Title	On the Universality of Jordan Centers for Estimating Infection Sources in Tree Networks
Author(s)	Luo, Wuqiong; Tay, Wee Peng; Leng, Mei
Citation	Luo, W., Tay, W. P., & Leng, M. (2017). On the Universality of Jordan Centers for Estimating Infection Sources in Tree Networks. IEEE Transactions on Information Theory, 63(7), 4634-4657.
Date	2017
URL	http://hdl.handle.net/10220/43475
Rights	© 2017 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: [http://dx.doi.org/10.1109/TIT.2017.2698504].

On the Universality of Jordan Centers for Estimating Infection Sources in Tree Networks

Wuqiong Luo, Wee Peng Tay, *Senior Member, IEEE* and Mei Leng

***Abstract*—Finding the infection sources in a network when we only know the network topology and infected nodes, but not the rates of infection, is a challenging combinatorial problem, and it is even more difficult in practice where the underlying infection spreading model is usually unknown a priori. In this paper, we are interested in finding a source estimator that is applicable to various spreading models, including the Susceptible-Infected (SI), Susceptible-Infected-Recovered (SIR), Susceptible-Infected-Recovered-Infected (SIRI), and Susceptible-Infected-Susceptible (SIS) models. We show that under the SI, SIR and SIRI spreading models and with mild technical assumptions, the Jordan center is the infection source associated with the most likely infection path in a tree network with a single infection source. This conclusion applies for a wide range of spreading parameters, while it holds for regular trees under the SIS model with homogeneous infection and recovery rates. Since the Jordan center does not depend on the infection, recovery and reinfection rates, it can be regarded as a *universal* source estimator. We also consider the case where there are $k > 1$ infection sources, generalize the Jordan center definition to a k-Jordan center set, and show that this is an optimal infection source set estimator in a tree network for the SI model. Simulation results on various general synthetic networks and real world networks suggest that Jordan center-based estimators consistently outperform the betweenness, closeness, distance, degree, eigenvector, and pagerank centrality based heuristics, even if the network is not a tree.**

***Index Terms*—Infection source estimation, universal source estimator, Jordan center, SIRI model, SIS model.**

I. Introduction

We define an infection to be a property that can be spread probabilistically from one node to another in a network. Examples of infection spreading include a rumor or a piece of news spreading in a social network, a contagious disease spreading in a community, and a computer virus spreading on the Internet. Various models have been developed to describe the spreading process of an infection. In this paper, we consider only discrete time stochastic spreading models. The two simplest models are the Susceptible-Infected (SI) model [1]–[4] and the Susceptible-Infected-Recovered (SIR) model [5], [6]. In the SI model, a susceptible node becomes infected probabilistically at each time step, while an infected node retains the infection forever once it is infected. In the SIR model, an infected node can recover from an infection with a given probability at each time step, upon which it gains immunity from further infections.

With increasing interconnectedness of the world, both physically and online, prompt identification and isolation of infection sources is crucial in many practical applications in limiting the damage caused by the infection, and dealing with the aftermath effectively. Therefore, the problem of infection sources estimation has attracted immense interest from the research community after the pioneering work of [7], which investigates the problem of identifying a single infection source in the SI model. The reference [8] considers single source estimation with a priori knowledge of the set of suspect nodes, while [9] investigates the use of multiple infection spreading instances to identify a source. These methods are based on variants of the distance or rumor centrality of the network graph. We have also developed procedures to identify a source with limited observations of the set of infected nodes [10], and to identify multiple infection sources in [11]. All these works adopt the SI model. Identification of a single infection source in the SIR model was considered in [12], [13], which showed that the Jordan center[1] gives the optimal estimator associated with a most likely infection path. Infection source estimators using a dynamic message passing (DMP) approach [14], and the belief propagation (BP) approach [15] have also been developed for the SIR model. These two approaches however require significant a priori knowledge of the infection spreading process like the infection and recovery rates of each node in the network.

The SI and SIR models have been widely adopted in the literature due to their simplicity, but these models do not adequately reflect many practical situations in which an infected node recovers and becomes infected again at some future time through either a relapse or reinfection. If an individual recovers from a disease such as bovine tuberculosis or human herpes virus, he may later experience a relapse and exhibit infection symptoms again [16]–[19]. The spread of such diseases are often modeled using a Susceptible-Infected-Recovered-Infected (SIRI) model [17]–[19]. On the other hand, if an individual recovers from a disease such as gonorrhea [20], he does not acquire any immunity from his previous infection and may later become reinfected with the same disease. These types of diseases are often modeled using a Susceptible-Infected-Susceptible (SIS) model [21]–[23]. A further example of SIRI and SIS type of infection spreading is rumor spreading in an online social network, as

Part of this work was presented at the 1st IEEE Global Conference on Signal and Information Processing, Austin, TX, December 2013. This work was supported in part by the Singapore Ministry of Education Academic Research Fund Tier 2 grants MOE2013-T2-2-006 and MOE2014-T2-1-028.

W. Luo was with the Nanyang Technological University, Singapore, and is currently with Allianz SE, Singapore Branch. W. P. Tay is with the Nanyang Technological University, Singapore. M. Leng is with the Temasek Laboratories@NTU, Singapore. E-mail: wluo1@e.ntu.edu.sg, wptay@ntu.edu.sg, lengmei@ntu.edu.sg.

[1]The Jordan center of the infected node set is the node in the network with the smallest maximum distance to any observed infected node.

monitored by an external agency that does not have access to the full database of the social network. An individual in the network may post a rumor, remove it, and repost the rumor subsequently. If the external agency only has access to a limited set of the most recent postings of each user (for example, due to storage constraints), then trying to identify the source of the rumor based purely on the time-stamps of the rumor posts will lead to an erroneous result.

To the best of our knowledge, finding infection sources under the SIRI and SIS models have not been investigated. Moreover, all the existing works assume that the underlying infection spreading model is known, and in most cases, the infection and recovery rates of each node are also known. This knowledge may be difficult to obtain in practice. For example, when a new type of infectious disease breaks out, the spreading characteristics of the disease is usually unclear before its epidemiology is determined. Therefore, it would be highly desirable if a source estimator can be shown to be robust, under a reasonable non-trivial statistical criterion, to the underlying spreading mechanism, and *universal* to a wide range of parameters governing the spreading process. Indeed, it is unclear that such an estimator even exists for the SI, SIR, SIRI and SIS models.

In this paper, we adopt the most likely infection path (MLIP) criterion of [10], [12] to find the optimal infection source estimator. Finding optimal source estimators is in general NP-hard, and proving the optimality of an estimator is also in general very challenging, with similar results in the current literature restricted to tree networks and the SI or SIR spreading models [7], [10], [12], [13]. Therefore, any hope of obtaining theoretical optimality guarantees is restricted to special classes of networks. Our work is a small step towards finding optimal source estimators for the more general SIRI and SIS models. Our main contributions are the following:

(i) For an infection spreading from a single source under the SI, SIR, and SIRI models,[2] and over an infinite tree network in which nodes may have different infection and recovery probabilities, we show that the Jordan center of the observed infected node set is an optimal infection source estimator under the MLIP criterion and under some mild technical assumptions. Our result corroborates that in [12], [13], which shows that the Jordan center is the optimal source estimator for the SIR model under assumptions slightly different from ours (cf. Section II for a detailed discussion), and that in [10], which gives the same result for the case where the infection spreading follows the SI model, but only a limited set of infected nodes are observed.

(ii) We show that if an infection spreads according to the SIS model over an infinite regular tree in which all nodes have the same infection and recovery probabilities, then the Jordan center is again the optimal infection source estimator under the MLIP criterion.

(iii) We introduce the concept of a k-Jordan center set, and show that if an infection spreads from $k > 1$ sources in an infinite tree network where nodes may have different infection probabilities, and in accordance to the SI model, then the k-Jordan center set is an optimal estimator of the infection source set under the MLIP criterion. A heuristic procedure was proposed in [11] to determine multiple infection sources in the SI model based on the single source maximum likelihood (ML) estimator for regular trees, but not shown to be optimal. Simulation results suggest that our estimator outperforms that in [11] in terms of the average error distance.

(iv) We extend the Jordan center-based estimators above heuristically to general graph networks, and perform extensive simulations to verify the performance of our estimators. We perform infection spreading simulations on random trees, part of the Facebook network, and the western states power grid network of the United States. In our simulation results, the Jordan center-based estimators consistently achieve the lowest average error distance compared to the betweenness, closeness, distance, degree, eigenvector, and pagerank centrality based heuristics.

Finding the Jordan center does not require knowledge of each node's infection and recovery probabilities. Therefore, our result in item (i) shows that the Jordan center is a universal source estimator for the SI, SIR and SIRI models, under a wide range of spreading parameters. In contribution (ii), we show that the Jordan center is also optimal for the SIS model in regular tree networks. Although we are not able to show that this is true for general graphs and for multiple infection sources, our simulation results suggest that Jordan center-based source estimators outperform many other source estimators, which similarly do not require knowledge of the underlying infection spreading parameters, regardless of which of the four considered infection spreading models is used. This is somewhat surprising since the SI, SIR, SIRI, and SIS spreading mechanisms are quite different from each other. Note that although [14] and [15] have reported better source detection rates in numerical experiments using the DMP and BP approaches respectively, these methods require the knowledge of the underlying infection spreading parameters, and are applicable only to the SIR model. There is also a lack of theoretical results on the optimality of the DMP and BP approaches, and extending them to the SIRI and SIS models is highly non-trivial [24]. We hope that the insights derived from our current work will inform future design of better source estimators in the case where the exact values of infection parameters are unknown.

The rest of this paper is organized as follows. In Section II, we present our system model, assumptions and problem formulation. In Section III, we show, under some technical conditions, that the Jordan center is an optimal source estimator for tree networks when there is a single infection source. In Section IV, we derive an estimator for tree networks when there is an infection spreading from multiple sources under the SI model. In Section V, we heuristically extend the proposed

[2]By setting the recovery probability and relapse probability in the SIRI model to zero, we obtain the SI and SIR models, respectively. However, in this paper, for clarity and due to some differences in the assumptions we make under each of these models, we explicitly differentiate the SIRI model from the SI and SIR models.

estimators to general graphs and propose heuristic algorithms to find them. We present simulation results in Section VI to verify the effectiveness of the proposed estimators. Finally we conclude and summarize in Section VII.

II. Problem Formulation

In this section, we present our system model, assumptions, and various notations used throughout this paper. We also describe the most likely infection path criterion. A table summarizing the most commonly used notations is provided at the end of this section.

A. Infection Spreading Model

We model the underlying network over which an infection spreads as an undirected graph $G = (V, E)$, where V is the set of nodes and E is the set of edges. Two nodes connected by an edge are called neighbors or neighboring nodes. Suppose that an infection starts spreading from one or more source nodes. In most of this paper, we will assume a single source node, and extend this to multiple source nodes for the SI model described below. We adopt a discrete time spreading model in which time is divided into discrete slots, and the states of the nodes in the graph G follow a Markov process with probability measure $\mathbb{P}$. Our goal is to infer the infection sources from observations of the infected nodes at a particular point in time. We consider the following four discrete time infection spreading models.

1) *SI model*: In the SI model, each node takes on one of 3 possible states: *susceptible* (**s**), *infected* (**i**) and *non-susceptible* (**n**). At any time slot, if a node is infected, we say that it is in state **i**. The set of uninfected nodes that have infected neighbors are in state **s**, and are called susceptible nodes. In the SI model, an infected node remains infected forever, and a susceptible node becomes infected probabilistically in the next time slot. All other nodes are in state **n**, and are called non-susceptible nodes. A non-susceptible node has probability zero of becoming infected in the next time slot.
2) *SIR model*: In the SIR model, the possible node states are *susceptible* (**s**), *infected* (**i**), *non-susceptible* (**n**), and *recovered* (**r**). The only difference with the SI model is that an infected node in state **i** in a time slot may recover to state **r** in the next time slot with a positive probability. A recovered node then stays in the recovered state **r** forever. In other words, a recovered node will never become infected again.
3) *SIRI model*: The possible nodes states in the SIRI model are the same as for the SIR model. The difference from the SIR model is that a recovered node (in state **r**) may become infected again at a future time slot with a positive probability. This infection relapse is spontaneous, and can take place even if the node does not have any infected neighbors. Here, we reserve the state **s** for those nodes that have infected neighbors and have never been infected before.
4) *SIS model*: In the SIS model, the possible node states are *susceptible* (**s**), *infected* (**i**) and *non-susceptible* (**n**). This model describes a more complicated spreading process where once an infected node recovers from the infection (with a positive probability), it immediately becomes a susceptible node (if it has at least one infected neighbor) or non-susceptible node (if it does not have any infected neighbor). There is therefore no *recovered* state in this model.

For any node $v \in V$, we let $p_{\mathbf{s}}(v)$, $p_{\mathbf{i}}(v)$ and $p_{\mathbf{r}}(v)$ be the probability for v to be in state **i** in the next time slot conditioned on v being susceptible, infected, or recovered in the current time slot, respectively. These probabilities characterize different infection spreading models, and we assume that they satisfy the following Assumptions 1–4. Let $\alpha = \min_{u \in V} p_{\mathbf{s}}(u)$ and $\beta = \max_{u \in V} p_{\mathbf{s}}(u)$. For simplicity, we assume that $p_{\mathbf{s}}(v)$, $p_{\mathbf{i}}(v)$ and $p_{\mathbf{r}}(v)$ do not change over time slots for each v, although all our results and proofs (with slight modifications) are still valid if these probabilities are time-varying as long as Assumptions 1–4 hold over all time slots.

Assumption 1. *Under the SI model, for every $v \in V$, we have*

$$\beta \leq \frac{\alpha}{(1-\alpha)^2}. \tag{1}$$

See Fig. 1 for the region where (α, β) satisfies the inequality (1). For example, if $\alpha \geq 0.382$, then (1) holds since its right hand side is greater than 1. In the inequality (1), we assume that the infection probabilities at each node in the network does not differ drastically for the SI model. This is required because in this work, we do not assume knowledge of the exact infection rates at each node. Therefore, if part of the network has nodes that are much easier to infect than other nodes, then any estimator with no knowledge of the infection rates will result in a highly biased result, which may not do better on average than making random choices for the infection sources. We provide an example in Fig. 3 to show that Assumption 1 is a necessary condition for Theorem 1 to hold for the SI model.

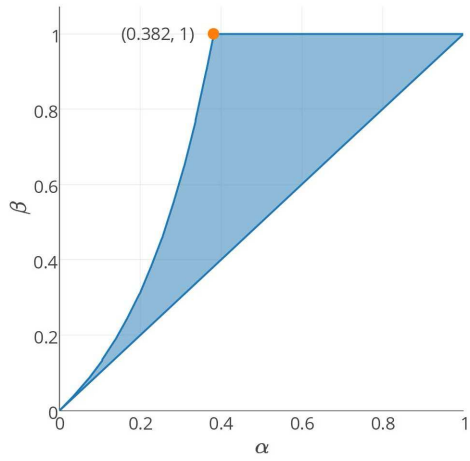


Fig. 1. Illustration of the region where (α, β) satisfies (1).

Assumption 2. *Under the SIR model, for every $v \in V$, we have*

$$0 \leq p_{\mathbf{i}}(v) \leq \sqrt{\frac{\alpha}{\beta}}. \tag{2}$$

The reference [12] assumes that $p_{\mathbf{s}}(v)$ is the same for every $v \in V$, which implies that $\alpha = \beta$, and (2) then reduces to the trivial condition $0 \leq p_{\mathbf{i}}(v) \leq 1$. It also assumes that $p_{\mathbf{i}}(v)$ is the same for every $v \in V$. Therefore, the setup in [12] is a special case of the problem studied in this paper. On the other hand, the reference [13] considers the SIR model under a heterogeneous setting, where an infection is transmitted across each edge (u, v) with probability $p(u, v)$ so that $p_{\mathbf{s}}(v) = 1 - \prod_{u \in N_v}(1 - p(u, v))$, where N_v is the set of infected neighbors of node v at the beginning of the current time slot. However, since [13] considers undirected graphs with $p(u, v) = p(v, u)$ for all edges (u, v) (see Fig. 3 for a counterexample if edge infection probabilities are not symmetric), no additional assumptions are required to show that the Jordan center is an optimal estimator under the MLIP criterion for an infinite tree, where each node has degree at least 2. In a social network, the strength of influence might not be symmetric between each pair of friends. Therefore, we do not make this assumption.

Assumption 3. *Under the SIRI model, for every $v \in V$, we have*

$$\frac{\beta - \alpha}{1 - \alpha} \leq p_{\mathbf{i}}(v) \leq \sqrt{\frac{\alpha}{\beta}}, \tag{3}$$

$$1 - \sqrt{\frac{\alpha}{\beta}} \leq p_{\mathbf{r}}(v) \leq \min\left\{1, \sqrt{\frac{\alpha}{\beta}} \frac{p_{\mathbf{i}}(v)}{1 - p_{\mathbf{i}}(v)}\right\}. \tag{4}$$

See Fig. 2 for the region of (α, β) that makes (3) feasible. Note that if $\alpha = \beta$, (3) reduces to $0 \leq p_{\mathbf{i}}(v) \leq 1$, and (4) reduces to $0 \leq p_{\mathbf{r}}(v) \leq \min\left\{1, \frac{p_{\mathbf{i}}(v)}{1 - p_{\mathbf{i}}(v)}\right\}$. Inequality (4) implies that a node does not easily relapse into an infected state (i.e., small $p_{\mathbf{r}}$) if it recovers quickly (i.e., small $p_{\mathbf{i}}$). This is intuitively appealing as it corresponds to the case where if an infected node has a low probability of staying infected in the next time slot, then it is unlikely for the node to relapse into the infection once it has recovered. A practical example is: it is hard to re-convince someone to believe a rumor if he already has a reason to reject the rumor.

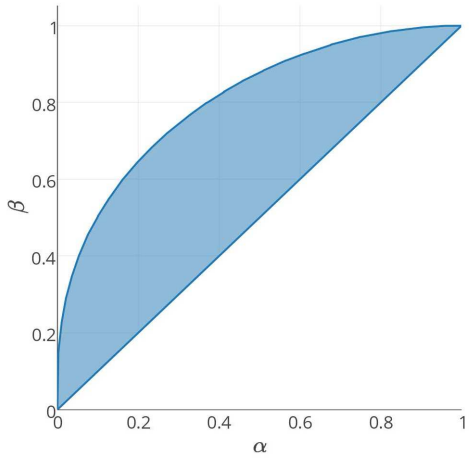


Fig. 2. Illustration of the region where (α, β) satisfies (3).

Assumption 4. *Under the SIS model, for every $v \in V$, we have*

$$p_{\mathbf{s}}(v) = p_{\mathbf{s}},$$
$$p_{\mathbf{i}}(v) = p_{\mathbf{i}},$$
$$0 \leq p_{\mathbf{s}} \leq p_{\mathbf{i}} \leq 1. \tag{5}$$

Inequality (5) helps us to avoid the case where an infection spreads very fast (i.e., large $p_{\mathbf{s}}$) and infected nodes also recover relatively quickly (i.e., $p_{\mathbf{i}} < p_{\mathbf{s}}$) from happening. In such cases, infected nodes close to sources are likely to have recovered by the time we observe the state of the network, while there may be a significant set of infected nodes at a distance away from the source. Therefore, trying to estimate the source nodes will result in a large bias.

In Assumptions 1-3 for the SI, SIR and SIRI models, the infection, recovery, and relapse probabilities can vary between different nodes. We call such networks *heterogeneous*. On the other hand, in Assumption 4, the infection and recovery probabilities are the same for all nodes in the network. We call such networks *homogeneous*.

B. Most Likely Infection Path Sources Estimator

In this subsection, we present the MLIP statistical criterion that we adopt to find the optimal infection sources in this paper. We focus on the single source formulation in the following description, and extend to the multiple sources case in Section IV. The following exposition and definitions follow mainly from [10], and is repeated here for completeness.

Let $\mathbf{X}(u, t)$ be a random variable denoting the state of a node u in time slot t. At time 0, suppose that there is a single infected node $s^* \in V$, which we call the infection source. Let $\mathbf{X}^t = \{\mathbf{X}(u, \tau) : u \in V, 1 \leq \tau \leq t\}$ be the collection of the states of all nodes in V from time 1 to t. A realization $X^t = \{X(u, \tau) : u \in V, 1 \leq \tau \leq t\}$ of $\mathbf{X}^t$ is an *infection path*. At time t, we observe the set of nodes that are currently infected. The observed set of infected nodes is denoted $V_{\mathbf{i}}$ and is assumed to be non-empty. We assume that the elapsed time t is unknown. We say that an infection path X^t is *consistent* with $V_{\mathbf{i}}$ if we have $X(u, t) = \mathbf{i}$ for all $u \in V_{\mathbf{i}}$ and no other nodes in V is infected in X^t. Conditioned on s being the infection source, we let $\mathcal{X}_s$ be the set of all possible infection paths consistent with $V_{\mathbf{i}}$, and $\mathcal{T}_s$ be the set of the corresponding feasible elapsed times.

We want to estimate the infection source based only on knowledge of $V_{\mathbf{i}}$ and the underlying graph G. Finding the ML estimator for a single infection source in the SI model for a general graph network is a #P-complete problem [7]. (Note that [7] considers a spreading model in which the propagation time of the infection across an edge has exponential distribution with rate 1. Due to the memoryless property of the exponential distribution, the problem of estimating the source in [7]'s model can be reduced to the problem of estimating the source in a discrete time spreading model where time is discretized into unit intervals, and the probability of an infection spreading across an edge in each time slot is $1 - e^{-1}$. Therefore, under the discrete time spreading model, finding the ML estimator is also a #P-complete problem.) We consider

instead an alternative statistical criterion first proposed by [12], and given by

$$\hat{s} \in \arg \max_{\substack{s \in V \\ t \in \mathcal{T}_s, X^t \in \mathcal{X}_s}} \mathbb{P}(\mathbf{X}^t = X^t \mid s^* = s). \tag{6}$$

The basic idea behind (6) is to estimate the source as the node associated with a *most likely infection path* out of all possible infection paths that are consistent with $V_{\mathbf{i}}$. The search of a most likely infection path depends not only on the elapsed time but also on the structure of the underlying graph. Even at a given elapsed time, the number of consistent paths cannot be calculated easily, and the most likely infection paths are not unique. Solving (6) directly involves searching over both $\mathcal{T}_s$ and $\mathcal{X}_s$, whose size increases exponentially fast with the number of nodes. In order to derive insights into an optimal source estimator for (6), we first consider the network with a single source in Section III. With the utilization of some properties of the elapsed times, we reduce the objective function to a simpler formulation and derive an estimator for all four considered infection spreading models. In Section IV, we generalize the idea to a tree network with multiple sources for the SI model.

C. Some Notations and Definitions

In this subsection, we list some notations and definitions that we use throughout this paper. We refer the reader to a summary of basic notations given in Table I.

For a given tree network A with v being the root, we assign directions to each edge of A so that all edges point towards v. For any $u \in A$, let $\mathrm{pa}(u)$ be the parent node of node u (i.e., the node with an incoming edge from u), and $\mathrm{ch}(u)$ be the set of child nodes of u in A (i.e., the set of nodes with outgoing edges to u).

For any infection path X^t, a subset $J \subset V$, and $0 \leq i \leq j \leq t$, let $X^t(J,[i,j])$ be the states of nodes in J from time slots i to j in the infection path X^t. To avoid cluttered expressions, we abuse notations and let

$$P_s\left(X^t(J,[i,j])\right) \triangleq \mathbb{P}\Big(\mathbf{X}^t(J,[i,j]) = X^t(J,[i,j]) \mid s^* = s\Big).$$

Therefore, $P_s(X^t)$ represents the probability of X^t conditioned on s being the source and t being the elapsed time. Moreover, when we want to remind the reader of the state of a node u at a specific time in the conditional probability $P_s(X^t)$, we use the notation $P_s(X(u,i) = a)$, where $a \in \{\mathbf{i}, \mathbf{s}, \mathbf{r}, \mathbf{n}\}$ is the state of u at time i.

Definition 1 (Most likely infection paths)**.** *For any $s \in V$ and any feasible elapsed time $t \in \mathcal{T}_s$, we say that an infection path X^t is most likely for (s,t) if $X^t \in \arg\max_{\tilde{X}^t \in \mathcal{X}_s} P_s(\tilde{X}^t)$. Moreover, an infection path X^t is called a most likely infection path if there exists some $s \in V$, and $t \in \mathcal{T}_s$ such that*

$$P_s(X^t) = \max_{u \in V, r \in \mathcal{T}_u, Y^r \in \mathcal{X}_u} P_u(Y^r).$$

Definition 2 (Jordan center)**.** *For any node $s \in V$, let its infection range be*

$$\bar{d}(s, V_{\mathbf{i}}) \triangleq \max_{u \in V_{\mathbf{i}}} d(s,u).$$

Any node in G with minimum infection range is called the Jordan center of $V_{\mathbf{i}}$.

Finally, in several of our proofs, we need to differentiate between subtrees that have infected nodes or not.

Definition 3 (Uninfected subtree and infected subtree)**.** *Suppose that v is the infection source. For any node u, we say that $T_u(v;G)$ is an* uninfected subtree *if* [3]

$$T_u(v;G) \bigcap V_{\mathbf{i}} = \emptyset;$$

and we say that $T_u(v;G)$ is an infected subtree *if*

$$T_u(v;G) \bigcap V_{\mathbf{i}} \neq \emptyset.$$

III. Single Source Estimation for Trees

In this section, we show that a Jordan center of the infected node set $V_{\mathbf{i}}$ is an optimal infection source estimator universally applicable for infection spreading under the SI, SIR, and SIRI models for trees, and the SIS model for regular trees. The Jordan center has previously been shown to be optimal estimators for SI infection spreading [10] and for the SIR model [12], [13], but under different technical assumptions.

As noted in Section I, proving optimality results for infection source estimators is in general challenging. In most of this paper, we restrict ourselves to the following specific graph networks depending on the infection spreading model. We say that a tree is an *infinite tree* if every node in it has degree at least two.

Assumption 5. *For an infection spreading according to the SI, SIR or SIRI models, the underlying graph G is an infinite tree. For an infection spreading according to the SIS model, the underlying graph G is a regular infinite tree, i.e., every node has the same degree.*

For the SI, SIR, and SIRI models, Assumption 5 is adopted to avoid boundary effects. Consider the extreme case where a source node has only one neighbor. Then, the infection can spread away from the source in only one direction. In this case, any estimator based only on the graph topology is expected to perform badly. In the SIS model, a recovered node is the same as a susceptible node, which leads to more complex evolution of the node states in the network as compared to the SIRI model in which the state evolution of a recovered node becomes independent from the rest of the network. To simplify the problem, we restrict to regular trees for the SIS model in Assumption 5. The problem of finding optimal source estimators for the SIS model in more general network topologies remains open.

A. Most Likely Elapsed Time

We assume no knowledge of the elapsed time when the set of nodes $V_{\mathbf{i}}$ is observed. Suppose that $v \in V$ is the source, then the feasible set of all elapsed times is given by $\mathcal{T}_v = [\bar{d}(v, V_{\mathbf{i}}), +\infty)$, where the lower bound is the minimum amount of time required for the infection to spread from v to

[3] See Table I for the definition of $T_u(v;G)$.

TABLE I
SUMMARY OF NOTATIONS

s^*	the true infection source
$G=(V,E)$	the underlying graph network
α	$\min_{u\in V} p_{\mathbf{s}}(u)$
β	$\max_{u\in V} p_{\mathbf{s}}(u)$
H_v	the minimum connected subgraph of G that contains $V_{\mathbf{i}}$ and the node v
$\lvert A\rvert$	the number of elements in A if A is a set, or the number of nodes in A if A is a graph
$V(u,i)$	the set of nodes i hops away from node u
$T_u(v;A)$	the subtree rooted at node u of the tree A, with the first link of the path from u to v in A removed
$d(s,u)$	the length of the shortest path between s and u in the graph G (i.e., the distance between them)
t_s	a most likely elapsed time conditioned on s being the infection source
$\mathcal{X}_s$	the set of all possible infection paths consistent with $V_{\mathbf{i}}$ conditioned on s being the source
$\mathcal{T}_s$	the set of the feasible elapsed time corresponding to $\mathcal{X}_s$

all the nodes in $V_{\mathbf{i}}$. It is obviously computationally inefficient to search over all elapsed times. In Proposition 1, we show how to find a *most likely elapsed time* t_v that maximizes the probability of observing $V_{\mathbf{i}}$.

Proposition 1. *Suppose that Assumptions 1–4 hold, $v\in V$ is the infection source, and a non-empty set of infected nodes $V_{\mathbf{i}}$ is observed. For an infection under the SI, SIR, SIRI or SIS model in a network satisfying Assumption 5, we have for any $t\in\mathcal{T}_v$, and any two most likely infection paths X^t for (v,t) and Y^{t+1} for $(v,t+1)$,*

(a) $P_v(Y^{t+1})\le\delta P_v(X^t)$, where $\delta=(1-\alpha)^2, \sqrt{\frac{\alpha}{\beta}}, \sqrt{\frac{\alpha}{\beta}}$ and 1 for the SI, SIR, SIRI and SIS model, respectively; and

(b) conditioned on v being the infection source, a most likely elapsed time is given by

$$t_v=\bar{d}(v,V_{\mathbf{i}}).$$

The proof of Proposition 1 is provided in Appendix A. Proposition 1(b) shows a universal property that is robust to the underlying infection spreading models: a most likely elapsed time t_v is the infection range of v (cf. Definition 2). Moreover, Proposition 1(a) shows that a most likely elapsed time should be as small as possible. This result is intuitive. Consider the conditional probability

$$P_v(X^t)=\prod_{u\in V,\tau\in[1,t]}P_v(u,\tau),$$

where the value of each term in the product on the right hand side is at most 1. When t decreases, there are less terms in the product, which in turn increases the value of $P_v(X^t)$.

Following Proposition 1, the problem in (6) is now reduced to

$$\hat{s}\in\arg\max_{\substack{v\in V,t_v=\bar{d}(v,V_{\mathbf{i}})\\ X^{t_v}\in\mathcal{X}_v}}P_v(X^{t_v}).$$

After the most likely elapsed time has been identified, we can now proceed to find the source node associated with the most likely infection path.

B. Source Associated With the Most Likely Infection Path

In this subsection, we derive the source estimator associated with a most likely infection path for all four considered infection spreading models, under specific graph networks. Although Proposition 1 gives a most likely elapsed time t_v conditioned on a node $v\in V$ being the infection source, it is still difficult to count the number of infection paths that are consistent with $V_{\mathbf{i}}$, not to mention finding the most likely infection path for (v,t_v). Therefore, instead of directly looking for the most likely infection path, we first consider the conditional probabilities $P_v(X^{t_v})$ and $P_u(Y^{t_u})$ of two infection paths, where v and u are a pair of neighboring nodes, X^{t_v} is a most likely infection path for (v,t_v), and Y^{t_u} is a most likely infection path for (u,t_u). We then show that if v has a smaller infection range, $P_v(X^{t_v})$ is not less than $P_u(Y^{t_u})$. Upon establishing this neighboring node relationship, we can find a path on which the infection range of each node is decreasing, and the conditional probability of the most likely infection path is non-decreasing. This in turn implies that the Jordan center of $V_{\mathbf{i}}$ is the source estimator we are looking for. The neighboring node relationship is summarized in Proposition 2, the proof of which is provided in Appendix B.

Proposition 2. *Suppose that $V_{\mathbf{i}}$ is non-empty. For an infection process under the SI, SIR, SIRI or SIS model satisfying Assumptions 1-5, and for any pair of neighboring nodes u and v, we have*

$$P_v(X^{t_v})\ge P_u(Y^{t_u}),\ \text{if } t_v<t_u,$$

where X^{t_v} and Y^{t_u} are most likely infection paths for (v,t_v) and (u,t_u) respectively.

We note that Proposition 1 and Proposition 2 match Proposition 2 and Lemma 4 in [10], respectively. Then following the same proof as Theorem 1 in [10], we have the following result.

Theorem 1. *Suppose that $V_{\mathbf{i}}$ is non-empty. For an infection process under the SI, SIR, SIRI or SIS model satisfying Assumptions 1–4, respectively, and Assumption 5 holds, a Jordan center of $V_{\mathbf{i}}$ is an optimal source estimator for* (6).

Theorem 1 shows that for regular infinite trees, a Jordan center is an optimal source estimator, regardless of which of the four considered infection spreading model the infection is following. This is a somewhat surprising result since the four infection spreading models are fundamentally different. The "universality" of the Jordan center makes it highly desirable in practice, where the underlying infection spreading model is

usually unknown a priori. A distributed linear time complexity algorithm has been proposed in [10] to find the Jordan center in a tree, which makes timely estimation of the infection source possible.

In Fig. 3, we provide an example to show that Assumption 1 is a necessary condition for Theorem 1 to hold for the SI infection process. Similar examples can be used to show that Assumptions 2, 3 and 4 are necessary for Theorem 1 to hold for the SIR, SIRI and SIS infection process, respectively.

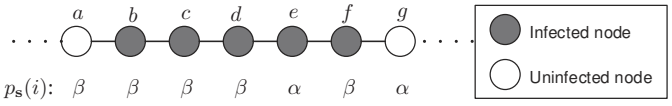


Fig. 3. An example that shows Assumption 1 is a necessary condition for Theorem 1 to hold for the SI infection process. Suppose the infection process follows the SI model. If node d (Jordan center of $V_{\mathbf{i}}$) is the infection source, for an infection path X^2 with elapsed time 2, we have $P_d(X^2) = \beta^3\alpha$. On the other hand, if node e is the infection source, for an infection path Y^3 with elapsed time 3 and in which node f is infected in the first time slot, we have $P_e(Y^3) = \beta^4(1-\alpha)^2$. For d to be the optimal MLIP estimator, we require $P_d(X^2) \geq P_e(Y^3)$, which in turn requires $\beta \leq \alpha/(1-\alpha)^2$.

IV. Multiple Sources Estimation for SI Infection Spreading in Trees

In this section, we restrict our discussion to an infection spreading under the SI model. Since the optimal single infection source estimator has been shown to be the Jordan center of $V_{\mathbf{i}}$ in Section III, we consider here the case where there are $k > 1$ infection sources, i.e., $S^* = \{s_1^*, s_2^*, \cdots, s_k^*\}$. Then the most likely infection path based sources estimation problem becomes

$$\hat{S} \in \arg \max_{\substack{S \subset V, |S|=k \\ t \in \mathcal{T}_S, X^t \in \mathcal{X}_S}} \mathbb{P}(\mathbf{X}^t = X^t \mid S^* = S). \tag{7}$$

The definitions in Section II-C are similarly generalized to the case of k infection sources by replacing s with S in the definitions. In particular, we generalize the Jordan center definition to k-Jordan infection center set.

Definition 4 (k-Jordan center set)**.** *The infection range of a set of source nodes $S = \{s_1, s_2, \cdots, s_k\}$ is defined as*

$$\bar{d}(S, V_{\mathbf{i}}) = \max_{u \in V_{\mathbf{i}}} \min_{s_i \in S} d(s_i, u).$$

The set of k nodes in G with minimum infection range is called the k-Jordan center set of $V_{\mathbf{i}}$.

Without loss of generality, we assume that the minimum subgraph B of G that contains $V_{\mathbf{i}}$ is connected, otherwise the same estimation procedure can be applied to each component of B. We first show a similar result as that in Proposition 1. The proof of Proposition 3 is provided in Appendix C.

Proposition 3. *Suppose that the underlying network G is an infinite tree, the infection sources are $S = \{s_1, s_2, \cdots, s_k\}$, and the set of observed infected nodes $V_{\mathbf{i}}$ is non-empty. For an infection spreading under the SI model, any most likely infection path X^t for (S, t) has the following properties:*
(a) $P_S(X^t)$ is non-increasing in $t \in \mathcal{T}_S$; and
(b) conditioned on S being the infection sources, a most likely elapsed time for X^t is given by

$$t_S = \bar{d}(S, V_{\mathbf{i}}).$$

In the following, we show how to transform the k sources estimation problem to an equivalent single source estimation problem, then we can use Theorem 1 to find the optimal multiple sources estimator. We first introduce the definition of *super node graph*. See Fig. 4 for an illustration of the super node graph construction.

Definition 5 (Super node graph)**.** *Suppose that G is an infinite tree. Given a set $S = \{s_1, s_2, \cdots, s_k\} \subset V$, where $k > 1$, and any infection path X^t conditioned on S being the infection sources, the super node graph $\tilde{G}(S, X^t)$ is constructed using the following procedure for each $\tau = 0, 1, \ldots, t$:*

- *Starting at $\tau = 0$, we initialize $A_i = \{s_i\}$ for each $i = 1, \ldots, k$.*
- *For each $\tau = 1, \ldots, t$, consider every node $v \in V_{\mathbf{i}}$ that becomes susceptible at time τ in X^t for the first time. Let N_v be the set of neighboring nodes of v that is infected at time $\tau - 1$. We choose a node $u \in N_v$ uniformly at random, and include v and the edge (u, v) in the component A_i that u belongs to.*
- *Based on the resulting graph $\mathcal{A} = \bigcup_{i=1}^k A_i$, the super node graph $\tilde{G}(S, X^t)$ is constructed by considering all infection sources as a single virtual node, which we call a super node and denote as Supernode(S).*

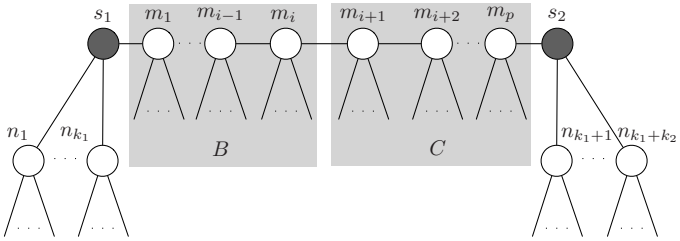
In summary, we trace the infection path X^t and assign each infected node to the tree A_i if its infection comes from s_i, with ties broken randomly. This then partitions the infection graph G into disjoint trees rooted at each $s_i \in S$. The trees are connected together to form the super node graph by treating S as a single "super node".

Given any infection path X^t following the SI model, it can be shown that (with probability one) the conditional probability $P_S(X^t)$ is the same for G and any corresponding $\tilde{G}(S, X^t)$ as defined in Definition 5. Consider any node v with $|N_v| > 1$ and assume v becomes susceptible at time slot t_1 and becomes infected at time slot t_2. Then $P_S(X^t(v, [1, t])) = (1 - p_{\mathbf{s}})^{t_2 - t_1 - 1} p_{\mathbf{s}}$, regardless of the number of infected neighbors v has as long as there is at least one infected neighbor.[4] We formally present this result in the following lemma.

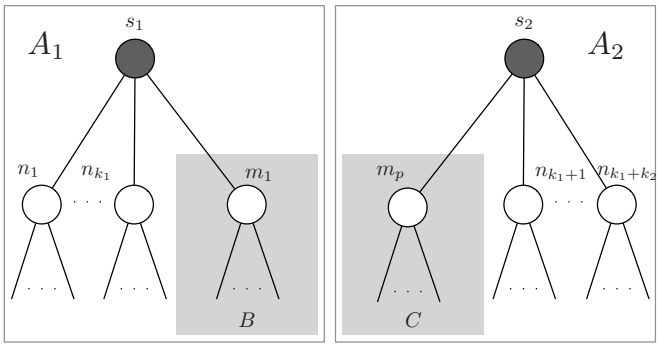
Lemma 1. *Let $S = \{s_1, s_2, \cdots, s_k\} \subset V$, where $k > 1$. Given any infection path X^t conditioned on S being the infection sources, $P_S(X^t)$ is the same for both G and any corresponding $\tilde{G}(S, X^t)$ as defined in Definition 5.*

Following Lemma 1, instead of searching for a most likely infection path for S in G, we can now search for a most likely infection path for Supernode(S) in a corresponding super node graph $\tilde{G}(S, X^t)$. In this way, we transform the k sources estimation problem to an equivalent single source estimation problem. As discussed in Section III, Theorem 1

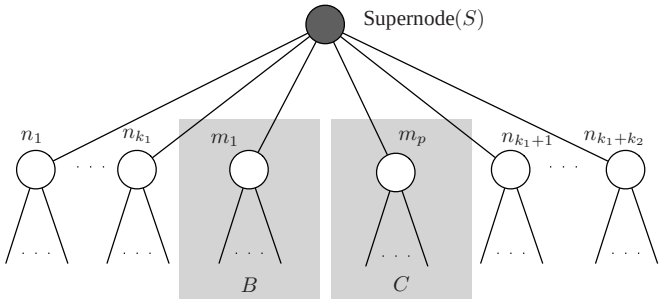
[4]This property does not hold for an infection following the SIR, SIRI or SIS model, where some infected neighbors of v may recover after t_1 and $P_S(X^t(v, [1, t]))$ may change if we remove some edges connecting v.



(a) Given any infection path X^t, suppose m_i becomes susceptible at time τ_1 in X^t for the first time. Suppose m_{i+1} becomes susceptible at time $\tau_2 > \tau_1$ in X^t for the first time when m_{i+2} becomes infected, while m_i stays susceptible at time τ_2. Then nodes in B belong to component A_1 and nodes in C belong to component A_2.



(b) Partition G into $\mathcal{A} = A_1 \bigcup A_2$.



(c) Constructed super node graph $\tilde{G}(S, X^t)$.

Fig. 4. Illustration of the construction of the super node graph $\tilde{G}(S, X^t)$ from an infinite tree G with $S = \{s_1, s_2\}$ being the infection sources.

shows that a Jordan center of the infected node set is an optimal single source estimator. Therefore, our objective is to find a set of k nodes S, where Supernode(S) is a Jordan center of the infected node set in $\tilde{G}(S, X^t)$. We show in the following lemma that k-Jordan center set is the solution.

Lemma 2. *Suppose that G is an infinite tree and the set of infected nodes $V_\mathbf{i}$ is non-empty. Given any infection path X^t consistent with $V_\mathbf{i}$ under the SI model, if $S = \{s_1, s_2, \cdots, s_k\}$ is the k-Jordan center set of $V_\mathbf{i}$ in G, then Supernode(S) is a Jordan center of $V_\mathbf{i}$ in any corresponding super node graph $\tilde{G}(S, X^t)$.*

The proof of Lemma 2 is provided in Appendix D. The following theorem follows immediately from Lemma 2 and Theorem 1.

Theorem 2. *Suppose that G is an infinite tree and there are $k > 1$ infection sources. For an infection in the SI model, a k-Jordan center set of $V_\mathbf{i}$ is an optimal source set estimator for* (7)*.*

Theorem 2 is consistent with Theorem 1 for an infection in the SI model. Due to the difficulty described in footnote 4, the multiple-sources estimation problem remains an open problem for more complicated infection spreading models including SIR, SIRI and SIS models. To verify the robustness of the proposed estimators, we conduct extensive simulations on both trees and general networks for SI, SIR, SIRI and SIS models in Section VI.

V. Source Estimation for General Graphs

In this section, we consider the case where the underlying network G is a general graph. Inspired by the robustness of Jordan center estimators in tree networks, we heuristically extend them to general graphs. We first review an algorithm, proposed in [12], that finds the Jordan center for $k = 1$. We then propose a heuristic algorithm to find the k-Jordan center set for $k > 1$.

A. Single Jordan Center Estimation Algorithm

A simple algorithm was proposed in [12] to find the Jordan center of $V_\mathbf{i}$ when there is a single source and the underlying network is a general graph. Let any node in $V_\mathbf{i}$ broadcast a message containing its own identity. The first node that receives messages from every node in $V_\mathbf{i}$ declares itself as a Jordan center and the algorithm terminates. We call this algorithm the Single Jordan Center estimation algorithm (SJC), with a computational complexity of $O(|V||E|)$.

B. Multiple Jordan Center Set Estimation Algorithm

When k is greater than 1, it is usually impractical to use exhaustive search methods to find the k-Jordan center set as the number of possible k-Jordan center sets is $\binom{|V|}{k}$. Therefore, we propose a heuristic algorithm to find an approximate k-Jordan center set when there are $k > 1$ sources and the underlying network is a general graph, which we call the Multiple Jordan Center set estimation algorithm (MJC). MJC starts with randomly selecting a set of k nodes $\hat{S}^0 = \{s_i^0\}_{i=1}^k$ as the initial guess, and then utilizes an iterative two-step optimization approach. Specifically, in iteration l, let $\hat{S}^l = \{s_i^l\}_{i=1}^k$ be the k-Jordan center set estimate. We perform the following two steps at each iteration l:

- **Partition step**. In this step, MJC partitions $V_\mathbf{i}$ into k sets $M_1, M_2, \cdots, M_k$ such that for all $v \in M_i$, $d(s_i^{l-1}, v) \leq d(s_j^{l-1}, v)$ if $i \neq j$. We call M_i the *Voronoi set* corresponding to s_i^{l-1}. To do this, let each s_i^{l-1} broadcast a message. The broadcasting process terminates when each node $v \in V_\mathbf{i}$ receives at least one message from a node in $\hat{S}^{l-1}$. In the broadcasting process, each node $v \in V_\mathbf{i}$ learns the distance between itself and the nearest nodes in $\hat{S}^{l-1}$. We choose a nearest node in $\hat{S}^{l-1}$ at random, and add v to the Voronoi set corresponding to this node.
- **Re-optimization step**. In this step, MJC updates each estimate s_i^{l-1} in the Voronoi sets M_i. For each Voronoi

set M_i, run SJC to find the Jordan center of M_i and set it as the new estimate s_i^l.

MJC terminates when $\max_{1\le i\le k} d(s_i^{l-1}, s_i^l) \le \eta$ for some predetermined small positive value η or when the number of iterations reach a predetermined positive number MaxIter. For the partition step in each iteration, the computation complexity is dominated by the broadcasting process, with a computational complexity of $O(k|E|)$. For the re-optimization step in each iteration, the computational complexity is $O(|V||E|)$ due to SJC. Therefore, the overall computational complexity for MJC is $O(\text{MaxIter}\cdot|V||E|)$. We show in the following proposition that the infection range is non-increasing over the iterations of MJC. The proof of Proposition 4 is provided in Appendix E.

Proposition 4. *Suppose that G is a general graph and there are $k>1$ infection sources. The infection range (cf. Definition 4) of is non-increasing over the iterations of MJC, i.e.,*

$$\bar{d}(\hat{S}^l, V_{\mathbf{i}}) \le \bar{d}(\hat{S}^{l-1}, V_{\mathbf{i}}).$$

VI. Simulation Results

In this section, we present simulation results using both synthetic and real world networks to evaluate the performance of the proposed estimators. We simulate infection spreading under the SI, SIR, SIRI and SIS models in both homogeneous and heterogeneous networks, and for single and multiple infection sources.

A. Single Infection Source

When there is a single infection source, we use the following six common centrality measures and random guessing as benchmarks to compare with our estimator. The first four definitions are the same as those in [10], and are repeated here for the convenience of the reader.

(i) The betweenness center (BC) is defined as

$$\text{BC} \triangleq \arg\max_{v\in G} \sum_{i,j\in V_{\mathbf{i}}, i\neq j\neq v} \frac{\sigma_{ij}(v)}{\sigma_{ij}},$$

where σ_{ij} is the number of shortest paths between node i and node j, and $\sigma_{ij}(v)$ is the number of those shortest paths that contain v.

(ii) The closeness center (CC) is defined as

$$\text{CC} \triangleq \arg\max_{v\in G} \sum_{i\in V_{\mathbf{i}}, i\neq v} \frac{1}{d(v,i)}.$$

(iii) The distance center (DisC) is defined as

$$\text{DisC} \triangleq \arg\min_{v\in G} \sum_{i\in V_{\mathbf{i}}} d(v,i).$$

For trees, the DisC is the same as the rumor center defined in [7], and it is shown in [7] that the DisC is the ML estimator for regular trees under the SI model with a single source.

(iv) Let H denote the minimum connected subgraph of G that contains $V_{\mathbf{i}}$. The degree center (DegC) is defined as

$$\text{DegC} \triangleq \arg\max_{v\in H} |N_H(v)|,$$

where $N_H(v)$ is the set of neighbors of v in H, and $|N_H(v)|$ is defined to be the degree centrality of v.

(v) The eigenvector centrality function of H is a function $\text{EC}: H \mapsto \mathbb{R}$ such that for any node v in H, we have

$$\text{EC}(v) = \frac{1}{\lambda} \sum_{i\in N_H(v)} \text{EC}(i),$$

where $N_H(v)$ is the set of neighbors of v in H, and λ is a constant. Then the eigenvector center (EC) is the node in H with maximum eigenvector centrality.

(vi) The pagerank centrality of any node v in H is defined as

$$\text{PC}(v) = d \sum_{i\in N_H(v)} \frac{\text{PC}(i)}{|N_H(i)|} + \frac{1-d}{|H|},$$

where $N_H(v)$ is the set of neighbors of v in H, and d is a damping factor in $(0,1)$. Then the pagerank center (PC) is the node in H with maximum pagerank centrality.

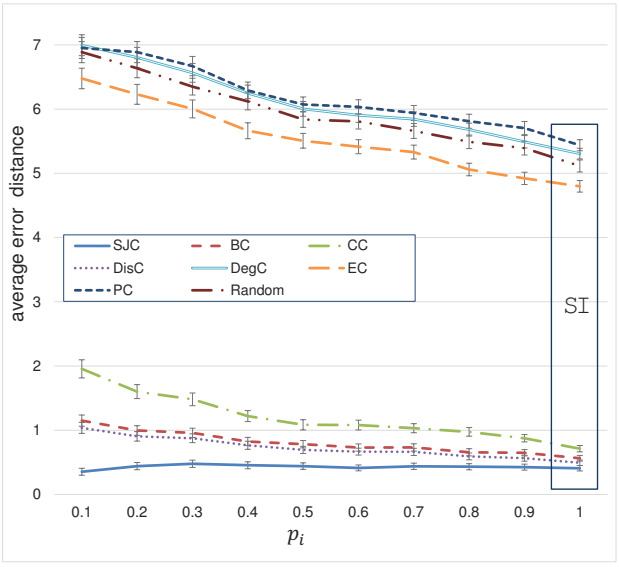
(vii) The random guess estimator randomly selects a node in H as the source estimator.

We evaluate the performance of our proposed estimator on three kinds of networks: random tree networks where the degree of every node is randomly chosen from $[3,5]$, a small part of the Facebook network with 4039 nodes [25] and the western states power grid network of the United States [26]. We consider both homogeneous and heterogeneous networks. In the homogeneous networks, we vary the recovery and relapse probabilities to demonstrate the impact of these spreading parameters on the performance of the proposed estimator. In the heterogeneous networks, we evaluate the robustness of the proposed estimator on a wide range of randomly generated spreading parameters. In the following, we describe the four different simulation experiments.

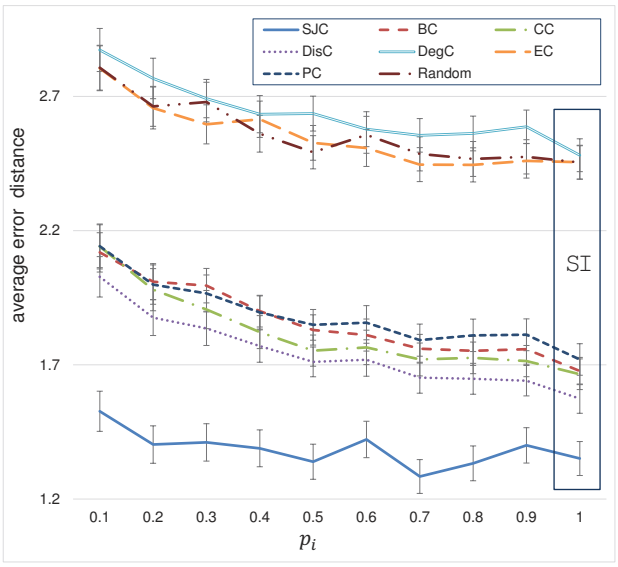
1) SI and SIRI models in homogeneous networks: For every $v\in V$, we let $p_{\mathbf{s}}(v)=p_{\mathbf{s}}$, $p_{\mathbf{i}}(v)=p_{\mathbf{i}}$ and $p_{\mathbf{r}}(v)=p_{\mathbf{r}}$, where the infection probabilities are set as follows: $p_{\mathbf{s}}$ is randomly chosen from $[0,1]$, $p_{\mathbf{i}}$ is set to be $0.1, 0.2, \cdots, 1$, respectively, and $p_{\mathbf{r}}$ is randomly chosen from $[0, \min\{1, \frac{p_{\mathbf{i}}}{1-p_{\mathbf{i}}}\}]$. For each kind of network and each value of $p_{\mathbf{i}}$, we perform 1000 simulation runs. In each simulation run, we randomly pick a node as the infection source and simulate the infection using the above parameters. The spreading terminates when the number of infected nodes is greater than 100. We then run SJC on the observed infected nodes to estimate the infection source and compare the result with the benchmarks.

The error distance is the number of hops between the estimated and the actual infection source, and is shown in Fig. 5. We see that the proposed estimator performs consistently better than the benchmarks for all considered networks. The random guess estimator actually performs better or comparable with some estimators like DegC and EC as these estimators only capture the local "connectivity" of a node, instead of the topological information inherent in the network.

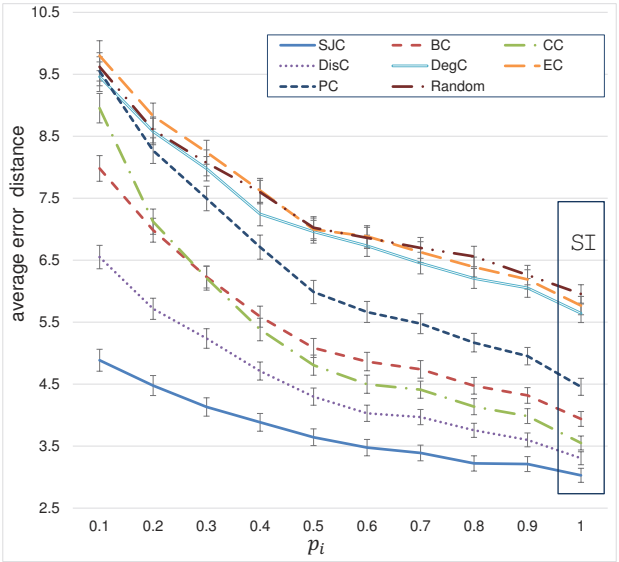
2) SIR and SIRI models in homogeneous networks: The infection probabilities are set as follows: $p_{\mathbf{s}}$ is randomly chosen from $[0,1]$, $p_{\mathbf{i}}$ is randomly chosen from $[0.5,1]$, and $p_{\mathbf{r}}$ is set to be $0, 0.1, \cdots, 1$. We compare the performances



(a) Random trees.



(b) Facebook network.



(c) Power grid network.

Fig. 5. Average error distances with error bars of 95% confidence interval for various networks and different values of $p_{\mathbf{i}}$ in homogeneous networks. The underlying infection follows the SIRI model for all values of $p_{\mathbf{i}}$ and the infection follows the SI model when $p_{\mathbf{i}} = 1$.

in Fig. 6. We see that our proposed estimator again performs consistently better than the benchmarks.

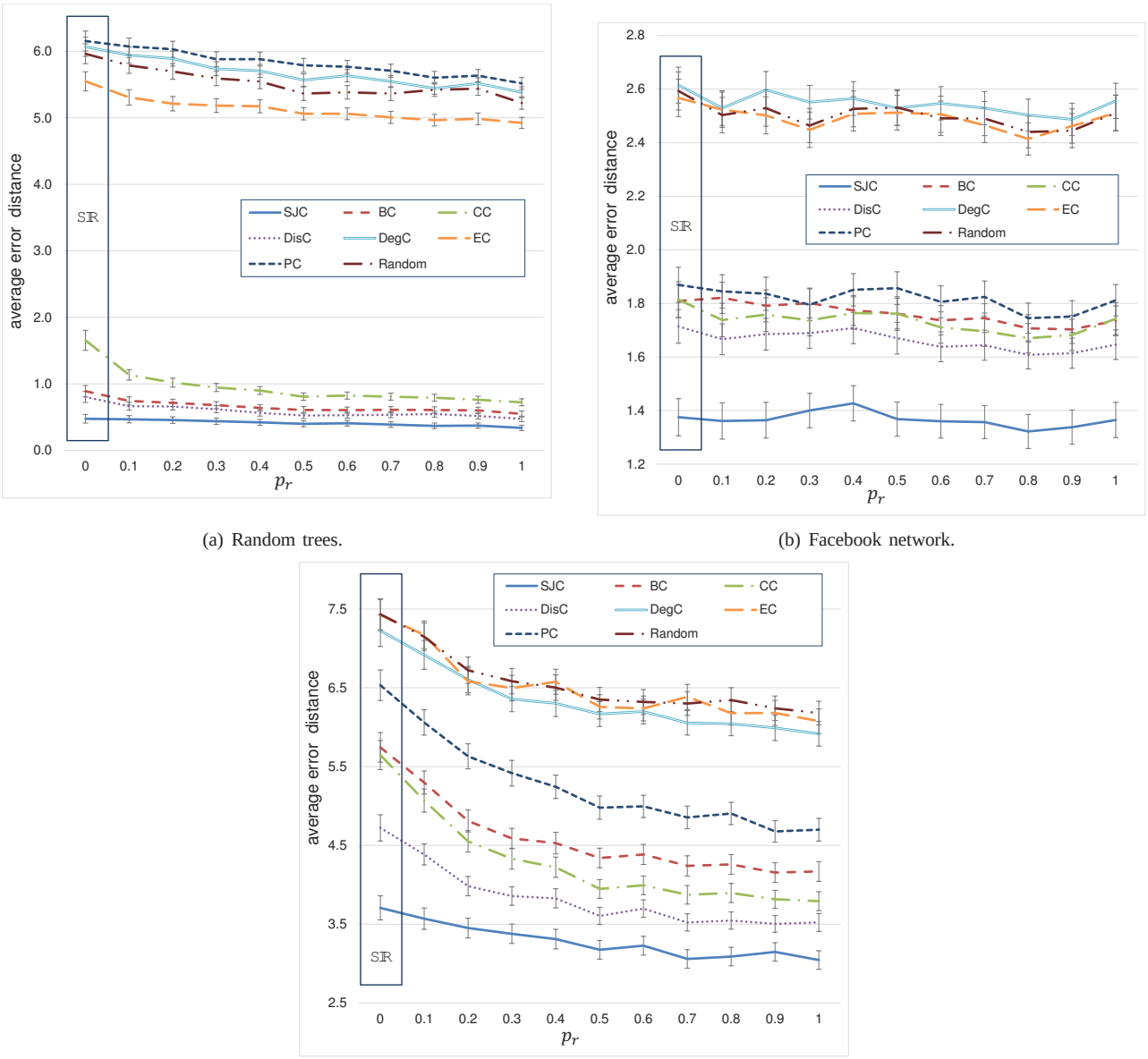
3) SIS model in homogeneous networks: We consider the SIS model where $p_{\mathbf{i}}$ is set to be $0.5, 0.6, \cdots, 1$, respectively, and $p_{\mathbf{s}}$ is randomly chosen from $[0, p_{\mathbf{i}}]$. In Fig. 7, we observe that our proposed estimator always results in smaller average error distances than the benchmarks for all considered networks.

4) SI, SIR, SIRI, and SIS models in heterogeneous networks: In this experiment, we drop the Assumptions 1-4 and randomly choose the infection probabilities $p_{\mathbf{s}}(v)$, $p_{\mathbf{i}}(v)$, $p_{\mathbf{r}}(v)$ from $[0, 1]$ for any node v. We then run simulations under the SI, SIR, SIRI, and SIS models, and compare the performances in Fig. 8. We see that SJC outperforms all the benchmarks.

B. Multiple Infection Sources

In this subsection, we consider the cases where $k = 2$ or $k = 3$ infection sources exist, respectively. By finding betweenness center, closeness center or distance center of each Voronoi set in the re-optimization step of MJC, we heuristically find multiple betweenness center set (MBC), multiple closeness center set (MCC) or multiple distance center set (MDisC) estimators, respectively. We also extend DegC, EC and PC by finding the k nodes in H with largest degree centralities, eigenvector centralities, pagerank centralities, respectively. Finally, for random guessing, we randomly pick K nodes in H as the estimator. We use MBC, MCC, MDisC, DegC, EC, PC and random guessing as comparison benchmarks.

For the SI, SIR, SIRI and SIS models, we randomly choose



(a) Random trees.

(b) Facebook network.

(c) Power grid network.

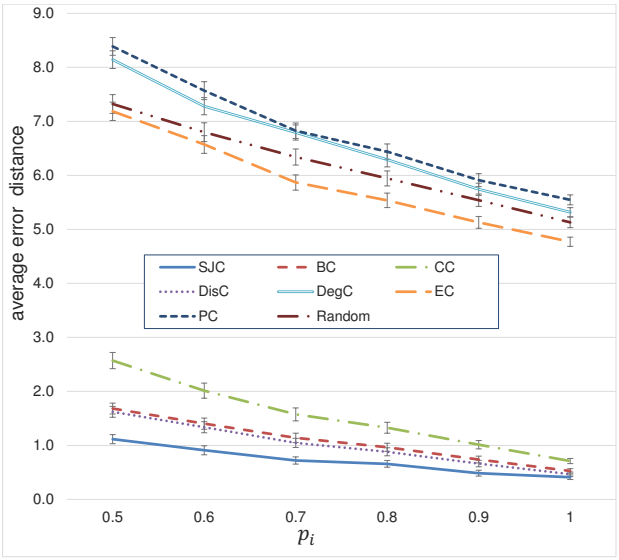
Fig. 6. Average error distances with error bars of 95% confidence interval for various networks and different values of $p_{\mathbf{r}}$ in homogeneous networks. The underlying infection follows the SIRI model for all values of $p_{\mathbf{r}}$ and the infection follows the SIR model when $p_{\mathbf{r}} = 0$.

the corresponding infection probabilities $p_{\mathbf{s}}(v)$, $p_{\mathbf{i}}(v)$, $p_{\mathbf{r}}(v)$ from $[0, 1]$ for any node v. For each value of k, each kind of network and each infection spreading model, we perform 1000 simulation runs. In each simulation run, we randomly pick k nodes as the infection sources and simulate the infection using the above mentioned spreading model. The spreading terminates when the number of infected nodes is greater than 100. We then run MJC on the observed infected nodes to estimate the infection sources and compare the result with the benchmarks.

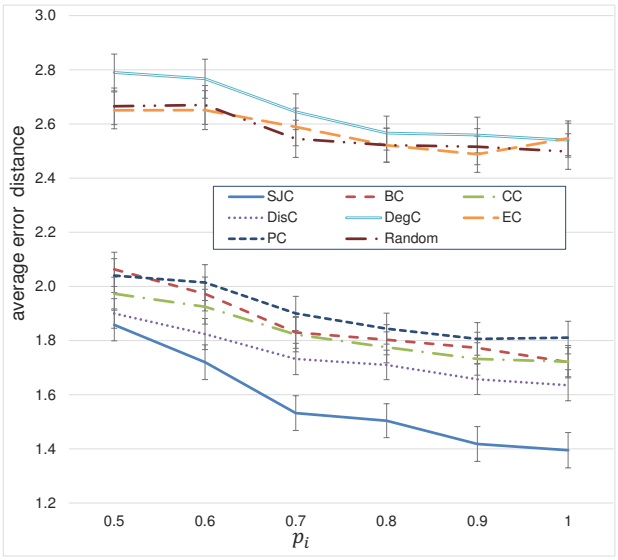
To quantify the performance of each algorithm, we first match the estimated with the actual sources so that the sum of the error distances between each estimated source and its match is minimized. Then the mean error distance is the average of the error distances for all matched pairs, and is shown in Fig. 9 and Fig. 10 for $k = 2$ and $k = 3$, respectively. We see that the proposed estimator performs better than the benchmarks for all considered networks under all considered infection spreading models.

VII. Conclusion

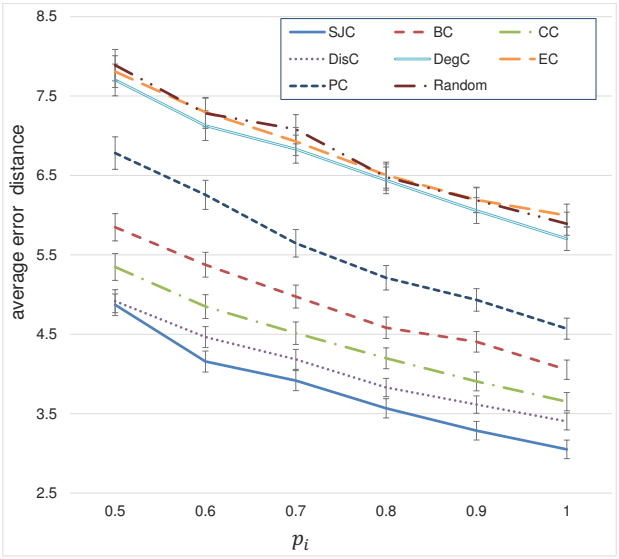
We have investigated the problem of estimating infection sources in a network for the SI, SIR, SIRI and SIS infection spreading models. For the case where a single infection source exists in an infinite tree network and under some technical assumptions, we have shown that the Jordan center of the infected node set is a universal infection source estimator for the SI, SIR, SIRI or SIS model. When there exists more than one infection sources in a tree network, we have shown that the k-Jordan center set is an optimal infection source set



(a) Random trees.



(b) Facebook network.



(c) Power grid network.

Fig. 7. Average error distances with error bars of 95% confidence interval for various networks and different values of $p_{\mathbf{i}}$ in homogeneous networks under the SIS model.

estimator for the SI model. Simulations have been conducted on random trees, part of the Facebook network and the western states power grid network of the United States. The results suggest that our estimators perform consistently better than the betweenness, closeness, distance, degree, eigenvector, and pagerank centrality based heuristics.

Appendix A
Proof of Proposition 1

For any $t \in \mathcal{T}_v$, consider any most likely infection path Y^{t+1} for $(v, t+1)$. To show claim (a), it suffices to construct an infection path $\tilde{X}^t$ for (v, t) such that

$$P_v(Y^{t+1}) \leq \delta P_v(\tilde{X}^t), \tag{8}$$

since $P_v(\tilde{X}^t) \leq P_v(X^t)$.

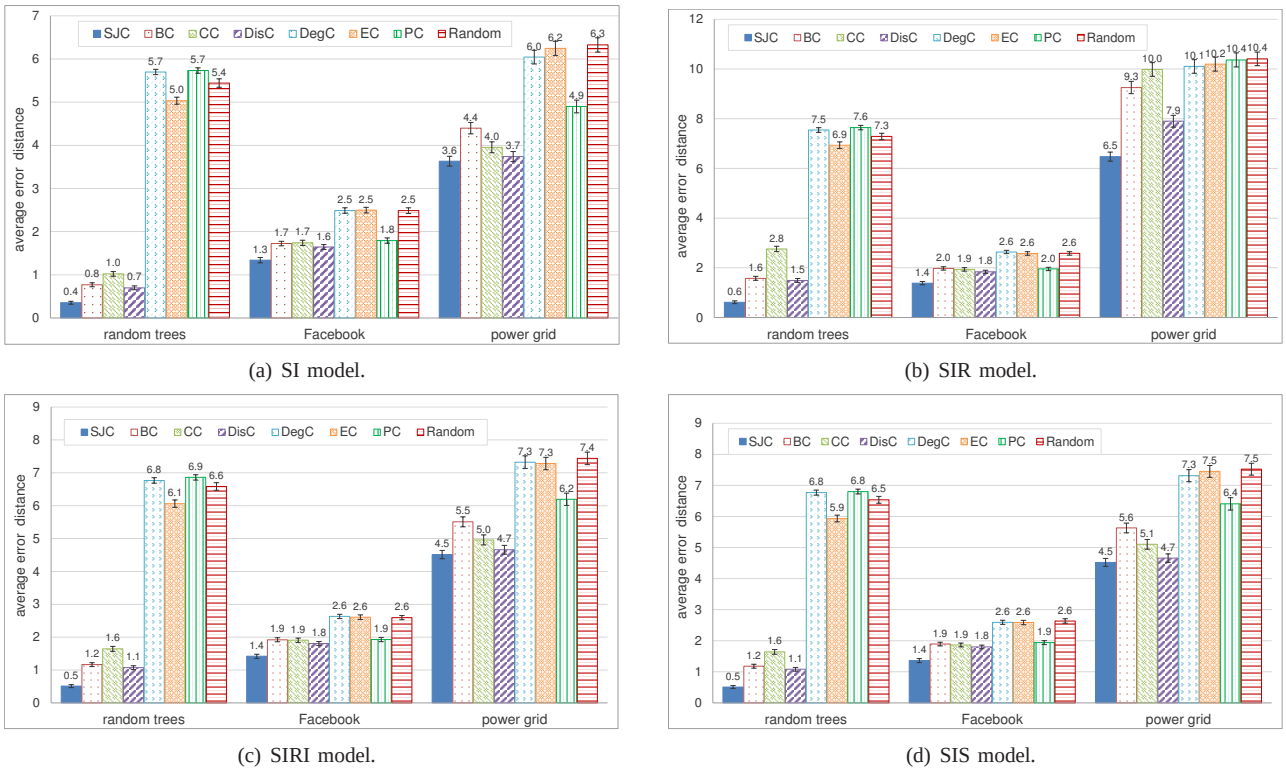
A. SI model

We first focus on any neighboring node u of v and consider $T_u(v; G)$. We claim that there exists an infection path $\tilde{X}^t$ such that

$$P_v(Y^{t+1}(T_u(v; G), [1, t+1])) \leq (1-\alpha) P_v(\tilde{X}^t(T_u(v; G), [1, t])). \tag{9}$$

We can see that $T_u(v; G)$ is either an uninfected subtree or infected subtree. In the following, we consider these two cases in order.

Suppose that $T_u(v; G)$ is an uninfected subtree. We have



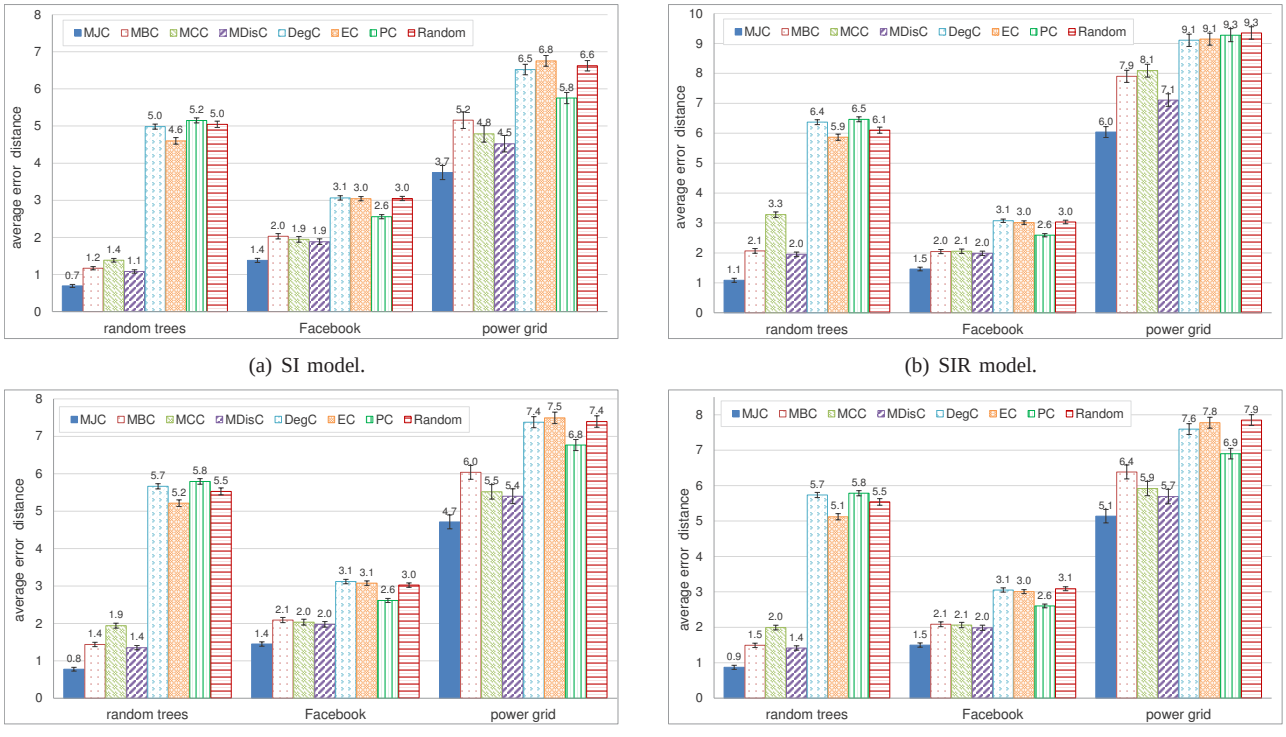
(a) SI model.

(b) SIR model.

(c) SIRI model.

(d) SIS model.

Fig. 8. Average error distances with error bars of 95% confidence interval for various networks under the SI, SIR, SIRI and SIS models in heterogeneous networks.



(a) SI model.

(b) SIR model.

(c) SIRI model.

(d) SIS model.

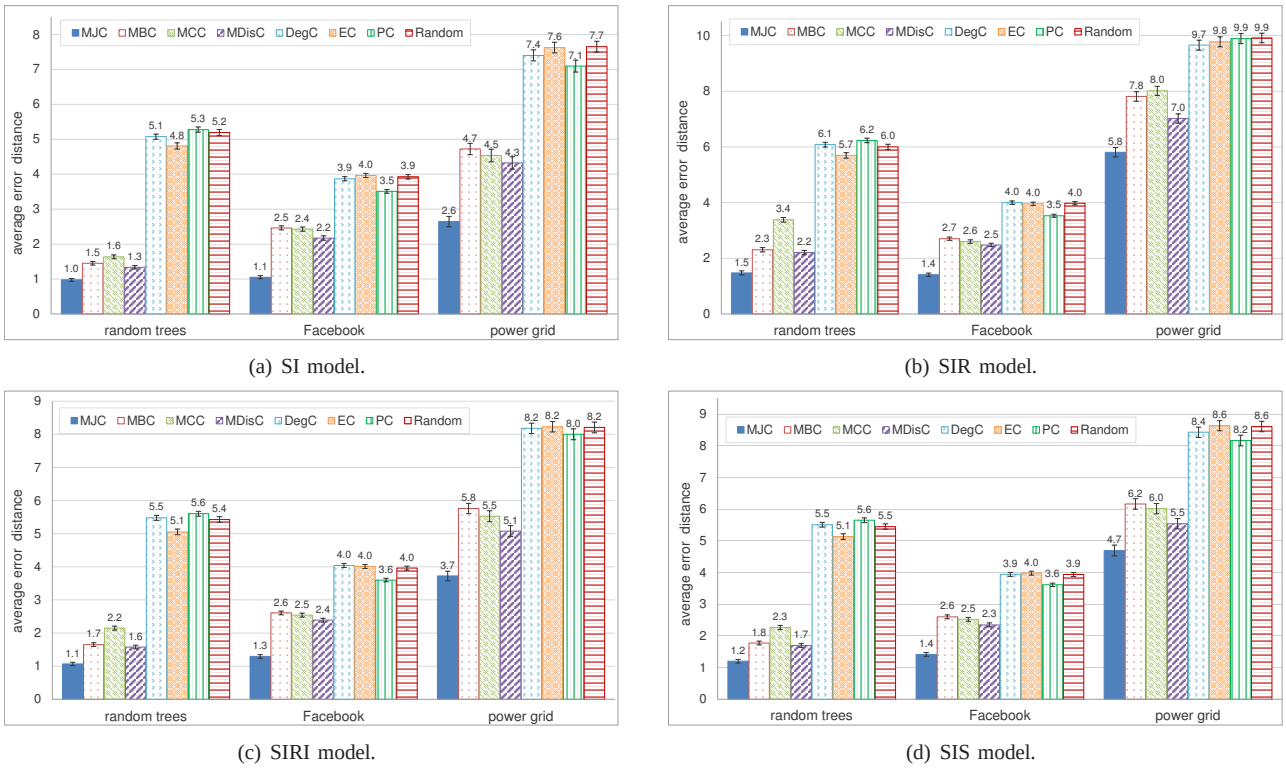
Fig. 9. Average mean error distances with error bars of 95% confidence interval for various networks under different infection spreading models when there are two infection sources.

for any $\tilde{X}^t$

$$\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))} = \frac{(1-p_{\mathbf{s}}(u))^{t+1}}{(1-p_{\mathbf{s}}(u))^t} \leq 1-\alpha. \quad (10)$$

Suppose that $T_u(v;G)$ is an infected subtree.

If $Y^{t+1}(u,1) = \mathbf{s}$, we let $\tilde{X}^t(T_u(v;G),[1,t]) =$



(a) SI model.

(b) SIR model.

(c) SIRI model.

(d) SIS model.

Fig. 10. Average mean error distances with error bars of 95% confidence interval for various networks under different infection spreading models when there are three infection sources.

$Y^{t+1}(T_u(v;G),[2,t+1])$, yielding

$$\begin{aligned}&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\&=\frac{P_v(Y^{t+1}(u,1)=\mathbf{s})P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\&=1-p_{\mathbf{s}}(u)\\&\leq 1-\alpha.\end{aligned}$$

If $Y^{t+1}(u,1)=\mathbf{i}$, we show (9) by mathematical induction on $\bar{d}(v,V_{\mathbf{i}})$.

Basis step: Suppose that $\bar{d}(v,V_{\mathbf{i}})=1$**.**

We let $\tilde{X}^t(u,1)=\mathbf{i}$. After it gets infected at time slot 1, node u serves as the infection source of the subtree $T_u(v;G)$ with the infection starting at time 1. From the assumption $\bar{d}(v,V_{\mathbf{i}})=1$, it follows that $T_w(u;G)$ is an uninfected subtree for any $w\in V(u,1)\bigcap T_u(v;G)$. Then following (10), we have

$$\begin{aligned}&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\&=\frac{P_v(Y^{t+1}(u,1)=\mathbf{i})P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^{t+1}(u,1)=\mathbf{i})P_v(\tilde{X}^t(T_u(v;G),[2,t]))}\\&=\prod_{w\in V(u,1)\bigcap T_u(u;G)}\frac{P_v(Y^{t+1}(T_w(u;G),[2,t+1]))}{P_v(\tilde{X}^t(T_w(v;G),[2,t]))}\\&\leq(1-\alpha)^{|V(u,1)\bigcap T_u(v;G)|}\\&\leq 1-\alpha,\end{aligned}$$

where the last inequality follows from Assumption 5. This completes the proof for the basis step.

Inductive step: Assume (9) holds for $\bar{d}(v,V_{\mathbf{i}})\leq n-1$, where $n\geq 2$. We want to show that (9) also holds for $\bar{d}(v,V_{\mathbf{i}})=n$.

Assume $\bar{d}(v,V_{\mathbf{i}})=n$ and let $\tilde{X}^t(u,1)=\mathbf{i}$. After it becomes infected at time slot 1, node u serves as the infection source of the subtree $T_u(v;G)$ with the infection starting at time 1. Since $\bar{d}(u,V_{\mathbf{i}}\bigcap T_u(v;G))\leq n-1$, from the induction assumption and for any $w\in V(u,1)\bigcap T_u(v;G)$, we can find a $\tilde{X}^t$ such that

$$\begin{aligned}&P_v(Y^{t+1}(T_w(u;G),[2,t+1]))\\&\leq(1-\alpha)P_v(\tilde{X}^t(T_w(u;G),[2,t])).\end{aligned}$$

We then have,

$$\begin{aligned}&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\&=\frac{P_v(Y^{t+1}(u,1)=\mathbf{i})P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^{t+1}(u,1)=\mathbf{i})P_v(\tilde{X}^t(T_u(v;G),[2,t]))}\\&=\prod_{w\in V(u,1)\bigcap T_u(u;G)}\frac{P_v(Y^{t+1}(T_w(u;G),[2,t+1]))}{P_v(\tilde{X}^t(T_w(v;G),[2,t]))}\\&\leq(1-\alpha)^{|V(u,1)\bigcap T_u(v;G)|}\\&\leq 1-\alpha,\end{aligned}$$

where the last inequality follows from Assumption 5. This completes the proof for the inductive step, and the claim is now proved.

By constructing $\tilde{X}^t$ to satisfy (9) for all $u \in V(v,1)$, we have

$$\begin{aligned}\frac{P_v(Y^{t+1})}{P_v(\tilde{X}^t)} &= \prod_{u\in V(v,1)} \frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\ &\le (1-\alpha)^{|V(v,1)|}\\ &\le (1-\alpha)^2,\end{aligned}$$

where the last inequality follows from Assumption 5. This completes the proof of claim (a) for the SI model.

B. SIR and SIRI model

We first present a property of the SIRI model in Lemma A.1.

Lemma A.1. *Suppose that $v \in V$ is the infection source and v has only one neighboring node u. Suppose that the set of observed infected nodes $V_{\mathbf{i}}$ is non-empty. Consider an infection under the SIRI model and suppose Assumptions 3 and 5 hold. For any $t \in \mathcal{T}_v$ and any most likely infection path Y^{t+1} for $(v, t+1)$, there exists an infection path $\tilde{X}^t$, such that*

(a) $P_v(Y^{t+1}(v,[1,t+1])) \le \sqrt{\frac{\alpha}{\beta}} P_v(\tilde{X}^t(v,[1,t]))$;

(b) $P_v(Y^{t+1}(T_u(v;G),[1,t+1])) \le P_v(\tilde{X}(T_u(v;G),[1,t]))$; *and*

(c) $P_v(Y^{t+1}) \le \sqrt{\frac{\alpha}{\beta}} P_v(\tilde{X}^t)$.

The proof of Lemma A.1 is provided in Appendix F. Lemma A.1 shows that, in the SIRI model, a most likely elapsed time t_v should be as small as possible when the source has only one neighboring node. We now extend this result to prove Proposition 1(a) for the SIRI model where v has more than one neighboring node.

In the SIRI model, since v is the source node, $\tilde{X}^t(v,[1,t])$ is independent of the states of other nodes. Furthermore, for any pair of neighboring nodes u and u' of v, the states of $T_u(v;G)$ and $T_{u'}(v;G)$ are independent conditioned on the states of node v. Therefore, by applying Lemma A.1 to v and each of its neighboring nodes, we have an infection path $\tilde{X}^t$ such that

$$\begin{aligned}&\frac{P_v(Y^{t+1})}{P_v(\tilde{X}^t)}\\ &= \frac{P_v(Y^{t+1}(v,[1,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))}\\ &\quad\cdot \frac{\prod_{u\in N_v(1)} P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{\prod_{u\in N_v(1)} P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\ &\le \sqrt{\frac{\alpha}{\beta}}.\end{aligned}$$

This completes the proof of claim (a) for the SIRI model. The proof of claim (a) for the SIR model is similar to that of the SIRI model, and we omit it here to avoid repetition.

C. SIS model

For the SIS model, a node can become infected, recover, and then be reinfected again for multiple times by the observation time. We characterize the time when a node is first infected (first infection time) in the following lemma, whose proof is provided in Appendix G. Recall that H_v is the minimum connected subgraph of G that contains $V_{\mathbf{i}}$ and v.

Lemma A.2. *Suppose that $v \in V$ is the infection source and a non-empty set of infected nodes $V_{\mathbf{i}}$ is observed. Suppose the infection follows the SIS model and Assumption 4 and 5 hold. Then, for any $t \in \mathcal{T}_v$, there exists a most likely infection path X^t for (v,t), such that, for any $u \in H_v \backslash \{v\}$, the first infection time $t_{int}(u)$ of u in X^t is given by*

$$t_{int}(u) = t - \bar{d}(u, T_u(v;H_v)). \tag{11}$$

Lemma A.2 enables us to calculate the first infection time of each node in H_v in a most likely infection path under the SIS model. Moreover, it shows that given the elapsed time, a most likely infection path for a node v is given by a path whose nodes "resist" the infection, and each node becomes infected only at the latest possible time. Therefore, intuitively the most likely elapsed time t_v should be as small as possible to minimize the time that nodes "resist" the infection spreading, so as to maximize the probability of the infection path.

Since $t \in \mathcal{T}_v$, we have $t \ge \bar{d}(v, V_{\mathbf{i}})$. For any $u \in V(v,1)$, from Lemma A.2, we have that the first infection time $t_{int}(u)$ of u in Y^{t+1} is given by

$$\begin{aligned}t_{int}(u) &= t+1-\bar{d}(u,T_u(v;H_v))\\ &\ge \bar{d}(v,V_{\mathbf{i}})+1-\bar{d}(u,T_u(v;H_v))\\ &\ge 2.\end{aligned} \tag{12}$$

We claim that $Y^{t+1}(v,1) = \mathbf{i}$. Otherwise, v and all its neighboring nodes are not infected at time 1 because of (12). Because the infection can propagate at most 1 hop away from v at time 1, all nodes are uninfected at time 1, and the infection propagation process stops. This contradicts the assumption that the set of observed infected nodes $V_{\mathbf{i}}$ is non-empty. Then, following Lemma A.2, we can let $\tilde{X}^t(V,[1,t]) = Y^{t+1}(V,[2,t+1])$, yielding

$$\begin{aligned}&\frac{P_v(Y^{t+1})}{P_v(\tilde{X}^t)}\\ &= \frac{P_v(Y^{t+1}(v,1))P_v(Y^{t+1}(V(v,1),1))P_v(Y^{t+1}(V,[2,t+1]))}{P_v(\tilde{X}^t(V,[1,t]))}\\ &= p_{\mathbf{i}}(1-p_{\mathbf{s}})^{|V(v,1)|}\\ &\le 1.\end{aligned}$$

This completes the proof of claim (a) for the SIS model.

It is easy to see that $\delta \le 1$ for all considered infection spreading models and claim (b) now follows from claim (a), and the proof of Proposition 1 is complete.

APPENDIX B
PROOF OF PROPOSITION 2

We first review the following topological property shown in [10].

Lemma B.1. *Suppose a non-empty set of infected nodes $V_{\mathbf{i}}$ is observed over G. For a pair of neighboring nodes u and v, if $\bar{d}(v,V_{\mathbf{i}}) < \bar{d}(u,V_{\mathbf{i}})$, we have*

(a) $l \in T_v(u; H_v \bigcup H_u)$*, for all* $l \in \arg\max_{x \in V_{\mathbf{i}}} d(u,x)$*; and (b)* $\bar{d}(v, V_{\mathbf{i}}) = \bar{d}(u, V_{\mathbf{i}}) - 1$*, and there exists* $l \in T_v(u; H_v \bigcup H_u)$ *such that* $d(v,l) = \bar{d}(v, V_{\mathbf{i}})$.

To prove Proposition 2, it suffices to construct an infection path $\tilde{X}^{t_v}$ with source node v, and show that $P_v(\tilde{X}^{t_v}) \geq P_u(Y^{t_u})$. Let $t_{int}(v)$ be the first infection time of node v in the infection path Y^{t_u} with source node u. We first show that $t_{int}(v) = 1$. Since u is the infection source, the infection can propagate at most $t_u - t_{int}(v)$ hops away from node v within the subtree $T_v(u; H_v \bigcup H_u)$. From Lemma B.1(b), if $t_{int}(v) > 1$, we have $t_v = t_u - 1 > t_u - t_{int}(v)$, a contradiction. Therefore, we must have $t_{int}(v) = 1$ in the infection path Y^{t_u}.

For the SI, SIR and SIRI models, we let $\tilde{X}^{t_v}(T_v(u;G),[1,t_v]) = Y^{t_u}(T_v(u;G),[2,t_u])$, yielding

$$\frac{P_u(Y^{t_u}(T_v(u;G),[1,t_u]))}{P_v(\tilde{X}^{t_v}(T_v(u;G),[1,t_v]))} \tag{13}$$

$$= \frac{P_u(Y^{t_u}(v,1) = \mathbf{i}) P_u(Y^{t_u}(T_v(u;G),[2,t_u]))}{P_v(\tilde{X}^{t_v}(T_v(u;G),[1,t_v]))}$$

$$= p_{\mathbf{s}}(v). \tag{14}$$

Let $\tilde{X}^{t_v}(u,1) = \mathbf{i}$ and u can be seen as the infection source of the subtree $T_u(v;G)$ with the infection starting at time 1. For the SI model, applying (9) twice, we have

$$\frac{P_u(Y^{t_u}(T_u(v;G),[1,t_u]))}{P_v(\tilde{X}^{t_v}(T_u(v;G),[1,t_v]))} \tag{15}$$

$$= \frac{P_u(Y^{t_u}(T_u(v;G),[1,t_u]))}{P_v(\tilde{X}^{t_v}(u,1) = \mathbf{i}) P_v(\tilde{X}^{t_v}(T_u(v;G),[2,t_v]))}$$

$$\leq \frac{(1-\alpha)^2}{p_{\mathbf{s}}(u)}. \tag{16}$$

Multiplying (14) by (16), we obtain

$$\begin{aligned}\frac{P_u(Y^{t_u})}{P_v(\tilde{X}^{t_v})} &\leq \frac{p_{\mathbf{s}}(v)\cdot(1-\alpha)^2}{p_{\mathbf{s}}(u)} \\ &\leq \frac{\beta(1-\alpha)^2}{\alpha} \\ &\leq 1,\end{aligned}$$

where the last inequality follows from (1).

For the SIR and SIRI models, applying Lemma A.1 twice to u and each of its neighboring nodes in $T_u(v;G)$, we have

$$\frac{P_u(Y^{t_u}(T_u(v;G),[1,t_u]))}{P_v(\tilde{X}^{t_v}(T_u(v;G),[1,t_v]))} \tag{17}$$

$$= \frac{P_u(Y^{t_u}(T_u(v;G),[1,t_u]))}{P_v(\tilde{X}^{t_v}(u,1) = \mathbf{i}) P_v(\tilde{X}^{t_v}(T_u(v;G),[2,t_v]))}$$

$$\leq \frac{\alpha}{\beta \cdot p_{\mathbf{s}}(u)}. \tag{18}$$

Multiplying (14) by (18), we have

$$\frac{P_u(Y^{t_u})}{P_v(\tilde{X}^{t_v})} \leq \frac{p_{\mathbf{s}}(v)}{\beta} \cdot \frac{\alpha}{p_{\mathbf{s}}(u)} \leq 1.$$

This completes the proof of Proposition 2 in the SI, SIR and SIRI models.

We next consider the SIS model. Following Lemma A.2, we have $Y^{t_u}(V(u,1)\backslash\{v\},1) = \mathbf{s}$ and we can let $\tilde{X}^{t_v}(V,[1,t_v]) = Y^{t_u}(V,[2,t_u])$. Moreover, following similar arguments as the worst case in (27), we have that $Y^{t_u}(u,1) \neq \mathbf{i}$, yielding

$$\begin{aligned}&\frac{P_u(Y^{t_u})}{P_v(\tilde{X}^{t_v})} \\ &= P_u(Y^{t_u}(v,1)) P_u(Y^{t_u}(u,1)) P_u(Y^{t_u}(V(u,1)\backslash\{v\},1)) \\ &\quad \cdot \frac{P_u(Y^{t_u}(V,[2,t_u]))}{P_v(\tilde{X}^{t_v}(V,[1,t_v]))} \\ &= p_{\mathbf{s}}(1-p_{\mathbf{i}})(1-p_{\mathbf{s}})^{|V(u,1)\backslash\{v\}|} \\ &\leq 1.\end{aligned}$$

This completes the proof of Proposition 2 in the SIS model. The proof of Proposition 2 is now complete.

Appendix C
Proof of Proposition 3

We extend the notation of subtree as follows. For any graph A, any node $v \in A$ and a set of nodes $S \subset A$, let $T_v(S;A)$ be the subtree of A rooted at node v with the first link in the path from v to each element in S removed. Moreover, for any set of nodes $M \subset A$, let $T_M(S;A) = \bigcup_{v \in M} T_v(S;A)$.

For any $t \in \mathcal{T}_S$, consider any most likely infection path Y^{t+1} for $(S, t+1)$. To show claim (a), it suffices to construct an infection path $\tilde{X}^t$ for (S,t) such that

$$P_S(Y^{t+1}) \leq P_S(\tilde{X}^t). \tag{19}$$

We start with the case where $k = 2$ and show (19) by mathematical induction on $d(s_1, s_2)$.

Basis step (i): The inequality (19) holds for $d(s_1,s_2) = 1$.

The states of $T_{s_1}(S;G)$ and $T_{s_2}(S;G)$ are independent. We can treat s_1 and s_2 as the infection source of $T_{s_1}(S;G)$ and $T_{s_2}(S;G)$, respectively. Then following Proposition 1, we can find a $\tilde{X}^t$ such that

$$\begin{aligned}\frac{P_S(Y^{t+1})}{P_S(\tilde{X}^t)} =& \frac{P_S(Y^{t+1}(T_{s_1}(S;G),[1,t+1]))}{P_S(\tilde{X}^t(T_{s_1}(S;G),[1,t]))} \\ &\cdot \frac{P_S(Y^{t+1}(T_{s_2}(S;G),[1,t+1]))}{P_S(\tilde{X}^t(T_{s_2}(S;G),[1,t]))} \\ \leq& 1.\end{aligned}$$

Basis step (ii): The inequality (19) holds for $d(s_1,s_2) = 2$.

Denote the common neighboring node of s_1 and s_2 to be u. Consider the following two possible cases of Y^{t+1}.

Case 1: $Y^{t+1}(u,1) = \mathbf{i}$.

We let $\tilde{X}^t(u,1) = \mathbf{i}$, conditioning on which the states of $T_{s_1}(S;G)$, $T_{s_2}(S;G)$ and $T_u(S;G)$ are independent. Moreover, u can be seen as the infection source of $T_u(S;G)$ with

the infection starting at time 1. Then following Proposition 1, we can find a $\tilde{X}^t$ such that

$$\begin{aligned}\frac{P_S(Y^{t+1})}{P_S(\tilde{X}^t)} =& \frac{P_S(Y^{t+1}(T_{s_1}(S;G),[1,t+1]))}{P_S(\tilde{X}^t(T_{s_1}(S;G),[1,t]))} \\ & \cdot \frac{P_S(Y^{t+1}(T_{s_2}(S;G),[1,t+1]))}{P_S(\tilde{X}^t(T_{s_2}(S;G),[1,t]))} \\ & \cdot \frac{P_S(Y^{t+1}(u,1))}{P_S(\tilde{X}^t(u,1))} \\ & \cdot \frac{P_S(Y^{t+1}(T_u(S;G),[2,t+1]))}{P_S(\tilde{X}^t(T_u(S;G),[2,t]))} \\ \leq & 1.\end{aligned}$$

Case 2: $Y^{t+1}(u,1) = \mathbf{s}$.

We let $\tilde{X}^t(T_u(S;G),[1,t]) = Y^{t+1}(T_u(S;G),[2,t+1])$. Then following Proposition 1, we can find a $\tilde{X}^t$ such that

$$\begin{aligned}&\frac{P_S(Y^{t+1})}{P_S(\tilde{X}^t)} \\ &= \frac{P_S(Y^{t+1}(T_{s_1}(S;G),[1,t+1]))}{P_S(\tilde{X}^t(T_{s_1}(S;G),[1,t]))} \\ &\quad \cdot \frac{P_S(Y^{t+1}(T_{s_2}(S;G),[1,t+1]))}{P_S(\tilde{X}^t(T_{s_2}(S;G),[1,t]))} \\ &\quad \cdot \frac{P_S(Y^{t+1}(u,1))P_S(Y^{t+1}(T_u(S;G),[2,t+1]))}{P_S(\tilde{X}^t(T_u(S;G),[1,t]))} \\ &\leq 1 - p_{\mathbf{s}}(u) \\ &\leq 1.\end{aligned}$$

Inductive step: If (19) holds for $d(s_1,s_2) \leq n$, then (19) also holds for $d(s_1,s_2) = n+1$, where $n \geq 2$.

Let $\rho(v,u)$ be the path between two nodes v and u. Denote the neighboring node of s_1 and s_2 in $\rho(s_1,s_2)$ to be u_1 and u_2, respectively. Consider the following four possible cases of Y^{t+1}.

Case 1: $Y^{t+1}(u_1,1) = \mathbf{i}$ and $Y^{t+1}(u_2,1) = \mathbf{i}$.

We let $\tilde{X}^t(u_1,1) = \mathbf{i}$ and $\tilde{X}^t(u_2,1) = \mathbf{i}$. Then u_1 and u_2 can be seen as the pair of infection sources of $T_{\rho(u_1,u_2)}(S;G)$ with the infection starting at time 1. Moreover, we have $d(u_1,u_2) = d(s_1,s_2) - 2 = n-1$. Then by induction assumption, we can find a $\tilde{X}^t$ such that

$$\begin{aligned}&P_S(Y^{t+1}(T_{\rho(u_1,u_2)}(S;G),[2,t+1])) \\ &\leq P_S(\tilde{X}^t(T_{\rho(u_1,u_2)}(S;G),[2,t])).\end{aligned}$$

Then by Proposition 1, we can find a $\tilde{X}^t$ such that (19) holds.

Case 2: $Y^{t+1}(u_1,1) = \mathbf{i}$ and $Y^{t+1}(u_2,1) = \mathbf{s}$.

We let $\tilde{X}^t(u_1,1) = \mathbf{i}$ and $\tilde{X}^t(u_2,1) = \mathbf{s}$. Then u_1 and s_2 can be seen as the pair of infection sources of $T_{\rho(u_1,u_2)} \bigcup \{s_2\}$ with the infection starting at time 1. Moreover, we have $d(u_1,u_2) = d(s_1,s_2) - 2 = n-1$. Then by induction assumption, we can find a $\tilde{X}^t$ such that

$$\begin{aligned}&P_S(Y^{t+1}(T_{\rho(u_1,u_2)}(S;G),[2,t+1])) \\ &\leq P_S(\tilde{X}^t(T_{\rho(u_1,u_2)}(S;G),[2,t])).\end{aligned}$$

Then by Proposition 1, we can find a $\tilde{X}^t$ such that (19) holds.

Case 3: $Y^{t+1}(u_1,1) = \mathbf{s}$ and $Y^{t+1}(u_2,1) = \mathbf{i}$.

Following similar arguments as that in Case 2, we can find a $\tilde{X}^t$ such that (19) holds.

Case 4: $Y^{t+1}(u_1,1) = \mathbf{s}$ and $Y^{t+1}(u_2,1) = \mathbf{s}$.

We let $\tilde{X}^t(T_{\rho(u_1,u_2)}(S;G),[1,t]) = Y^{t+1}(T_{\rho(u_1,u_2)}(S;G),[2,t+1])$. Then following Proposition 1, we can find a $\tilde{X}^t$ such that

$$\begin{aligned}&\frac{P_S(Y^{t+1})}{P_S(\tilde{X}^t)} \\ &= \frac{P_S(Y^{t+1}(T_{s_1}(S;G),[1,t+1]))}{P_S(\tilde{X}^t(T_{s_1}(S;G),[1,t]))} \\ &\quad \cdot \frac{P_S(Y^{t+1}(T_{s_2}(S;G),[1,t+1]))}{P_S(\tilde{X}^t(T_{s_2}(S;G),[1,t]))} \\ &\quad \cdot P_S(Y^{t+1}(u_1,1))P_S(Y^{t+1}(u_2,1)) \\ &\quad \cdot \frac{P_S(Y^{t+1}(T_{\rho(u_1,u_2)}(S;G),[2,t+1]))}{P_S(\tilde{X}^t(T_{\rho(u_1,u_2)}(S;G),[1,t]))} \\ &\leq (1-p_{\mathbf{s}}(u_1))(1-p_{\mathbf{s}}(u_2)) \\ &\leq 1.\end{aligned}$$

This completes the proof for the inductive step. By the spirit of mathematical induction, we have shown that (19) holds for $k=2$. When $k>2$, similar arguments can be applied to each pair of source nodes, and this completes the proof of claim (a).

We show that $\mathcal{T}_S = [\bar{d}(S,V_{\mathbf{i}}),+\infty)$. Consider any node $l \in V_{\mathbf{i}}$ such that $d(S,l) = \bar{d}(S,V_{\mathbf{i}})$. The infection can propagate at most one hop further from any source node in one time slot. If $t < \bar{d}(S,V_{\mathbf{i}})$, the infection can not reach node l. Claim (b) now follows from claim (a), and the proof of Proposition 3 is complete.

Appendix D
Proof of Lemma 2

We first show that the value of the minimum infection range in $\tilde{G}(S,X^t)$ can not be less than $\bar{d}(S,V_{\mathbf{i}})$. Assume there is a super node Supernode(S') in $\tilde{G}(S,X^t)$ that is associated with a set of k nodes $S' \subset V$ such that, $\bar{d}(\text{Supernode}(S'),V_{\mathbf{i}}) < \bar{d}(S,V_{\mathbf{i}})$. Then it is implied that $\bar{d}(S',V_{\mathbf{i}}) < \bar{d}(S,V_{\mathbf{i}})$, which contradicts with the assumption that S is a k-Jordan center set.

We then show that Supernode(S) is a Jordan center of $V_{\mathbf{i}}$ in the transformed super node graph $\tilde{G}(S,X^t)$, i.e., Supernode(S) has the minimum infection range in $\tilde{G}(S,X^t)$. In other words, we want to show that $d(\text{Supernode}(S),v) \leq \bar{d}(S,V_{\mathbf{i}})$ for any node $v \in V_{\mathbf{i}}$. From Definition 5, it suffices to show that $d(s_i,v) \leq \bar{d}(S,V_{\mathbf{i}})$ for any node $v \in A_i$, where $i \in \{1,2,\cdots,k\}$. Suppose that there exists a node $v \in A_i$ such that $d(s_i,v) \geq \bar{d}(S,V_{\mathbf{i}})+1$. Then the first infection time $t_{int}(v)$ of v in X^{t_S} is

$$\begin{aligned}t_{int}(v) &\geq d(s_i,v) \\ &\geq \bar{d}(S,V_{\mathbf{i}})+1,\end{aligned}$$

because the infection can spread at most one hop further from s_i in one time slot. Following Proposition 3(b), we have that $t_S = \bar{d}(S,V_{\mathbf{i}}) < t_{int}(v)$, a contradiction. Therefore we have

$d(s_i, v) \leq \bar{d}(S, V_{\mathbf{i}})$ for any $v \in A_i$, where $i \in \{1, 2, \cdots, k\}$. This competes the proof of Lemma 2.

APPENDIX E
PROOF OF PROPOSITION 4

For a node $u \in V_{\mathbf{i}}$, suppose that $\hat{s}_j^{l-1}$ is a nearest node in $\hat{S}^{l-1}$ to u, and M_j be the Voronoi set corresponding to $\hat{s}_j^{l-1}$. Following Definition 4, it suffices to show that

$$d(\hat{s}_j^l, u) \leq \bar{d}(\hat{S}^{l-1}, V_{\mathbf{i}}),$$

where $\hat{s}_j^l$ is the Jordan center of M_j. We have $d(\hat{s}_j^l, u) \leq \max_{z \in V_{\mathbf{i}} \cap M_j} d(\hat{s}_j^l, z) \leq \max_{z \in V_{\mathbf{i}} \cap M_j} d(\hat{s}_j^{l-1}, z) \leq \bar{d}(\hat{S}^{l-1}, V_{\mathbf{i}})$. The proof of Proposition 4 is now complete.

APPENDIX F
PROOF OF LEMMA A.1

We first show the following property of the SIRI model in a network with only one node.

Lemma F.1. *Suppose that G has only one node v. For any $t \in [1, +\infty)$, consider any two most likely infection paths X^t for (v, t) and Y^{t+1} for $(v, t+1)$ under the SIRI model. Assume Assumption 3 holds. We have*

$$\frac{P_v(Y^{t+1})}{P_v(X^t)} \leq \sqrt{\frac{\alpha}{\beta}}. \tag{20}$$

Proof: Given any most likely infection path Y^{t+1} for $(v, t+1)$ with $t \in [1, +\infty)$, it suffices to construct another infection path $\tilde{X}^t$ for (v, t) such that

$$\frac{P_v(Y^{t+1})}{P_v(\tilde{X}^t)} \leq \sqrt{\frac{\alpha}{\beta}}. \tag{21}$$

Let $\tilde{X}^t(v, [1, t]) = Y^{t+1}(v, [2, t+1])$, we have three cases for Y^{t+1} which are discussed in the following.

Case 1: If $Y^{t+1}(v, 1) = \mathbf{i}$, we have

$$\begin{aligned}\frac{P_v(Y^{t+1})}{P_v(\tilde{X}^t)} &= \frac{P_v(Y^{t+1}(v,1))P_v(Y^{t+1}(v,[2,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))} \\ &= p_{\mathbf{i}}(v) \\ &\leq \sqrt{\frac{\alpha}{\beta}},\end{aligned}$$

where the last inequality holds from (3).

Case 2: If $Y^{t+1}(v, 1) = \mathbf{r}$ and $Y^{t+1}(v, 2) = \mathbf{i}$, we have

$$\begin{aligned}\frac{P_v(Y^{t+1})}{P_v(\tilde{X}^t)} &= \frac{P_v(Y^{t+1}(v,[1,2]))P_v(Y^{t+1}(v,[3,t+1]))}{P_v(\tilde{X}^t(v,1))P_v(\tilde{X}^t(v,[2,t]))} \\ &= \frac{(1-p_{\mathbf{i}}(v))p_{\mathbf{r}}(v)}{p_{\mathbf{i}}(v)} \\ &\leq \sqrt{\frac{\alpha}{\beta}},\end{aligned}$$

where the last inequality holds from (4).

Case 3: If $Y^{t+1}(v, 1) = \mathbf{r}$ and $Y^{t+1}(v, 2) = \mathbf{r}$, we have

$$\begin{aligned}\frac{P_v(Y^{t+1})}{P_v(\tilde{X}^t)} &= \frac{P_v(Y^{t+1}(v,[1,2]))P_v(Y^{t+1}(v,[3,t+1]))}{P_v(\tilde{X}^t(v,1))P_v(\tilde{X}^t(v,[2,t]))} \\ &= \frac{(1-p_{\mathbf{i}}(v))(1-p_{\mathbf{r}}(v))}{1-p_{\mathbf{i}}(v)} \\ &\leq \sqrt{\frac{\alpha}{\beta}},\end{aligned}$$

where the last inequality holds from (4).

We see that (21) holds for all three possible cases. The proof for Lemma F.1 is now complete. ■

We note that $T_u(v; G)$ is either an uninfected subtree or infected subtree. In the following, we prove these two cases separately.

A. Proof of Lemma A.1 for Uninfected Subtree

If $T_u(v; G)$ is an uninfected subtree, we can easily see that $\mathcal{T}_v = [1, +\infty)$. It is clear that claim (c) follows from claim (a) and (b). In the following, we prove claim (a) and (b) by mathematical induction on the elapsed time t.

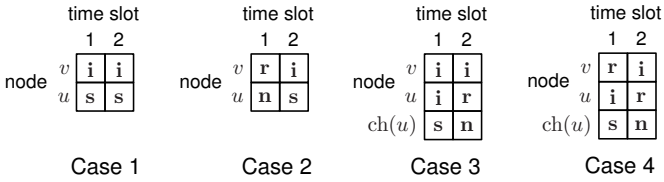


Fig. 11. Illustration of four possible cases for Y^2, where we omit the states for any node that only have non-susceptible state. We have $Y^2(v, [1, 2]) = p_{\mathbf{i}}(v)^2$, $(1 - p_{\mathbf{i}}(v))p_{\mathbf{r}}(v)$, $p_{\mathbf{i}}(v)^2$, or $(1 - p_{\mathbf{i}}(v))p_{\mathbf{r}}(v)$ for four cases respectively. Moreover, we have $X^2(T_u(v; G), [1, 2]) = (1 - p_{\mathbf{s}}(u))^2$, $1 - p_{\mathbf{s}}(u)$, $p_{\mathbf{s}}(u)(1 - p_{\mathbf{i}}(u)) \prod_{w \in \mathrm{ch}(u)}(1 - p_{\mathbf{s}}(w))$, or $p_{\mathbf{s}}(u)(1 - p_{\mathbf{i}}(u)) \prod_{w \in \mathrm{ch}(u)}(1 - p_{\mathbf{s}}(w))$ for four cases respectively.

Basis step: $t = 1$**.**

If $v \in V_{\mathbf{i}}$, we let $\tilde{X}^1(v, 1) = \mathbf{i}$ and $\tilde{X}^1(u, 1) = \mathbf{s}$, then $P_v(\tilde{X}^1(v, 1)) = p_{\mathbf{i}}(v)$ and $P_v(\tilde{X}^1(T_u(v; G), 1)) = P_v(\tilde{X}^1(u, 1)) = 1 - p_{\mathbf{s}}(u)$. As shown in Figure 11, there are four possible cases for Y^2. Following Assumption 5, we have

$$\begin{aligned}\frac{p_{\mathbf{s}}(u)(1-p_{\mathbf{i}}(u))\prod_{w\in\mathrm{ch}(u)}(1-p_{\mathbf{s}}(w))}{1-p_{\mathbf{s}}(u)} &\leq \frac{(1-p_{\mathbf{i}}(u))(1-\alpha)}{1-\beta} \\ &\leq 1,\end{aligned} \tag{22}$$

where the last inequity holds from (3). Then following (3), (4)

and (22), we have

$$\begin{aligned}
&\frac{P_v(Y^2(v,[1,2]))}{P_v(\tilde{X}^1(v,1))}\\
&=\frac{\max\left\{p_{\mathbf{i}}(v)^2,(1-p_{\mathbf{i}}(v))p_{\mathbf{r}}(v)\right\}}{p_{\mathbf{i}}(v)}\\
&\leq\sqrt{\frac{\alpha}{\beta}},\\
&\frac{P_v(Y^2(T_u(v;G),[1,2]))}{P_v(\tilde{X}^1(T_u(v;G),1))}\\
&=\max\{(1-p_{\mathbf{s}}(u))^2,1-p_{\mathbf{s}}(u)\\
&\qquad,p_{\mathbf{s}}(u)(1-p_{\mathbf{i}}(u))\prod_{w\in\mathrm{ch}(u)}(1-p_{\mathbf{s}}(w))\}/(1-p_{\mathbf{s}}(u))\\
&=1.
\end{aligned}$$

If $v\notin V_{\mathbf{i}}$, we have $\tilde{X}^1(v,1)=\mathbf{r}$ and $\tilde{X}^1(u,1)=\mathbf{n}$, then $P_v(\tilde{X}^1(v,1))=1-p_{\mathbf{i}}(v)$ and $P_v(\tilde{X}^1(T_u(v;G),1))=P_v(\tilde{X}^1(u,1))=1-p_{\mathbf{s}}(u)$. Change the states of node v at time slot 2 for all four cases in Figure 11 from infected to recovered. Then following (3), (4) and (22), we have

$$\begin{aligned}
&\frac{P_v(Y^2(v,[1,2]))}{P_v(\tilde{X}^1(v,1))}\\
&=\frac{\max\left\{p_{\mathbf{i}}(v)(1-p_{\mathbf{i}}(v)),(1-p_{\mathbf{i}}(v))(1-p_{\mathbf{r}}(v))\right\}}{1-p_{\mathbf{i}}(v)},\\
&\leq\sqrt{\frac{\alpha}{\beta}},\\
&\frac{P_v(Y^2(T_u(v;G),[1,2]))}{P_v(\tilde{X}^1(T_u(v;G),1))}\\
&=\frac{\max\left\{1-p_{\mathbf{s}}(u),p_{\mathbf{s}}(u)(1-p_{\mathbf{i}}(u))\prod_{w\in\mathrm{ch}(u)}(1-p_{\mathbf{s}}(w))\right\}}{1-p_{\mathbf{s}}(u)}\\
&=1.
\end{aligned}$$

This completes the proof for the basis step.

Inductive step: assume claim (a) and (b) hold for $t=\tau-1$, where $\tau\geq 2$. We want to show that claim (a) and (b) also hold for $t=\tau$.

Assume $t=\tau$ and consider the following six possible cases for Y^{t+1}.

Case 1: $Y^{t+1}(v,1)=\mathbf{i}$ and $Y^{t+1}(u,1)=\mathbf{s}$.

Let $\tilde{X}^t(V,[1,t])=Y^{t+1}(V,[2,t+1])$. Then following (3), we have

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(v,[1,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))}\\
&=\frac{P_v(Y^{t+1}(v,1)=\mathbf{i})P_v(Y^{t+1}(v,[2,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))}\\
&=p_{\mathbf{i}}(v)\\
&\leq\sqrt{\frac{\alpha}{\beta}},\\
&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\
&=\frac{P_v(Y^{t+1}(u,1)=\mathbf{s})P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\
&=1-p_{\mathbf{s}}(u)\\
&\leq 1.
\end{aligned}$$

Case 2: $Y^{t+1}(v,1)=\mathbf{i}$ and $Y^{t+1}(u,1)=\mathbf{i}$.

Let $\tilde{X}^t(v,[1,t])=Y^{t+1}(v,[2,t+1])$ and $\tilde{X}^t(u,1)=\mathbf{i}$. In this case, the states of v do not depend on the states of any other nodes, therefore, it can be seen as the infection source of a graph containing only itself with the infection starting at time 1. Then by Lemma F.1, we can find a $\tilde{X}^t$ such that

$$\frac{P_v(Y^{t+1}(v,[2,t+1]))}{P_v(\tilde{X}^t(v,[2,t]))}\leq\sqrt{\frac{\alpha}{\beta}}.$$

We then have

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(v,[1,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))}\\
&=\frac{P_v(Y^{t+1}(v,1)=\mathbf{i})P_v(Y^{t+1}(v,[2,t+1]))}{P_v(\tilde{X}^t(v,1)=\mathbf{i})P_v(\tilde{X}^t(v,[2,t]))}\\
&\leq\sqrt{\frac{\alpha}{\beta}}.
\end{aligned}$$

After it gets infected at time 1, node u serves as the infection source of $T_u(v;G)$ with the infection starting at time 1. By the induction assumption, we can find a $\tilde{X}^t$ such that

$$\frac{P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[2,t]))}\leq\sqrt{\frac{\alpha}{\beta}}\leq 1.$$

The following inequality then holds,

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\
&=\frac{P_v(Y^{t+1}(u,1)=\mathbf{i})P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^t(u,1)=\mathbf{i})P_v(\tilde{X}^t(T_u(v;G),[2,t])}\\
&\leq 1.
\end{aligned}$$

Case 3: $Y^{t+1}(v,1)=\mathbf{r}$, $Y^{t+1}(v,2)=\mathbf{i}$ and $Y^{t+1}(u,1)=\mathbf{n}$.

Let $\tilde{X}^t(V,[1,t]) = Y^{t+1}(V,[2,t+1])$. Following (4), we have

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(v,[1,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))}\\
&= \frac{P_v(Y^{t+1}(v,1)=\mathbf{r})P_v(Y^{t+1}(v,2)=\mathbf{i})}{P_v(\tilde{X}^t(v,1)=\mathbf{i})}\\
&\quad\cdot\frac{P_v(Y^{t+1}(v,[3,t+1]))}{P_v(\tilde{X}^t(v,[2,t]))}\\
&= \frac{(1-p_{\mathbf{i}}(v))p_{\mathbf{r}}(v)}{p_{\mathbf{i}}(v)}\\
&\le \sqrt{\frac{\alpha}{\beta}},\\
&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\
&= \frac{P_v(Y^{t+1}(u,1)=\mathbf{n})P_v(Y^{t+1}(u,2)=\mathbf{s})}{P_v(\tilde{X}^t(u,1)=\mathbf{s})}\\
&\quad\cdot\frac{P_v(Y^{t+1}(T_u(v;G),[3,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[2,t]))}\\
&= 1.
\end{aligned}$$

Case 4: $Y^{t+1}(v,1)=\mathbf{r}$, $Y^{t+1}(v,2)=\mathbf{i}$ and $Y^{t+1}(u,1)=\mathbf{i}$.

Let $\tilde{X}^t(v,[1,t]) = Y^{t+1}(v,[2,t+1])$ and $\tilde{X}^t(u,1)=\mathbf{i}$. Following (4), we have

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(v,[1,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))}\\
&= \frac{P_v(Y^{t+1}(v,1)=\mathbf{r})P_v(Y^{t+1}(v,2)=\mathbf{i})}{P_v(\tilde{X}^t(v,1)=\mathbf{i})}\\
&\quad\cdot\frac{P_v(Y^{t+1}(v,[3,t+1]))}{P_v(\tilde{X}^t(v,[2,t]))}\\
&= \frac{(1-p_{\mathbf{i}}(v))p_{\mathbf{r}}(v)}{p_{\mathbf{i}}(v)}\\
&\le \sqrt{\frac{\alpha}{\beta}}.
\end{aligned}$$

Following the same arguments as that in case 2, we can find a $\tilde{X}^t$ such that

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\
&= \frac{P_v(Y^{t+1}(u,1)=\mathbf{i})P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^t(u,1)=\mathbf{i})P_v(\tilde{X}^t(T_u(v;G),[2,t]))}\\
&\le 1.
\end{aligned}$$

Case 5: $Y^{t+1}(v,1)=\mathbf{r}$, $Y^{t+1}(v,2)=\mathbf{r}$ and $Y^{t+1}(u,1)=\mathbf{n}$.

Let $\tilde{X}^t(V,[1,t]) = Y^{t+1}(V,[2,t+1])$. Following (4), we have

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(v,[1,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))}\\
&= \frac{P_v(Y^{t+1}(v,1)=\mathbf{r})P_v(Y^{t+1}(v,2)=\mathbf{r})}{P_v(\tilde{X}^t(v,1)=\mathbf{r})}\\
&\quad\cdot\frac{P_v(Y^{t+1}(v,[3,t+1]))}{P_v(\tilde{X}^t(v,[2,t]))}\\
&= \frac{(1-p_{\mathbf{i}}(v))(1-p_{\mathbf{r}}(v))}{1-p_{\mathbf{i}}(v)}\\
&\le \sqrt{\frac{\alpha}{\beta}},\\
&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\
&= \frac{P_v(Y^{t+1}(u,1)=\mathbf{n})P_v(Y^{t+1}(u,2)=\mathbf{n})}{P_v(\tilde{X}^t(u,1)=\mathbf{n})}\\
&\quad\cdot\frac{P_v(Y^{t+1}(T_u(v;G),[3,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[2,t]))}\\
&= 1.
\end{aligned}$$

Case 6: $Y^{t+1}(v,1)=\mathbf{r}$, $Y^{t+1}(v,2)=\mathbf{r}$ and $Y^{t+1}(u,1)=\mathbf{i}$.

Let $\tilde{X}^t(v,[1,t]) = Y^{t+1}(v,[2,t+1])$ and $\tilde{X}^t(u,1)=\mathbf{i}$. Following the same arguments as that in case 2, we can find a $\tilde{X}^t$ such that

$$\frac{P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[2,t]))} \le 1.$$

Then following (4), we have

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(v,[1,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))}\\
&= \frac{P_v(Y^{t+1}(v,1)=\mathbf{r})P_v(Y^{t+1}(v,2)=\mathbf{r})}{P_v(\tilde{X}^t(v,1)=\mathbf{r})}\\
&\quad\cdot\frac{P_v(Y^{t+1}(v,[3,t+1]))}{P_v(\tilde{X}^t(v,[2,t]))}\\
&= \frac{(1-p_{\mathbf{i}}(v))(1-p_{\mathbf{r}}(v))}{1-p_{\mathbf{i}}(v)}\\
&\le \sqrt{\frac{\alpha}{\beta}},\\
&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\
&= \frac{P_v(Y^{t+1}(u,1)=\mathbf{i})P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^t(u,1)=\mathbf{i})P_v(\tilde{X}^t(T_u(v;G),[2,t]))}\\
&\le 1.
\end{aligned}$$

Therefore, we have shown that claim (a) and (b) hold for all six possible cases. This completes the proof for the inductive step. The proof of Lemma A.1 for uninfected subtree is now complete.

B. Proof of Lemma A.1 for Infected Subtree

If $T_u(v;G)$ is an infected subtree, we can see that $\mathcal{T}_v = [\bar{d}(v,V_{\mathbf{i}}),+\infty)$. We prove claim (a) and (b) for infected subtree by mathematical induction on $\bar{d}(v,V_{\mathbf{i}})$.

Basis step: $\bar{d}(v,V_{\mathbf{i}})=1$**.**

For any $t \geq 1$, we consider any most likely infection path Y^{t+1} for $(v, t+1)$. In the following, six possible cases for Y^{t+1} are discussed in order.

Case 1: $Y^{t+1}(v,1) = \mathbf{i}$ and $Y^{t+1}(u,1) = \mathbf{s}$.

Let $\tilde{X}^t(V,[1,t]) = Y^{t+1}(V,[2,t+1])$. Then following (3), we have

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(v,[1,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))} \\
&= \frac{P_v(Y^{t+1}(v,1)=\mathbf{i})P_v(Y^{t+1}(v,[2,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))} \\
&= p_{\mathbf{i}}(v) \\
&\leq \sqrt{\frac{\alpha}{\beta}}, \\
&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))} \\
&= \frac{P_v(Y^{t+1}(u,1)=\mathbf{s})P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))} \\
&= 1 - p_{\mathbf{s}}(u) \\
&\leq 1.
\end{aligned}$$

Case 2: $Y^{t+1}(v,1) = \mathbf{i}$ and $Y^{t+1}(u,1) = \mathbf{i}$.

Let $\tilde{X}^t(v,[1,t]) = Y^{t+1}(v,[2,t+1])$ and $\tilde{X}^t(u,1) = \mathbf{i}$, following (3), we have

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(v,[1,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))} \\
&= \frac{P_v(Y^{t+1}(v,1)=\mathbf{i})P_v(Y^{t+1}(v,[2,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))} \\
&= p_{\mathbf{i}}(v) \\
&\leq \sqrt{\frac{\alpha}{\beta}}.
\end{aligned}$$

After it gets infected at time slot 1, node u serves as the infection source of the subtree $T_u(v;G)$ with the infection starting at time 1. From the assumption $\bar{d}(v, V_{\mathbf{i}}) = 1$, it follows that $T_w(u;G)$ is an uninfected subtree for any $w \in V(u,1) \bigcap T_u(v;G)$. Then by Lemma A.1(c) for uninfected subtree, we can find a $\tilde{X}^t$ such that

$$\begin{aligned}
\frac{P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[2,t]))} &\leq \sqrt{\frac{\alpha}{\beta}} \\
&\leq 1. \qquad (23)
\end{aligned}$$

We then have,

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))} \\
&= \frac{P_v(Y^{t+1}(u,1)=\mathbf{i})P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^t(u,1)=\mathbf{i})P_v(\tilde{X}^t(T_u(v;G),[2,t]))} \\
&\leq 1.
\end{aligned}$$

Case 3: $Y^{t+1}(v,1) = \mathbf{r}$, $Y^{t+1}(v,2) = \mathbf{i}$ and $Y^{t+1}(u,1) = \mathbf{n}$.

Let $\tilde{X}^t(V,[1,t]) = Y^{t+1}(V,[2,t+1])$. Following (4), we have

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(v,[1,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))} \\
&= \frac{P_v(Y^{t+1}(v,1)=\mathbf{r})P_v(Y^{t+1}(v,2)=\mathbf{i})}{P_v(\tilde{X}^t(v,1)=\mathbf{i})} \\
&\quad \cdot \frac{P_v(Y^{t+1}(v,[3,t+1]))}{P_v(\tilde{X}^t(v,[2,t])} \\
&= \frac{(1-p_{\mathbf{i}}(v))p_{\mathbf{r}}(v)}{p_{\mathbf{i}}(v)} \\
&\leq \sqrt{\frac{\alpha}{\beta}}, \\
&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))} \\
&= \frac{P_v(Y^{t+1}(u,1)=\mathbf{n})P_v(Y^{t+1}(u,2)=\mathbf{s})}{P_v(\tilde{X}^t(u,1)=\mathbf{s})} \\
&\quad \cdot \frac{P_v(Y^{t+1}(T_u(v;G),[3,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[2,t])} \\
&= 1.
\end{aligned}$$

Case 4: $Y^{t+1}(v,1) = \mathbf{r}$, $Y^{t+1}(v,2) = \mathbf{i}$ and $Y^{t+1}(u,1) = \mathbf{i}$.

Let $\tilde{X}^t(v,[1,t]) = Y^{t+1}(v,[2,t+1])$ and $\tilde{X}^t(u,1) = \mathbf{i}$. Following the same arguments as that in case 2, we can find a $\tilde{X}^t$ such that (23) holds. Then following (4), we have

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(v,[1,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))} \\
&= \frac{P_v(Y^{t+1}(v,1)=\mathbf{r})P_v(Y^{t+1}(v,2)=\mathbf{i})P_v(Y^{t+1}(v,[3,t+1]))}{P_v(\tilde{X}^t(v,1)=\mathbf{i})P_v(\tilde{X}^t(v,[2,t]))} \\
&= \frac{(1-p_{\mathbf{i}}(v))p_{\mathbf{r}}(v)}{p_{\mathbf{i}}(v)} \\
&\leq \sqrt{\frac{\alpha}{\beta}}, \\
&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))} \\
&= \frac{P_v(Y^{t+1}(u,1)=\mathbf{i})P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^t(u,1)=\mathbf{i})P_v(\tilde{X}^t(T_u(v;G),[2,t]))} \\
&\leq 1.
\end{aligned}$$

Case 5: $Y^{t+1}(v,1) = \mathbf{r}$, $Y^{t+1}(v,2) = \mathbf{r}$ and $Y^{t+1}(u,1) = \mathbf{n}$.

Let $\tilde{X}^t(V,[1,t]) = Y^{t+1}(V,[2,t+1])$. Then following (4),

we have

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(v,[1,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))}\\
&=\frac{P_v(Y^{t+1}(v,1)=\mathbf{r})P_v(Y^{t+1}(v,2)=\mathbf{r})}{P_v(\tilde{X}^t(v,1)=\mathbf{r})}\\
&\quad\cdot\frac{P_v(Y^{t+1}(v,[3,t+1]))}{P_v(\tilde{X}^t(v,[2,t]))}\\
&=\frac{(1-p_{\mathbf{i}}(v))(1-p_{\mathbf{r}}(v))}{1-p_{\mathbf{i}}(v)}\\
&\leq\sqrt{\frac{\alpha}{\beta}},\\
&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\
&=\frac{P_v(Y^{t+1}(u,1)=\mathbf{n})P_v(Y^{t+1}(u,2)=\mathbf{n})}{P_v(\tilde{X}^t(u,1)=\mathbf{n})}\\
&\quad\cdot\frac{P_v(Y^{t+1}(T_u(v;G),[3,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[2,t]))}\\
&=1.
\end{aligned}$$

Case 6: $Y^{t+1}(v,1)=\mathbf{r}$, $Y^{t+1}(v,2)=\mathbf{r}$ and $Y^{t+1}(u,1)=\mathbf{i}$.

Let $\tilde{X}^t(v,[1,t])=Y^{t+1}(v,[2,t+1])$ and $\tilde{X}^t(u,1)=\mathbf{i}$. Following the same arguments as that in case 2, we can find a $\tilde{X}^t$ such that (23) holds. Then following (4), we have

$$\begin{aligned}
&\frac{P_v(Y^{t+1}(v,[1,t+1]))}{P_v(\tilde{X}^t(v,[1,t]))}\\
&=\frac{P_v(Y^{t+1}(v,1)=\mathbf{r})P_v(Y^{t+1}(v,2)=\mathbf{r})}{P_v(\tilde{X}^t(v,1)=\mathbf{r})}\\
&\quad\cdot\frac{P_v(Y^{t+1}(v,[3,t+1]))}{P_v(\tilde{X}^t(v,[2,t]))}\\
&=\frac{(1-p_{\mathbf{i}}(v))(1-p_{\mathbf{r}}(v))}{1-p_{\mathbf{i}}(v)}\\
&\leq\sqrt{\frac{\alpha}{\beta}},\\
&\frac{P_v(Y^{t+1}(T_u(v;G),[1,t+1]))}{P_v(\tilde{X}^t(T_u(v;G),[1,t]))}\\
&=\frac{P_v(Y^{t+1}(u,1)=\mathbf{i})P_v(Y^{t+1}(T_u(v;G),[2,t+1]))}{P_v(\tilde{X}^t(u,1)=\mathbf{i})P_v(\tilde{X}^t(T_u(v;G),[2,t]))}\\
&\leq 1.
\end{aligned}$$

We have shown that claim (a) and (b) hold for all six possible cases. This completes the proof for the basis step.

Inductive step: assume claim (a) and (b) hold for $\bar{d}(v,V_{\mathbf{i}})\leq n-1$, where $n\geq 2$. We want to show that claim (a) and (b) also hold for $\bar{d}(v,V_{\mathbf{i}})=n$.

Assume $\bar{d}(v,V_{\mathbf{i}})=n$ and consider any most likely infection path Y^{t+1} for $(v,t+1)$, where $t\geq n$. We first show that (23) in case 2 also holds in the inductive step. For case 2, we have $Y^{t+1}(v,1)=\mathbf{i}$ and $Y^{t+1}(u,1)=\mathbf{i}$. Let $\tilde{X}^t(v,[1,t])=Y^{t+1}(v,[2,t+1])$ and $\tilde{X}^t(u,1)=\mathbf{i}$, after it gets infected at time slot 1, node u will serve as the infection source of the subtree $T_u(v;G)$ with the infection starting at time 1. Since $\bar{d}(u,V_{\mathbf{i}}\bigcap T_u(v;G))\leq n-1$, from the induction assumption, we can find a $\tilde{X}^t$ such that (23) holds. From the same arguments as that in the basis step, it follows that claim (a) and (b) hold for all six possible cases. This completes the proof for the inductive step. By the spirit of mathematical induction, the proof of Lemma A.1 for infected subtree is now complete. This completes the proof of Lemma A.1.

Appendix G
Proof of Lemma A.2

Let d be the degree of any node in G. Fix the elapsed time to be t and consider any most likely infection path X^t for (v,t). Given any $u\in H_v\backslash\{v\}$, we first show that

$$t_{int}(u)\in[d(v,u),t-\bar{d}(u,T_u(v;H_v))]. \tag{24}$$

Firstly, it is easy to see that any node in H_v has been infected at least once due to the assumption that the underlying network G is a tree, otherwise, the infection can not reach the leaf nodes of H_v. We now consider the lower bound of $t_{int}(u)$ in (24). Since the infection can spread at most one hop away from v in each time slot, the earliest time for u to get the infection is $d(v,u)$. Then we consider the upper bound of $t_{int}(u)$ in (24). After node u gets infected for the first time at $t_{int}(u)$, the infection can spread at most $t-t_{int}(u)$ hops away from u. Consider a node w such that $d(u,w)=\bar{d}(u,T_u(v;H_v))$. In order for the infection to reach node w, $t-t_{int}(u)\geq d(u,w)$. So $t_{int}(u)\leq t-d(u,w)=t-\bar{d}(u,T_u(v;H_v))$. The proof for (24) is now complete.

Suppose that there exists a node $u\in H_v\backslash\{v\}$ such that the first infection time $t_{int}(u)$ of u in X^t is less than $t-\bar{d}(u,T_u(v;H_v))$. To prove Lemma A.2, following (24), it suffices to show that we can construct another infection path $\tilde{X}^t$ for (v,t) that occurs with at least the same probability as X^t, where the first infection time of u in $\tilde{X}^t$ is $\tilde{t}(u)=t-\bar{d}(u,T_u(v;H_v))$.

We let the states of $G\backslash(T_u(v;G)\bigcup\{\mathrm{pa}(u)\})$ in $\tilde{X}^t$ to be the same as those in X^t, i.e.

$$\begin{aligned}
&\tilde{X}^t(G\backslash(T_u(v;G)\bigcup\{\mathrm{pa}(u)\}),[1,t])\\
&=X^t(G\backslash(T_u(v;G)\bigcup\{\mathrm{pa}(u)\}),[1,t]).
\end{aligned}$$

Let $\tilde{X}^t(\mathrm{pa}(u),[1,t_{int}(u)-1])=X^t(\mathrm{pa}(u),[1,t_{int}(u)-1])$ and $A=T_u(v;G)\bigcup\{\mathrm{pa}(u)\}\bigcup V(\mathrm{pa}(u),1)$. It suffices to show that

$$P_v(\tilde{X}^t(A,[t_{int}(u),t]))\geq P_v(X^t(A,[t_{int}(u),t])). \tag{25}$$

We show (25) by mathematical induction on $\bar{d}(u,T_u(v;H_v))$.

Basis step: show (25) holds for $\bar{d}(u,T_u(v;H_v))=0$.

Let B denote the set of nodes $V(\mathrm{pa}(u),1)\backslash\{u\}$. Consider a time slot $\tau<\tilde{t}(u)$ where $\tilde{X}^t(\mathrm{pa}(u),\tau)=\mathbf{i}$. We show the worst case for $\tilde{X}^t$ at time $\tau+1$.

If $X^t(\mathrm{pa}(u),\tau)=\mathbf{i}$, we have

$$P_v(\tilde{X}^t(B,\tau+1))=P_v(X^t(B,\tau+1). \tag{26}$$

If $X^t(\mathrm{pa}(u),\tau)\neq\mathbf{i}$, we have

$$\frac{P_v(\tilde{X}^t(B,\tau+1))}{P_v(X^t(B,\tau+1)}\geq(1-p_{\mathbf{s}})^{d-1}, \tag{27}$$

where the equality holds when every node in B is susceptible in $\tilde{X}^t$ and non-susceptible in X^t at time τ. By (26) and (27), we can see that the worst case for $\tilde{X}^t$ at time $\tau+1$ is that $X^t(\mathrm{pa}(u),\tau)\neq \mathbf{i}$ and $X^t(B,\tau)=\mathbf{n}$.

We divide the time interval $[t_{int}(u),t]$ into three parts: $t_{int}(u)$, $[t_{int}(u)+1,\tilde{t}(u)-1]$ and $\tilde{t}(u)$, where $\tilde{t}(u)=t-\bar{d}(u,T_u(v;H_v))=t$.

Part 1: time $\tau=t_{int}(u)$.

Since node u is infected for the first time at time slot $t_{int}(u)$ in X^t, we know that node $\mathrm{pa}(u)$ must be infected at time $t_{int}(u)-1$, which in turn suggests that $\tilde{X}^t(\mathrm{pa}(u),\tau-1)=X^t(\mathrm{pa}(u),\tau-1)=\mathbf{i}$, yielding

$$P_v(\tilde{X}^t(B,\tau))=P_v(X^t(B,\tau)). \tag{28}$$

We let $\tilde{X}^t(\mathrm{pa}(u),\tau)=\mathbf{i}$ and consider the worst case in (27). Following (5) and (28), we have

$$\begin{aligned}
&\frac{P_v(\tilde{X}^t(A,t_{int}(u)))}{P_v(X^t(A,t_{int}(u)))}\\
&=\frac{P_v(\tilde{X}^t(\mathrm{pa}(u),t_{int}(u)))P_v(\tilde{X}^t(u,t_{int}(u)))}{P_v(X^t(\mathrm{pa}(u),t_{int}(u)))P_v(X^t(u,t_{int}(u)))}\\
&\quad\cdot\frac{P_v(\tilde{X}^t(B,t_{int}(u)))}{P_v(X^t(B,t_{int}(u)))}\\
&\geq\frac{p_{\mathbf{i}}(1-p_{\mathbf{s}})}{(1-p_{\mathbf{i}})p_{\mathbf{s}}} &(29)\\
&\geq 1. &(30)
\end{aligned}$$

Part 2: time $\tau\in[t_{int}(u)+1,t-1]$.

We first consider the case that at least one node in $T_u(v;G)$ is infected at time τ. We let $\tilde{X}^t(\mathrm{pa}(u),\tau)=\mathbf{i}$ and consider the worst case, i.e., $X^t(\mathrm{pa}(u),\tau-1)\neq\mathbf{i}$ and $X^t(B,\tau-1)=\mathbf{n}$. We then have

$$P_v(\tilde{X}^t(A,\tau))\geq p_{\mathbf{i}}(1-p_{\mathbf{s}})^d.$$

Since $X^t(\mathrm{pa}(u),\tau-1)\neq\mathbf{i}$ and at least one node in $T_u(v;G)$ is infected at time τ, there must exist a node $w\in T_u(v;G)$, s.t., $X^t(w,\tau-1)=\mathbf{i}$. Consider any neighboring node z of w. If $X^t(z,\tau-1)=\mathbf{i}$, due to the fact that $X^t(\mathrm{pa}(u),\tau-1)\neq\mathbf{i}$ and the assumption that G is an infinite tree, we can always find a node $y\in T_z(w;G)\bigcap(T_u(v;G)\bigcup\{\mathrm{pa}(u)\})$, s.t., $X^t(y,\tau-1)=\mathbf{s}$. If $X^t(z,\tau-1)=\mathbf{s}$, following similar arguments as the worst case in (27), we can see that $X^t(z,\tau)\neq\mathbf{i}$. We then have

$$P_v(X^t(T_z(w;G)\bigcap(T_u(v;G)\bigcup\{\mathrm{pa}(u)\}),\tau))\leq 1-p_{\mathbf{s}},$$

for any neighboring node z of w. Moreover, we have at least one node in $T_u(v;G)$ being infected at time τ, yielding

$$\begin{aligned}
P_v(X^t(A,\tau))&\leq\max\{p_{\mathbf{i}},p_{\mathbf{s}}\}(1-p_{\mathbf{s}})^d\\
&=p_{\mathbf{i}}(1-p_{\mathbf{s}})^d\\
&\leq P_v(\tilde{X}^t(A,\tau)). &(31)
\end{aligned}$$

We then consider the case that no node in $T_u(v;G)$ is infected at time τ. Without loss of generality, we assume τ is the earliest time after $t_{int}(u)$ that no node in $T_u(v;G)$ is infected. We let $\tilde{X}(\mathrm{pa}(u),\tau)=X^t(\mathrm{pa}(u),\tau)$ and consider the worst case for $\tilde{X}^t(\mathrm{pa}(u),\tau)$. If $X^t(\mathrm{pa}(u),\tau)=\mathbf{i}$, we have

$$\begin{aligned}
\frac{P_v(\tilde{X}^t(A,\tau))}{P_v(X^t(A,\tau))}&\geq\frac{p_{\mathbf{i}}(1-p_{\mathbf{s}})^d}{p_{\mathbf{s}}(1-p_{\mathbf{i}})(1-p_{\mathbf{s}})^{d-1}}\\
&\geq 1. &(32)
\end{aligned}$$

If $X^t(\mathrm{pa}(u),\tau)\neq\mathbf{i}$, we have

$$\begin{aligned}
\frac{P_v(\tilde{X}^t(A,\tau))}{P_v(X^t(A,\tau))}&\geq\frac{(1-p_{\mathbf{i}})(1-p_{\mathbf{s}})^d}{(1-p_{\mathbf{s}})(1-p_{\mathbf{i}})(1-p_{\mathbf{s}})^{d-1}}\\
&=1. &(33)
\end{aligned}$$

From (31)-(33), we have

$$P_v(\tilde{X}^t(A,[t_{int}(u)+1,\tau]))\geq P_v(X^t(A,[t_{int}(u)+1,\tau])). \tag{34}$$

We have now $\tilde{X}^t(V,\tau)=X^t(V,\tau)$. If there are other time slots after τ that no node in $T_u(v;G)$ is infected, we can apply the same arguments again. Then by (31) and (34), we have

$$P_v(\tilde{X}^t(A,[t_{int}(u)+1,t-1]))\geq P_v(X^t(A,[t_{int}(u)+1,t-1])). \tag{35}$$

Part 3: time $\tau=t$.

If $\mathrm{pa}(u)\in V_{\mathbf{i}}$, we have

$$\begin{aligned}
\frac{P_v(\tilde{X}^t(A,t))}{P_v(X^t(A,t))}&\geq\frac{p_{\mathbf{i}}p_{\mathbf{s}}(1-p_{\mathbf{s}})^{d-1}}{p_{\mathbf{s}}p_{\mathbf{i}}(1-p_{\mathbf{s}})^{d-1}}\\
&=1. &(36)
\end{aligned}$$

Then (25) holds from (30), (35) and (36).

If $\mathrm{pa}(u)\notin V_{\mathbf{i}}$, we have

$$\begin{aligned}
\frac{P_v(\tilde{X}^t(A,t))}{P_v(X^t(A,t))}&\geq\frac{(1-p_{\mathbf{i}})p_{\mathbf{s}}(1-p_{\mathbf{s}})^{d-1}}{(1-p_{\mathbf{s}})p_{\mathbf{i}}(1-p_{\mathbf{s}})^{d-1}}\\
&=\frac{(1-p_{\mathbf{i}})p_{\mathbf{s}}}{p_{\mathbf{i}}(1-p_{\mathbf{s}})}. &(37)
\end{aligned}$$

Then (25) holds from (29), (35) and (37). This competes the proof of the basis step.

Inductive step: assume (25) holds for $\bar{d}(u,T_u(v;H_v))\leq n$, where $0\leq n\leq\bar{d}(v,H_v)-1$. Show (25) also holds for $\bar{d}(u,T_u(v;H_v))=n+1$.

We divide the time interval $[t_{int}(u),t]$ into four parts: $t_{int}(u)$, $[t_{int}(u)+1,\tilde{t}(u)-1]$, $\tilde{t}(u)$ and $[\tilde{t}(u)+1,t]$, where $\tilde{t}(u)=t-\bar{d}(u,T_u(v;H_v))$. For any node $w\in\mathrm{ch}(u)$, we have $\bar{d}(w,T_w(v;H_v))\leq\bar{d}(u,T_u(v;H_v))-1=n$. By induction assumption, we have that node w get infected for the first time at $t(w)=t-\bar{d}(w,T_w(v;H_v))$ in X^t, which in turn suggests that $X^t(u,\tilde{t}(u))=\mathbf{i}$. For the time range $[\tilde{t}(u)+1,t]$, we let $\tilde{X}^t(A,[\tilde{t}(u)+1,t])=X^t(A,[\tilde{t}(u)+1,t])$, yielding

$$P_v(\tilde{X}^t(A,[\tilde{t}(u)+1,t]))=P_v(X^t(A,[\tilde{t}(u)+1,t])).$$

For the first three parts, following similar arguments as in the basis step, we have

$$P_v(\tilde{X}^t(A,[t_{int}(u)+1,\tilde{t}(u)]))\geq P_v(X^t(A,[t_{int}(u)+1,\tilde{t}(u)])).$$

We can now conclude that (25) holds for the inductive step. By the spirit of mathematical induction, (25) holds and the proof of Lemma A.2 is now complete.

REFERENCES

[1] W.-J. Bai, T. Zhou, and B.-H. Wang, "Immunization of susceptible-infected model on scale-free networks," *Physica A: Statistical Mechanics and its Applications*, vol. 384, no. 2, pp. 656–662, 2007.

[2] A. Wu and Y. Wang, "Role of diffusion in an epidemic model of mobile individuals on networks," *The European Physical Journal B*, vol. 85, no. 8, pp. 1–6, 2012.

[3] Y. Shang, "Mixed SI(R) epidemic dynamics in random graphs with general degree distributions," *Applied Mathematics and Computation*, vol. 219, no. 10, pp. 5042–5048, 2013.

[4] Y.-F. Chou, H.-H. Huang, and R.-G. Cheng, "Modeling information dissemination in generalized social networks," *IEEE Commun. Lett.*, vol. 17, no. 7, pp. 1356–1359, 2013.

[5] N. Bailey, *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, 1975.

[6] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

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

[8] W. Dong, W. Zhang, and C. W. Tan, "Rooting out the rumor culprit from suspects," in *Proc. IEEE International Symposium on Information Theory*, 2013.

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

[10] W. Luo, W. P. Tay, and M. Leng, "How to identify an infection source with limited observations," *IEEE J. Sel. Top. Sign. Proces.*, vol. 8, no. 4, pp. 586–597, 2014.

[11] ——, "Identifying infection sources and regions in large networks," *IEEE Trans. Signal Process.*, vol. 61, no. 11, pp. 2850–2865, 2013.

[12] K. Zhu and L. Ying, "Information source detection in the SIR model: a sample path based approach," in *Proc. Information Theory and Applications Workshop*, 2013.

[13] ——, "A robust information source estimator with sparse observations," *Computational Social Networks*, vol. 1, no. 1, 2014.

[14] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, "Inferring the origin of an epidemic with a dynamic message-passing algorithm," *arXiv*, 2013. [Online]. Available: http://arxiv.org/abs/1303.5315

[15] F. Altarelli, A. Braunstein, L. Dall'Asta, A. Lage-Castellanos, and R. Zecchina, "Bayesian inference of epidemics on networks via belief propagation." *arXiv*, 2013. [Online]. Available: http://arxiv.org/abs/1307.6786

[16] S. M. Blower, T. C. Porco, and G. Darby, "Predicting and preventing the emergence of antiviral drug resistance in HSV-2," *Nature Medicine*, vol. 4, pp. 673 – 678, 1998.

[17] P. V. D. Driessche and X. Zou, "Modeling relapse in infectious diseases," *Mathematical Biosciences*, vol. 207, no. 1, pp. 89–103, 2007.

[18] V. D. L. Cruz, "On the global stability of infectious diseases models with relapse," *Abstraction & Application*, vol. 9, pp. 50–61, 2013.

[19] P. Georgescu and H. Zhang, "A lyapunov functional for a SIRI model with nonlinear incidence of infection and relapse," *Applied Mathematics and Computation*, vol. 219, no. 16, pp. 8496 – 8507, 2013.

[20] H. W. Hethcote and J. A. Yorke, *Gonorrhea Transmission Dynamics and Control*, ser. Lecture Notes in Biomathematics. Springer-Verlag, 1984.

[21] H. Hethcote, "Qualitative analyses of communicable disease models," *Mathematical Biosciences*, vol. 28, no. 3-4, pp. 335–356, 1976.

[22] L. J. Allen, "Some discrete-time SI, SIR, and SIS epidemic models," *Mathematical Biosciences*, vol. 124, no. 1, pp. 83–105, 1994.

[23] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.

[24] W. Hu, W. P. Tay, A. Harilal, and G. Xiao, "Network infection source identification under the SIRI model," *arXiv:1410.2995*, 2014.

[25] J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks," in *Proc. Neural Information Processing Systems Conference*, 2012.

[26] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks." *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

Wuqiong Luo (S'12 M'15) received the B.Eng. degree in Electrical and Electronic Engineering with First Class Honours from Nanyang Technological University, Singapore in 2010. He obtained the PhD degree in Electrical and Electronic Engineering from Nanyang Technological University, Singapore in 2015. He is currently a Senior Data Scientist at Allianz SE, Singapore Branch. His current research interest are in statistical learning using big data.

Dr. Luo was coawarded the Best Student Paper Award at the 46th Asilomar Conference on Signals, Systems, and Computers.

Wee Peng Tay (S'06 M'08 SM'14) received the B.S. degree in Electrical Engineering and Mathematics, and the M.S. degree in Electrical Engineering from Stanford University, Stanford, CA, USA, in 2002. He received the Ph.D. degree in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008. He is currently an Assistant Professor in the School of Electrical and Electronic Engineering at Nanyang Technological University, Singapore. His research interests include distributed inference and signal processing, sensor networks, social networks, information theory, and applied probability.

Dr. Tay received the Singapore Technologies Scholarship in 1998, the Stanford University President's Award in 1999, the Frederick Emmons Terman Engineering Scholastic Award in 2002, and the Tan Chin Tuan Exchange Fellowship in 2015. He is a coauthor of the best student paper award at the Asilomar conference on Signals, Systems, and Computers in 2012, and the IEEE Signal Processing Society Young Author Best Paper Award in 2016. He is currently an Associate Editor for the IEEE Transactions on Signal Processing, serves on the MLSP TC of the IEEE Signal Processing Society, and is the chair of DSNIG in IEEE MMTC. He has also served as a technical program committee member for various international conferences.

Mei Leng (S'07 M'10) received the B.Eng. degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2005 and the Ph.D. degree from The University of Hong Kong in 2011.

She was a research fellow at the School of Electrical and Electronic Engineering, Nanyang Technological University from 2011 to 2014. She is currently a Research Scientist at TL@NTU. Her current research interests include tracking and navigation algorithm design and implementation, cooperative and distributed algorithms, statistical signal processing and machine learning with applications to wireless sensor networks and wireless communication systems.