An Algorithmic Framework for Estimating Rumor Sources with Different Start Times

Feng Ji, Wee Peng Tay, *Senior Member, IEEE*, and Lav R. Varshney, *Senior Member, IEEE*

Abstract

We study the problem of identifying multiple rumor or infection sources in a network under the susceptible-infected model, and where these sources may start infection spreading at different times. We introduce the notion of an abstract estimator, which given the infection graph, assigns a higher value to each vertex in the graph it considers more likely to be a rumor source. This includes several of the single-source estimators developed in the literature. We introduce the concepts of a quasi-regular tree and a heavy center, which allows us to develop an algorithmic framework that transforms an abstract estimator into a two-source joint estimator, in which the infection graph can be thought of as covered by overlapping infection regions. We show that our algorithm converges to a local optimum of the estimation function if the underlying network is a quasi-regular tree. We further extend our algorithm to more than two sources, and heuristically to general graphs. Simulation results on both synthetic and real-world networks suggest that our algorithmic framework outperforms several existing multiple-source estimators, which typically assume that all sources start infection spreading at the same time.

Index Terms

Rumor source, infection source, multiple source estimation, SI model, quasi-regular tree

I. Introduction

Online social networks have grown immensely in recent decades. More and more users are obtaining news and other information from social networks [1]–[7]. Information is also being propagated across networks with increasing speed due to increases in network connectivity. For example, a recent report [8] shows there are now 1.59 billion Facebook users and the number of degrees of separation between them is only 3.57 on average. Therefore, if a rumor is posted by several individuals on the network, a significant proportion of the network population can be "infected" by it in a short period of time. If the rumor leads to reputation or economic loss, a law enforcement agency may want to identify the network members who started the rumor. Similarly in the epidemiology of infectious diseases that spread through social contacts, it is important to identify *patient zero(s)*.[1] There is also growing interest in network science to trace the sources and spread of ideas [9], which may have several independent origins [10], [11]. All of these real-world problems can be placed in the framework of rumor source detection, which we now describe.

Suppose a rumor spreads in a network, possibly starting at different sources and times. After the rumor has spread for a certain amount of time, we observe the members in the network infected by the rumor. Our task is to infer the rumor source(s) from the infected network. This is called the *rumor source detection* problem.

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.

F. Ji and W. P. Tay are with the School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore (e-mail: jifeng@ntu.edu.sg, wptay@ntu.edu.sg). L. R. Varshney is with the Coordinated Science Laboratory and the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA (e-mail: varshney@illinois.edu).

[1]Outbreaks of the same disease can have independent sources. For example, the Ebola virus first emerged in the Democratic Republic of Congo (then Zaire) in 1976 near the Ebola River, but also nearly simultaneously the same year in South Sudan (then Sudan). In 2014, an outbreak in the Democratic Republic of Congo followed an outbreak in West Africa. Detailed virological and epidemiological analysis performed later shows all four outbreaks had distinct patient zeros.

Mathematically, the problem can be described as follows. Let $G = (V, \mathcal{E})$ be a network graph with set of vertices V and set of edges $\mathcal{E}$. A rumor starts to spread from several independent sources $s_1, \ldots, s_n \in V$ at different times. Although there are various spreading models, in this paper, we consider the homogeneous *Susceptible-Infected (SI)* spreading, in which the rate at which a rumor propagates across each edge in the network is the same, and any infected vertex cannot recover from the infection. Our framework allows us to consider both the continuous-time diffusion model in [12], [13] and the discrete-time diffusion model in [14], [15]. In the continuous-time model, the time taken for the rumor to propagate across each edge is independent and identically distributed as a continuous distribution with exponential tail behavior [12], [13]. In the discrete-time model, time is discretized into slots and the probability of the rumor propagating from an infected node to a suspectible neighbor is the same in each time slot and for all infected nodes [14], [15]. At a certain time instance, an observation of all infected vertices is made. The collection of all infected vertices and the edges among them gives a subgraph I of G. We call I the *infection graph*. We wish to estimate $s_1, \ldots, s_n$ based only on the information in I. Several existing works on infection spreading in a network focus on the diffusion process (see *e.g.* [16]–[19]); we are looking at the "inverse problem."

The single rumor source detection problem ($n = 1$) has been studied extensively under various assumptions using a variety of algorithmic techniques, *e.g.* [12], [14], [15], [20]–[31], to name a few. The problem proves to be challenging even in this situation. One of the pioneering works in rumor source estimation, [12], proposed to find a node that maximizes an estimator function called rumor centrality. Herein, we shall follow this optimization-based strategy as a guideline in the multiple sources problem we consider. Note that several single-source estimators can be cast in the same optimization framework. We would like to mention that apart from an estimator based approach for the single source, there are also discussions using time-stamp information [20], [25], [27], belief propagation [21] and a dynamic message passing method [24]. In this paper, we mainly work with estimators that assume only limited knowledge about the infection process, including a snapshot observation of the infected nodes at a particular point in time, and the network topology. This differs from [21] and [24] (with appropriate extension to estimate multiple sources) as these Bayesian approaches assume additional knowledge of the spreading process in order to construct a probability model.

There have been prior attempts to tackle the multiple sources detection problem, for example [23], [30], [32]–[34]. Let us briefly recall one of the key geometric ideas used in [30]: make a partition[2] of I and find a source in each partition. The drawbacks of this approach are: (a) it is possible that infected vertices from different sources can merge, and performing source estimation independently in each partition may result in a biased estimate (see Figure 1); and (b) different sources may start infection spreading at different times. A successful approach should take both of these issues into account. The first drawback can be mitigated in the absence of the second: algorithms proposed for when infections start at the same time at all sources and the infection size is large are proven to be asymptotically correct for the class of geometric trees [23]. However, all previous works on multiple sources detection assume that sources start their infection spreading at the same time, which limits their practical application.

In this paper, we develop a theoretical framework to estimate rumor sources, given an observation of the infection graph I and the number of rumor sources. We do not assume rumor spreading starts at the same time at every source. For easier explanation, we first study the two-source identification problem in a tree network under the SI model and then generalize our framework to multiple sources and for all graph types. Our approach combines probabilistic, combinatorial, and geometric inference. Compared to other approaches, we replace the concept of partition by that of *covering*, namely, we allow large overlap between infection regions assigned to different infection sources (see Figure 1). To achieve this, we introduce the notion of *heavy center* to describe the subgraph that would have been infected by a single infection source if the infection is deterministic with an unknown rate. Thus the resulting infection graph can be interpreted to contain a deterministic *center* region and stochastic extensions hanging from

[2]For a review of theory on Voronoi partitions and power diagrams, see [35]–[37].

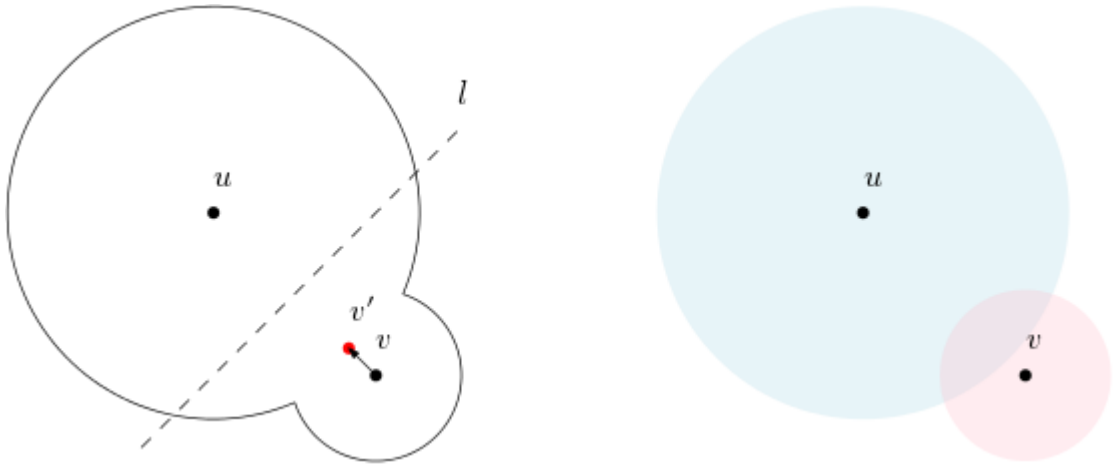

Fig. 1. Schematic illustration of the difference between a partition scheme and a covering scheme we introduce in the paper. Suppose u and v are sources. A partition scheme (left) may give two regions separated by the line l; however, some parts that are apparently infected by u are allocated to v. If one applies a single-source detection algorithm in the region of v, the estimated position of the source v may be shifted away from its true location to v' (the red node). Therefore, we propose a new scheme (right) by allowing overlap, while making sure the regions are "correctly" allocated to both nodes u (the blue region) and v (the pink region).

this center due to the randomness of the infection process. (This center is called *heavy* in the sense that it is a region instead of a single vertex. More details are presented in Section III.) Intuitively, we want the region infected by every source to be: (a) combinatorially/probabilistically optimal; as well as (b) geometrically feasible.

The rest of the paper is organized as follows. Section II introduces the notion of quasi-regular tree. This is the basic graph model adopted in the paper for source estimation. Section III is the core of the paper. It introduces the notions of covering, heavy center, and the joint source estimators. It demonstrates how to combine geometry, probability, and combinatorics in handling the two-source identification problem in a tree. Section IV presents our algorithm and discusses its implementation and complexity, and Section V extends to general graphs. Section VI presents simulation results to verify the performance of our proposed approach. Finally, Section VII concludes.

II. The Basic Model: Quasi-regular Tree

In this section, we introduce quasi-regular trees as the basic model for discussion. Let T be a tree with V as the set of vertices and $\mathcal{E}$ as the set of edges. Given any two vertices u and v, denote the shortest path between them by $[u, v]$ and use (u, v), $(u, v]$, and so on to denote the path with the corresponding boundary vertex at the open side excluded. Let $d(u, v)$ be the length of the path between u and v in T. For any subset $S \subset T$, by abuse of notation, we write S for $S \cap V$ if no confusion arises. The size of $S \cap V$ is denoted $|S|$. In the following, we give a few elementary definitions that lead to the notion of quasi-regular trees.

Definition 1. Let $\mathbb{Z}^+$ denote the set of non-negative integers. For $x \in V$ and $r \in \mathbb{Z}^+$, write

$$D_T(x, r) = \{y \in T \mid d(x, y) \le r\}$$

as the *closed disc centered at x with radius r*.

Definition 2. Let $S \subset T$.

(a) The convex hull $\mathrm{conv}(S)$ of S is defined as the intersection of all connected subtrees of T containing S.

(b) If S is connected, it is called a subtree. If S is also finite, define the height of S as

$$\ell(S) = \max\{d(x, y) \mid x, y \in S\}.$$

For a fixed point x_0 of S, define the height based at x_0 as

$$\ell_{x_0}(S) = \max\{d(x_0, y) \mid y \in S\}.$$

(c) A subtree S' of height r of $S \subset T$ is called a *full subtree of height r in S* if for any $x \in S \backslash S'$, the convex hull of $S' \cup \{x\}$ has height greater than r.

In this paper, we assume that the number of rumor sources is known *a priori* (we have developed a method to estimate the number of sources in [38]). We develop an algorithmic framework that finds an appropriate infection region for each source (which can overlap with each other), if that source had been the only rumor source. Then, existing rumor source estimators like those in [12], [14], [15], [23] can be applied in each of these infection regions to identify the corresponding rumor source. We refer to any of these rumor source estimators as *abstract estimators*, as defined below.

Definition 3. An abstract estimator is a pair (E, e) such that:

(a) e is an estimation function that assigns a non-negative real number to each pair (T', v), where T' is a subtree of T and $v \in T'$; and

(b) E is a source estimator that assigns to each connected subtree $T' \subset T$ a subset of vertices $E(T')$ of T' such that

$$E(T') = \arg\max_{v \in T'} e(T', v).$$

Because of property (b) in Definition 3, it is enough to specify the estimation function e. We include E as part of the definition to simplify notation in the rest of the paper. The function $e(T', v)$ is an estimation function that assigns a higher value to a vertex v it considers more likely to be a rumor source, given that the infected nodes are T'. In the following, we give some examples to illustrate Definition 3.

Example 1. As we remarked earlier, to give a complete description of (E, e), it is enough to specify the estimation function e. Let T' be an observed infection tree.

(i) The maximum likelihood estimator (MLE): Let I denote the infection tree generated by the source s. The MLE estimator is given by

$$e_{MLE}(T', v) = \mathbb{P}\left(I = T' \,|\, s = v\right). \tag{1}$$

It is widely adopted in theory as a probabilistic rumor source estimator; however, the computation is costly even for the single-source problem. It is described in several references, including [12], [23].

(ii) The Jordan center estimator:

$$e_J(T', v) = \left(\max_{v' \in T'} d(v, v')\right)^{-1}. \tag{2}$$

The set of nodes $E_J(T') = \arg\max_{v \in T'} e_J(T', v)$ are called the Jordan centers of T'. This is a combinatorial estimator that can be found with low computational complexity. For motivation and relation with e_{MLE}, see [30], [39]. Note here that in order to state the problem as a maximization problem, we take the reciprocal of the distance function. This is slightly different from the notion of the Jordan center estimator used in other works, though equivalent.

(iii) The rumor centrality estimator:

$$e'_{RC}(T', v) = \frac{|T'|!}{\prod_{u \in T'} |T'^v_u|}, \tag{3}$$

where T'^v_u is the subtree of T' rooted at u *pointing away* from v. Recall that this means that T'^v_u consists of nodes v' such that $u \in [v, v']$. Assuming homogeneous spreading in the SI model, the quantity $e'_{RC}(T', v)$ gives the number of different paths leading to T' given v as the source. It is introduced in [12] (where it is denoted by $R(T', v)$) and discussed in greater details in subsequent papers, such as [13].

As the factors $|T'|!$ and $|T_v^v| = |T'|$ are both independent of the vertex v, we may omit them from the formula. We therefore have the following equivalent estimator

$$e_{RC}(T', v) = \frac{1}{\prod_{u \in T' \setminus v} |T_u^v|}. \tag{4}$$

The estimator e_{RC} is a combinatorial approximation of e_{MLE}, for in the case of a regular tree, the source detection functions E_{MLE} and E_{RC} associated with e_{MLE} and e_{RC} satisfy: $E_{MLE} = E_{RC}$ (cf. [12]).

(iv) The weight and distance centrality:

$$e_W(T', v) = (\max_{u:\text{child of } v} |T_u^v|)^{-1}; \tag{5}$$

$$e_{DC}(T', v) = (\sum_{u \in T'} d(u, v))^{-1}. \tag{6}$$

Detailed discussion of these estimators can be found in [12] and [40]. Our presentation is slightly different from their original forms in view of Definition 3(c), which always casts the detection problem as a maximizing problem. The associated source detection function $E_W(T')$ is called the *centroid* of T'; while $E_{DC}(T')$ is called the *distance center* of T'. These two estimators are closely related to the rumor centrality estimator. To be more precise, in the case of general tree, they are essentially the same as it is proved in [12] and [40] (Theorem 2) that

$$E_{RC} = E_{DC} = E_W.$$

Although we do not discuss the centroid and the distance center in detail, we note that both fit well in the general framework of the paper. The discussion can be modified from that of the Jordan center and the rumor centrality estimators.

Now we introduce the abstract quasi-regularity condition as follows.

Definition 4. Given an estimator (E, e), an infinite tree T is called quasi-regular with respect to (w.r.t.) (E, e) if the following conditions hold:

(a) $v \in E(D_T(v, r))$, for all $v \in V$ and $r \in \mathbb{Z}^+$.

(b) Let $v \in V$ be a vertex. For every pair of vertices u and u' such that $d(u, v) = r$ and $d(u', v) = r + 1$, let $T' = \text{conv}(\{D_T(v, r) \backslash u\} \cup \{u'\})$. Then

$$e(D_T(v, r), v) > e(T', v).$$

Intuitively, in the rumor spreading problem, condition (a) in Definition 4 says that if the infection graph is a disc, then the center of the disc is identified by the estimator (E, e) as the rumor source. Condition (b) says that, roughly speaking, the estimator assumes the rumor spreads homogeneously at the same rate in all directions. These two properties qualitatively justify concepts we will introduce in subsequent sections. Let us provide some further insight into Definition 4.

Lemma 1. *The following statements are true for a tree T w.r.t. the rumor centrality estimator (E_{RC}, e_{RC}) described in Example 1 above.*

(i) Suppose in any disc $D_T(v, r) \subset T$, for any full subtrees T_1 and T_2 within the disc and having heights r_1 and r_2, respectively, with $r_1 < r_2$, one has

$$|T_1| \leq |T_2|.$$

Then condition (a) of Definition 4 holds true.

(ii) Let

$$\alpha_l = \inf \left\{ \frac{(|T_1| + 1)(|T_2| - 1)}{|T_1||T_2|} \;\middle|\; T_1, T_2 \text{ are full subtrees of height } l + 1 \right\}.$$

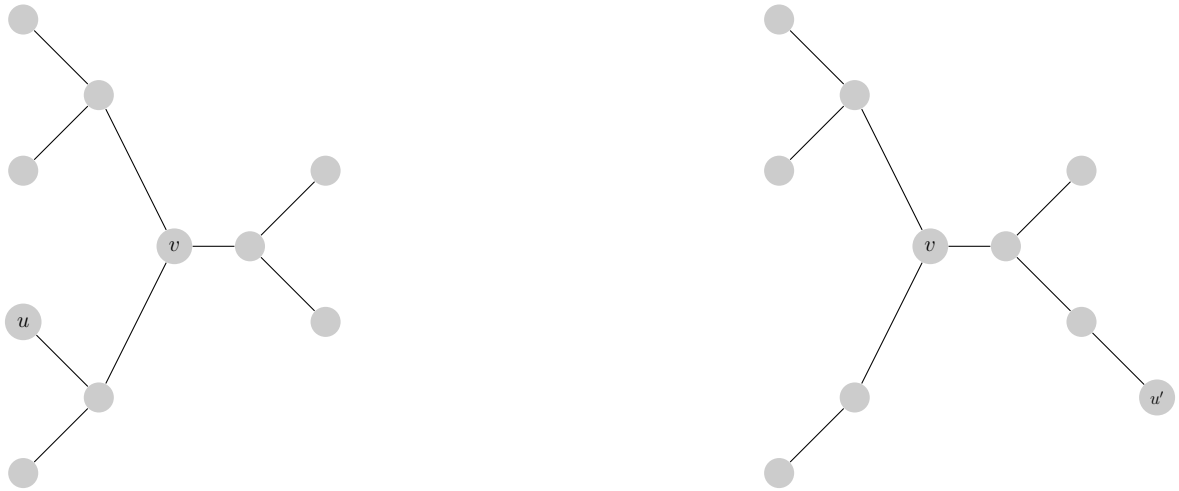

Fig. 2. An illustration of condition (b) of Definition 4. The condition requires that the estimation function e assigns a higher value to the subtree on the left.

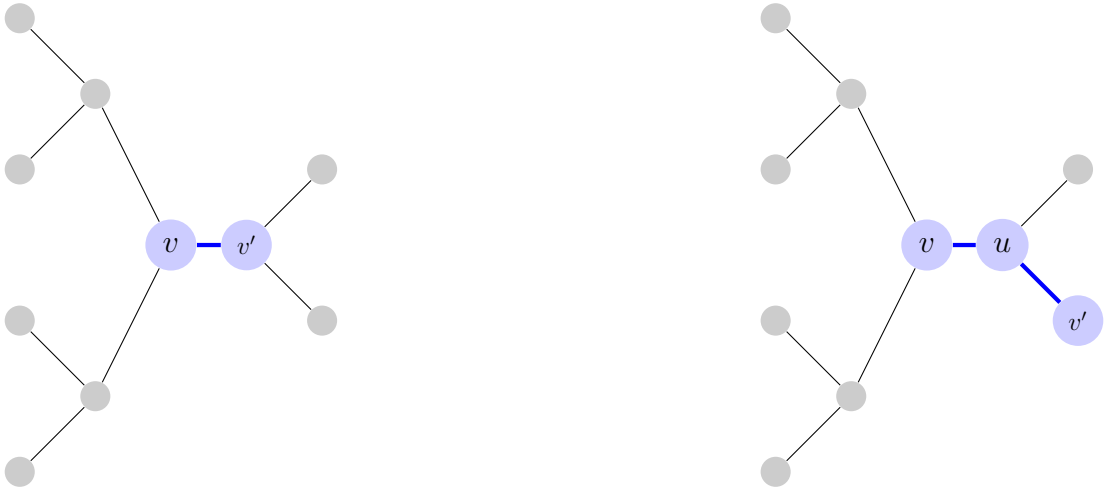

Fig. 3. Examples of trees in the proof of Lemma 1 (i). From (4), to prove (8), it suffices to consider only nodes $u \in [v, v']$. In the tree on the left, the reflection of v w.r.t. the midpoint of $[v, v']$ is v', while in the tree on the right, the reflection of u is itself.

If

$$\lim_{n\to\infty} \prod_{1\le l\le n} \alpha_l \ge 1/2, \tag{7}$$

then condition (b) of Definition 4 holds true.

In particular, T is quasi-regular w.r.t. the (E_{RC}, e_{RC})-estimator if the conditions of both (i) and (ii) are satisfied.

Proof:

(i) It suffices to prove that $e_{RC}(D_T(v,r), v) \ge e_{RC}(D_T(v,r), v')$ for all $v' \in D_T(v,r)$; or equivalently, following [12] (cf. Fig. 3),

$$\prod_{u\in[v,v']} |T_u^{v'}| \ge \prod_{u\in[v,v']} |T_u^{v}|. \tag{8}$$

For each vertex $u \in [v, v']$, let u' be the reflection of u w.r.t. the midpoint of $[v, v']$. We claim that $|T_{u'}^{v'}| \ge |T_u^v|$ for each $u \in [v, v']$. By definition, $|T_{u'}^{v'}|$ is the number of vertices of a full subtree of

height strictly larger than the height of the subtree corresponding to T_u^v. By the given condition in (i), $|T_{u'}^{v'}| \geq |T_u^v|$.

(ii) Taking $T_1 = T_2$, we see that $\alpha_l < 1$ for all l. The sequence of products $\prod_{1\leq l\leq n} \alpha_l$ is thus decreasing in l, and converges since it is lower-bounded by zero. Under the condition (7) we have $\prod_{1\leq l\leq n} \alpha_l > 1/2$ for each $l \geq 1$.

For a vertex v, let $S = D_T(v, r)$ and $S' = \mathrm{conv}(\{S\backslash u\} \cup \{u'\})$, with u, u' as given in Definition 4. Let $u'' \in S$ be the parent vertex of u' w.r.t. v (*i.e.*, the neighboring node of u' that is closer to v than u'). As u is removed, $u \neq u''$. We obtain

$$\begin{aligned}
\frac{e_{RC}(S, v)}{e_{RC}(S', v)} &= \frac{\prod_{y\in(v,u'']}(|T_y^v| + 1)\prod_{z\in(v,u]}(|T_z^v| - 1)}{\prod_{y\in(v,u'']}|T_y^v|\prod_{z\in(v,u]}|T_z^v|} \\
&= \frac{2\prod_{y\in(v,u'')}(|T_y^v| + 1)\prod_{z\in(v,u)}(|T_z^v| - 1)}{\prod_{y\in(v,u'')}|T_y^v|\prod_{z\in(v,u)}|T_z^v|} \\
&= 2\prod_{y\in(v,u''),z\in(v,u),d(v,y)=d(v,z)} \frac{(|T_y^v| + 1)(|T_z^v| - 1)}{|T_y^v||T_z^v|} \\
&\geq 2\prod_{1\leq l\leq r-1} \alpha_l \\
&> 1.
\end{aligned}$$

■

In Lemma 1, we have provided technical conditions under which a tree T is quasi-regular w.r.t. the rumor centrality estimator. It is easy to see that a regular tree is quasi-regular by our definition (see Example 2 (ii) below). There are however non-regular trees that can be quasi-regular.

In the following, we give specific examples of quasi-regular trees for the Jordan center estimator and the MLE.

Example 2. (i) For the Jordan center estimator (E_J, e_J), it is not hard to see from the definition that any tree is quasi-regular.

(ii) In the case of regular trees, $E_{MLE} = E_{RC}$ (cf. [12]). Condition (i) of Lemma 1 clearly holds for a regular tree T. Let us verify condition (ii) in the lemma for T. As any two full subtrees have the same size, we have

$$\alpha_l \geq \frac{(l+2)l}{(l+1)(l+1)} = \frac{l}{l+1}\frac{l+2}{l+1}.$$

It is easy to verify that $\lim_{n\to\infty}\prod_{1\leq l\leq n}\alpha_l \geq 1/2$. In conclusion, any regular tree is quasi-regular w.r.t. both the MLE and the rumor centrality estimators.

In the rest of the paper, we shall be concerned mainly with tree networks that are quasi-regular w.r.t. the (E_{RC}, e_{RC})-estimator or the (E_J, e_J)-estimator. For convenience, we call such trees *quasi-regular* without specifying the estimator unless otherwise mentioned.

III. Heavy Centers and the Two-Source Joint Rumor Centrality Estimator

In this section, we first define the concept of a heavy center, and use it to develop a two-source joint estimation framework. Our development in this section is based on a quasi-regular tree. Our proposed algorithm is generalized to more than two rumor sources and extended heuristically general graphs in Section V. In the following discussion, we let T be a quasi-regular tree w.r.t. an abstract estimator (E, e) and $I \subset T$ be the observed infected subtree.

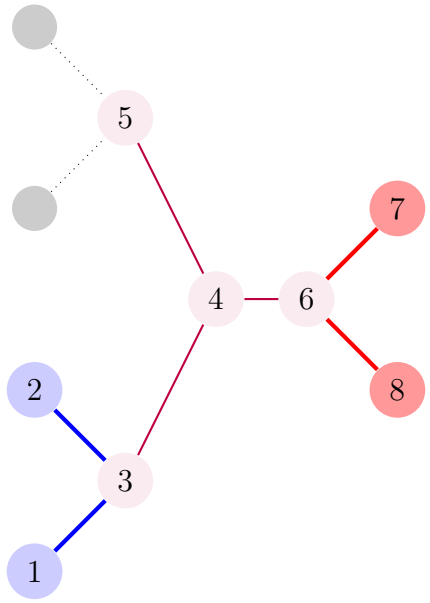

Fig. 4. An example of Definition 5. The tree T' consists of node 1 to node 8. Let $u = 3$ and $v = 6$ be two nodes. We note that $T_1 = \mathrm{conv}(\{1,2,3,4,5,6\}) = D_T(3,2) \subset T$, while $D_T(3,3)$ contains (gray) nodes outside T'. By definition, the heavy center at $u = 3$ is the subtree T_1. Similarly, the heavy center at $v = 6$ is $T_2 = \mathrm{conv}(\{3,4,5,6,7,8\})$. As $T' = T_1 \cup T_2$, we see that T_1 and T_2 form a 2-covering of T'. Moreover, this is not a partition as $T_1 \cap T_2 = \mathrm{conv}(\{3,4,5,6\})$ (the pink nodes).

A. *The General Setup*

We need the following definitions.

Definition 5. Let $T' \subset T$ be a subtree.

(a) A heavy center $h(v,r)$ of T' is a disc $D_T(v,r) \subset T'$ such that $D_T(v,r+1) \nsubseteq T'$. The latter condition is called the *maximality* condition. We may denote $h(v,r)$ by h or h_v for convenience if no confusion arises. It is important to note that the discs $D_T(v,r)$ and $D_T(v,r+1)$ are taken in the ambient tree T in order for the concept to be of any use.

(b) We say that the subtrees $T_1, T_2, \ldots, T_k$ form a *k-covering* of T' if $T' = \bigcup_{i=1}^{k} T_i$. In certain situations, if k is obvious from the context, we may just say "covering" for convenience.

Definition 6. Given subtrees $T_1 \subset T_2 \subset T$, define the contraction T_2/T_1 as the tree obtained from T_2 by replacing T_1 by a single point.[3] The contracted vertex is also denoted by T_1 if no confusion arises.

Definition 7. A *two-source joint estimator* (w.r.t. (E,e)) is a pair (E^2,e^2) such that there exists a function f (which depends on (E,e)) with the following properties:

(a) For each covering $\{T_1,T_2\}$ of I with heavy centers $h_1 \subset T_1$ and $h_2 \subset T_2$,

$$e^2(T_1,h_1;T_2,h_2) = f(e(T_1/h_1,h_1), e(T_2/h_2,h_2));$$

(b) $E^2(I) = \arg \max_{h_i \subset T_i, i=1,2} e^2(T_1,h_1;T_2,h_2)$, where the maximization is over all possible 2-coverings $\{T_1,T_2\}$ of I; and

(c) The function f satisfies the following property: if $x \geq x' \geq 0$ and $y \geq y' \geq 0$, then $f(x,y) \geq f(x',y')$.

In the following two subsections, we detail the two-source joint estimators w.r.t. (E_J,e_J) and (E_{RC},e_{RC}) respectively. In particular, we indicate in each case the function f being used.

[3]The term *contraction* and the corresponding notation are borrowed from topology; what we introduce here is nothing but the quotient space.

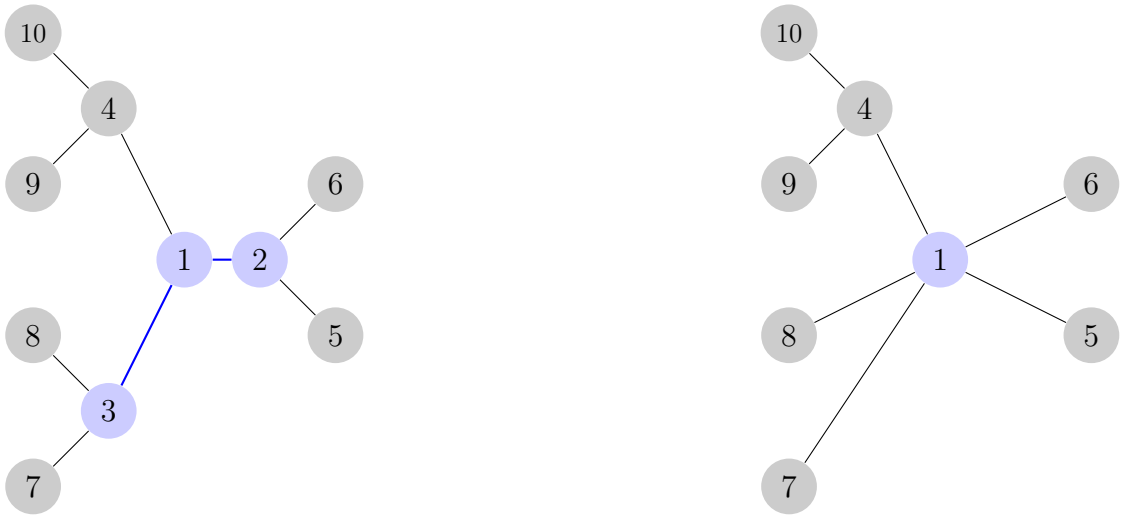

Fig. 5. An example of Definition 6. T_1 is the (full) tree on the left consisting of node 1 to node 10 and T_2 is the (blue) subtree consisting of node 1 to node 3. The contraction procedure shrinks T_2 to a single vertex, also labeled 1. The remaining vertices and connections are not changed. The resulting tree T_1/T_2 is shown on the right.

B. The Joint Jordan Center Estimator

Consider a subtree $I \subset T$ and suppose it has been infected by a single rumor source. The Jordan center estimator for the single rumor source identification problem [14], [15] finds the set of Jordan centers of I, *i.e.*, nodes v such that the maximum distance of v to all vertices of I is minimized. Let $\mathcal{J}(I)$ denote the set of Jordan centers. We now define the *heavy Jordan centers* $h^{\mathcal{J}}(I)$ as the collection of nodes $v \in I$ such that the maximal distance from all the vertices of I to h_v, the heavy center of v in I, is minimized. See Fig. 6 for an example. The following lemma gives a relationship between $\mathcal{J}(I)$ and $h^{\mathcal{J}}(I)$.

Lemma 2. *For any subtree $I \subset T$, the intersection $\mathcal{J}(I) \cap h^{\mathcal{J}}(I)$ is non-empty.*

Proof: Let u_0 be a vertex of $h^{\mathcal{J}}(I)$. If u_0 is already in $\mathcal{J}(I)$, then there is nothing to prove. Otherwise, we can always choose $v \in I$ with the following properties:

(i) $v \in \mathcal{J}(I)$;
(ii) each $u \in [u_0, v)$ is not in $\mathcal{J}(I)$.

Let $u' \in [u_0, v)$ be the neighbor of v on the path between v and u_0. Suppose that for all $w \in T_v^{u_0}$, $d(w, v) < \max_{v' \in I} d(v', v)$. Then, we have $d(w, u') \leq \max_{v' \in I} d(v', v)$ for all $w \in I$, which implies that $u \in \mathcal{J}(I)$, a contradiction to property (ii) of v. Therefore, there exists a node $w \in T_v^{u_0}$ such that $d(w, v) = \max_{v' \in I} d(v', v)$. Let $h_v = D_T(v, r)$ be the heavy center of v in I. Every vertex $v' \in h_v \cap T_v^{u_0}$ has equal distance $d(v', u_0) = r + d(v, u_0)$ to u_0. On the other hand, for each vertex $v' \in h_v \cap T_{u'}^{v}$, we have $d(v', u_0) < r$. Therefore, $h_v \subset h_{u_0}$ the heavy center of u_0 in I. Moreover, h_v shares the same boundary with h_{u_0} in $T_v^{u_0}$. Therefore, the maximal distance of points of I from h_{u_0} and h_v agree. Hence, $v \in h^{\mathcal{J}}(I)$ and lies in $\mathcal{J}(I) \cap h^{\mathcal{J}}(I)$. ∎

A simple consequence of this is the following: for each vertex u which is a Jordan center, there is always a vertex $v \in h^{\mathcal{J}}(I)$ such that $d(u, v) \leq 1$. To see this, let $v = u_0 \in \mathcal{J}(I) \cap h^{\mathcal{J}}(I)$. As the distance between two vertices of $\mathcal{J}(I)$ is at most 1, so is the distance between v and u. This observation allows us to find canonical members of the set $h^{\mathcal{J}}(I)$, which contains more than one element in general.

We now describe our joint Jordan center estimator. Intuitively speaking, suppose $u, v \in I$ are the rumor sources. Let h_u and h_v be the heavy centers with disc centers u and v respectively. We want the maximum distance from the remaining vertices of I to either h_u or h_v, whichever is nearer, as small as possible. This is exactly the idea of *infection range* introduced in [30, Definition 4].

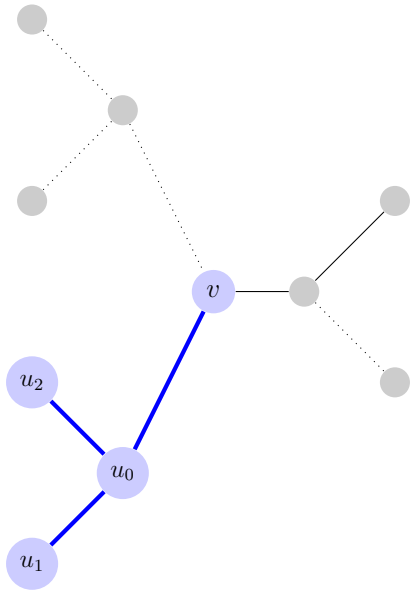

Fig. 6. A subtree I with 6 vertices is indicated by the solid line segments. The heavy centers of u_0, u_1, and u_2 in I are the same tree, which is indicated in blue. The vertex u_0 is in $h^{\mathcal{J}}(I)$ but not in $\mathcal{J}(I)$. The procedure described in the proof of Lemma 2 finds the node v in $h^{\mathcal{J}}(I) \cap \mathcal{J}(I)$.

Therefore we are looking for a covering T_1 and T_2 of I with heavy centers $h_1 \subset T_1$ and $h_2 \subset T_2$ such that $\min\{e_J(T_1/h_1, h_1), e_J(T_2/h_2, h_2)\}$ is maximized. Notice our definition of e_J given in Example 1 is slightly different from the one given in [30].

In the language of Section III-A, we choose the joint Jordan center estimator (E_J^2, e_J^2) with e_J^2 given by

$$e_J^2(T_1, h_1; T_2, h_2) = f(e_J(T_1/h_1, h_1), e_J(T_2/h_2, h_2))$$

where $f(x, y) = \min(x, y)$. It is clear f satisfies property (c) of Definition 7.

C. The Joint Rumor Centrality Estimator

To describe the joint rumor centrality estimator, we begin with a lemma.

Lemma 3. *Suppose we are given covering and heavy centers $h_i \subset T_i, i = 1, 2$. The number of different paths leading to T_1 and T_2 jointly, starting from h_1 and h_2 respectively, is given by the following formula*

$$e'_{RC}(T_1/h_1, h_1) e'_{RC}(T_2/h_2, h_2) \frac{(|T_1/h_1| + |T_2/h_2| - 2)!}{(|T_1/h_1| - 1)!(|T_1/h_1| - 1)!}.$$

Temporarily denote this quantity by $e^u_{RC}(T_1, h_1; T_2, h_2)$.

Proof: Given T_1, T_2, h_1, h_2, we form an auxiliary tree T' with root v that branches in two directions: T_1/h_1 with h_1 identified with v and T_2/h_2 with h_2 identified with v. This is the same as gluing the two contractions together at a single vertex.

Now it is easy to see that the quantity we wish to calculate is the same as $e'_{RC}(T', v)$ and it suffices to compute this estimator.

The most direct way is to apply the definition:

$$\begin{aligned} e'_{RC}(T', v) &= \frac{(|T'| - 1)!}{\prod_{y \neq v \in T'} |T'_y|} = (|T'| - 1)! \frac{e'_{RC}(T_1/h_1, h_1)}{(|T_1/h_1| - 1)!} \frac{e'_{RC}(T_2/h_2, h_2)}{(|T_2/h_2| - 1)!} \\ &= e'_{RC}(T_1/h_1, h_1) e'_{RC}(T_2/h_2, h_2) \frac{(|T_1/h_1| + |T_2/h_2| - 2)!}{(|T_1/h_1| - 1)!(|T_1/h_1| - 1)!}. \end{aligned}$$

One recognizes the last fraction in the expression is just a binomial term and so the result can also be obtained from a standard combinatorial argument. ■

The quantity $e^u_{RC}(T_1, h_1; T_2, h_2)$ may not, however, be a good estimator, as the sizes of the contractions $T_i/h_i, i = 1, 2$ depend on the choices of $h_i, i = 1, 2$ even if we fix $T_i, i = 1, 2$. In order to make a fair comparison, we need to normalize $e^u_{RC}(T_1, h_1; T_2, h_2)$.

The quasi-regular condition predicts that the spreading tends to fill the infection region *layer-by-layer*. Therefore, we expect that given $h_i = D_T(v_i, r_i) \subset T_i, i = 1, 2$, the vertices $T_i \backslash h_i$ are distributed at the distance $r_i + 1$ away from v_i. If this is the case, a calculation similar to the lemma suggests to choose

$$e^n_{RC}(T_1, h_1; T_2, h_2) = \frac{|T_1/h_1| + |T_2/h_2| - 2)!}{((|T_1/h_1| - 1)!)^2((|T_2/h_2| - 1)!)^2}$$

as the normalization factor. To measure the deviation from the putative optimal, we take the quotient

$$\frac{e^u_{RC}(T_1, h_1; T_2, h_2)}{e^n_{RC}(T_1, h_1; T_2, h_2)};$$

and this yields the two-source joint estimator.

Definition 8. For a 2-covering of I with two heavy centers $h_i \subset T_i, i = 1, 2$, define the joint rumor centrality estimator

$$e^2_{RC}(T_1, h_1; T_2, h_2) = \frac{e'_{RC}(T_1/h_1, h_1)}{(|T_1/h_1| - 1)!} \frac{e'_{RC}(T_2/h_2, h_2)}{(|T_2/h_2| - 1)!}.$$

Define the joint source detection function

$$E^2_{RC}(I) = \underset{h_i \subset T_i, i=1,2}{\arg\max}\ e^2_{RC}(T_1, h_1; T_2, h_2).$$

By the definition of the single-source estimator e_{RC} and the construction of the contractions, we also have the following formula;

$$e^2_{RC}(T_1, h_1; T_2, h_2) = \prod_{u \in T_1 \backslash h_1} \frac{1}{|T^{v_1}_u|} \prod_{v \in T_2 \backslash h_2} \frac{1}{|T^{v_2}_v|} = e_{RC}(T_1/h_1, h_1) e_{RC}(T_2/h_2, h_2). \tag{9}$$

From this formula, we see qualitatively that we essentially have to satisfy two things: (a) minimize the number of vertices not contained in the union of the two heavy centers; (b) these vertices should distribute evenly around the boundaries of the heavy centers. These agree with the goal of our task.

We notice that the estimator e^2_{RC} decays quickly with increase in the size of I. Therefore in actual implementation using this estimator, one may use $\log(e^2_{RC})$ instead.

In the language of Section III-A, we choose the joint rumor centrality estimator (E^2_{RC}, e^2_{RC}) with e^2_{RC} given by

$$e^2_{RC}(T_1, h_1; T_2, h_2) = f(e_{RC}(T_1/h_1, h_1), e_{RC}(T_2/h_2, h_2))$$

where $f(x, y)$ is simply xy (recall e_{RC} in Example 1 (iii) is equivalent to e'_{RC}). It is clear that f satisfies property (c) of Definition 7.

IV. Joint Source Detection (JSD) Algorithm

In this section, we give a two-step heuristic algorithm in the two-source identification problem (for a given abstract estimator), called the *joint source detection (JSD) algorithm*. We give an overview of the proposed algorithm, and defer specific implementation details to Section V as the implementation depends on the choice of abstract estimator. We then prove that the algorithm converges, under a mild condition on the abstract estimator. In the following discussion, let T be a fixed quasi-regular tree, I the observed infected subtree and (E^2, e^2) a joint two-source estimator associated with the single source

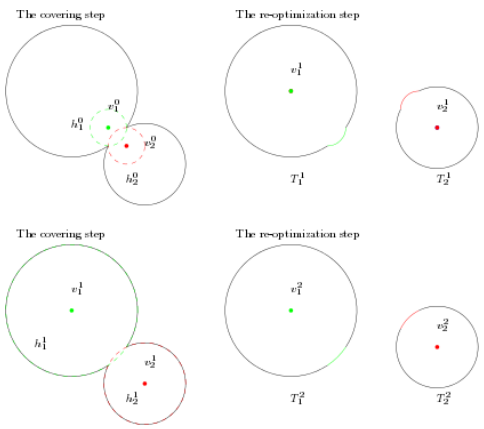

Fig. 7. A schematic illustration of the JSD algorithm. Suppose in the l-th iteration, we obtain v_1^l and v_2^l (top-left). We fist find the associated heavy centers h_1^l and h_2^l (top-right). We perform the covering step and obtain two (possibly overlapping) regions T_1^{l+1} bounded by the green curve and T_2^{l+1} bounded by the red curve (bottom-left). Within each region, we identify the new (estimated) source, and this completes the re-optimization step and hence one iteration (bottom-right).

estimator (E, e) (recall the definition given in Section III-A). A simple illustration of the scheme is given in Figure 7.

Initialize with a covering $\{T_1^0, T_2^0\}$ of I with heavy centers $h_i^0 \subset T_i^0$, $i = 1, 2$. Let v_i^0 be the disc center of h_i^0, for $i = 1, 2$. We perform the following two steps at each iteration $l \geq 0$.

Step 1: the covering step.

From the previous iteration, we have $h_i^l \subset T_i^l, i = 1, 2$. Enlarge h_i^l with center v_i^l, if necessary, so that they are both heavy centers of I. Reassign the vertices of $I \backslash \{h_1^l \cup h_2^l\}$ to h_1^l, h_2^l to get a covering $\{T_1^{l+1}, T_2^{l+1}\}$ so that $e^2(T_1^{l+1}, h_1^l; T_2^{l+1}, h_2^l)$ is maximized.

We remark here that h_1^l and h_2^l are not always heavy centers of I as the maximality condition may not hold. Therefore we first make them heavy centers. We shall justify this in Theorem 1 below.

Step 2: the re-optimization step.

In each T_i^{l+1} obtained from Step 1, find

$$h_i^{l+1} = \underset{h_i' \text{: heavy center of } T_i^{l+1}}{\arg\max} e(T_i^{l+1}/h_i', h_i').$$

Set v_i^{l+1} to be the disc center of h_i^{l+1}.

The JSD algorithm terminates if $\max\{d(v_i^l, v_i^{l+1}), i = 1, 2\} \leq \eta$ for some predetermined positive value η or a fixed number of iterations are completed.

Theorem 1. The JSD algorithm w.r.t. the abstract estimator (E, e) converges to a local optimum of e^2 if the following holds true: for any $T_1 \subset T_2 \subset T_3$, if T_1 contracts to a node labeled as T_1 in T_3/T_1 and T_2 contracts to a node labeled as T_2 in T_3/T_2, then $e(T_3/T_1, T_1) \leq e(T_3/T_2, T_2)$ (see Figure 8).

Proof: It suffices to show that in both Step 1 and Step 2 of the JSD algorithm, the joint estimator e^2 is improved.

Consider Step 1. Let $h_i^{l'}$ be the heavy center of I having v_i^l (the disc center of h_i^l) as the disc center. Form a covering of I as $T_i^{l'} = h_i^{l'} \cup T_i^l$, which is a tree as it is the union of two trees with nontrivial intersection. Moreover, $T_i^{l'}/h_i^{l'}$ is the contraction of $h_i^{l'} \cap T_i^l$ in T_i^l. Clearly, we have the following inclusions: $h_i^l \subset h_i^{l'} \cap T_i^l \subset T_i^l$. Consequently, the following inequality holds due to the condition of the theorem and Property (c) of Definition 7:

$$\begin{aligned} e^2(T_1^l, h_1^l; T_2^l, h_2^l) &= f(e(T_1^l/h_1^l, h_1^l), e(T_2^l/h_2^l, h_2^l)) \\ &\leq f(e(T_1^{l'}/h_1^{l'}, h_1^{l'}), e(T_2^{l'}/h_2^{l'}, h_2^{l'}) = e^2(T_1^{l'}, h_1^{l'}; T_2^{l'}, h_2^{l'}). \end{aligned}$$

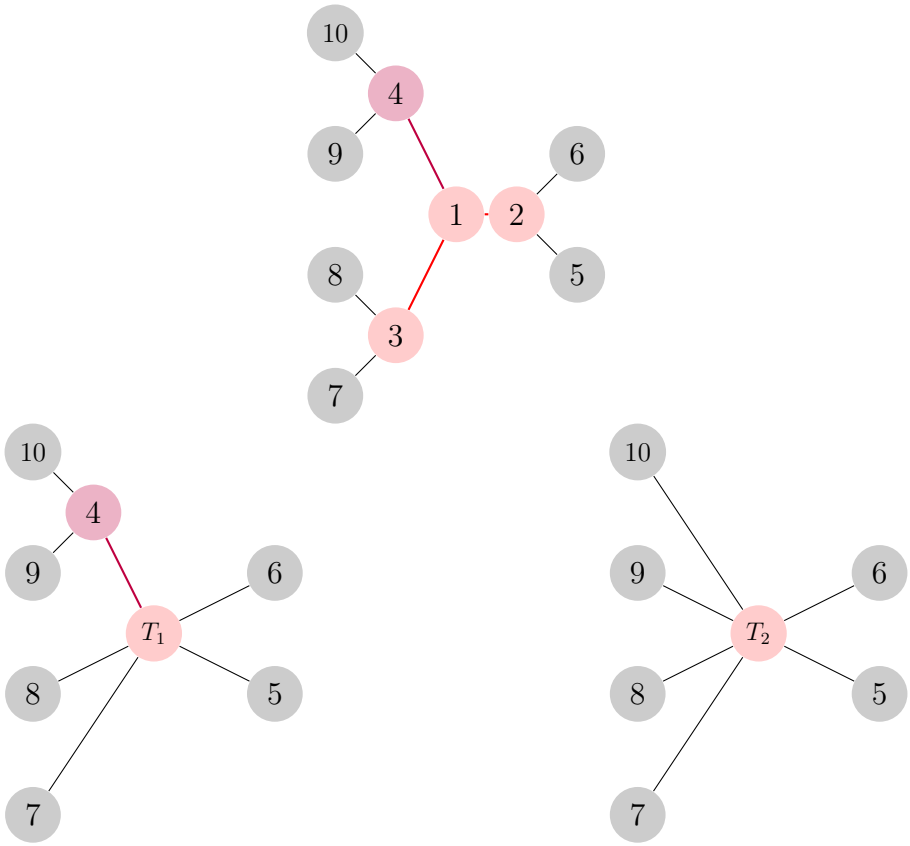

Fig. 8. An example of the condition in Theorem 1. T_3 contains node 1 to node 10, T_2 contains node 1 to node 4; and T_1 contains node 1 to node 3 (top). T_3/T_1 is shown on bottom left and T_3/T_2 is shown on bottom right. One observes $e_J(T_3/T_2, 1) = 1 \geq e_J(T_3/T_1, 1) = 1/2$ and $e_{RC}(T_3/T_2, 1) = 1 \geq e_{RC}(T_3/T_1, 1) = 1/3$.

The value on the right-side of the inequality is further improved in Step 1; and the claim follows.

For Step 2, we use Property (c) in Definition 7 again. As f is non-decreasing in the two variables, therefore to maximize e', it is enough to maximize e in each component of the covering. Therefore, e' is improved in Step 2. The proof is now complete. ■

Notice that in Step 1 of the JSD algorithm, we essentially fix $h_i, i = 1, 2$ and improve e^2 by changing the covering, whereas in Step 2 we fix $T_i, i = 1, 2$ and improve e^2 by changing the heavy centers. This is just a discrete version of gradient descent. It can be easily verified that both e_J and e_{RC} satisfy the condition given in Theorem 1 above. Therefore, applying the JSD algorithm for each of these estimators yields a local optimal estimate of the two sources, w.r.t. the respective estimator function e^2.

V. Implementation and Extensions

In this section, we describe in detail the implementation of the JSD algorithm, in which (E^2, e^2) is chosen to be either the joint Jordan center estimator (E^2_J, e^2_J) or the joint rumor centrality estimator (E^2_{RC}, e^2_{RC}). Let T be a quasi-regular tree and $I \subset T$ be the observed infected subtree.

Suppose the degree of each vertex of T is at least 3. Then an easy calculation gives the following bound on the size of a disc $T' = D_T(x, r)$ with radius r: $|T'| \geq 1 + 3(2^{r-1} - 1)$, which implies that its height $\ell(T') = 2r = O(\ln |T'|)$. If we assume that I is made up of two discs, then the size of $\ell(I)$ is of order $O(\ln |I|)$. This simple observation is used in computing the complexity of the covering step.

A. The Joint Jordan Center Estimator (E^2_J, e^2_J)

The covering step. According to the last section, we first enlarge h^l_i to heavy centers of I. We check

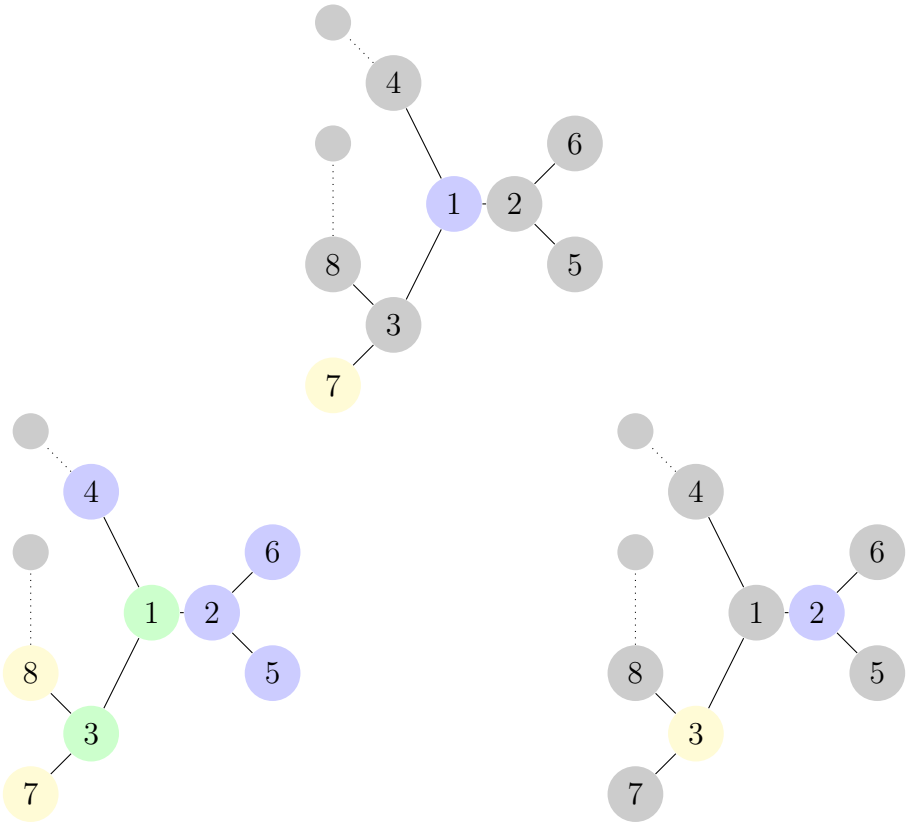

Fig. 9. An illustration of the JSD algorithm with the (E_J^2, e_J^2)-estimator. Suppose I contains node 1 to node 8 connected by solid line segments and $v_1^0 = 1, v_2^0 = 7$ (top). In the covering step (bottom left), we first identify $h_1^1 = \text{conv}(\{1, 3, 7, 8\})$ and $h_2^1 = \text{conv}(\{1, 2, 3, 4\})$. The remaining nodes 5 and 6 are assign to hs_2^1, forming $T_2^1 = \text{conv}(\{1, 2, 3, 4, 5, 6\})$. On the other hand, $T_1^1 = h_1^1$. In the re-optimization step (bottom right), in T_1^1 and T_2^1, node 3 and node 2 are heavy Jordan centers respectively. Therefore $v_1^1 = 2$ and $v_2^1 = 3$. It is easy to check that no more changes occur in the next iteration.

whether all the children of ∂h_i^l are in I or not. If not, stop the procedure; otherwise enlarge the radius of h_i^l by 1 and repeat this procedure. Clearly, the complexity is $O(|I|)$.

If we follow the covering step described in the previous section strictly, we need to assign the nodes between h_1^l and h_2^l to the two heavy centers so that e_J^2 is maximized. To simplify the algorithm heuristically, we are content in assigning the nodes to the nearest heavy center, either h_1^l or h_2^l. We achieve this by broadcasting a message from each heavy center, and assigning each of the remaining vertices of $I \backslash (h_1 \cup h_2)$ to a nearest heavy center. The complexity is again $O(|I|)$. Notice that the main difference of this algorithm with the one given in [30] is that when the two heavy centers are of unequal size, the one with larger radius is given significantly more vertices, instead of a symmetric partition.

The re-optimization step. In each T_i^l, according to the theory, we should find a heavy Jordan center $v \in h^{\mathcal{J}}(T_i^l)$. By Lemma 2 and the remarks below the lemma, there is a member of $h^{\mathcal{J}}(T_i^l)$ that is also a Jordan center and stays at most distance 1 away from any other Jordan center. Therefore in this step, we find a Jordan center instead. We can apply the algorithm described in [30], which we briefly recall.

Let each vertex in T_i^l broadcast a message containing its own identity. The first vertex that receives the message from each vertex is the Jordan center v_i^{l+1} selected by the algorithm. The complexity of going through the algorithm for both T_1^l and T_2^l once is $O(|I|)$.

Finally, h_i^{l+1} is obtained by expanding v_i^{l+1} to a heavy center in T_i^l. The complexity is again $O(|I|)$, and so is the complexity of the entire algorithm. A simple example is given in Figure 9.

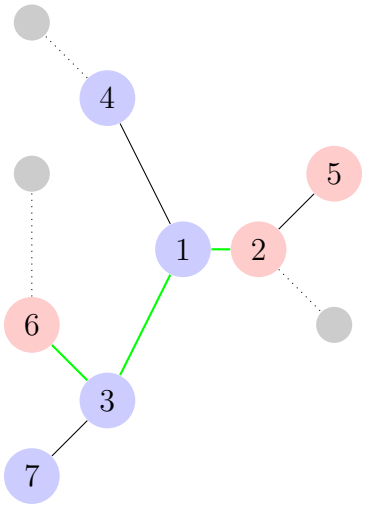

Fig. 10. An illustration of Case 2 of the covering step when using the (E^2_{RC}, e^2_{RC})-estimator. Suppose at some iteration l, we have $h^l_1 = \{6\}$ and $h^l_2 = \text{conv}(\{2, 5\})$ (the red nodes). In this case, $P = (6, 2)$ (the green path together with node 3 and node 1). Therefore $T_P = \text{conv}(\{1, 3, 4, 7\})$ (the blue nodes). A simple calculation shows that node 3 and node 7 should be assigned to h^l_1, giving $T^{l+1}_1 = \text{conv}(\{3, 6, 7\})$. Node 1 and node 4 should be assigned to h^l_2, giving $T^{l+1}_2 = \text{conv}(\{1, 2, 4, 5\})$.

B. The Joint Rumor Centrality Estimator (E^2_{RC}, e^2_{RC})

One can always first perform the message-passing algorithm of [12] to obtain $|T^v_u|$ for each pair $u, v \in I$ with complexity $O(|I|)$. Therefore we assume this step has been performed.

The covering step. Suppose $h^l_i, i = 1, 2$ are both heavy centers of I. Let P be the unique open path connecting h^l_1 and h^l_2. By *open* we mean the intersection of the path with h^l_1 and h^l_2 removed. There are two cases to consider.

Case 1: P is empty. In this case, each vertex $v \in I$ is connected to one of the two heavy centers, say h^l_i, by a direct path without passing through the other heavy center. Place v in T^{l+1}_i (i is the index of the heavy center determined as in the previous line). A message-passing algorithm as in the previous subsection will work. The complexity is $O(|I|)$.

Case 2: P is non-empty (see Figure 10 for a simple example). We first find P with complexity $O(|I|)$ (briefly, this can be done by expanding h^l_1 until it reaches h^l_2). Write T_P for the union of all the subtrees of I rooted at a vertex of P, pointing away (see Example 1(iii) for the meaning) from both h^l_1 and h^l_2.

For the vertices in $I \backslash \{h^l_1 \cup h^l_2 \cup T_P\}$, they are assigned to T^{l+1}_i the same as the steps in case 1, with $O(|I|)$ complexity.

For convenience, call the vertices of P as $v_0, v_1, \ldots, v_n$. For each vertex $v_j \in P, 0 \le j \le n$, we record the size $T^{h^l_1}_{v_i}$ and $T^{h^l_2}_{v_i}$. If $n = 2$, just set $j_0 = v_1$. Otherwise, for each $v_j, 1 \le j \le n-2$, compute the product

$$\epsilon_{v_j} = \prod_{1 \le k \le j} \frac{1}{|T^{h^l_1}_{v_k}| - |T^{h^l_1}_{v_{j+1}}|} \prod_{j+1 \le k \le n-1} \frac{1}{|T^{h^l_2}_{v_k}| - |T^{h^l_2}_{v_j}|}.$$

Let j_0 be the index j with the largest ϵ_{v_j} value. According to (9), all the vertices $v_j, 1 \le j \le j_0$ and their rooted subtrees are placed in T^{l+1}_1 and $T^{l+1}_2 = \{I \backslash T^{l+1}_1\} \cup h^l_2$ and the algorithm is complete. In this final stage, the complexity is $O(\ell(I)^2)$.

In summary, the overall complexity is $O(\max\{|I|, \ell(I)^2\})$. According to our remark at the beginning of this section, in many cases (for example, the degree of each vertex is at least 3), this is just $O(|I|)$.

The re-optimization step. For simplicity, we apply the most straightforward algorithm. For each vertex $v \in T^l_i$, we first let v grow to become a heavy center h_v. The complexity is $O(|I|)$.

Apply the (first half of) the message-passing algorithm of [12] to T^l_i / h_v to find $e_{RC}(T^l_i / h_v, h_v)$. The complexity is again $O(|I|)$. Therefore the overall complexity is $O(|I|^2)$ if we make a comparison over all vertices.

The overall complexity is therefore $O(|I|^2)$. The re-optimization step is the dominating step.

C. Heuristic Extension and Generalization

In this subsection, we show how our two-source JSD framework can be extended to general graphs and more than two sources. We also discuss some possible further generalizations.

1) Heuristic extension to general graphs: When we have a general graph, there are various possible ways to extend our method using the *breadth-first search (BFS) tree heuristic* (see [12]). We propose the following way.

For the covering step of the JSD algorithm, in each iteration l, given v_i^l, $i = 1, 2$ from the previous iteration, we first find the BFS trees rooted at these centers by starting expanding from them simultaneously. In doing so, we obtain two connected components $C_i^l, i = 1, 2$ with $v_i^l \in C_i^l$. Choose any edge $e^l \in G$ such that the two end nodes of e are in C_1^l and C_2^l respectively. We can perform the covering step on the union $C_1^l \cup C_2^l \cup e$, which is a connected tree.

For the re-optimization step, we perform the optimization on the BFS trees constructed in the covering step. For a general single-source estimator (E^2, e^2), we follow the idea of [12, Section 2.7]. To be more precise, suppose T_i^l is given and $v \in T_i^l$. We first find the BFS tree of T_i^l at v, denoted by $T_{i,v}^l$. After which, we can find the heavy center h_v of $T_{i,v}^l$; and compute $e(T_{i,v}^l/h_v, h_v)$. The vertex v with the largest $e(T_{i,v}^l/h_v, h_v)$ is used in the next iteration. The procedure can be modified and simplified on a case by case basis. For example, if we use the Jordan center estimator (E_J^2, e_J^2), we can follow [30, Section V] in view of Lemma 3 and the discussion thereafter.

2) Generalization to $k > 2$ sources: If there are k sources for a fixed $k > 2$, we replace the two-source estimator e^2 by a k-source estimator e^k. The only essential change one has to make is to replace the two-variable function f in Definition 7 by the corresponding k-variable function. For example, in the case of the (E_{RC}^2, e_{RC}^2)-estimator, we let f be the product function on k-variables, *i.e.*, $f((x_1, \ldots, x_k)) = x_1 x_2 \cdots x_k$. If we use the (E_J^2, e_J^2)-estimator, we can still let f be the $\max$ function, *i.e.*, $f((x_1, \ldots, x_k)) = \max(\{x_1, \ldots, x_k\})$.

Our algorithm is divide-and-conquer in nature and therefore can be generalized, *mutatis mutandis*, to the k-source situation. We only need to replace the 2-covering step by a k-covering step, which maximizes e^k instead of e^2. The re-optimization step remains the same since it is performed w.r.t. each cover. Theorem 1 generalized to e^k again holds.

3) Further generalizations: There is potential to generalize our method to an even broader scheme. First of all, the combinatorial estimator (E_J^2, e_J^2) or (E_{RC}^2, e_{RC}^2) can be replaced by a general statistical estimator, by the same procedure described in the paper, as long as all requirements for an abstract estimator are satisfied (cf. Definition 7 and Theorem 1). Such estimators may include those that utilize observations of timestamps or prior information about possible suspects for the sources (for example, [20] and the follow-up works [25] and [27]).

In this paper, our proposed heavy center h_v at a vertex $v \in I$ is the largest disc contained fully in I centered at v. However, one can arbitrarily define h_v to be a connected subgraph of I containing v satisfying a few given conditions. Once such a general notion of "heavy center" is given, we can apply the same procedure described in the paper (using covering and contraction) to develop an associated joint source detection algorithm. For example,[4] instead of requiring h_v to be the largest disc, one can require h_v to be the largest disc of I with at most $1 - \Gamma$ fraction of nodes not in I, where $\Gamma \in (0, 1]$. Such a generalization adds more flexibility and may yield better results in specific cases. On the other hand, it may also require case-by-case study of the specific spreading pattern on a given network. Simulation results on this generalization are provided in Section VI-D.

[4]This example is suggested to us by one of the reviewers.

VI. Simulation Results

In this section, we present simulation results to verify the performance of the JSD algorithm for the (E_J^2, e_J^2) and (E_{RC}^2, e_{RC}^2) estimators. We compare its performance with the Multiple Jordan Center (MJC) algorithm described in [30], and the Clustering and Localization/Clustering and Reverse Infection (CL-CRI) algorithm described in [34].

In our simulations, we adopt either a continuous-time or discrete-time spreading model. For the continuous-time model, we assume that the propagation time of the rumor across each edge in the graph is distributed independently according to an exponential distribution with unit rate. To simulate an infection subgraph, we randomly generate a permitted permutation using the exponential spreading model with unit rate (as defined in Section II.C of [12]) of a fixed size starting from different sources at different times. For the discrete-time model, we assume that at each discrete time slot, a susceptible node (i.e., a node that has at least one infected neighbor) has probability $1/2$ of being infected by each of its infected neighbors independently. Moreover, multiple nodes can become infected in the same time slot.

In general, the gradient descent style algorithm can be sensitive to the initialization. There are several initialization schemes one can choose. One can use the scheme developed in [34], which chooses the initial nodes as far away from each other as possible. For fairness, we shall use this scheme when we compare performance among different algorithms. To mention some other possible schemes, we can choose a node v near the center of the infected graph, and the second one between v and a node furthest away from v. Another approach is to perform a rough source estimation using the method in [38], which also estimates the number of sources. We terminate the algorithm once a fixed number of iterations are completed.

A. Comparison between the errors on different sources

In the first set of simulations, we apply our JSD algorithm on a 3-regular tree with 766 nodes (9 layers), synthetic scale-free graphs [41] with 750 nodes, and synthetic small world graphs [42] with 750 nodes. We run the algorithm for both the (E_J^2, e_J^2)-estimator and the (E_{RC}^2, e_{RC}^2)-estimator, which we label as JSD-J and JSD-RC, respectively.

In each of the random network ensembles, we generate a connected infected subgraph containing 20% of vertices of a full graph using the continuous-time model, and two of them are the sources. As we do not assume the two sources start the infection process at the same time, and we record this time difference as a reference (the horizontal axis in Figure 11).

Once the infected subgraph is generated, we record the positions of the two sources as s_1 and s_2. Applying the JSD algorithm described in Section V, we obtain the positions of the two estimated sources $\hat{s}_1$ and $\hat{s}_2$. The true sources s_1, s_2 are optimally paired with the estimated sources $\hat{s}_1, \hat{s}_2$ to get a minimum total estimation error. This procedure is repeated 100 times, and the average estimation error is then computed and plotted against the difference in infection start times of the two sources. In the graphs, we plot the average error of the two sources for both the JSD-J algorithm (green curves) and JSD-RC algorithm (black curves). For the algorithm that performs better, we also include the blue and red curves giving the errors of the two sources respectively.

Simulation results of a reasonable algorithm should have the following properties.

(a) The error/time difference graph should display an increasing trend. If the time difference is large, the second source has a very small impact on the final outcome; therefore it may be hard to distinguish the second source from a probabilistic fluctuation.

(b) The average estimation error of the first source should be smaller than that of the second source. The obvious reason is that the first source is more influential than the second source. Therefore the first source is easier to identify.

Our simulation results (Figure 11) display these features. Moreover, in all three cases, the errors of the algorithm are always within 20% of the average diameters of the observed infected graph I.

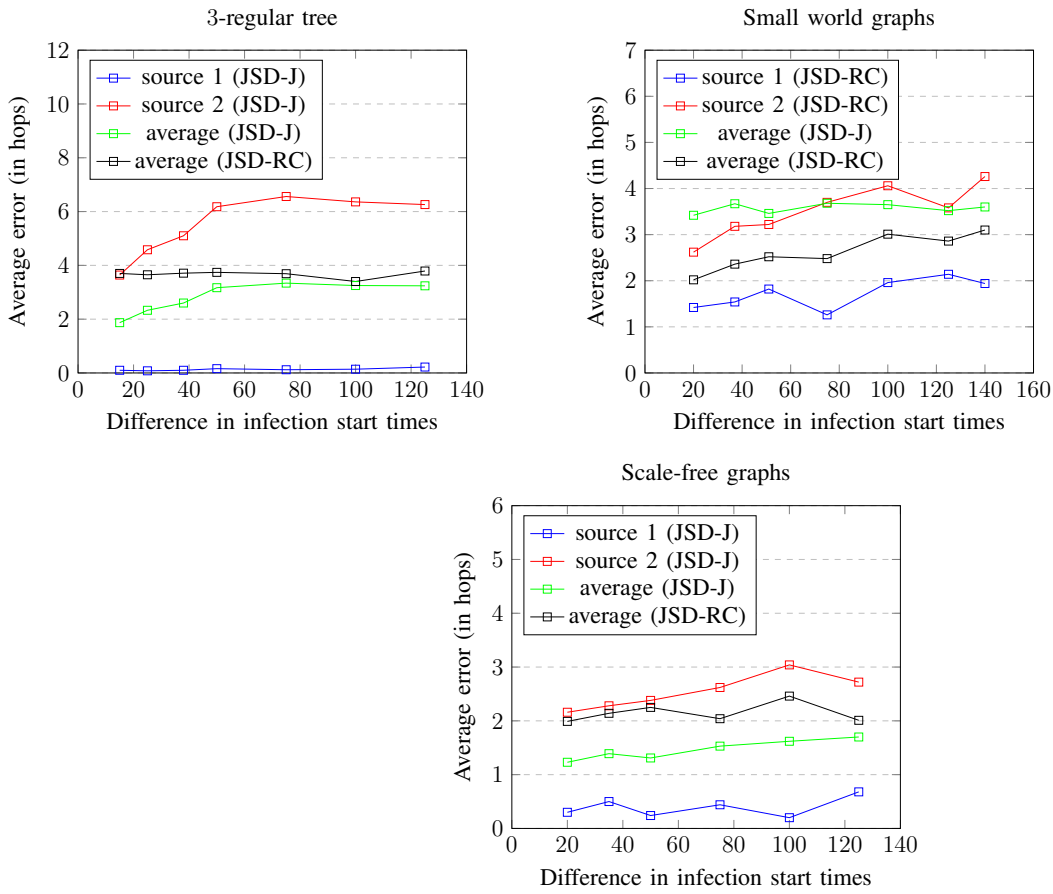

Fig. 11. Performance of the JSD algorithm for the (E_J^2, e_J^2) and (E_{RC}^2, e_{RC}^2) estimators.

The JSD-RC algorithm does not perform as well as the JSD-J algorithm for regular and scale-free graphs. In these cases, there are many local optimums and the JSD-RC algorithm may be converging to nodes further away from the true sources.

B. Comparison with MJC and CL-CRI; and varying time difference

We also compare the performance of the JSD algorithm with the MJC algorithm [30] and the CL-CRI algorithm [34]. The MJC algorithm and the CL-CRI algorithm are among the most successful algorithms in multiple sources detection when the sources have the same infection starting time. As we commented earlier, if the time difference is large, it is hard to differentiate infection from the second source. We therefore make comparisons only for time differences that are not too large in each of the three graph types.

In the regular and scale-free cases, we use the JSD-J algorithm; while in the small world case, we apply the JSD-RC algorithm. For convenience, we just call both the JSD algorithm in the following. The results are summarized in Figure 12. We see that in the case of 3-regular tree, the JSD algorithm performs significantly better than both the MJC algorithm and the CL-CRI algorithm. In the case of small world graphs, the JSD algorithm performs much better than the MJC algorithm and slightly better than the CL-CRI algorithm. While in the scale-free case, the JSD performs much better than the CL-CRI algorithm and slightly better than the MJC algorithm.

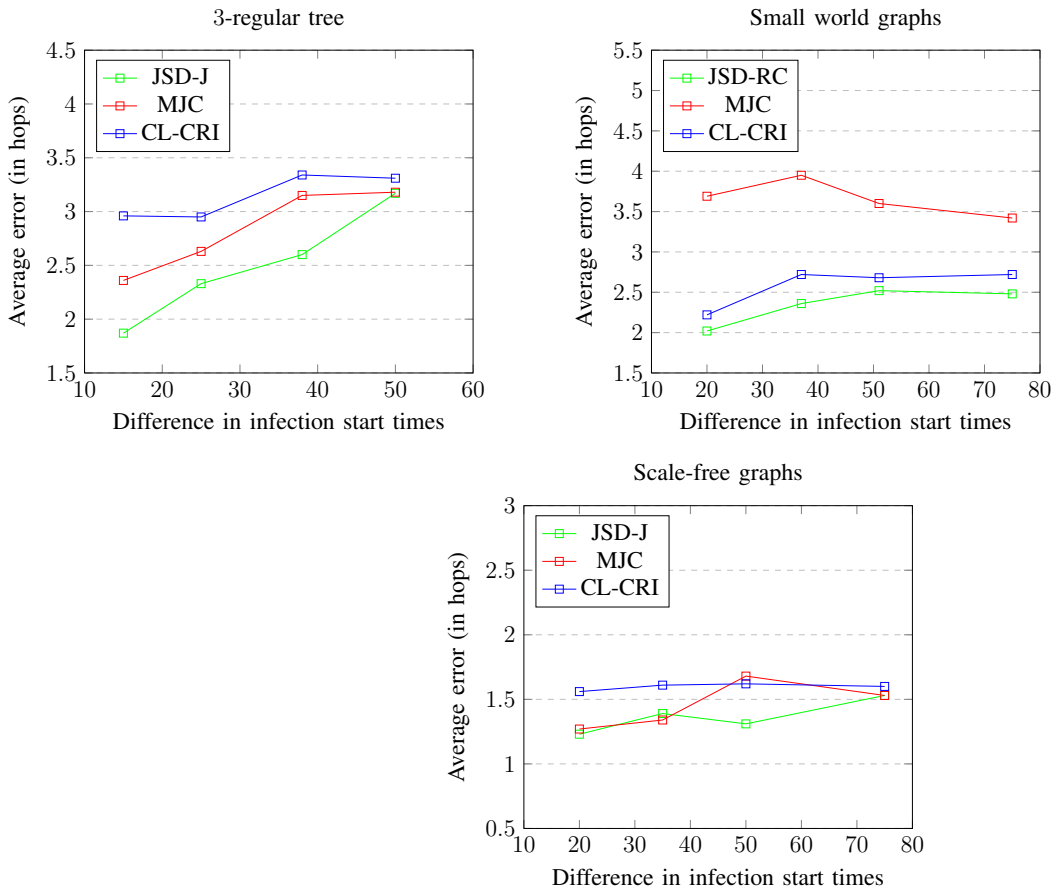

Fig. 12. The red curves depict the average estimation error of the MJC algorithm; the green curves depict the average estimation error of the JSD algorithm; and the blue curves depict the average estimation error of the CL-CRI algorithm.

C. More than two sources on large networks

In the next set of simulations, we consider more than two sources. The networks we use include regular tree, synthetic scale-free graphs, as well as Facebook and email networks from the Stanford Network Analysis Project, [43], [44]. Each graph contains $3000-5000$ nodes and $20\%-25\%$ of them are infected, depending on the type of graph. Moreover, we choose the time difference between two successive infections uniformly with mean d; and randomly over a small time interval of size s (the values of d and s are displayed in Figure 13). In each plot in Figure 13, we use a to denote the average distance between pairs of vertices for the corresponding graph. Except for the difference in diffusion models, we keep all the rest of the settings unchanged for the same graph type.

For the first set of simulations, we use the continuous-time model. The simulations show that our method performs better in many cases. For the real-world graphs, we see that the average error trends lower with more sources. This is because these real-world networks are more dense than the synthetic networks (as evidenced by the smaller average pairwise distance). The error distance for each source is thus smaller, and adding more sources actually increases the chance that an estimated source is close to one of the true sources. In the synthetic cases where the average pairwise distances are larger (in particular for the regular tree), sources with later infection times incur a larger error distance. In conducting the simulations, we also monitor the rate of convergence of the algorithm. Although we do not have theoretical results on the convergence rate, experiments suggest that our method converges within 5 iterations on average.

We next test our algorithms on the same networks as above using a discrete-time diffusion model. From Figure 14, we see that our JSD-J algorithm again outperforms both the MJC and CL-CRI methods in

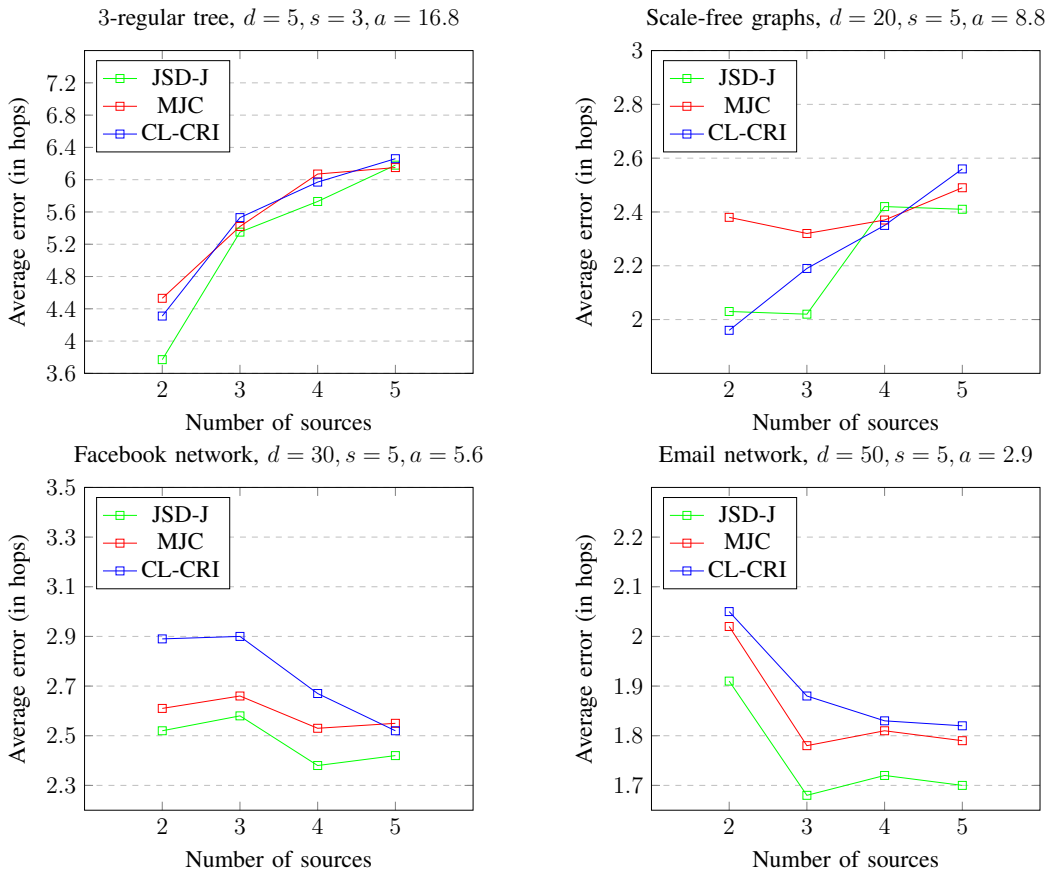

Fig. 13. The continuous-time model. The red curves depict the average estimation error of the MJC algorithm; the green curves depict the average estimation error of the JSD algorithm; and the blue curves depict the average estimation error of the CL-CRI algorithm.

most situations.

D. Γ-*heavy centers*

As we mentioned in Section V-C, we can further generalize the notion of heavy center by considering h_v with at most $1-\Gamma$ fraction of nodes not in I, where $\Gamma \in (0, 1]$. We call such an h_v a Γ-*heavy center*. The heavy center considered in Definition 5(a) is the special case when $\Gamma = 1$. Although we do not have theoretical results for this generalized heavy center, we run simulations to test the performance for various values of Γ on both the Facebook and Email networks, and under the continuous-time and discrete-time models. We use the same simulation settings as Section VI-C. The results are summarized in Figures 15.

From Figure 15, we see that the dependency of the estimation error performance on the Γ-value depends on the type of network and the number of sources. In general, $\Gamma = 1$ does not guarantee the best performance, although the performance in terms of average estimation error is comparable across different values of Γ. In some cases however, smaller Γ values can produce noticeable improvements. The question of how to choose Γ requires further investigation.

VII. Conclusion

We have proposed a general framework and algorithm to perform multiple-sources identification based on estimators that have been developed for single-source identification. Our algorithm is designed to perform estimation when the sources may start their infections at different times. We showed that our proposed algorithm converges if the underlying network is a quasi-regular tree. We also showed specifically how to construct the proposed multiple-source estimator based on the single-source Jordan center estimator

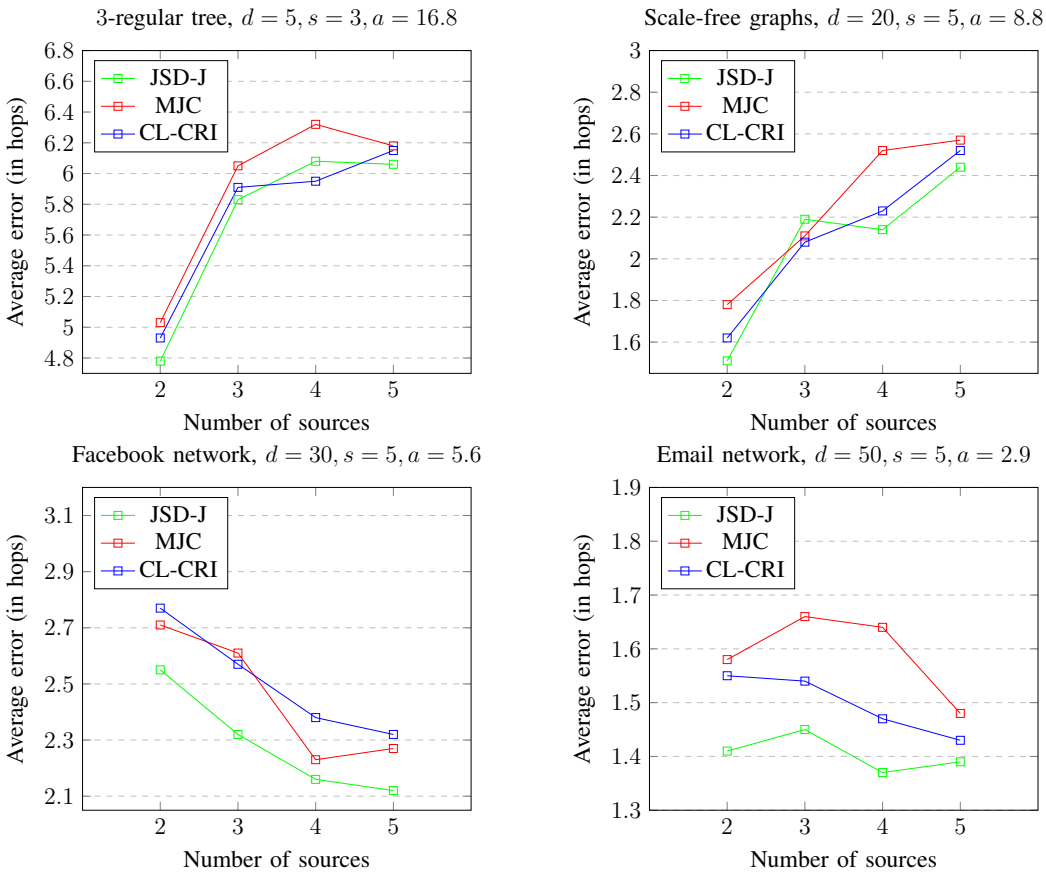

Fig. 14. The discrete-time model. The red curves depict the average estimation error of the MJC algorithm; the green curves depict the average estimation error of the JSD algorithm; and the blue curves depict the average estimation error of the CL-CRI algorithm.

and rumor centrality estimator. Simulations suggest that our proposed framework improves the estimation accuracy compared to other multiple-sources identification methods when sources have different infection start times.

REFERENCES

[1] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on Twitter," in *Proc. ACM Int. Conf. Web Search Data Min.*, Feb. 2011, pp. 65–74.

[2] S. Aral and D. Walker, "Identifying influential and susceptible members of social networks," *Science*, vol. 337, no. 6092, pp. 337–341, Jul. 2012.

[3] W. P. Tay, "The value of feedback in decentralized detection," *IEEE Trans. Inf. Theory*, vol. 58, no. 12, pp. 7226–7239, Dec. 2012.

[4] A. Mitchell, J. Kiley, J. Gottfried, and E. Guskin, "The role of news on Facebook: Common yet incidental," Pew Research Center, Tech. Rep., Oct. 2013. [Online]. Available: http://www.journalism.org/2013/10/24/the-role-of-news-on-facebook

[5] D. W. Soh, W. P. Tay, and T. Q. S. Quek, "Randomized information dissemination in dynamic environments," *IEEE/ACM Trans. Netw.*, vol. 21, no. 3, pp. 681–691, Jun. 2013.

[6] J. Ho, W. P. Tay, T. Q. S. Quek, and E. K. P. Chong, "Robust decentralized detection and social learning in tandem networks," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5019–5032, Oct. 2015.

[7] W. P. Tay, "Whose opinion to follow in multihypothesis social learning? A large deviations perspective," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 2, pp. 344–359, Mar. 2015.

[8] S. Edunov, C. Diuk, I. Filiz, S. Bhagat, and M. Burke. (2016, Feb.) Three and a half degrees of separation. Research at Facebook blog. [Online]. Available: https://research.facebook.com/blog/three-and-a-half-degrees-of-separation/

[9] A. Pentland, *Social Physics: How Good Ideas Spread–The Lessons from a New Science*. New York: Penguin Press, 2014.

[10] W. F. Ogburn and D. Thomas, "Are inventions inevitable? A note on social evolution," *Polit. Sci. Q.*, vol. 37, no. 1, pp. 83–98, Mar. 1922.

[11] R. K. Merton, "Singletons and multiples in scientific discovery: A chapter in the sociology of science," *Proc. Am. Philos. Soc.*, vol. 105, no. 5, pp. 470–486, Oct. 1961.

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

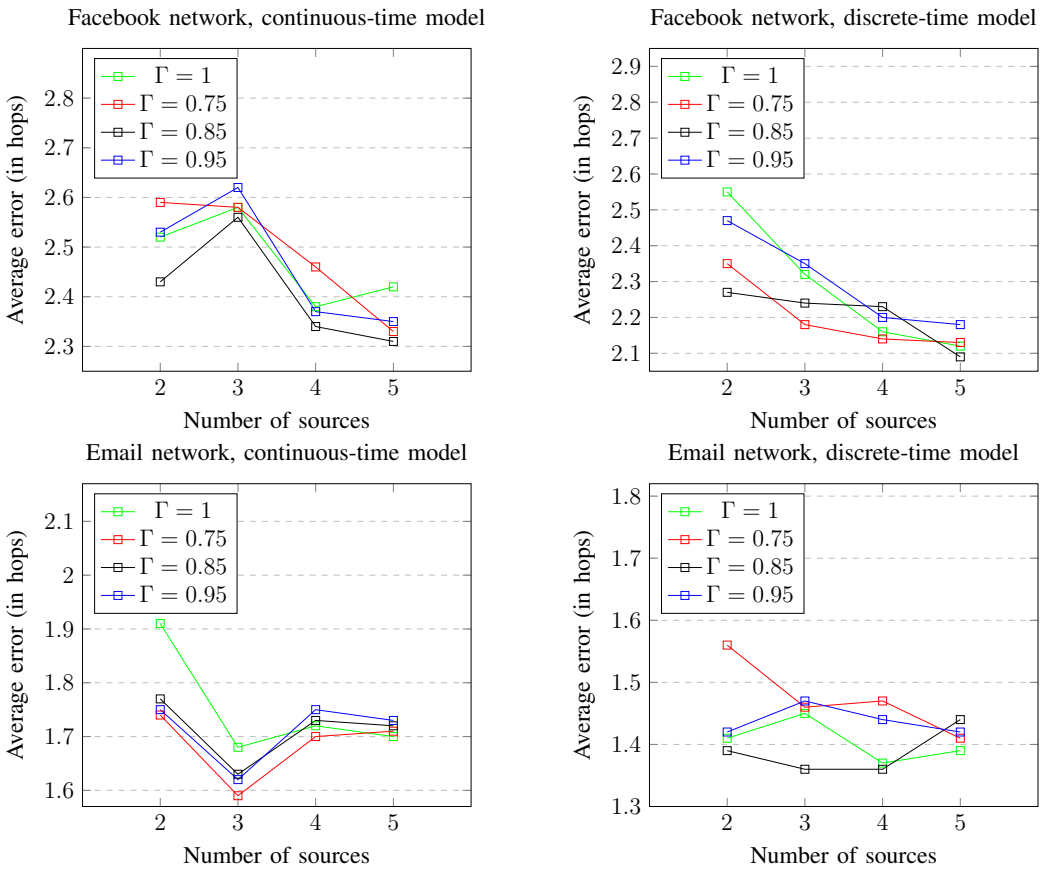

Fig. 15. Average estimation error in real networks for different Γ.

[13] ——, "Rumor centrality: A universal source detector," in *Proc. ACM SIGMETRICS Int. Conf. Meas. Model. Comput. Syst.*, Jun. 2012, pp. 199–210.

[14] K. Zhu and L. Ying, "A robust information source estimator with sparse observations," *Computational Social Networks*, vol. 1, no. 3, Dec. 2014.

[15] W. Luo, W. P. Tay, and M. Leng, "How to identify an infection source with limited observations," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 4, pp. 586–597, Aug. 2014.

[16] C. Moore and M. E. J. Newman, "Epidemics and percolation in small-world networks," *Phys. Rev. E*, vol. 61, no. 5, pp. 5678–5682, 2000.

[17] M. E. J. Newman, "Spread of epidemic disease on networks," *Phys. Rev. E*, vol. 66, no. 1, p. 016128, Jul. 2002.

[18] P. D. O'Neill, "A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods," *Math. Biosci.*, vol. 180, no. 1-2, pp. 103–114, 2002.

[19] A. Ganesh, L. Massoulié, and D. Towsley, "The effect of network topology on the spread of epidemics," in *Proc. IEEE Infocom*, vol. 2, Jan. 2005, pp. 1455–1466.

[20] P. C. Pinto, P. Thiran, and M. Vetterli, "Locating the source of diffusion in large-scale networks," *Phys. Rev. Lett.*, vol. 109, no. 6, p. 068702, Aug. 2012.

[21] F. Altarelli, A. Braunstein, L. Dall'Asta, A. Lage-Castellanos, and R. Zecchina, "Bayesian inference of epidemics on networks via belief propagation," *Phys. Rev. Lett.*, vol. 112, no. 11, p. 118701, Mar. 2014.

[22] W. Dong, W. Zhang, and C. W. Tan, "Rooting out the rumor culprit from suspects," in *Proc. IEEE Int. Symp. Inf. Theory*, July 2013, pp. 2671–2675.

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

[24] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, "Inferring the origin of an epidemic with a dynamic message-passing algorithm," *Phys. Rev. E*, vol. 90, p. 012801, Jul. 2014.

[25] A. Louni and K. P. Subbalakshmi, "A two-stage algorithm to estimate the source of information diffusion in social media networks," in *Proc. IEEE Conf. Comput. Commun. Workshops*, Apr. 2014, pp. 329–333.

[26] W. Hu, W. P. Tay, A. Harilal, and G. Xiao, "Network infection source identification under the SIRI model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 1712–1716.

[27] A. Louni, A. Santhanakrishnan, and K. P. Subbalakshmi, "Identification of source of rumors in social networks with incomplete information," in *Proc. ASE Int. Conf. Social Comput.*, Aug. 2015.

[28] Z. Wang, W. Dong, W. Zhang, and C. W. Tan, "Rooting our rumor sources in online social networks: The value of diversity from multiple observations," *IEEE J. Sel. Top. Signal Proces.*, vol. 9, no. 4, pp. 663–677, June 2015.

[29] W. Luo, W. P. Tay, and M. Leng, "Rumor spreading and source identification: A hide and seek game," *IEEE Trans. Signal Process.*, vol. 64, no. 16, pp. 4228–4243, 2016.

[30] ——, "On the universality of Jordan centers for estimating infection sources in tree networks," *IEEE Trans. Inf. Theory*, 2014, submitted. [Online]. Available: http://arxiv.org/abs/1411.2370

[31] K. Zhu and L. Ying, "Information source detection in networks: Possibility and impossibility results," in *Proc. IEEE INFOCOM*, Apr. 2016.

[32] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila, "Finding effectors in social networks," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Jul. 2010, pp. 1059–1068.

[33] B. A. Prakash, J. Vreeken, and C. Faloutsos, "Spotting culprits in epidemics: How many and which ones?" in *Proc. IEEE Int. Conf. Data Min.*, Dec. 2012, pp. 11–20.

[34] Z. Chen, K. Zhu, and L. Ying, "Detecting multiple information sources in networks under the SIR model," in *Proc. Annu. Conf. Inf. Sci. Syst.*, Mar. 2014.

[35] R. Klein, "Abstract Voronoi diagrams and their applications," in *Computational Geometry and its Applications*, ser. Lecture Notes in Computer Science, H. Noltemeier, Ed. Berlin: Springer, 1989, vol. 333, pp. 148–157.

[36] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd ed. Wiley, 2000.

[37] F. Aurenhammer, "Power diagrams: properties, algorithms and applications," *SIAM J. Comput.*, vol. 16, pp. 78–96, 1987.

[38] F. Ji, W. P. Tay, and L. R. Varshney, "Estimating the number of infection sources in a tree," in *Proc. IEEE Global Conference on Signal and Information Processing*, Dec. 2016.

[39] K. Zhu and L. Ying, "Information source detection in the SIR model: A sample-path-based approach," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 408–421, Feb. 2016.

[40] C. W. Tan, P. D. Yu, C. K. Lai, W. Y. Zhang, and H. L. Fu, "Optimal detection of influential spreaders in online social networks," in *Proc. Annu. Conf. Inf. Sci. Syst.*, Mar. 2016, pp. 145–150.

[41] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.

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

[43] J. Leskovec and J. McAuley, "Learning to discover social circles in ego networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Cambridge, MA: MIT Press, 2012, pp. 539–547.

[44] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. Knowl. Discovery Data*, vol. 1, no. 1, Mar. 2007.