} \\ &= c_1 \cdot \frac{\Gamma(n-1)\Gamma(1 + \frac{1}{d-2} + n)\Gamma(3 + \frac{1}{d-2})}{\Gamma(n/2)\Gamma(2 + \frac{1}{d-2} + n)\Gamma(2 + \frac{1}{d-2})} \\ &= c_1 \cdot \frac{2 + \frac{1}{d-2}}{1 + \frac{1}{d-2} + n} \cdot \frac{\Gamma(n-1)}{\Gamma(n/2)} \end{aligned}$$

where c_1 is constant. The above result shows the ratio will larger than 1 when n is large enough. So we conclude that,

$$\frac{P(v_c|G_n)}{P(v_e|G_n)} > \frac{P_{v_c}^{v_e}(G_n,n) \cdot m_{v_c}^{v_e}(G_n,n)}{P(v_e|G_n)} > 1$$

when n is sufficiently large. □

B. Proof of Lemma 2

To prove Lemma 2, we need another lemma as following.

Lemma 3. Let X_i and Y_i be two non-strictly increasing sequence of numbers for $i = 1..n$, with properties $\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$ and $\exists h$ such that $X_i \geq Y_i$ when $i \leq h$, $X_i \leq Y_i$ when $i > h$. Then,

$$\sum_{i=1}^{L} X_i \geq \sum_{i=1}^{L} Y_i \text{ for } L = 1...n.$$

Proof. $\sum_{i=1}^{L} X_i \geq \sum_{i=1}^{L} Y_i$ is trivial when $L \leq h$. To contrary, suppose $\exists L_2 > h$ such that $\sum_{i=1}^{L_2} X_i < \sum_{i=1}^{L_2} Y_i$. Then we have $\sum_{i=1}^{n} X_i < \sum_{i=1}^{n} Y_i$, since $X_i \leq Y_i$ when $i > h$. This is a contradiction. □

Now, we can prove Lemma 2.

Proof. Consider the subgraph $G'_n = G_n \backslash \{v_e\}$ first. For simplicity, let $X_i = m_{v_c}^{p_{D-1}}(G'_n, i)$ and $Y_i = m_{v_p}^{p_{D-1}}(G'_n, i)$, then we have

$$\begin{aligned} X_D + X_{D+1} + ... + X_n &= R(v_c, G'_n) \\ Y_D + Y_{D+1} + ... + Y_n &= R(v_c, G'_n) \end{aligned}$$

Observe that $m_{v_c}^{v_e}(G_n, u) = \sum_{i=D}^{u-1} X_i$ and $m_{v_p}^{v_e}(G_n, u) = \sum_{i=D}^{y-1} Y_i$, for $u = D+1, D+2...n$. By Lemma 1, we know that both X_i and Y_i are non-strictly increasing sequence and by Lemma 3, we conclude that $\sum_{i=1}^{u} X_i \geq \sum_{i=1}^{u} Y_i$, that is, $m_{v_c}^{v_e}(G_n, u) \geq m_{v_p}^{v_e}(G_n, u)$, for $u = D+1, D+2...n$. □

We can extend the result of Lemma 2 to a more general case.

Lemma 4. G_n is the infected subgraph with a unique center v_c and an end vertex v_e under the regular tree G. Let $distance(v_c, v_e)$ denoted as D. v_p is any neighbor of v_c satisfies $distance(v_p, v_e) > D$. Then,

$$m_{v_c}^{v_e}(G_n, i) \geq m_{v_p}^{v_e}(G_n, i)$$

for $D + 1 \leq i \leq n$.

C. Proof of Theorem 2

Proof. Let G_n be a tree of size n. Suppose that v_c is the centroid in G_n and v_e is the end vertex in G_n. Assume that $d(v_c, v_e) = d$. Then, the path P from v_c to v_e is shown in Figure 5, where $v_1 = v_c$ and $v_{d+1} = v_e$. Define $D_i = \{ v \in G_n |$ the path from v to v_e contains $v_i \}$. Given $i = 1, 2, ..., d$, let $v \in D_i$, then we have $R(v, G_n) \leq R(v_i, G_n)$ (Suppose not, then v_1 is not the centroid, and we get a contradiction.) Therefore, each vertex in D_i has less spreading orders than v_i and are farther away (in hops) from v_e than v_i does. By Lemma 4, we deduce that $P(G_n|v_i) \geq P(G_n|v)$. Hence, the vertex with the maximum probability to spread the rumor to all the nodes in G_n is located on the path from v_1 to v_e. □

D. Proof of Theorem 3

Proof. Consider the ratio $\frac{P(v_c|G_n)}{P(v_e|G_n)}$,

$$\begin{aligned} \frac{P(G_n|v_c)}{P(G_n|v_e)} &= \frac{\sum_{k=2}^{n} m_{v_c}^{v_e}(G_n, k) \cdot P_{v_c}^{v_e}(G_n, k)}{R(v_e, G_n) \cdot P_{v_e}^{v_e}(G_n, 1)} \\ &\leq \frac{\sum_{k=2}^{n} m_{v_c}^{v_e}(G_n, k) \cdot P_{v_c}^{v_e}(G_n, 2)}{R(v_e, G_n) \cdot P_{v_e}^{v_e}(G_n, 1)} \\ &= \frac{R(v_c, G_n) \cdot P_{v_c}^{v_e}(G_n, 2)}{R(v_e, G_n) \cdot P_{v_e}^{v_e}(G_n, 1)} \\ &= \frac{(n-1)! \cdot \frac{1}{d} \cdot \frac{1}{d-1} \cdot \frac{1}{2d-3} \cdots}{(n-2)! \cdot \frac{1}{1} \cdot \frac{1}{d-1} \cdot \frac{1}{2d-3} \cdots} \\ &= \frac{n-1}{d} < 1 \end{aligned}$$

which implies that v_e is the ML-estimator of G_n. □