Epidemic Source Detection in Contact Tracing Networks: Epidemic Centrality in Graphs and Message-Passing Algorithms

Pei-Duo Yu†, Chee Wei Tan† and Hung-Lin Fu∗
Chung Yuan Christian University†, City University of Hong Kong‡, National Yang Ming Chiao Tung University∗
peiduoyu@cycu.edu.tw, cheewtan@gmail.com, hlfu@math.nctu.edu.tw

Abstract—We study the epidemic source detection problem in contact tracing networks modeled as a graph-constrained maximum likelihood estimation problem using the susceptible-infected model in epidemiology. Based on a snapshot observation of the infection subgraph, we first study finite degree regular graphs and regular graphs with cycles separately, thereby establishing a mathematical equivalence in maximal likelihood ratio between the case of finite acyclic graphs and that of cyclic graphs. In particular, we show that the optimal solution of the maximum likelihood estimator can be refined to distances on graphs based on a novel statistical distance centrality that captures the optimality of the nonconvex problem. An efficient contact tracing algorithm is then proposed to solve the general case of finite degree-regular graphs with multiple cycles. Our performance evaluation on a variety of graphs shows that our algorithms outperform the existing state-of-the-art heuristics using contact tracing data from the SARS-CoV 2003 and COVID-19 pandemics by correctly identifying the superspreaders on some of the largest superspreading infection clusters in Singapore and Taiwan.

I. INTRODUCTION

The COVID-19 coronavirus pandemic has revealed severe deficiencies in public health protection [1]. As the COVID-19 disease is highly contagious and wide-ranging with long incubation periods (transmission rate of 3-5 persons within 6 feet), it becomes necessary to track down all the infected persons and their recent contacts once an outbreak has occurred. It is often necessary to account for the initial source of the outbreak (e.g., identification of superspreaders [2, 3]) so that public health is resilient against further outbreaks or to understand the underlying cause of secondary transmissions. Public health authorities worldwide employ contact tracing to address these problems [1, 4–10].

In general, manual contact tracing is a complex and tedious process: Once a person has been diagnosed as infected, public health authorities fan out to trace the recent contacts of this person for the purpose of monitoring or quarantine. This process repeats if one of those contacts exhibits symptoms until all the contacts who have been exposed are out of circulation. The COVID-19 coronavirus epidemic has overwhelmed most contact tracing capabilities due to the speed and scale of infection [7]. For contact tracing to be efficient, especially to identify superspreaders in recurrent large-scale outbreaks during the COVID-19 pandemic, efficient and scalable contact tracing algorithms need to be developed [1, 4–10].

Recently, there are public health protection schemes that employ mobile software technologies to use wireless signals (e.g., Bluetooth) to collect data on social contact connectivity for digital contact tracing in epidemiology [4, 10, 11]. These data lead to contact tracing networks that are essentially large graphs generated by stochastic processes. From an epidemiology perspective, all the infected persons in a contact tracing graph are potential candidates for tracking purposes as well as identification of Patient Zero or superspreaders [2, 3]. A fundamental question in digital contact tracing is how to unravel stochastic spreading processes to find the initial outbreak source quickly, accurately and reliably with high confidence by exploiting the topological and statistical properties of contact tracing networks.

The spreading of epidemics and rumors share much in common as stochastic processes in mathematical epidemiology [12, 13]. Formulating the source detection as solving a maximum likelihood estimation problem was first studied in [14] using a network centrality called the rumor centrality to optimally solve a special case of degree-regular tree graph with countably infinite number of vertices and assuming a SI (susceptible-infectious) spreading model. There were a number of problem extensions subsequently, e.g., random increasing trees in [15], probabilistic sampling in [16], star graph topology in [17], multiple sources or observations in [18–20], Markov chain Monte Carlo based algorithms in [21] and probabilistic characterization of infection network boundaries in [22]. Source estimation for other spreading models include the Susceptible-Infected-Recovered model in [23] and a random branching process for irregular trees in [15, 24]. Related work on proactive network protection against epidemics include [25–27], e.g., a maximum likelihood algorithm in [27] that complements the work in [26]. How to optimally place observations concerning the efficiency and the cost and design message-passing algorithms were studied in [25]. Probabilistic inference methods that use machine learning techniques like graph neural networks and deep learning for epidemiology have been studied in [1, 10, 28, 29].

To the best of our knowledge, there is no prior work that solves the problem in [14] when the network has cycles or
finite graph boundaries even under the SI spreading model. This more general problem is computationally challenging due to the presence of irregular vertices (vertices without susceptible neighbors) and cycles. Irregular vertices are practical for modeling actual graph connectivity or users who are quarantined after being infected. In addition, the presence of cycles cannot be ignored. In essence, unlike the tree graphs, the cycles allow the stochastic spreading process to spread through multiple alternate paths, thus increasing the likelihood of those vertices on the cycle. As such, the presence of cycles and irregular vertices introduce nontrivial irregular effects that significantly shape the way that virus spread as well as the performance of detecting the original source on the infected graph. To be exact, existing algorithms in the literature, e.g., [14], [30]–[32], are no longer optimal even with the presence of a single cycle in degree-regular pseudo-tree graphs.

In this paper, we address the outstanding issues of epidemic source detection for the general cases when there are irregular vertices and cycles. In particular, we demonstrate that the likelihood ratios between vertices of the infection graph provide insights to locating the most likely source, which leads to a new network centrality connecting graph-theoretic structures with estimation performance and enables low-complexity algorithms using the same spreading model as those studied in the literature [14], [16], [30], [31], [33].

A. Our Contributions

The main contributions are summarized as follows:

- We consider a general network (i.e., finite large network with cycles) for epidemic source detection, and analytically characterize how graph distances between vertices in an infection graph and a single cycle or irregular vertex affect the likelihood of each vertex being the source.
- For a finite size degree-regular tree, we characterize the globally optimal maximum likelihood estimator based on graph distances and propose a scalable message-passing algorithm to compute this solution in linear time.
- For a degree-regular graph with cycles, we prove that its epidemic center can be equal to the epidemic center of any of its spanning trees. We propose a polynomial-time algorithm to find the epidemic center of a graph with a single cycle by characterizing the optimal maximum likelihood estimator based on the graph distances.
- We combine the results in finite-size degree regular graphs and unicyclic graphs and propose a polynomial-time distance-based algorithm to find the source estimator. For synthetic data, we conduct simulations on two finite-size regular graphs with cycles which are grid graphs and circulant graphs. We evaluate the performance by comparing the results with the BFS heuristic rumor centrality [14], showing that the error (number of hops) of our source estimator is at least 50% smaller than BFS rumor center. Using real-world contact tracing data (due to SARS-CoV 2003 and COVID-19 pandemic), we apply our algorithm on some of the largest superspreading infection clusters in Singapore and Taiwan, correctly identifying the superspreader in this cluster.

II. PRELIMINARIES ON VIRUS SPREADING MODEL

We model a contact network by an undirected graph $G = (V, E)$, where the set of vertices $V$ represents the vertices in the underlying network, and the set of edges $E$ represents the links between the vertices. We shall assume that $V$ is countably finite (this is the crucial departing point from the previous assumption of infinite graph in the literature [14], [18], [30]–[32]). In this paper, we use the Susceptible-Infected (SI) model in [12], [13], [34] to model virus spreading. Vertices that infected by the virus are called infected vertices and otherwise they are susceptible vertices. The spreading is initiated by a single vertex $v^* \in V$ that we call the source. Once a vertex is infected, it stays infected and can in turn infect its susceptible neighbors. The virus can be spread from vertex $i$ to vertex $j$ if and only if there is an edge between them (i.e., $(i, j) \in E$). Let $\tau_{ij}$ be the spreading time from $i$ to $j$, which are random variables that are independently and exponentially distributed with parameter $\lambda$ (without loss of generality, let $\lambda = 1$). Let $S$ denote the set of all susceptible vertices that have at least one infected neighbor, i.e., those vertices in $S$ might be infected in the near future. In the real world there are some people that are more likely to be infected by the virus, and some are more likely to spread the virus to others. We can assume that each person has two parameters say $R_i$ and $R_u$ which are corresponding to the rate of being infected and the rate of spreading the virus to others respectively. Due to the memoryless property of the exponential distribution. We have the fact that each newly infected vertex $v$ is randomly chosen from $S$ with the probability that

$$P(\text{v is infected}) \propto R_i \cdot \sum_u R_u,$$

where each $u$ is an infected neighbor of $v$.$R_i^c$ denote the infected rate of $v$ and $R_u$ denote the spreading rate of $u$. Hence, the probability of a vertex $v_a$ being infected in the next time period is defined as

$$P(\text{\text{v}_a is infected}) = \frac{R_i^v \cdot \sum_u R_u}{\sum_{v \in S} (R_i^v \cdot \sum_u R_u)};$$

where $u_a$ and $u$ represents each infected neighbor of $v_a$ and $v$ respectively.

Now, we have a random spreading model over an underlying finite graph $G$. We can view $G$ as a contact network [35]–[38] where nodes represent individuals and edges represent social contacts such as two individuals have been in the same place or in close contact (within about 6 feet). Contact tracing is a process to identify the people who have been in contact with an infected individual [10]. Hence, an epidemic contact tracing network can be seen as a connected subgraph of a human contact network containing infected people. Let $G_n$ be a subgraph of order $n$ of $G$, that models a snapshot observation of the spreading when there are $n$ infected vertices, i.e., $G_n$ is a contact tracing network in $G$ and $|G_n| = n$. In the following, we shall call $G_n$ an infected subgraph of $G$ for brevity. We denote the actual source in $G_n$ as $v^*$. The epidemic source detection problem is thus to find $v^*$ given this observation of $G_n$. For example, in Fig. [1] the infected
Infinite-size

No

No

No

On the path from $v_c$ to $v_{ir}$

On the path from $v_c$ to $C_h$

Theorems 1, 2

Theorems 3, 4

TABLE I

We summarize the main results of this paper with a comparison with prior art in [14] in this table. Assume the underlying network $G$ is a degree-regular tree. We denote the irregular vertex as $v_{ir}$ and the epidemic center as $v_c$. Note that, the case that $G_n$ containing an irregular vertex can be treated mathematically as a special case of $G_n$ containing a single cycle, since an irregular vertex can be treated as a size one cycle which is on the boundary of the graph. Therefore, the logical flow of this paper is such that we first consider the irregular effect caused by an irregular vertex and then analysis to graphs with cycles.

In particular, for a $d$-regular tree, we have [14]:

$$P(\sigma_i|v) = \prod_{k=1}^{n-1} \frac{1}{dk - 2(k-1)}.$$  (5)

Now, if the spreading has not reached the irregular vertices, then $P(\sigma_i|v) = P(\sigma_j|v)$ for all $\sigma_i, \sigma_j \in M(v,G_n)$. By combining (4) and (5), we have

$$P(G_n|v) = \sum_{\sigma_i \in M(v,G_n)} P(\sigma_i|v) = |M(v,G_n)| \cdot P(\sigma|v) \quad \forall \sigma_i \in M(v,G_n)$$

$$= |M(v,G_n)| \cdot \prod_{k=1}^{n-1} \frac{1}{dk - 2(k-1)},$$

which means that $P(G_n|v)$ is proportional to $|M(v,G_n)|$. If we treat the rooted tree as a partially ordered set, then a spreading order is a linear extension corresponding to this poset. Hence, the quantity $|M(v,G_n)|$ is actually the number all linear extensions of the poset $G_n$ rooting at $v$. The authors in [14] called $|M(v,G_n)|$ rumor centrality, which is crucial to solving the maximum likelihood estimation for degree-regular trees. We denote the vertex having the maximum $|M(v,G_n)|$ among all vertices in $G_n$ as $v_c$, that is

$$v_c = \arg\max_{v \in G_n} |M(v,G_n)|.$$  (6)

In particular, $v_c$ is called the rumor center when $G_n$ is a tree in [14] and the authors in [39], [40] established its equivalence to the graph distance center and tree centroid. When $G_n$ is a general graph, rumor center is only defined on the BFS spanning tree of $G_n$. To avoid confusion, in this paper we call the quantity $|M(v,G_n)|$ epidemic centrality and $v_c$ epidemic center for a general graph.

**Definition II.1.** For a given infection subgraph $G_n$ over the underlying graph $G$, $\hat{v}$ is an maximum likelihood estimator for the epidemic source in $G_n$, i.e., $P(G_n|\hat{v}) = \max_{v \in G_n} P(G_n|v)$.

In the remaining part of this section, we review the maximum likelihood estimation problem of this epidemic source in the simplest case: regular-tree networks. By Bayes’ theorem, $P(G_n|v)$ is the probability that $v$ is the actual epidemic infection source that leads to observing $G_n$. Now, let $\sigma_i$ be the possible spreading order starting from $v$, and let $M(v,G_n)$ be the collection of all $\sigma_i$ when $v$ is the source in $G_n$. Then, we have

$$P(G_n|v) = \sum_{\sigma_i \in M(v,G_n)} P(\sigma_i|v).$$  (4)

In particular, for a $d$-regular tree, we have [14]:

$$P(\sigma_i|v) = \prod_{k=1}^{n-1} \frac{1}{dk - 2(k-1)}.$$  (5)

Now, if the spreading has not reached the irregular vertices, then $P(\sigma_i|v) = P(\sigma_j|v)$ for all $\sigma_i, \sigma_j \in M(v,G_n)$. By combining (4) and (5), we have

$$P(G_n|v) = \sum_{\sigma_i \in M(v,G_n)} P(\sigma_i|v) = |M(v,G_n)| \cdot P(\sigma|v) \quad \forall \sigma_i \in M(v,G_n)$$

$$= |M(v,G_n)| \cdot \prod_{k=1}^{n-1} \frac{1}{dk - 2(k-1)},$$

which means that $P(G_n|v)$ is proportional to $|M(v,G_n)|$. If we treat the rooted tree as a partially ordered set, then a spreading order is a linear extension corresponding to this poset. Hence, the quantity $|M(v,G_n)|$ is actually the number all linear extensions of the poset $G_n$ rooting at $v$. The authors in [14] called $|M(v,G_n)|$ rumor centrality, which is crucial to solving the maximum likelihood estimation for degree-regular trees. We denote the vertex having the maximum $|M(v,G_n)|$ among all vertices in $G_n$ as $v_c$, that is

$$v_c = \arg\max_{v \in G_n} |M(v,G_n)|.$$  (6)

In particular, $v_c$ is called the rumor center when $G_n$ is a tree in [14] and the authors in [39], [40] established its equivalence to the graph distance center and tree centroid. When $G_n$ is a general graph, rumor center is only defined on the BFS spanning tree of $G_n$. To avoid confusion, in this paper we call the quantity $|M(v,G_n)|$ epidemic centrality and $v_c$ epidemic center for a general graph.

**Definition II.2.** For a graph $G$ and a vertex $v \in G$, the distance centrality of $v$ can be defined as

$$\text{distance centrality of } v = \sum_{u \in G} d(u,v),$$

where $d(u,v)$ is the shortest path distance, i.e., the number of edges along the shortest path from $u$ to $v$. The distance center of a graph is the vertex with the smallest distance centrality.
III. TREES WITH A SINGLE DEGREE-ONE IRREGULAR VERTEX

In this section we study the effect on the maximum likelihood due to an irregular vertex in a regular tree network. Assuming that the original underlying network is a regular tree, this irregular vertex is a vertex with different number of neighborhoods compared to other vertices. For example, in a regular tree $G$ of bounded size, if the infected subgraph $G_n$ contains a leaf $v$ of $G$, then $v$ is an irregular vertex in $G$ since it has no other neighbors except its parent vertex. This models a person under quarantine in a contact tracing network with fewer neighborhoods than other people.

We consider the case where the degree of the irregular vertex is less than that of all other vertices. Note that a degree-one irregular vertex is a leaf of both $G$ and $G_n$. Now, assume that the infection subgraph $G_n$ only has a single irregular vertex denoted as $v_{ir}$, where $\deg(v_{ir}) < d$ in $G$. Take Fig. 1 for example, we have $v_{ir} = v_5$ since $\deg(v_{ir}) = 2$ and other vertices are of degree 3 in $G$. In the following, we study how $v_{ir}$ affects the maximum-likelihood estimation performance.

In particular, we compare this single irregular vertex special case with a naive prediction that assumes an underlying infinite graph. This illustrates that ignoring the irregular effect in the finite graph ultimately leads to a wrong estimate and thus requires an in-depth analysis and new epidemic source detection algorithm design for the general case of finite graphs.

A. Impact of the Irregularity On $P(G_n|v)$

Example 1. Let $G$ be a infinite 3-regular tree except $v_5$ has a different degree, and $G_0 = G$ is a subgraph of $G$ shown in Fig. 1. Consider $P(G_0|v_1)$ and with a spreading order $\sigma : v_1 \rightarrow v_2 \rightarrow v_5 \rightarrow v_3 \rightarrow v_4 \rightarrow v_6$, we have $P(\sigma|v_1) = (1/3) \cdot (1/4) \cdot (1/4) \cdot (1/5) \cdot (1/6)$. Had $v_5$ not been the irregular vertex, then $P(\sigma|v_1) = (1/3) \cdot (1/4) \cdot (1/5) \cdot (1/6) \cdot (1/7)$. This demonstrates that the order at which the virus spreads to the irregular vertex $v_5$ is important when computing $P(\sigma|v_1)$. In particular, $P(G_0|v_1) \approx 0.0149$. We also have $P(G_0|v_2) \approx 0.0114$. Now, observe that $v_1$ and $v_2$ are two vertices with the largest epidemic centrality among all vertices in $G_0$, but $P(G_0|v_1) > P(G_0|v_2)$, and thus $\hat{v} = v_1$. Note that, the likelihood of $v_5$ being the source is also greater than that of $v_2$ even $v_2$ has a greater epidemic centrality.

Example 1 reveals some interesting properties of irregular effects due to even a single irregular vertex:

- $P(\sigma_1|v)$ increases with how soon the irregular vertex appears in $\sigma$ (as ordered from left to right of $\sigma$).
- When there is at least one irregular vertex in $G_n$, then $P(G_n|v)$ is no longer proportional to $|M(v, G_n)|$.

This means that $P(\sigma_i|v)$ is no longer a constant for each $i$, and depends on the position of the irregular vertex in each spreading order. We proceed to compute $P(\sigma_i|v)$ as follows. For brevity of notation, let $v_{ir}$ be the irregular vertex and let

$$M_{v_{ir}}^v(G_n, k) = \{ \sigma \in M(v, G_n) | \sigma(v_{ir}) = k \}$$

where $M_{v_{ir}}^v(G_n, k)$ is the set of all the spreading orders starting from $v$ and with $v_{ir}$ at the $k$th position, and its size is the combinatorial object of interest:

$$m_{v_{ir}}^v(G_n, k) = |M_{v_{ir}}^v(G_n, k)|.$$ (7)

Let $D$ be the distance (in terms of number of hops) from $v$ to $v_{ir}$. Then we have

$$|M(v, G_n)| = \sum_{k=D+1}^{n-t_{v_{ir}}+1} m_{v_{ir}}^v(G_n, k).$$ (8)

Now, (8) shows that $M(v, G_n)$ can be decomposed into $M_{v_{ir}}^v(G_n, k)$ for $k = D+1, D+2, \ldots, n-t_{v_{ir}}+1$. This decomposition allows us to handle the irregular effect due to the different position of the irregular vertex in each spreading order. For example, in Table I we list out all $m_{v_{ir}}^v(G_n, k)$ for each possible starting vertex $v$ and the position $k$ of $v_{ir}$. Let $P_{v_{ir}}^v(G_n, k)$ be the corresponding probability for each $k$. We can rewrite $P(G_n|v)$ for the case with single irregularity as:

$$P(G_n|v) = \sum_{k=D+1}^{n-t_{v_{ir}}+1} m_{v_{ir}}^v(G_n, k) \cdot P_{v_{ir}}^v(G_n, k).$$ (9)

Thus, solving (3) means finding the vertex $\hat{v}$ that solves

$$P(G_n|\hat{v}) = \max_{v_i \in G_n} P(G_n|v_i).$$ (10)

Since $P(G_n|v)$ is no longer proportional to $|M(v, G_n)|$, we now describe how to compute $P(G_n|v)$ in $G_n$ over an
underlying $d$-regular tree except the degree of $v_{ir}$ is $d'$. First, consider $P_{v}^{ir}(G_n, k)$ and let $z_d(i) = (i-1)(d-2)$, then
\[
P_{v}^{ir}(G_n, k) = \prod_{i=1}^{k-1} \frac{1}{d+z_d(i)} \cdot \prod_{i=k-1}^{n-2} \frac{1}{d+z_d(i)+(d'-2)}.
\]
(11)

where the first factor of $P_{v}^{ir}(G_n, k)$ in (11) is the probability that $k$ vertices are infected once the virus reaches the irregular vertex, i.e., $v_{ir}$ is the $k$th vertex infected in $G_n$, and the second factor is the probability that all remaining $n-k$ vertices are infected thereafter. On the other hand, the value of $m_{vi}^{ir}(G_n, k)$ in (7) is dependent on the network topology, and thus there is no closed-form expression in general (though when $G_n$ is a line, a closed-form expression for $m_{vi}^{ir}(G_n, k)$ is given in (12)). We now use a special case, line graph, to demonstrate how an irregular vertex and network topology affect the probability $P(G_n|v)$.

B. Analytical Characterization of Likelihood Function

Suppose $G$ is a finite degree-regular tree and $G_n$ is a line graph with a single irregular vertex due to the bounded size of $G$. Without loss of generality, suppose $n$ is odd to ensure a unique $v_i$ and $n = 2t+1$ for some $t$. We label all the vertices in $G_n$ from 1 to $2t+1$ and assume that $v_{2t+1}$ is the end vertex, i.e., irregular vertex with degree 1. To compute $P(G_n|v_i)$ for $v_i \in G_n$ from (7) and (11), we already have $P_{v_i}^{ir}(G_n, k)$, so we need to compute $m_{vi}^{ir}(G_n, k)$. The enumeration of $m_{vi}^{ir}(G_n, k)$ can be accomplished in polynomial-time complexity with a path-counting message-passing algorithm (see, e.g., Chapter 16 in [41]). In particular, we have a closed-form expression for $m_{vi}^{ir}(G_n, k)$ given by:

\[
m_{vi}^{ir}(G_n, k) = \begin{cases} \frac{k-2}{k-n+i-1}, & \text{if } i \neq n, \end{cases}
\]
(12)

when $i \neq n$, leading to an analytical formula for $P(G_n|v_i)$:

\[
P(G_n|v_i) = \begin{cases} \prod_{l=1}^{n-1} \frac{1}{z_d(l) + 1}, & \text{if } i = n; \\
\sum_{k=n-i+1}^{n} \frac{k-2}{k-n+i-1} \cdot P_{v_i}^{ir}(G_n, k), & \text{otherwise}, \end{cases}
\]
(13)

where $P_{v_i}^{ir}(G_n, k)$ is given in (11).

In (13), we suppose that $n$ is odd. Using (13), let us numerically compute $P(G_n|v_i)$ for all $v_i$ in Fig. 2. where $G$ is a 4-regular tree and $G_n$ is a line graph with a single irregular vertex $v_{ir} = v_n$ as boundary values for $n = 7, 8, 9, 10$. The $x$-axis is the vertex $v_i$ where $i = 1, 2, \ldots, 10$, and the $y$-axis plots $P(v_i = v^*|G_n)$. Fig. 2 illustrates that the influence due to the irregular vertex on $P(v_i = v^*|G_n)$ dominates that of the epidemic center when $n = 7, 8, 9$. However, the situation reverses when $n = 10$ (i.e., the epidemic center on $P(G_n|v_i)$ is dominant thereafter).

Theorem 1. Suppose $G$ is a $d$-regular graph ($d > 2$) with finite order. If $G_n$ is a line-graph with a single irregular vertex at one end of the line graph, then there is a constant $j$ such that $P(G_n|v_c) > P(G_n|v_{ir})$ when $n > j$.

Remark: When $n$ increases, i.e., the distance between $v_c$ and $v_{ir}$ increases, then the location of $d$ in $G_n$ converges to the neighborhood of the epidemic center.

Example 2. To verify Theorem 1 we plot $P(G_n|v_i)$ for a line graph $G_n$ with $G$ being a finite 4-regular graph in Fig. 2. Clearly, we have $j = 9$.

Theorem 1 implies that, for any $d$-regular underlying graph, when $G_n$ is a line graph with a single irregular vertex, the influence of the irregular vertex $v_{ir}$ on $P(v_i|G_n)$ decreases monotonically as $n$ grows. In fact, this reduces to the special case in (20), when $n$ goes to infinity asymptotically, i.e., $\hat{v}$ is the ML estimation of the source.

C. Optimality Characterization of Likelihood Estimate

Example 1 reveals that in addition to the spreading order and the location of the epidemic center on $P(G_n|v_i)$, the dependence of $P(G_n|v_i)$ on $v_{ir}$ also affects the likelihood probability $P(G_n|v)$. We can conclude that there are two factors affect the likelihood of $v$ being the source, one is the distance $d(v, v_{ir})$ and the other one is the epidemic centrality $|M(v, G_n)|$. Intuitively, if we consider two vertices in $G_n$ say $v_a$ and $v_b$ where $|M(v_a, G_n)| > |M(v_b, G_n)|$ and $d(v_a, v_{ir}) < d(v_b, v_{ir})$, then the above observation may lead us to $P(G_n|v_a) > P(G_n|v_b)$. We formalize this optimality result that characterizes the probabilistic inference performance between any two vertices and the location of $\hat{v}$ in $G_n$ with a single irregular vertex $v_{ir}$ in the next theorem.

Lemma 1. Let $G$ be a tree of size $n$, for any three vertices say $v_a$, $v_{ir}$ and $v_{ir}$, if one of the following conditions is satisfied
1) $|M(v_a, G_n)| \geq |M(v_b, G_n)|$ and $d(v_a, v_{ir}) < d(v_b, v_{ir})$
2) $|M(v_b, G_n)| > |M(v_a, G_n)|$ and $d(v_a, v_{ir}) \leq d(v_b, v_{ir})$,
then we have

\[
\sum_{i=d(v_a, v_{ir})+1}^{k} m_{vi}^{ir}(G_n, k) \geq \sum_{i=d(v_a, v_{ir})+1}^{k} m_{vi}^{ir}(G_n, i),
\]

Fig. 2. $P(G_n|v)$, where $G_n$ is a line graph with a single irregular vertex $v_1$ over an underlying 4-regular finite graph.
for all possible \( k \).

Remark: Lemma \( \text{I} \) applies for any positive degree of \( v_{ir} \). We can verify Lemma \( \text{I} \) by Fig. 1. Let \( v_a = v_1, v_b = v_2 \) and \( v_{ir} = v_5 \), then \( |M(v_1,G_3)| = |M(v_2,G_3)| = 20 \) and \( d(v_1,v_5) < d(v_2,v_5) \) which satisfies the first condition of Lemma \( \text{I} \). Hence, the partial sum of \( |M(v_1,G_3)| \) is always greater or equal to the partial sum of \( |M(v_2,G_3)| \).

**Theorem 2.** Let \( G \) be a \( d \)-regular tree with a single irregular vertex and \( G_n \subseteq G \) be a subtree of \( G \) with a single irregular vertex \( v_{ir} \) in \( G_n \), where \( \text{deg}(v_{ir}) < d \). Then, the maximum likelihood estimator \( \hat{v} \) with maximum probability \( P(G_n|v) \) is located on the path from the \( v_c \) to \( v_{ir} \).

The above theorem can be immediately deduced from Lemma \( \text{I} \). In addition, we can leverage Lemma \( \text{I} \) in the case when \( \text{deg}(v_{ir}) \) is greater than \( d \) to reduce the search space of the optimization problem \( \text{II} \). We can conclude only that the ML estimator is on the path from \( v_c \) to \( v_{ir} \), which is illustrated in Fig. 1 thus narrowing down the search for the ML estimator to a vertex on this path.

### IV. Trees with Multiple Degree-one Irregular Vertices

In this section, we consider the case when \( G_n \) has more than one irregular vertex (naturally, this also means \( d > 2 \) in \( G \) ruling out the trivial case of \( G \) being a line). The key insight from the single irregular vertex analysis still holds: Once the virus reaches an irregular vertex in \( G \), \( \hat{v} \) can be located near this very first infected irregular vertex. In addition, the algorithm design approach is to decompose the graph into subtrees to narrow the search for the maximum-likelihood estimate solution. To better understand the difficulty of solving the general case, we start with a special case: The entire finite underlying network is infected, i.e., \( G_n = G \), then \( P(G_n|v) = 1/n \) for each vertex in \( G_n \), as each vertex is equally likely to spread the virus to all the other vertices in \( G \) to yield \( G_n = G \). In this case, \( P(G_n|\hat{v}) \) is exactly the minimum detection probability. So the bound of \( P(G_n|\hat{v}) \) given in previous study are not suitable for the case with irregular vertices. Therefore, when simulating the virus spreading in a network, we will set an upper bound \( n/k \) of the number of irregular vertices where \( k \) is some integer greater than 1, once the number of irregular vertices in \( G_n \) reaches to \( \lceil n/k \rceil \), then we will stop the spreading process.

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

In Section III we have shown that, when \( G_n \) is sufficiently large, the effect of the single irregular vertex on \( P(v|G_n) \) for each vertex \( v \) on the line graph \( G_n \) is dominated by \( v_c \). Now, we study the effect of multiple degree-one irregular vertices on a class of graphs whose topology is richer than the line graph in Section III. In particular, as shown in Fig. 1 we add degree-one irregular vertices to \( v_2 \), so that when \( G \) is \( d \)-regular, there will be at most \( d - 1 \) degree-one irregular vertices in \( G_n \). We call this the broom graph. We can compute \( P(v|G_n) \) by extending the result in Section III. Let \( P_{v_{i1},v_{i2},...,v_{ik}}(\{h_1,h_2,...,h_k\},G_n) \) be the probability of the spreading order starting from \( v_i \) with the irregular vertex set \( \{u_1,u_2,...,u_k\} \) and their position set \( \{h_1,h_2,...,h_k\} \) in this spreading order. We do not assume that \( h_i \) is the position of \( u_i \), as it can be the position of any irregular vertex in \( G_n \). The probability \( P_{v_{i1},v_{i2},...,v_{ik}}(\{h_1,h_2,...,h_k\},G_n) \) can be obtained by the same analysis in III. To compute \( m_{v_{i1},v_{i2},...,v_{ik}}^{v_{ir}}(\{h_1,h_2,...,h_k\},G_n) \), we first consider the line-shaped part of \( G_n \), i.e., the part \( \{v_{1},v_{2},...,v_{2t}\} \), say \( G_{2t} \). From the previous discussion, we have

\[
m_{v_{i1}}^{v_{ir}}(G_{2t},j) = \binom{2t-i}{j},
\]

and for each spreading order that \( v_{2t} \) lies on the \( j \)th position, the degree-one irregular vertices \( u_1,u_2,...,u_k \) can be placed on any position after the \( j \)th position. So for each spreading order in \( m_{v_{i1}}^{v_{ir}}(G_{2t},j) \), there are \( k! \cdot \binom{n-k-j+1}{k} \) corresponding spreading orders in \( G_n \). Thus, we have

\[
m_{v_{i1},...,v_{ik}}(G_n,\{h_1,\ldots,h_k\}) = k! \sum_{j=2t-i+1}^{h_i-1} \binom{j-2}{2t-i-1}.
\]

(14)

With \( P_{v_{i1},v_{i2},...,v_{ik}}(v,c) \) and \( m_{v_{i1},v_{i2},...,v_{ik}}(v,c) \), we can now compute the probability \( P(v|G_n) \) by going through all possible \( \{h_1,h_2,...,h_k\} \). Fig. 5 shows that even though there are five irregular vertices, the effect of \( v_c \) on \( P(v|G_n) \) eventually dominates that of the irregular vertices as \( n \) grows from 37 to 39. These result implies that: When there are more degree-one irregular vertices in \( G_n \), \( n \) needs to be sufficiently large to offset the effect of irregular vertices, i.e., for the transition.
phenomenon to take place. For other $d$ and $n$ in the broom graph, as shown in the proof of Theorem 1, we can prove this in the same way to conclude that, if we fix the number of irregular vertex, the probability $P(v_c|G_n)$ will be greater than $P(v_{ir}|G_n)$ when $n$ is large enough.

### B. Message-passing Algorithm

We propose a message-passing algorithm to find $\hat{v}$ on the finite regular tree $G$ by leveraging the key insights derived in the previous sections. We summarize these features as follows:

1) If there is only a single irregular vertex $v_{ir}$ in $G_n$, then $\hat{v}$ is located on the path from $v_c$ to $v_{ir}$.
2) If $G_n = G$, then for all $v_i \in G_n$, $P(G_n|v_i) = 1/n$.
3) If $G_n$ has $q$ degree-one irregular vertices, then there exists an $n'$ such that, if $n > n'$, then $P(G_n|v_c) > \max_{1 \leq i \leq k} P(G_n|v_{ir})$. Furthermore, $n'$ increases as $q$ increases.
4) If two vertices $v_1$ and $v_2$ are on the symmetric position of $G_n$, then $P(G_n|v_1) = P(G_n|v_2)$. For example, $u_1, u_2, \ldots, u_k$ are topologically symmetric in Fig. 4.

In particular, Feature 1 is the optimality result related to decomposing $G_n$ into subtrees to search for $\hat{v}$. The subtree $t_{ML}$ in $G_n$ corresponds to first finding the decomposed subtree containing $v_c$ and the likelihood estimate needed for Theorem 2 to apply. Then, Features 3 and 4 identify $\hat{v}$ on a subtree $t_{ML}$ of $G_n$ as Theorem 2 only pinpoints the relative position of $\hat{v}$.

Algorithm 1 first finds the epidemic center of $G_n$ and then determines the number of irregular vertices corresponding to each branch of the epidemic center $v_c$. The final step is to collect vertices on the subtree where $\hat{v}$ is, leading to a subtree of $G_n$ denoted as $t_{ML}$. Observe that each step requires $O(n)$ computational time complexity and $t_{ML}$ in a graph with multiple irregular vertices is akin to the path from $v_c$ to the irregular vertex in an infected subgraph with a single irregular vertex in Section III. Finally, we obtain a set $\kappa$ containing the parent vertices of the leaves of $t_{ML}$ and $v_c$.

### Algorithm 1 Message-passing algorithm to compute $\hat{v}$ for $G_n$ with multiple irregular vertices

**Input:** $G_n, \kappa = \{\}$

1. Compute $v_c$ of $G_n$.
2. Choose $v_c$ as the root of a tree and use a message-passing algorithm to count the number of irregular vertices on each branch of this rooted tree.
3. Starting from $v_c$, and at each hop choose the child with the maximum number of irregular vertices (if there were more children with the same maximal number of irregular vertices, then choose all of them). This tree traversal yields a subtree $t_{ML}$ rooted at $v_c$.

**Output:** $\kappa = \{\text{parent vertices of leaves of } t_{ML}, v_c\}$

Next, we use the example in Fig. 6 to illustrate Algorithm 1. Let $G_{19}$ be the network in Fig. 6 with the six degree-one irregular vertices depicted as more shaded. Suppose $v_c$ is determined by the end of Step 1. Then, Step 2 enumerates the number of irregular vertices at each branch of the subtrees connected to $v_c$, and these numbers are then passed iteratively from the leaves to $v_c$. These messages correspond to the numerical value on the edges in Fig. 6. The message in Step 2 is an upward (leaf-to-root) message. Step 3 is a message passing procedure from $v_c$ back to the leaves, which is a downward message, and the message is the maximum of number of irregular vertices in each branch. For example, the message from $v_c$ to $\text{child}(v_c)$ is $\max\{1, 2, 3\}$ which is 3.

Lastly, the second part of Step 3 collects those vertices whose upward message = downward message. For example, the left hand-side child of $v_c$ is first added to $t_{ML}$, and then $v_i$ is added to $t_{ML}$, and finally, the two leaves on the left hand-side is added to $t_{ML}$. Observe that $t_{ML}$ must be connected.

### C. Simulation Results for Finite General Tree Networks

We simulate the virus spreading on finite regular tree networks and general tree networks with $|G| = 1000$ and $|G_n| = 100$. Each vertex in a general tree network is not larger than a positive integer $d_{max}$. The construction of $G$ is a random branching process in which we start with a single
Table III

| $d_m$ | $|\kappa|$ | Algo. 1 | JC | DA |
|-------|----------|--------|----|----|
| 3     | 6.34     | 1.44   | 3.26 | 4.55 | 2.93 |
| 4     | 5.65     | 1.50   | 2.64 | 3.52 | 2.18 |
| 5     | 4.05     | 1.48   | 2.36 | 3.18 | 2.05 |
| 6     | 3.72     | 1.40   | 2.32 | 2.94 | 2.00 |

Table IV

| $d_m$ | $|\kappa|$ | Algo. 1 | JC | DA |
|-------|----------|--------|----|----|
| 4     | 2.82     | 3.07   | 3.34 | 4.51 | 3.02 |
| 5     | 2.6     | 2.59   | 3.00 | 3.96 | 2.59 |
| 6     | 2.46     | 2.46   | 2.85 | 3.87 | 2.51 |
| 7     | 2.45     | 2.33   | 2.61 | 3.55 | 2.40 |

vertex $v_1$, and then randomly pick an integer, say $i$, from 1 to $d_m - 1$ to be the number of children of $v_1$, and then assign $v_2$ to $v_{i+1}$ to be the neighborhood of $v_1$. Recursively applying these steps generates a finite tree $G$ with five thousand vertices whose maximum degree is not larger than $d_m$.

We simulate a thousand times the spread of the virus on $G$ by picking $v^*$, the true source, uniformly on $G$, and compare the average performance of Algorithm 1, a naive heuristic that simply uses the rumor centrality (RC) [14], Jordan centrality (JC) [23], and a spectral based method called Dynamical Age (DA) [42]. To fairly compare these algorithms, when Algorithm 1 yields a set with $|\kappa|$ vertices, then other methods find a set of $|\kappa|$ vertices having the top $|\kappa|$ maximum score for all $v$ of $G_n$. Obviously, the size of the solution set $|\kappa|$ depends on the topology of $G_n$ in each run of the simulation, and thus is not a constant in general over that thousand times. To quantify the performance of these two algorithms, let us define the error function of a vertex set $\eta$:

$$\text{error}(\eta) = \min\{d(v, v^*) \mid \forall \eta \in \eta\}.$$ 

This is the smallest number of hops between $v^*$ and the nearest vertex in the set $\eta$. As shown in Table IV, we can observe that the number of vertices in $\kappa$ is surprisingly small, moreover, $|\kappa|$ is decreasing as $d_m$ increases.

V. PSEUDO-TREES WITH A CYCLE

**Definition V.1.** A pseudo-tree is a connected graph with equal number of vertices and edges, i.e., a tree plus an edge that creates a cycle.

In this section, we consider the special case where $G$ is a degree-regular graph, and $G_n$ has only a single cycle, i.e., $G_n$ is a pseudo-tree. We denote the cycle as $C_h$ where $h$ is the size (number of vertices on the cycle) of the cycle. Here, we call those vertices on $C_h$ cycle vertices. Assume $v$

![Image](https://example.com/image.png)

**Fig. 7.** $G_6$ is an infected subgraph with a single cycle $C_3$ containing three cycle vertices $v_1$, $v_2$ and $v_3$. We can partition $G_6$ into three subtrees say $t_{v_1} = \{v_1, v_4, v_7\}$, $t_{v_2} = \{v_2, v_5\}$, $t_{v_3} = \{v_3\}$.

**Table V**

| $\sigma_i$ | Spreading Order | $P(\sigma_i|G_6)$ |
|-----------|-----------------|------------------|
| $\sigma_1$ | $v_4 \rightarrow v_1 \rightarrow v_3 \rightarrow v_2 \rightarrow v_5 \rightarrow v_7$ | $\frac{2}{1200}$ |
| $\sigma_2$ | $v_4 \rightarrow v_1 \rightarrow v_2 \rightarrow v_5 \rightarrow v_3 \rightarrow v_7$ | $\frac{2}{2400}$ |
| $\sigma_3$ | $v_4 \rightarrow v_1 \rightarrow v_2 \rightarrow v_5 \rightarrow v_7 \rightarrow v_3$ | $\frac{2}{2520}$ |

is a cycle vertex, then we define $t_v$ to be the subtree rooted at $v$ in $G_n$. Take Fig. 7 for example, $t_{v_1}$ is the subtree that contains $v_1$, $v_4$, and $v_7$. In this section, we study how a cycle affects the probability $P(v|G_n)$ when $G_n$ contains a cycle $C_h$. To generalize the analysis in [14], we should intuitively assume that the probability of being infected is proportional to the number of infected neighborhoods. With this assumption, the analysis in [14] will not change, but we can consider the case that two infected vertices have a common susceptible neighborhood, i.e., there is a cycle in $G_n$.

A. Impact of a Single Cycle On $P(G_n|v)$

**Example 3.** Consider the infected subgraph $G_6 \subset C$ as shown in Fig. 7 where $G_6 = \{v_1, v_2, v_3, v_4, v_5, v_7\}$ and there is a 3-cycle in $G_6$. Consider a spreading order $\sigma_1 \in M(v_4, G_6)$, where $\sigma_1 : v_4 \rightarrow v_1 \rightarrow v_3 \rightarrow v_2 \rightarrow v_5 \rightarrow v_7$. We have $P(\sigma_1|v_4) = (1/3) \cdot (1/4) \cdot (2/5) \cdot (1/4) \cdot (1/5) = 2/1200$. Note that when $v_1$ and $v_2$ are infected, $v_3$ has two infected neighborhoods which implies that the probability of $v_3$ being infected in the next time period is twice higher than $v_5$, $v_7$ and $v_8$. In particular, there are three possible values for $P(\sigma_i|v_4)$ as shown in Table V for all $\sigma_i \in M(v_4, G_6)$. Moreover, we observe that the denominators are different in Table V due to sharing a common neighbor in the presence of a cycle. We call this property the cycle effect.

Example 3 reveals some interesting properties of the cycle effect due to a single cycle:

1. $P(\sigma_i|v)$ increases with how soon the last cycle vertex appears in $\sigma_i$ (as ordered from left to right of $\sigma_i$). For example, the last cycle vertex on $\sigma_1$ is $v_2$, and is $v_3$ on $\sigma_2$ and $\sigma_3$.
2. When there is a cycle in $G_n$, then $P(G_n|v)$ is no longer proportional to $|M(v, G_n)|$.
3. For each $\sigma_i$, there are actually two corresponding permitted spreading orders due to the cycle.

The first property shows that $P(\sigma_i|v)$ is dependent on the position of the last cycle vertex in each spreading order. Note
that the cycle effect is similar to the irregular effect, and the main difference lies in that all cycle vertices may cause the cycle effect instead of only one irregular vertex may cause the irregular effect. We proceed to compute \( P(\sigma_i|v) \) as follows. For brevity of notation, let \( v_i \) denote the last cycle vertex.

**Definition V.2.** We let the distance from a vertex \( v \) to the cycle \( C_h \) denoted by \( d(v, C_h) \) be defined by the minimum value of distances from \( v \) to all cycle vertices on \( C_h \). That is,

\[
d(v, C_h) = \min_{v_j \in C_h} \{d(v, v_j)\}.
\]

Take Fig. 7 for example, let \( C_3 = \{v_1, v_2, v_3\} \), then \( d(v_1, C_3) = d(v_2, v_1) = 2 \) and \( d(v_3, C_3) = 1 \).

**Remark:** For each \( \sigma \in |M(v, G_n)| \), \( v_i \) can be any vertex on the cycle \( C_h \) except the vertex \( v' \) with the distance \( d(v, v') = d(v, v_i) \). Hence, there are \( h - 1 \) possible \( v_i \).

From previous observations, we have

\[
|M(v, G_n)| = 2 \cdot \sum_{k=d(v, C_h)+h}^{n-t_v+1} m_v^{\sigma}(G_n, k)
\]

since the position of \( v_i \) on the spreading order ranges from \( d(v, C_h) + h \) to \( n - t_v + 1 \). For example, in Table [V] we can see that \( v_1 = v_2 \) is the 4th element on \( \sigma_1 \) and \( v_1 = v_3 \) is the 6th element on \( \sigma_3 \) whence the order 4 comes from \( d(v_4, C_3) + h = 1 + 3 \), and the order 6 comes from \( n - t_v + 1 = 6 - 1 + 1 \). Finally, the multiplication with \( 2 \) is due to the third property.

Now, we can rewrite \( P(G_n|v) \) for \( G_n \) with a cycle as:

\[
P(G_n|v) = \sum_{k=d(v, C_h)+h}^{n-t_v+1} m_v^{\sigma}(G_n, k) \cdot P_v^{\sigma}(G_n, k),
\]

and our goal is to find the vertex \( \hat{v} \) that achieves

\[
P(G_n|\hat{v}) = \max_{v_i \in G_n} P(G_n|v_i).
\]

Since \( P(G_n|v) \) is not proportional to \( |M(v, G_n)| \), we should compute \( P(G_n|v) \) by considering each part \( m_v^{\sigma}(G_n, k) \) and their corresponding probability \( P_v^{\sigma}(G_n, k) \). Let \( z_d(i) = (i-1)(d-2) \), then

\[
P_v^{\sigma}(G_n, k) = 2 \cdot \prod_{i=1}^{k-1} \frac{1}{d + z_d(i)} \cdot \prod_{i=k}^{n-2} \frac{1}{d + z_d(i) - 1}.
\]

The first factor in (18) is the probability that \( k \) vertices are infected where the \( d \)th infected vertex is \( v_i \), and the second factor is the probability of that all remaining \( n - k \) vertices being infected thereafter. The \(-1\) in the denominator of the second factor and the coefficient \( 2 \) at the front are due to the common neighbor in a cycle. Note that multiplying by \( 2 \) at the front makes no difference when computing \( P(G_n|v) \) for each \( v \in G_n \). From [16], we see that the number of spreading orders and the corresponding position of \( v_i \) affect \( P(G_n|v) \).

**B. Computing \(|M(v, G_n)|\) for \( v \) on \( G_n\)**

In this section, we focus on computing \(|M(v, G_n)|\). To compute \(|M(v, G_n)|\), we can leverage the message-passing algorithm in [14] if \( G_n \) is a tree. Observe that for each infected vertex in \( G_n \), it is infected by one of its infected neighbors (even if it has two infected neighbors), so the actual infecting route is a spanning tree of \( G_n \) instead of a graph with cycle. Hence, the number of all spreading orders on a graph \( G_n \) with a cycle can be computed as

\[
|M(v, G_n)| = \sum_{1 \leq i \leq h} |M(v, T_i)|,
\]

where \( T_i \) is the spanning tree of \( G_n \), for \( i = 1, 2, \ldots, h \). If \( G_n \) contains a \( C_h \), then the time complexity of computing \( M(v, G_n) \) for \( v \in G_n \) is \( O(hn) \). Since \( G_n \) has \( h \) spanning trees, and for each spanning tree, we can use the message-passing algorithm in [14] that has \( O(n) \) time complexity.

**C. Epidemic Center on Uni-cyclic \( G_n\)**

In this section, we propose a theorem and a lemma to characterize the location of epidemic center in \( G_n \). Instead of computing \(|M(v, G_n)|\) for all \( v \in G_n \) in each spanning tree, we leverage some analytical results to find \( v_c \). Let \( t_i \equiv t_{v_i} \) be defined as above, and we slightly abuse the notation of the subtree size \( |t_i| \) as \( t_i \).

**Theorem 3.** Let \( G \) be a degree regular graph, and \( G_n \) be a subgraph of \( G \) with a single cycle \( C_h = \{v_1, v_2, \ldots, v_h\} \). The epidemic center \( v_c \) of \( G_n \) satisfies one of the following conditions:

1. Each connected component of \( G_n \setminus \{v_c\} \) is of size less or equal to \( n/2 \).
2. \( v_c \) is a cycle vertex and \( v_i \leq n/2 \) for \( i = 1, 2, \ldots, h \).

**Remark:** Note that \( v \) being epidemic center of \( G_n \) does not mean that each connected component of \( G_n \setminus \{v\} \) is of size less or equal to \( n/2 \). (Had \( G_n \) been a tree, then this is true [14].) However, condition 1 is the sufficient condition of a vertex being the epidemic center even in a general graph.

We can locate \( v_c \) of \( G_n \) by the first condition in Theorem 3 however, if the first condition is not satisfied, then \( v_c \) is on the cycle. In the following, we proceed to consider the case that \( v_c \) is a cycle vertex. Let \( T_j \) denote the spanning tree of \( G_n \) which is constructed by \( G_n \setminus \{v_j, v_j+1\} \), for \( j = 1, 2, \ldots, h-1 \) and \( T_h = G_n \setminus \{v_h, v_1\} \). Note that \((v_h, v_1)\) and \((v_j, v_j+1)\) for \( j = 1, 2, \ldots, h-1 \) are cycle edges of \( C_h \).
Proposition 1. Let \( v_i \) be a cycle vertex and assume that \( |M(v_i, T_p)| = r \), where \( r \) and \( p \) are integers and \( 1 \leq p \leq h \). Then for \( 1 \leq q \leq h \), we have
\[
\frac{|M(v_i, T_q)|}{|M(v_i, T_p)|} = \prod_{j \in C_h, j \neq i} T_{p,j}^{q,j} / \prod_{k \in C_h, k \neq i} T_{q,k}^{i,k},
\]
where \( T_{p,j}^{q,j} \) is the subtree of the spanning tree \( T_p \).

The ratio \( |M(v_i, T_p)| / |M(v_i, T_q)| \) is proportional to their branch size in \( T_p \) if \( v_i \) and \( v_j \) are adjacent. Now, for the same vertex \( v_i \), but in different spanning tree say \( T_p \) and \( T_x \), we can also derive the ratio \( |M(v_i, T_p)| / |M(v_i, T_q)| \) from (20).

Hence, if we assume \( |M(v_i, T_1)| = r \), then we can derive \( |M(v, T)| \) for all \( v \in C_h \) and \( T \) is a spanning tree of \( G_n \) in terms of \( r \) and \( t_i \), where \( t_i \) is the subtree size as shown in Fig. 8. The time complexity to find \( v_i \) is \( O(n + h^2) \) in the worst case, where \( h \) is the size of the cycle in \( G_n \).

Example 4. Take \( C_3 \) for example, \( |M(v_1, T_1)| / |M(v_1, T_2)| = t_3 / (t_2 + t_3) \). Assume \( |M(v_1, T_1)| = r \), by (20) we can conclude that
\[
|M(v_1, G_n)| = 2(t_2 + t_3) / t_3 r; \\
|M(v_2, G_n)| = 2(t_2 + t_3) / t_2 r; \\
|M(v_3, G_n)| = 2(t_2 + t_3) / t_1 r,
\]
which implies
\[
|M(v_1, G_n)| : |M(v_2, G_n)| : |M(v_3, G_n)| = t_1 : t_2 : t_3.
\]

By Theorem 5 and Proposition 1, we conclude that if \( G_n \) contains a \( C_3 \), then \( v_c \) of \( G_n \) is either a vertex that satisfies the first condition or a vertex \( v_i \) on \( C_3 \) with \( t_i = \max_{1 \leq j \leq 3} t_j \).

Lastly, we combine Lemma 1 and Theorem 5 to characterize the location of the ML estimator on regular pseudo-tree.

Theorem 4. Let \( G_n \) be defined as in Theorem 3. The optimal solution to (21) is either on the path from the epidemic center of \( G_n \) to the cycle or on the cycle.

Remark: Theorem 4 is a combination of Theorem 3 and 2. Besides, Theorem 4 generalizes the results in [43], [44].

VI. ALGORITHM FOR FINITE DEGREE REGULAR GRAPH WITH CYCLES

This section proposes a novel distance-based algorithm to solve the epidemic source detection problem on a degree regular graph with cycles. From Theorem 2, we can deduce that the likelihood of a vertex is greater if its distance to those irregular vertices and cycles is smaller. Hence, the maximum likelihood estimator should lie on the smallest induced subgraph containing three specific vertices: the epidemic center, the vertex closest to all cycles, and the vertex closest to all irregular vertices. Note that a degree on irregular vertex can be treated mathematically as a size one cycle. Hence we combine the irregular effect and the cycle effect in our algorithm.

Definition VI.1. We say is a cycle is a minimum cycle if there is no path between any two non-consecutive cycle vertices except the path along the cycle.

Fig. 9. The infected subgraph \( G_{10} \) contains ten nodes colored in grey. The epidemic centralities of \( v_8 \) and \( v_{10} \) are the same, however, \( v_{10} \) is the MLE of the true source in \( G_{10} \). Since the irregular effect caused by a small cycle is greater than that of large cycle.

Since a vertex \( v \) can be contained in multiple different-size cycles, we only take the minimum cycle that contains \( v \) into consideration in our algorithm. Let \( C(v) \) denote the size of the minimum cycle containing \( v \). If \( v \) is not in any cycle and \( deg(v) > 1 \), then we set \( C(v) = \infty \), otherwise \( C(v) = 1 \). Note that when \( deg(v) = 1 \), \( v \) is regarded as a size 1 cycle.

Base on Theorem 2,4, we heuristically define the weight \( w_u(v) \) of a vertex \( v \) as
\[
w_u(v) = \frac{C(v)}{C(v) + 1}. \tag{21}
\]

Since we define the minimum cycle to be the vertex with minimum distance centrality (cf. Definition II.2), we can design a weight such that the location of the ML estimator tends to be close to vertices with “small weights”. This is also motivated by the fact that the likelihood of a vertex \( v \) being the source is greater if \( v \) has a larger epidemic centrality and is closer to those irregular vertices or cycles (cf. Theorem 2 and Theorem 3). The definition of SDC(\( v, G_n \)) is a distance-based centrality. Furthermore, the definition of \( w_u(v) \) reveals that the irregular effect caused by a small cycle is greater than that caused by a large cycle which can be observed from Table VI and 15. This definition implies that a vertex within a smaller cycle has a smaller weight which contributes “more” to SDC(\( v, G \)) while a vertex not in any cycle has weight 1 which contributes “less” to SDC(\( v, G \)). Fig. 9 illustrates such an example of a regular graph \( G \) and the infected subgraph \( G_{10} \) containing two different-sizes cycles, say \( C_3 \) and \( C_4 \). Note that, \( v_{10} \) and \( v_8 \) have the same epidemic centrality, however, we have \( P(G_{10}|v_{10}) > P(G_{10}|v_8) \) since \( v_{10} \) is closer to the smaller cycle than \( v_8 \).

Definition VI.2. Given a \( d \)-regular graph \( G \) and vertex \( v \) of \( G \), we define the statistical distance centrality of \( v \), SDC(\( v, G \)) as the summation of the weighted distance from \( v \) to all other vertices in \( G \). Hence, the statistical distance centrality of \( v \) in \( G \) is defined by
\[
SDC(v, G_n) = \sum_{u \in G_n} w_u \cdot d(v, u). \tag{22}
\]

The vertex \( v_s \) with the minimum value for SDC(\( v, G \)) is called the statistical distance center.

Algorithm 2 is based on the idea of message passing. Let \( lv(v) \) denote the level of \( v \) in a BFS tree. In Step 2, for a given root \( v_r \), we start a message-passing procedure in a BFS traversal to send a downward message containing level
Algorithm 2 Statistical Distance-based Contact Tracing (SCT)

Input: $G_n$

Step 1: For each vertex $v$, compute the size $C(v)$ of the minimum cycle containing $v$, and set $w_v = C(v) + \tau$.

Step 2: For each vertex $v$, compute $SDC(v, G_n)$.

Step 3: Let $\hat{v} = \arg\min_{v \in G} SDC(v, G)$.

information to other vertices in the BFS tree. Upon receipt of this information, each leaf $v_l$ sends back an upward message containing $w_{v_l} \cdot l(v_l)$ to its parent. Each internal vertex $v_{in}$ sends an upward message, containing the summation of all message from its children plus $w_{v_{in}} \cdot l(v_{in})$, to its parent.

In the following, we provide a time complexity analysis of Algorithm 2. For Step 1, the worst case time complexity is $O(|C_{\min}| \cdot |E(G_n)|)$, where $|C_{\min}|$ is the number of all minimum cycles in $G_n$. Since $G$ is a $d$-regular graph, each vertex in $G_n$ is contained in at most $d$ minimum cycles which implies $C_{\min} \leq d \cdot n$. The worst case time complexity for the Step 2 in the Algorithm 2 is $O(n^3)$, since the BFS traversal for each vertex takes $O(n + |E(G_n)|)$. Hence, the worst case time complexity of Algorithm 2 is $O(d \cdot n^3)$. In comparison with the BFS heuristic approach in [14], it applies the BFS traversal for each vertex and compute their epidemic centrality which ends up with worst time complexity $O(n^3)$.

A. Experimental Performance Evaluation

We provide simulation results on different finite graphs with cycles, the first two simulations are conducted on synthetic graphs such as finite size grid graph and circulant graphs. Both synthetic graphs are regular graphs with cycles except those vertices on the boundary of the grid graph. In synthetic graphs, we first simulate the virus spreading in a given network based on the model described in Section II, to construct the infected subgraph $G_n$. Then, we apply Algorithm SCT to compute the source estimator.

We conduct the other four experiments on real-world SARS-CoV2003 and COVID-19 contact tracing networks in Singapore and Taiwan. If we can identify the connection between any two confirmed cases in real-world contact tracing networks, we denote the connection (or contact) as an edge. However, when the number of confirmed cases is too large to record details of contact information, we can only have information about the geographical footprint for some confirmed cases. In this situation, we also denote those visited places as vertices, and we add an edge between a confirmed case and a place if the confirmed case had visited the place.

Since $G$ is unknown in practice and contact tracing networks are infected subgraphs $G_n$, we assume that $G$ is a regular graph with a few irregular vertices, and apply Algorithm SCT to the contact tracing networks to compute the source estimator. We use the graph distance from the actual source to the estimator to evaluate its performance.

1) Grid Graph: Disease spreading on grid graph is often considered under different spreading rules and models [34], [46]. Hence, we select grid graph to be one of the testing synthetic networks. Simulation results are in Table VI and one of the error distributions is in Fig. 10. We can observe from Table VI that the statistical distance based algorithm outperforms the BFS heuristic in [14]. Moreover, the average error is increasing as the number of irregular vertices is increasing which again reveals the fact that the likelihood is evened out to those irregular vertices.

2) Circulant Graph: A circulant graph $G = (N; S)$ is a class of graphs which can be defined by its vertex set $V(G) = Z$ and a set $S$. The edge set is defined by $E(G) = \{(v_i, v_{i+j}) \mid i-j \in S\}$. Hence, if $S$ is a subset of $[1, N/2]$ then $G$ is $(2|S|)$-regular. Note that a circulant graph $G$ is connected if and only if $S$ generates the integer group $Z_N$ and we only consider the connected circulant graph. In the simulation, we fix the size of $N$ and $S$ and randomly choose integers from the interval $[1, N/2]$ to form the set $S$.

3) SARS-CoV2003 Contact Tracing Network in Taiwan: We reconstruct the contact tracing network data of SARS-CoV2003 and COVID-19 in Taiwan in Section IV, and we simulate the virus spreading in such a network to obtain the infected subgraph $G_n$. Then, we apply Algorithm SCT to compute the source estimator and compare the result with the BFS heuristic [14].

Fig. 10. Comparing the error distribution (in the number of hops) between Algorithm SCT and the BFS heuristic [14] in a finite grid graph with $|G| = 10000$ and $|G_n| = 150$. In particular, the rate of the correct detection, i.e., error = 0, is 12.1% for Algorithm SCT and 2.6% for the BFS heuristic.

Table VI

| $n$  | $|C_4|$  | $|v_{irr}|$ | SCT | BFS heuristic [14] |
|------|---------|------------|-----|-------------------|
| 150  | 85.5    | 3.5        | 1.87| 3.79              |
| 300  | 199.5   | 7.5        | 2.33| 6.11              |
| 500  | 364.2   | 12.2       | 3.14| 8.37              |
| 800  | 625.0   | 19.1       | 4.23| 11.71             |

Table VII

| $d$  | #cycles    | SCT | BFS heuristic [14] |
|------|------------|-----|-------------------|
| 6    | 255.6      | 1.67| 2.75              |
| 8    | 198.0      | 1.45| 2.37              |
| 10   | 167.1      | 1.33| 2.07              |
| 12   | 147.7      | 1.24| 1.97              |
Fig. 11. Comparing the error distribution (in the number of hops) between Algorithm SCT and the BFS heuristic in [14] on a circulant graph with $|G| = 6000$ and $|G_n| = 400$. In particular, the rate of the correct detection, i.e., $error = 0$, is 19.1% for Algorithm SCT and 7.0% for the BFS heuristic.

Fig. 12. SARS-CoV2003 Contact Tracing Network in Taiwan. Each vertex represents either a confirmed case or a hospital. The orange and red vertices represent the source estimator determined by Algorithm SCT and the BFS heuristic respectively. The orange vertex is the Taipei City Hospital Heping Branch where the major outbreak occurs and the red vertex is a confirmed case (not the first case) who had been to this hospital.

CoV2003 Taiwan from a graph, which indicates Potential bridges among hospitals and households, in [47]. In the original data, there are four types of nodes which represent the confirmed case, suspected case, hospital, and area, respectively. Since cities or countries provide no information for personal contact, we delete all area nodes from the original data. In addition, we also delete all the nodes that represent suspected cases. We apply Algorithm SCT on this infected network and correctly identify the first place, Taipei Municipal Heping Hospital (now Taipei City Hospital Heping Branch), of cluster infection in April 2003 in Taiwan. In addition, the BFS heuristic approach chooses the red vertex, which represents a confirmed case (not the first case) who had been to this hospital.

4) COVID-19 Contact Tracing Network in Singapore, 2020 Mar-Apr: The contact tracing network is an unconnected network due to the asymptomatic carriers, so we focus on the largest connected subgraph (cluster), including several worker dormitories and a construction site. We apply Algorithm SCT on subgraphs of the contact tracing network in Singapore provided in [48]. In our computation, each vertex represents either an infected person or a place where the person had visited. An edge between two vertices implies that either a person has visited a place or two places have at least one common visitor. Here we omit the edge of person-person contact since most of the contact history can only be traced back to a place, not a single person. Hence, we treat each person-vertex as a leaf vertex connecting to a place. The first massive outbreak occurred at the beginning of April and peaked on April 20. Hence, we consider the infected subgraph after April 1. We define the source in the connected subgraph to be the first case in this connected subgraph. As far as we know, Case 655 attaching to Westlite Toh Guan (WTG) is the first case found in this subgraph on March 26. On April 3, the connected subgraph formed a 4-cycle, which is shown in Fig. 13. We can apply Algorithm SCT, which shows that WTG is the source estimator in this subgraph. After April 10, S11 Dorm (S11) becomes the new epidemic center due to the link between STL and S11 Dorm (S11). Note that S11 contains the second earliest case in this cluster and becomes the largest cluster in Singapore, which has more than two thousand cases confirmed in the middle of May.

5) COVID-19 Contact Tracing Network in Taiwan, 2021 Feb and 2021 May: We conduct experiments on two cluster infection in Taiwan recently. The first cluster infection originated from a northern Taiwan hospital on February. This network contains 18 tractable domestic cases. The reason we select this networked data is that the contact network is public information provided by Central Epidemic Command Center in Taiwan [49] and the relation between cases in this network is clearly defined. Note that if we apply the BFS heuristic, then both case 838 and case 856 have the same possibility to be the source estimator. However, case 838 is the vertex with
maximum statistical distance centrality, i.e., Algorithm SCT correctly identifies the first domestic case.

The second cluster infection is the latest cluster infection found at the beginning of May 2021. As the source of this cluster is unknown, we let the source be the person with the earliest symptom onset. We collect the data before May 14, 2021 from [49], and apply the 2-mode network model [47] to this cluster. Each vertex in this graph is either a workplace or a confirmed case. Both Algorithm SCT and the BFS heuristic identify the workplace, of the first case in this network. The contact tracing network is shown in Fig. [15] and the red vertex is the source estimator determined by both algorithms.

6) Simulation Results on Other Networks: In addition to the rumor centrality approach, we have selected two other approaches, Dynamical Age [42] and Jordan Centrality [23], to compare to Algorithm SCT. We use the average distance-error to measure the performance of each algorithm. The simulation results are shown in Table VIII. In each simulation, we repeat the following process: generate $G_n$, find estimators and compute errors for five hundred times in each type of network. All datasets in Table VIII are available at [50], [51] or can be generated by networkx [52]. Note that SCT works well in circulant graph and grid graph, since SCT is designed to solve the maximum likelihood estimation problem on finite regular graph with cycles. In addition to the selected algorithms, there are other approaches such as PTVA [26] and gradient maximum likelihood algorithm (GMLA) [27] perform well on source detection problem. However, PTVA and GMLA are incomparable with the methods listed in Table VIII. Since PTVA and GMLA require additional information such as relative time stamps observed by observers. Moreover, to fairly compare all algorithms, the number of the observers and their location in the network also need to be carefully considered which is out of scope of this article.

VII. CONCLUSION

We present an optimal contact tracing algorithm for epidemic source detection that finds application in identification of superspreaders in an epidemic (e.g., the COVID-19 pandemic). Our algorithm design leverages a statistical distance centrality based on solving a graph-constrained maximum likelihood problem for contact tracing graphs with irregular vertices and cycles. As a high-dimensional statistical problem, epidemic source detection requires a careful understanding of the interaction between stochastic spreading processes and topological features like graph distances, cycles and finite boundaries, allowing us to resolve an open problem of finite degree-regular graphs with cycles in [14]. Performance analysis demonstrated that our algorithm was near-optimal in correctly identifying superspreaders at some of the largest superspreading infection clusters using data from the SARS-CoV 2003 and COVID-19 pandemics in Singapore and Taiwan. It will be most interesting to study computational epidemiology through the lens of “network centrality as statistical inference” to design robust contact tracing algorithms using mathematical and data-driven methods (e.g., machine learning techniques like deep learning, see [1], [10], [29]) that can analyze past epidemic behaviors to fight against newly emerging epidemics.

VIII. ACKNOWLEDGEMENT

We would like to acknowledge collaborations and interactions on this topic with H. Vincent Poor, Cheng-Shang Chang, Guanrong Chen, Siya Chen and Weng Kee Wong.

APPENDIX A

PROOF OF THEOREM AND LEMMAS

A. Proof of Theorem 7

In this proof, we use the fact that

$$P(v_c|G_n) > P^v_{v_c}(G_n, n) \cdot m^v_{v_c}(G_n, n),$$

and consider the ratio between the lower bound of $P(v_c|G_n)$ and $P(v_c|G_n)$ to simplify the proof.

Proof. Let $G_n$ be the infected subgraph. Without loss of generality, we assume $n = 2t + 1$ and $d \geq 3$. For $v_c$, we have

$$P(v_c|G_n) = m^v_{v_c}(G_n, 1) \cdot P^v_{v_c}(G_n, 1)$$

$$= 1 \cdot P^v_{v_c}(G_n, 1)$$

$$= \prod_{i=0}^{2t-1} \frac{1}{1 + i(d - 2)} = \prod_{i=0}^{n-2} \frac{1}{1 + i(d - 2)}.$$
We first prove Lemma 2 and Lemma 1, which are tools to estimate the Barabási-Albert(5000,3) network. Note that in this case, since we have $t(v_{xa}, t(v_{xa}, v_{xb})) = 1$, we can rewrite the above inequality as
\[
\frac{t(v_{xa}, t(v_{xa}, v_{xb}))}{t(v_{xa}, v_{xb})} > \frac{t(v_{xa}, v_{xb})}{t(v_{xa}, v_{xb})}.
\]
Moreover, we can leverage the fact $n = t(v_{xa}, v_{xb}) + t(v_{xa}, v_{xb})$ to replace $t(v_{xa}, v_{xb})$ and $t(v_{xa}, v_{xb})$ in the above inequality. Hence, we have
\[
\frac{t(v_{xa}, v_{xb})}{n-t(v_{xa}, v_{xb})} > \frac{t(v_{xa}, v_{xb})}{n-t(v_{xa}, v_{xb})},
\]
which implies $t(v_{xa}, v_{xb}) > t(v_{xa}, v_{xb})$.

\[\square\]

**Proof.** To prove Lemma 2, we first consider the second condition, i.e., the case when $d(v_{xa}, v_{xb}) = d(v_{xa}, v_{xb}) = 1$. Note that in this case, $v_{xa}$ is not on the shortest path from $v_{xb}$ to $v_{xb}$ and vice versa.

We consider the ratio of $m(v_{xa}, v_{xb})$ to $m(v_{xb}, v_{xb})$ for all possible $i$. Note that
\[
m(v_{xa}, v_{xb}) = \left(\frac{n-i}{t(v_{xa}, v_{xb})-(i+1)}\right) M(v_{xa}, t(v_{xb})) \cdot M(v_{xb}, t(v_{xb})),
\]
where $v = v_{xa}, v_{xb}$. Since $M(v_{xa}, t(v_{xb}))$, $M(v_{xb}, t(v_{xb}))$ and $M(v_{xb}, t(v_{xb}))$ are fixed when $G_{xy}$ is given, we have
\[
\frac{m(v_{xa}, v_{xb})}{m(v_{xb}, v_{xb})} \propto \left(\frac{n-i}{t(v_{xa}, v_{xb})-(i+1)}\right) \propto \left(\frac{n-i}{t(v_{xa}, v_{xb})-(i+1)}\right).
\]
Since $G_{xy}$ is a tree and $d(v_{xa}, v_{xb}) = d(v_{xa}, v_{xb}) = 1$, we have $d(v_{xa}, v_{xb}) = 2$. By Lemma 2, we have $t(v_{xa}, v_{xb}) > t(v_{xb}, v_{xb})$, which implies $t(v_{xa}, v_{xb}) > t(v_{xb}, v_{xb})$. Hence, the ratio $m(v_{xb}, v_{xb})$ is an increasing sequence with respect to $i$.

Since $m(v_{xb}, v_{xb}) = m(v_{xb}, v_{xb})$ and $m(v_{xb}, v_{xb})$ is increasing with respect to $i$, we can conclude that $m(v_{xb}, v_{xb}) \geq m(v_{xb}, v_{xb})$, which leads to
\[
\sum_{i=d(v_{xa}, v_{xb})+1}^{k} m(v_{xb}, v_{xb}) \geq \sum_{i=d(v_{xa}, v_{xb})+1}^{k} m(v_{xb}, v_{xb}).
\]

**Proof.** To prove Lemma 2, we first consider the second condition, i.e., the case when $d(v_{xa}, v_{xb}) = d(v_{xa}, v_{xb}) = 1$. Note that in this case, $v_{xa}$ is not on the shortest path from $v_{xb}$ to $v_{xb}$ and vice versa.

We consider the ratio of $m(v_{xa}, v_{xb})$ to $m(v_{xb}, v_{xb})$ for all possible $i$. Note that
\[
m(v_{xa}, v_{xb}) = \left(\frac{n-i}{t(v_{xa}, v_{xb})-(i+1)}\right) M(v_{xa}, t(v_{xb})) \cdot M(v_{xb}, t(v_{xb})),
\]
where $v = v_{xa}, v_{xb}$. Since $M(v_{xa}, t(v_{xb}))$, $M(v_{xb}, t(v_{xb}))$ and $M(v_{xb}, t(v_{xb}))$ are fixed when $G_{xy}$ is given, we have
\[
\frac{m(v_{xa}, v_{xb})}{m(v_{xb}, v_{xb})} \propto \left(\frac{n-i}{t(v_{xa}, v_{xb})-(i+1)}\right) \propto \left(\frac{n-i}{t(v_{xa}, v_{xb})-(i+1)}\right).
\]
Since $G_{xy}$ is a tree and $d(v_{xa}, v_{xb}) = d(v_{xa}, v_{xb}) = 1$, we have $d(v_{xa}, v_{xb}) = 2$. By Lemma 2, we have $t(v_{xa}, v_{xb}) > t(v_{xb}, v_{xb})$, which implies $t(v_{xa}, v_{xb}) > t(v_{xb}, v_{xb})$. Hence, the ratio $m(v_{xb}, v_{xb})$ is an increasing sequence with respect to $i$.

Since $m(v_{xb}, v_{xb}) = m(v_{xb}, v_{xb})$ and $m(v_{xb}, v_{xb})$ is increasing with respect to $i$, we can conclude that $m(v_{xb}, v_{xb}) \geq m(v_{xb}, v_{xb})$, which leads to
\[
\sum_{i=d(v_{xa}, v_{xb})+1}^{k} m(v_{xb}, v_{xb}) \geq \sum_{i=d(v_{xa}, v_{xb})+1}^{k} m(v_{xb}, v_{xb}).
\]

**Proof.** To prove Lemma 2, we first consider the second condition, i.e., the case when $d(v_{xa}, v_{xb}) = d(v_{xa}, v_{xb}) = 1$. Note that in this case, $v_{xa}$ is not on the shortest path from $v_{xb}$ to $v_{xb}$ and vice versa.

We consider the ratio of $m(v_{xa}, v_{xb})$ to $m(v_{xb}, v_{xb})$ for all possible $i$. Note that
\[
m(v_{xa}, v_{xb}) = \left(\frac{n-i}{t(v_{xa}, v_{xb})-(i+1)}\right) M(v_{xa}, t(v_{xb})) \cdot M(v_{xb}, t(v_{xb})),
\]
where $v = v_{xa}, v_{xb}$. Since $M(v_{xa}, t(v_{xb}))$, $M(v_{xb}, t(v_{xb}))$ and $M(v_{xb}, t(v_{xb}))$ are fixed when $G_{xy}$ is given, we have
\[
\frac{m(v_{xa}, v_{xb})}{m(v_{xb}, v_{xb})} \propto \left(\frac{n-i}{t(v_{xa}, v_{xb})-(i+1)}\right) \propto \left(\frac{n-i}{t(v_{xa}, v_{xb})-(i+1)}\right).
\]
Since $G_{xy}$ is a tree and $d(v_{xa}, v_{xb}) = d(v_{xa}, v_{xb}) = 1$, we have $d(v_{xa}, v_{xb}) = 2$. By Lemma 2, we have $t(v_{xa}, v_{xb}) > t(v_{xb}, v_{xb})$, which implies $t(v_{xa}, v_{xb}) > t(v_{xb}, v_{xb})$. Hence, the ratio $m(v_{xb}, v_{xb})$ is an increasing sequence with respect to $i$.

Since $m(v_{xb}, v_{xb}) = m(v_{xb}, v_{xb})$ and $m(v_{xb}, v_{xb})$ is increasing with respect to $i$, we can conclude that $m(v_{xb}, v_{xb}) \geq m(v_{xb}, v_{xb})$, which leads to
\[
\sum_{i=d(v_{xa}, v_{xb})+1}^{k} m(v_{xb}, v_{xb}) \geq \sum_{i=d(v_{xa}, v_{xb})+1}^{k} m(v_{xb}, v_{xb}).
\]
On the other hand, we consider the case when $v_b$ is on the path from $v_b$ to $v_{ir}$, i.e., $\{v_a, v_b\} \in E(G)$ and $d(\{v_a, v_{ir}\}) = d(v_b, v_{ir}) + 1$.

For brevity, we denote the distance from $v_a$ to $v_{ir}$ as $D$, and relabel the vertices on the path from $v_p$ to $v_e$ as $u_0$ to $u_{D+1}$ which are shown in Fig. 4.

Again, we consider the ratio of $m_{u_{D+1}}(G_n, i)$ to $m_{u_{D+1}}(G_n, i)$ for all possible $i$. We can express $m_{u_{D+1}}(G_n, i)$ and $m_{u_{D+1}}(G_n, i)$ as follows:

$$m_{u_{D+1}}(G_n, i) = \sum_{1 < i_1 < \ldots < i_{D-1} < i} \left[ \prod_{l=1}^{D-1} \left( n - i_l - t_{u_{l+1}}^{\nu_{u_{l+1}}} \right) \right] \left( v_{u_{D+1}} - 1 \right) M_u,$$

where

$$M_u = M(u_1, t_{u_{D+1}}^{\nu_{u_{D+1}}}) \cdot \prod_{l=2}^{D+1} M(u_l, t_{u_{l}}^{\nu_{u_{l}}}).$$

We can express $m_{u_0}^{u_{D+1}}(G_n, i)$ in the same way. Since $M_u$ is fixed when $G_n$ is given, we have

$$m_{u_0}^{u_{D+1}}(G_n, i) = \sum_{1 < i_1 < \ldots < i_{D-1} < i} \left[ \prod_{l=1}^{D-1} \left( n - i_l - t_{u_{l+1}}^{\nu_{u_{l+1}}} \right) \right] \left( v_{u_{D+1}} - 1 \right) M_u,$$

which is a decreasing sequence with respect to $i$.

Note that when $i = D + 1$, we have $m_{u_0}^{u_{D+1}}(G_n, D + 1) > m_{u_0}^{u_{D+1}}(G_n, D + 1)$.

Now, we can prove the main statement of Lemma 1. To contrary, suppose there is an integer $k$ such that

$$\sum_{i = D+1}^{k} m_{v_0}^{v_i}(G_n, i) < \sum_{i = D+1}^{k} m_{v_0}^{v_i}(G_n, i).$$

Since the ratio of $m_{u_{D+1}}(G_n, i)$ to $m_{u_{D+1}}(G_n, i)$ is a decreasing sequence, we have for all $k' > k$,

$$\sum_{i = D+1}^{k'} m_{v_0}^{v_i}(G_n, i) < \sum_{i = D+1}^{k'} m_{v_0}^{v_i}(G_n, i).$$

This leads to $M(u_1, G_n) < M(u_0, G_n)$, which is a contradiction to the assumption. Hence, we can conclude that $\forall k$,

$$\sum_{i = D+1}^{k} m_{v_0}^{v_i}(G_n, i) \geq \sum_{i = D+1}^{k} m_{v_0}^{v_i}(G_n, i).$$

C. Proof of Theorem 3

Proof. Let $D = d(v_e, v_{ir})$ and we relabel vertices on the path from $v_e$ to $v_{ir}$ as $u_1, u_2, \ldots, u_{D+1}$. We can partition $V(G_n)$ as follows

$$V(G_n) = V(t_{v_e}^{nu} + V(t_{v_{ir}}^{nu}) + \sum_{i=2}^{D} V(t_i).$$

For any vertex $v$ in $V(t_{v_{ir}}^{nu})$ other than $v_{ir}$. If $d(v, v_{ir}) > 1$, then there is a neighbor of $v$, say $u$, such that $d(v, v_{ir}) > d(u, v_{ir})$. Since $M(u, G_n) > M(v, G_n)$, by the condition 1 in Lemma 1 we can conclude that

$$\sum_{i=d(u, v_{ir})+1}^{k} m_{v_{ir}}(G_n, i) \geq \sum_{i=d(u, v_{ir})+1}^{k} m_{v_{ir}}(G_n, i).$$

Since the probability $P(G_n \mid v)$ can be computed as follows

$$P(G_n \mid v) = \sum_{i=D+1}^{n-t_{v_{ir}}^{nu}+1} m_{v_{ir}}(G_n, i) \cdot P_{v_{ir}}^{nu}(G_n, i),$$

and $P_{v_{ir}}^{nu}(G_n, i)$ is a decreasing with respect to $i$. We can conclude that $P(G_n \mid u) > P(G_n \mid v)$. This procedure can be continued until we reach to $v_{ir}$. By the same argument, we can prove the case when $v$ is in other parts with the remaining unconvertible vertices being on the path from $v_e$ to $v_{ir}$. \[ \square \]

D. Proof of Theorem 3

Proof. Let $v \in G_n$ and $v$ is not a cycle vertex and each connected component of $G_n \setminus \{v\}$ is of size less or equal to $n/2$. Consider any given spanning tree $T_j$ of $G_n$, assume that $v$ is not the rumor center of $T_j$. Then, there is a vertex $u$ such that $t_{u}^{nu} > t_{v}^{nu}$ on $T_j$. Since $t_{u}^{nu} + t_{v}^{nu} = n$, we can conclude that the size of the connected component of $G_n \setminus \{v\}$ containing $u$ is greater than $n/2$, which contradicts to the assumption. Hence, we have the fact that $v$ is the rumor center on each spanning tree $T_j$, for $j = 1, 2, \ldots, h$, and $v$ is also the rumor center of $G_n$ from (19). \[ \square \]

REFERENCES

[1] Z. Fei, Y. Ryeznik, A. Sverdlov, C. W. Tan, and W. K. Wong, “An overview of healthcare data analytics with applications to the COVID-19 pandemic,” IEEE Transactions on Big Data, 2021.

[2] J. T. Kenper, “On the identification of superspreaders for infectious disease,” Mathematical Biosciences, vol. 48, no. 1-2, pp. 111–127, 1980.

[3] R. A. Stein, “Super-spreaders in infectious diseases,” International Journal of Infectious Diseases, vol. 15, no. 8, pp. 510–513, 2010.

[4] L. Ferretti, C. Wymant, M. Kendall, and et al, “Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing,” Science, vol. 368, no. 6491, pp. 316–329, 2020.

[5] M. Salathé, M. Kazandjieva, J. W. Lee, and et al, “A high-resolution human contact network for infectious disease transmission,” Proc. Natl. Acad. Sci. U.S.A., vol. 107, p. 22020–22025, 2010.

[6] S. Kojaku, L. Hebert-Dufresne, E. Mones, S. Lehmann, and Y.-Y. Ahn, “The effectiveness of contact tracing in emerging epidemics,” PloS One, vol. 1, 2006.

[7] D. Klinkenberg, C. Fraser, and H. Heesterbeek, “The effectiveness of backward contact tracing in networks,” Nature Physics, vol. 1, 2021.

[8] G. Cencetti, G. Santin, A. Longa, E. Pignani, A. Barrat, C. Cattuto, S. Lehmann, M. Salathe, and B. Lepri, “Digital proximity tracing on empirical contact networks for pandemic control,” Nature Communications, vol. 12, no. 1, pp. 1–12, 2021.

[9] W. J. Bradshaw, E. C. Alley, J. H. Huggins, A. L. Lloyd, and K. M. Esveldt, “Bidirectional contact tracing could dramatically improve COVID-19 control,” Nature Communications, vol. 12, no. 1, pp. 1–9, 2021.
[10] Y. Bengio, P. Gupta, T. Maharaj, N. Rahaman, M. Weiss, and et al, “Predicting infectiousness for proactive contact tracing,” in Proc. of Int. Conf. on Learning Representations, 2021.

[11] E. Yoneki and J. Crowcroft, “Epimap: Towards quantifying contact networks for understanding epidemiology in developing countries,” Adhoc Networks, vol. 13, p. 83–93, 2014.

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

[13] D. J. Daley and D. G. Kendall, “Epidemics and rumours,” Nature, vol. 204, no. 1118, 1964.

[14] D. Shah and T. Zaman, “Rumors in a network: Who’s the culprit?,” IEEE Trans. Information Theory, vol. 57, pp. 5163–5181, 2011.

[15] M. Fuchs and P.-D. Yu, “Rumor source detection for rumor spreading on random increasing trees,” Electronic Communications in Probability, vol. 20, no. 2, 2015.

[16] N. Karamchandani and M. Franceschetti, “Rumor source detection under probabilistic sampling,” in Proc. of IEEE Int. Symp. on Information Theory, 2015.

[17] S. Spencer and R. Srikant, “Maximum likelihood rumor source detection in a star network,” in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 2199–2203, Mar. 2016.

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

[19] B. A. Prakash, J. Vreeken, and C. Faloutsos, “Spotting culprits in observations,” in Proc. of IEEE Int. Conf. on Data Mining, pp. 11–20, Dec. 2012.

[20] Z. Wang, W. Dong, W. Zhang, and C. W. Tan, “Rooting out rumor sources in online social networks: The value of diversity from multiple observations,” IEEE Journal of Selected Topics in Signal Processing, vol. 9, pp. 663–677, Jun. 2015.

[21] A. Kumar, V. S. Borkar, and N. Karamchandani, “Temporally agnostic rumor-source detection,” IEEE Trans. on Signal and Information Processing over Networks, vol. 3, pp. 316–329, Jun. 2017.

[22] L. Zheng and C. W. Tan, “A probabilistic characterization of the rumor graph boundary in rumor source detection,” in IEEE Int. Conf. on Digital Signal Processing (DSP), pp. 765–769, Jul. 2015.

[23] K. Zhu and L. Ying, “Information source detection in the SIR model: A sample-path-based approach,” IEEE/ACM Trans. Netw., vol. 24, pp. 408–421, Feb. 2016.

[24] G. Fanti, P. Kairouz, and S. Oh, “Rumor source obfuscation on irregular trees,” in Proc. of ACM SIGMETRICS, 2016.

[25] P. Yu, C. W. Tan, and H. Fu, “Averting cascading failures in networked infrastructures: Poset-constrained graph algorithms,” IEEE Journal of Selected Topics in Signal Processing, vol. 12, pp. 733–748, Aug. 2018.

[26] P. C. Pinto, F. Thiran, and M. Vetterli, “Locating the source of diffusion in complex networks,” Phys. Rev. Lett., vol. 109, p. 068702, Aug 2012.

[27] R. Paluch, X. Lu, K. Suchecki, B. K. Szymański, and J. A. Holyst, “Fast and accurate detection of spread source in large complex networks,” Scientific Reports, vol. 8, Feb 2018.

[28] C. Murphy, E. Laurence, and A. Allard, “Deep learning of contagion dynamics on complex networks,” Nature Communications, vol. 12, no. 1, pp. 1–11, 2021.

[29] S. Chen, P.-D. Yu, C. W. Tan, and H. V. Poor, “Identifying the superspreader in proactive backward contact tracing by deep learning,” in Proc. of 56th Annu. Conf. on Information Sciences and Systems, 2022.

[30] D. Shah and T. Zaman, “Rumor centrality: a universal source detector,” in Proc. of ACM SIGMETRICS, 2012.

[31] W. Dong, W. Zhang, and C. W. Tan, “Rooting out the rumor culprit from online social networks,” in Proc. of IEEE Int. Symp. on Information Theory, 2013.

[32] 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.

[33] J. Khim and P. Loh, “Confidence sets for the source of a diffusion in regular trees,” IEEE Trans. Netw. Sci. Eng., vol. 4, no. 1, pp. 27–40, 2017.

[34] B. Bollobás, The Art of Mathematics: Coffee Time in Memphis. Cambridge University Press, 2006.

[35] M. Salathé, M. Kazandjieva, J. W. Lee, P. Lewis, M. W. Feldman, and J. H. Jones, “A high-resolution human contact network for infectious disease transmission,” Proceedings of the National Academy of Sciences, vol. 107, no. 51, pp. 22202–22205, 2010.

[36] R. Flaxman, Y. Zhang, K. M. Carley, O. Vagian, and H. V. Poor, “The effects of evolutionary adaptations on spreading processes in complex networks,” Proceedings of the National Academy of Sciences of the United States of America, vol. 117, no. 11, p. 5664, 2020.

[37] O. Yagan, A. Sridhar, R. Eletreby, S. Levin, J. Plotkin, and H. V. Poor, “Modeling and analysis of the spread of COVID-19 under a multiple-strain model with mutations,” Harvard Data Science Review, 2021.

[38] Y.-C. Chen, H.-c. Chen, and C.-C. King, “Social network analysis for modeling and source tracing of COVID-19,” Scientific Reports, vol. 13, p. 409–422, oct 1998.

[39] Y.-D. Chen, H.-c. Chen, and C.-C. King, “Social network analysis for contact tracing: Infectious Disease Informatics and Biosurveillance,” vol. 27, pp. 339–358, Oct. 2010.

[40] Hung-Lin Fu

[41] E. A. Hagberg, D. A. Schult, and P. J. Swart, “Exploring network structure, dynamics, and function using networkx,” in Proc. of 56th Annu. Conf. on Information Sciences and Systems, 2021.