Math. Model. Nat. Phenom. 18 (2023) 2
https://doi.org/10.1051/mmnp/2022048

Mathematical Modelling of Natural Phenomena
www.mmnp-journal.org

EPICENTER OF RANDOM EPIDEMIC SPANNING TREES ON FINITE GRAPHS*

Daniel R. Figueiredo[1] and Giulio Iacobelli[2,**]

Abstract. Epidemic source detection is the problem of identifying the network node that originated an epidemic from a partial observation of the epidemic process. The problem finds applications in different contexts, such as detecting the origin of rumors in online social media, and has been studied under various assumptions. Different from prior studies, this work considers an epidemic process on a finite graph that starts on a random node (epidemic source) and terminates when all nodes are infected, yielding a rooted and directed epidemic tree that encodes node infections (*i.e.*, a directed spanning tree of the graph with every edge directed away from the epidemic source). Assuming knowledge of the underlying graph and the *undirected* spanning tree (*i.e.*, the infection edges but not their directions), can the epidemic source be accurately identified? This work tackles this problem by introducing the *epicenter*, an efficient estimator for the epidemic source, and thus, the direction of every edge in the epidemic tree. When the underlying graph is vertex-transitive the epicenter can be computed in linear time and it coincides with the well-known distance center of the epidemic tree. Moreover, on a complete graph the epicenter is also the most likely estimator for the source. Finally, the accuracy of the epicenter is evaluated numerically on five different graph models and the performance strongly depends on the graph structure, varying from 31% (on complete graphs) to 13% (on sparse power-law graphs). However, for all graph models considered the epicenter exhibited an accuracy higher than the distance center, being three times more accurate on sparse power-law graphs.

Mathematics Subject Classification. 05C85, 05C80, 62Mxx, 05C05.

Received May 3, 2022. Accepted November 21, 2022.

1. Introduction

Network epidemics is a ubiquitous model to capture different diffusion processes on networks such as the spread of a disease in a population, viruses in computer networks, and *fake news* in online social networks [4, 16, 17, 19, 20]. Within this context, an important and general problem is determining the *epidemic source*: identifying the node or nodes that were first infected, that originated the epidemic, from a partial observation of the epidemic process [1, 2, 6, 10, 13, 21, 24–26]. This problem finds various applications such as identifying

* This work was partially funded by research grants awarded by CNPq (Brazil) and FAPERJ (Brazil).

Keywords and phrases: Random spanning trees, source identification, network epidemics, inference in networks.

[1] Department of Systems Engineering and Computer Science (PESC), Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil.
[2] Department of Statistical Methods (DME/IM), Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil.

** Corresponding author: giulio@im.ufrj.br

the node responsible for starting the dissemination of a *fake news* in an online social network, or identifying the individual first infected by a disease within a given population.

The problem of identifying the epidemic source has many flavors, as it strongly depends on the epidemic model and the information that is observed. One of the simplest epidemic model is the SI model where network nodes can either be susceptible (S) or infected (I). In this model, nodes do not recover and the only possible epidemic transition is from S to I. Starting form a single infected node (epidemic source) the infection spreads to other nodes through the edges of the network. Consider observing the set of infected nodes at a given time during the epidemic. Given the network and the set of infected nodes, can the epidemic source be identified?

The seminal work of Shah and Zaman introduced the notion of *rumor centrality* and characterized its effectiveness in identifying the epidemic source on trees [23, 24]. Their work considers a probabilistic SI epidemic model spreading on a known infinite tree (possibly random) and the set of infected nodes at a given time. They show that *rumor centrality* coincides with the Maximum Likelihood (ML) estimator for the epidemic source, and provide a linear time algorithm to compute it. Various subsequent works have followed this methodology and assumed the underlying graph to be infinite or very large in comparison to the observed infected nodes [3, 6, 9, 13].

In a different setting, the seminal work of Pinto *et al.* considered that the infection times of a small fraction of the nodes are observed along with knowledge of the underlying network [21]. However, multiple epidemic cascades (*i.e.*, independent realizations of the epidemic process) are observed. They provide a polynomial time algorithm to infer the epidemic source, studying the inherent tradeoffs in the model. Various subsequent works have taken this approach of observing multiple cascades to infer the epidemic source and other epidemic characteristics [2, 6, 18, 26].

Different from prior studies, this work considers an SI epidemic process on an arbitrary finite graph that starts on a random node (epidemic source) and terminates when all nodes are infected. The epidemic process generates a rooted and directed epidemic tree that contains all network nodes and encodes node infections (*i.e.*, a directed spanning tree of the graph with every edge directed away from the epidemic source). Given the network G and a single observation of the *unrooted* and *undirected* epidemic tree τ, can the epidemic source be identified? Note that the observation is a spanning tree of the network that encodes the edges through which nodes were infected, but not their direction or any other timing information. Moreover, the identification of the epidemic source also reveals the direction of all edges of the epidemic tree.

As a possible application for this model, consider a meme (*e.g.*, picture) spreading through a messaging service (*e.g.*, Whatsapp) where edges of the network are given by the users' contact lists. Imagine that an adversary obtains information between the first exchanges of users but with no timing information or the direction of the exchange. Can the adversary identify the source of the meme?

Another possible application concerns the study of *arborescence* which are rooted and directed spanning trees of graphs used in various optimization problems, such as distribution networks [11]. While the minimum cost root location problem asks for the best arborescence of a graph, the proposed formulation asks to determine the most likely arborescence of a graph when constructed by a random process, given its edges but not their direction.

Beyond the novel observation model, this work makes the following contributions:

- Propose a novel estimator for the epidemic source named *epicenter* (see, Def. 4.4) that leverages distances on both G and the spanning tree τ to reveal the epidemic source. The epicenter can always be computed in polynomial time and structural properties of G and τ allow to reduce its running time complexity.
- Show that when G is a vertex-transitive graph the *epicenter* can be computed in linear time (in the number of nodes). In such graphs, it is shown that the epicenter coincides with the distance center and rumor center of τ (under an exponential SI model). Moreover, when G is a complete graph, the epicenter is also the Maximum Likelihood (ML) estimator for the epidemic source.
- Evaluate, through simulations, the accuracy and other characteristics of the *epicenter* in identifying the epidemic source in five different graph models, as well as a direct comparison with the distance center of the epidemic tree. While the accuracy strongly depends on the graph structure, varying from 31% (on

complete graphs) to 13% (on sparse power-law graphs), the epicenter always exhibited an accuracy higher than the distance center, being three times more accurate on sparse power-law graphs.

The remainder of this paper is organized as follows. The related work is briefly addressed in Section 2. Section 3 presents the epidemic model and observation process. Section 4 presents the *epicenter* and some of its structural properties. The special case of vertex-transitive graphs is studied in Section 5. Section 6 presents a numerical evaluation of the epicenter on different graph models. Section 7 presents a short conclusion.

2. Related work

Identifying the source of an epidemic is a fundamental problem that has received attention from both theoretical and practical perspectives. The problem has direct applications such as determining the source of a rumor that spreads through an online social network or the source of a blackout that spreads through the power grid [10, 25].

Not surprising, various formulations of the problem have been investigated and their main differences concern the prior knowledge about the epidemic process and the observation model. The former determines the prior information that is available for the source identification, such as the network structure over which the epidemic unfolds, or the epidemic model and its parameters. This prior information does not depend on the realization of the epidemic. The latter determines what is observed with respect to the epidemic realization such as the infected nodes or the infection times (of a small fraction) up to a given time instant [10, 25].

Beyond the observation model, another fundamental aspect in the problem formulation is the number of epidemic realizations (cascades) observed. While some formulations rely on a single epidemic cascade, others consider multiple independent (and even dependent) cascades where results are often a function of the number of cascades (more cascades leading to better results). This formulation is often used in more difficult cases such as when the network structure is unknown or multiple epidemic sources co-exist in the network [6, 10, 15, 18, 28].

Part of the prior theoretical works assume that the underlying graph is an infinite tree (possibly random) and that infected nodes are observed after a long period of time (yielding a finite tree, a subgraph of the original infinite tree) [3, 24]. Indeed, the main theoretical contribution of [24] are lower and upper bounds for the asymptotic accuracy of the Maximum Likelihood Estimator (which can be computed in linear time) for the source when the observation time goes to infinity (or number of observed nodes goes to infinity). The main proof technique is casting the epidemic model as a generalized Pólya urn model and using known results from the latter to estimate the asymptotic accuracy of the estimator. Unfortunately, this approach cannot be used in the problem formulation considered in this work, for two reasons: (i) the underlying graph is not a tree (thus, the mapping to a Pólya urn is, in general, not possible); (ii) the underlying graph is finite (thus, asymptotic results are not available). While the subsequent work of Dong *et al.* considers a partial observation of the set of infected nodes and provides result for finite number of nodes, its proof technique also relies on Pólya urn models [3]. Indeed, prior works that consider a finite underlying graph assumes that the observation of the epidemic process occurs when a relatively small fraction of the network nodes are infected [6, 13, 27].

The observation model proposed in Kumar *et al.* consists of the set of infected nodes as well as a random fraction of directed edges from the epidemic tree [13]. While this observation model is related to the model here proposed, there are some fundamental differences: when a directed edge is observed, it is clear that the target node of that edge cannot be the epidemic source. Thus, as more edges are observed, less nodes are left as candidates for the epidemic source. In contrast, this work observes all epidemic edges, no direction or timing information is available, and thus, all nodes are candidates for the epidemic source.

A model related to SI epidemics are random tree growing processes: nodes arrive sequentially and connect to the existing tree according to some probabilistic rule (becoming part of the tree before the next node arrives). Two classic models are the uniform attachment and preferential attachment. Thus, given an observed tree of certain size, can the first node be identified? This problem is related to epidemic source identification and has been investigated mostly in theoretical grounds [1, 8, 14]. In this context, a common problem formulation asks for a set of tree nodes such that the epidemic source is within this set with probability at least $1 - \epsilon$. Interestingly,

a key result is that the size of this set does not depend on the network size, but only on ϵ [1]. Again, a key ingredient for proving such result is casting the tree construction process as a generalized Pólya urn model, and allowing the tree to grow to infinity. Differently from the model consider in this paper, the tree resulting from these random tree growing processes is not constrained by an underlying graph.

3. Random epidemic trees on graphs

In the classic SI epidemic model individuals of a population can either be susceptible (S) or infected (I), and the only possible epidemic transition is from S to I. The epidemic will unfold on an arbitrary undirected finite graph $G = (V, E)$ henceforth assumed connected, where the set V represents the individuals (nodes) with size $n = |V|$, and the set E the possibility of contagion (edges) with size $m = |E|$.[1]

The epidemic is described by a discrete time model and a partition of V into $S(t)$ and $I(t)$, representing the set of susceptible and infected nodes at time $t = 1, 2, \ldots$. Initially, *i.e.*, at time one, a single node is infected and $I(1) = \{v\}$, with $v \in V$. This node is called the *epidemic source* since the epidemic process will unfold from this node. We assume that a single node is infected at each time instant such that $|I(t+1) \setminus I(t)| = 1$ for $t = 1, \ldots, n-1$. Note that the epidemic process will eventually reach all nodes of G and, in particular $I(n) = V$.

The edges of G encode the possibility of contagion, in the sense that the epidemic unfolds through the edges. In particular, in order for a node to become infected, one of its neighbors must be infected (with the exception of the epidemic source). Therefore, only nodes that have an infected neighbor at time t can become infected at time $t+1$. Moreover, we assume that an infection event occurs through an specific edge: a node becomes infected by exactly one of its infected neighbors.

Let $C(t)$ denote the edge cut induced by the partition $I(t)$ and $S(t)$. Note that for each edge $e = \{u, v\} \in C(t)$, one node $u \in I(t)$ is infected and the other node $v \in S(t)$ is susceptible. Let $b_t = (u, v)$ be a directed edge corresponding to the edge that infected node v at time $t+1$, in the sense that $I(t+1) \setminus I(t) = \{v\}$. Thus, b_t provides the infection event at time t. As it turns out, the entire epidemic process is characterized by the epidemic source, denoted by r, and the sequence of directed edges $b_1, \ldots, b_{n-1}$, where b_1 is incident to r. This rooted sequence will be denoted as $b^r = (b_1, \ldots, b_{n-1})$, meaning that $I(1) = \{r\}$ and that b_1 is incident to r. Note that b^r induces one rooted and directed spanning tree of G, since a node is only infected once. Let (τ', r) denote the rooted and directed spanning tree induced by b^r. While b^r precisely constructs (τ', r), it is possible for (τ', r) to be constructed by other edge sequences.

The following probabilistic model for edge selection is considered, which also determines how the epidemic unfolds through the graph. Let e_t be a random variable denoting the edge chosen at time t by the epidemic process. We assume that e_t has a uniform distribution over $C(t)$, the edge cut at time t. In particular, $P(e_t = \{u, v\}) = 1/|C(t)|$, for all $\{u, v\} \in C(t)$, and is 0 otherwise. Note that the probability that a susceptible node is infected at time t is proportional to its number of infected neighbors at that time, and thus in general not uniform over $S(t)$.

The probability that a given rooted edge sequence $b^r = (b_1, \ldots, b_{n-1})$ is observed is simply the product of the edge cut sizes induced by the sequence. In particular, the set of infected nodes are given by $I_{b^r}(1) = \{r\}$ and, for $t = 2, \ldots, n$, $I_{b^r}(t) = \bigcup_{i=1}^{t-1} b_i$ which in turn can be used to determine $C_{b^r}(t)$. Thus,

$$P((e_1, \ldots, e_{n-1}) = b^r) = \prod_{t=1}^{n-1} \frac{1}{|C_{b^r}(t)|}.$$

The network epidemic model above is related to random tree growing processes models [5]. In the classic uniform attachment random tree model, a node at time t joins the tree connecting uniformly at random to one of the nodes in the existing tree. Note that this model is equivalent to the above epidemic model when G is a complete graph (nodes are relabeled by their infection times), since in complete graphs a susceptible node is infected by an infected node chosen uniformly at random.

[1] Throughout the paper, $|S|$ denotes the cardinality of a set S.

The model above is also related to the classic continuous time SI network epidemic model, where the time to infection of a node follows an exponential distribution with rate given by the number of infected neighbors [23]. Since time is continuous, only one node will be infected at any given time instant. Moreover, since time is exponentially distributed, the probability that a given node is the next to become infected can be shown to be exactly the same as in the above model. Thus, these two epidemic models are equivalent.

Problem formulation: Given a graph G, the proposed network epidemic model generates (τ', r), a random rooted and directed spanning tree induced by the random source r and random sequence b^r. Let τ be the *unrooted* and *undirected* spanning tree constructed from (τ', r) by removing the direction of every edge (as well as the root). Thus, τ encodes the infection edges but not the infection direction. We consider the following problem: given G and a single realization of τ, determine the epidemic source. Note that τ encodes the infection edges with no information concerning the infection direction or any other timing information.

While any node of τ can be the epidemic source, their probability of being the source varies and depends on τ. Intuitively, the structure of τ along with G provides evidence for nodes that are more likely to be the epidemic source. For example, the epidemic source is more likely to be at the "center" of τ when also considering G.

4. Epicenter and ML source estimator of epidemic trees

Given a graph G and a single realization of the epidemic tree τ (unrooted and undirected), the goal is to determine the epidemic source. Henceforth, let $V(\tau), E(\tau)$ and $|\tau|$, denote the set of nodes, set of edges and size, of the tree τ, respectively. Table 2 provides a summary of the main notation used throughout the paper.

Note that when fixing a possible root for τ, say v, there are a myriad of different edge sequences starting from v which could have generated the now rooted and directed tree τ_v. This motivates the following definition which establishes the conditions for an edge sequence to be capable of generating the tree τ starting from a node v:

Definition 4.1. Given $v \in V$, an *edge sequence* $b^v = (b_1, \dots, b_{n-1})$ rooted at v generates the rooted tree $\tau_v = (\tau, v)$ if for $t = 1, \dots, n-1$, $b_t = \{u_t, v_t\} \in E(\tau)$ and for $t = 2, \dots, n-1$, $|\bigcup_{i=1}^{t-1}\{u_i, v_i\} \cap \{u_t, v_t\}| = 1$.

Let $B(\tau_v)$ denote the set of all edge sequences rooted at v that generate the rooted tree τ_v. Note that $B(\tau_v)$ depends only on τ and v but not on the graph G from which τ was constructed. In what follow we present a couple of results on the size of $B(\tau_v)$.

Lets start with a notation that will be used quite extensively: given a rooted tree τ_v and a node $u \in V(\tau_v)$, we denote by τ_u^v the rooted subtree of τ_v dangling from node u (with respect to v), rooted at u. Specifically, if $u \neq v$, τ_u^v denotes the subtree rooted at u obtained by removing the edge connecting u to its parent with respect to v (*i.e.*, the neighbor of u in the unique path between u and v in τ_v); whereas, if $u = v$, $\tau_u^v = \tau_v$. The first lemma establishes a recursive formula for $|B(\tau_v)|$.

Lemma 4.2. *Let $|B(\tau_v)|$ be the number of edge sequences rooted at v which generate τ_v, and let N_v denote the set of neighbors of v in τ. The following recursion holds:*

$$|B(\tau_v)| = \prod_{u \in N_v} |B(\tau_u^v)| \, \frac{\left(\sum_{u \in N_v} |\tau_u^v|\right)!}{\prod_{u \in N_v} |\tau_u^v|!}, \qquad \text{if } N_v \neq \emptyset,$$

with $|B(\tau_v)| = 1$ if $N_v = \emptyset$ (this case accounts for when τ_v is the single node v).

The combinatorial argument used above is quite standard and a sketch of the proof can be found in the Appendix. Recursively applying Lemma 4.2, we obtain the following proposition.

Proposition 4.3. *The number of edge sequences rooted at v which generates the rooted spanning tree τ_v is given by*

$$|B(\tau_v)| = \frac{(n-1)!}{\prod_{u \neq v} |\tau_u^v|} .$$

Recall that given a tree τ generated by the random epidemic process, the probability that a node is the epidemic source of τ is node dependent. Specifically, the probability that τ was constructed from a root v can be computed by summing over all possible rooted edge sequences that generate τ_v (since they are all mutually exclusive). In particular, we have:

$$\begin{aligned} P_G(\tau|\text{root} = v) &= \sum_{b^v \in B(\tau_v)} P_G\big((e_1, \dots, e_{n-1}) = b^v\big) \\ &= \sum_{b^v \in B(\tau_v)} \prod_{t=1}^{n-1} \frac{1}{|C_{b^v}(t)|} , \end{aligned} \tag{4.1}$$

where the dependence on the underlying graph G is through the edge cut sizes. Applying Bayes rule we obtain the probability that a node is the epidemic source given the tree τ:

$$P_G(\text{root} = v|\tau) = \frac{P_G(\tau|\text{root} = v)P(\text{root} = v)}{P_G(\tau)} . \tag{4.2}$$

The above equation requires a prior for the epidemic source, namely $P(\text{root} = v)$, which is assumed to be uniform across nodes in V, *i.e.*, $P(\text{root} = v) = 1/n$ for all $v \in V$. Moreover, it also requires $P_G(\tau)$ which can be computed using the Law of Total Probability. More importantly, neither the prior nor $P_G(\tau)$ depend on the specific v and thus both can be treated as constants in equation (4.2). Thus, the Maximum Likelihood estimator (MLE) for the epidemic source, denoted by $r^*_{\text{ML}}(G, \tau)$, corresponds to

$$r^*_{\text{ML}}(G, \tau) = \arg\max_{v \in V} P_G(\text{root} = v|\tau) = \arg\max_{v \in V} P_G(\tau|\text{root} = v) , \tag{4.3}$$

where the last equality holds due to the uniform prior.

In general, maximizing equation (4.1) is computationally expensive, also due to the dependence on the underlying graph G (see discussion in Sect. 4.1). Intuitively, the structure of G and τ can be used more directly (in terms of computational complexity) to provide information about the chances that a given node is the epidemic source. Thus, a new estimator for the epidemic source is proposed, called *epicenter*:

Definition 4.4. Let $G = (V, E)$ be a graph and τ a spanning tree of G. The *epicenter* of τ in G is defined as

$$r^*_{\text{EPI}}(G, \tau) := \arg\min_{v \in V} \sum_{u \in V} \Big(d_\tau(v, u) - d_G(v, u) \Big) ,$$

where $d_G(u, v)$ denotes the graph distance between node u and v in G (with ties in $\arg\min$ broken uniformly at random).

Since $d_\tau(v, u) - d_G(v, u) \geq 0$, $\forall u, v \in V$, the epicenter of τ in G is the node that minimize the sum of those positive differences. Intuitively, the epicenter is the node that better aligns the tree τ in G in terms of distances. Note that, the epicenter can be rewritten as $r^*_{\text{EPI}}(G, \tau) = \arg\min_{v \in V} \Big(d_\tau(v) - d_G(v) \Big)$, where $d_G(v) := \sum_{u \in V} d_G(v, u)$ is the distance centrality of v in G (equivalently for τ) [16].

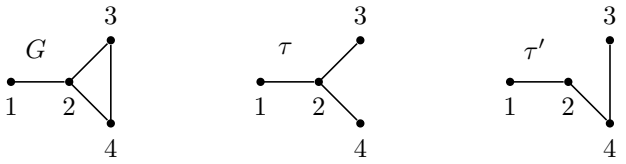


FIGURE 1. A graph G and two spanning trees τ and τ'.

Example 4.5. Let G, τ, τ' be as depicted in Figure 1. We have that $d_G(1) = 5, d_G(2) = 3$, $d_G(3) = d_G(4) = 4$, while $d_\tau(1) = 5, d_\tau(2) = 3$, $d_\tau(3) = d_\tau(4) = 5$ and $d_{\tau'}(1) = 6, d_{\tau'}(2) = 4$, $d_{\tau'}(3) = 6$, $d_{\tau'}(4) = 4$. Thus, the epicenter of τ in G is chosen uniformly from the nodes $\{1, 2\}$, whereas the epicenter of τ' in G is 4.

When the underlying graph is itself a tree, *i.e.*, $G = \tau$, the only possible epidemic tree is τ itself, which implies $d_\tau(v, u) - d_G(v, u) = 0$, $\forall u, v \in V$. Thus, whenever G is a tree, the epicenter does not provide any information concerning the epidemic source. Remarkably, the ML estimator r^*_{ML} in this scenario will be uniformly distributed on V, since $P_G(\tau|\mathrm{root} = v)$ will not depend on v. This follows because the epidemic source is chosen uniformly at random by the epidemic model.

4.1. Computing the epicenter

The epicenter of a spanning tree in a graph can be directly computed from its definition. In particular, one can simply compute the distance centrality for every node $v \in V$, both in $G = (V, E)$ and its spanning tree τ, compute the difference of the corresponding distances, and return the node with the smallest value.

However, given a particular structure for G and τ, it is reasonable that not every node $v \in V$ needs to be considered in this computation. Indeed, the following proposition (to be proven later) states that distance centrality of leaves of τ (*i.e.*, nodes with degree one in τ) are not required to determine the epicenter of τ in G.

Proposition 4.6. *Let $G = (V, E)$ be a graph and τ a spanning tree of G. Let $v \in V$ be a leaf of τ. Then,*

- *if v is not a leaf in G,*

$$v \neq r^*_{\mathrm{EPI}}(G, \tau) := \arg\min_{w \in V} \Big(d_\tau(w) - d_G(w) \Big) ,$$

- *if v is also a leaf in G,*

$$d_\tau(v) - d_G(v) = d_\tau(u) - d_G(u) ,$$

where, u is the unique neighbor of v in G (and also in τ).

The above proposition guarantees that a node which is a leaf in τ but not in G is never the epicenter. Moreover, a leaf in τ is an epicenter only if its parent in τ is an epicenter.

The computation of the epicenter is shown in Algorithm 1. Note that a breadth-first search (BFS) starting at v is sufficient to compute the distances from v to every node in G and τ. This has computational complexity $\Theta(m)$ and $\Theta(n)$, respectively. Computing the differences requires time $\Theta(n)$ (lines $9 - 11$). This process is repeated for every non-leaf node of τ, which is bounded above by n. In general, the number of leaves in τ can be arbitrary, and thus the complexity of Algorithm 1 is $O(nm)$.

Note that finding the MLE for the epidemic source as defined by equation (4.3) requires solving equation (4.1) for every node $v \in V$. A direct computation of equation (4.1) for a given node v would require iterating over all sequences in $B(\tau_v)$. While this number strongly depends on the structure of τ_v, it is likely to grow exponentially with the number of nodes for most trees. For example, for the root of a full binary tree with $k > 0$ levels and

Algorithm 1: *Epicenter of a spanning tree of a graph*

Input: $G = (V, E)$, τ `/* G and a spanning tree τ */`
Output: $r \in V$ `/* epicenter of τ in G */`

`/* initialization */`

epi $= \infty$

`/* main body */`

for $v \in V$ **do**
  if deg(τ, v) $==$ *1* **then**
    next
  end
  DistTree $=$ bfs(τ, v)
  DistGraph $=$ bfs(G, v)
  sum $= 0$
  for $u \in V$ **do**
    sum += DistTree[u] - DistGraph[u]
  end
  if sum $\leq$ epi **then**
    if sum $==$ epi **then**
      $C = C \cup \{v\}$
    else
      $C = \{v\}$
      epi $=$ sum
    end
  end
end
$C = C \cup \{v | v \in$ nei(G, u) & deg$(G, v) = 1, \forall u \in C\}$
center $=$ random.uniform (C)
return center

$n = 2^k - 1$ nodes, it can be shown that

$$|B(\tau_r)| = \frac{(2^k - 2)!}{\prod_{i=1}^{k-1} (2^{k-i} - 1)^{2^i}} = \Omega(2^n) \, .$$

Thus, computing the MLE directly is prohibitive (*i.e.*, exponential number of iterations) for most cases and the epicenter provides a much more efficient approach. Moreover, when G is vertex-transitive the epicenter can be computed even more efficiently, requiring only linear time in n (details in Sect. 5).

4.2. Some properties of the epicenter

This section presents auxiliary results which will be used to prove Proposition 4.6, and also Proposition 5.2 in the sequel.

Given a graph $G = (V, E)$ and two nodes $u, v \in V$, we denote by $\mathcal{P}_G(u, v)$ the set of shortest paths in G between u and v (shortest path may not be unique). Also, given a path $p \in \mathcal{P}_G(u, v)$ and a node w, we say that

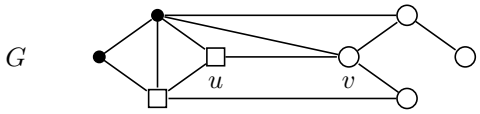


FIGURE 2. The nodes drawn as circles correspond to the set $V_v^u(G)$; the nodes drawn as squares correspond to the set $V_u^v(G)$; the filled nodes do not belong neither to $V_v^u(G)$ nor to $V_u^v(G)$. Note that $v \in V_v^u(G)$ (resp. $u \in V_u^v(G)$), that $V_v^u(G) \cap V_u^v(G) = \emptyset$, and that $V_v^u(G) \cup V_u^v(G) \subseteq V$, with V denoting the set of all nodes of G.

$w \in p$ if and only if the path p crosses the node w. Then, we define

$$V_v^u(G) := \{w \in V : \exists p \in \mathcal{P}_G(w,u) \text{ such that } p \ni v\} ,$$

the set of nodes w for which there exists a shortest path between w and u crossing v in G. Note that $v \in V_v^u(G)$ since the shortest path connecting v to u necessarily contains the node v. Moreover, $V_v^u(G) \cap V_u^v(G) = \emptyset \, , \forall G$ since if a shortest path from a node w to u crosses node v, it cannot be the case that a shortest path from w to v crosses u. Thus, we also have that $V_v^u(G) \cup V_u^v(G) \subseteq V$, with V denoting the set of all nodes of G. See Figure 2 for an example of the set $V_v^u(G)$. To avoid clutter, we shall remove the dependence from G and write V_v^u instead of $V_v^u(G)$, unless otherwise needed.

The following lemma relates the distance centrality of two neighboring nodes in a graph.

Lemma 4.7. *Let $G = (V, E)$ be a graph (connected) and $\{u, v\} \in E$. Then,*

$$d_G(v) + |V_v^u| = d_G(u) + |V_u^v| .$$

Corollary 4.8. *Let $G = (V, E)$ be a graph, τ a spanning tree of G, and $\{u, v\} \in E(\tau)$. It holds that*

$$d_\tau(v) - d_G(v) + |\tau_v^u| - |V_v^u| = d_\tau(u) - d_G(u) + |\tau_u^v| - |V_u^v| , \tag{4.4}$$

where we recall that τ_u^v denotes the subtree of τ corresponding to the connected component of u when the edge $\{u, v\}$ is removed from τ.

If we apply Lemma 4.7 to a tree τ, we obtain that $|V_v^u(\tau)| = |V(\tau_v^u)| = |\tau_v^u|$ and $|V_u^v(\tau)| = |V(\tau_u^v)| = |\tau_u^v|$. Therefore, by applying Lemma 4.7 to G and to a spanning tree τ of G, we obtain equation (4.4).

Proposition 4.6. Since v is a leaf in τ it has a unique neighbor (in τ), henceforth denoted by u, and $|\tau_v^u| = 1$ and $|\tau_u^v| = |V| - 1$. Moreover, $|V_v^u| \geq 1$, since $v \in V_v^u$, and it is also the case that $|V_u^v| \leq |V| - 1$. Therefore, using equation (4.4), we conclude that $|\tau_v^u| - |V_v^u| \leq 0$ and $|\tau_u^v| - |V_u^v| \geq 0$. To prove the first claim it suffices to show that $|\tau_u^v| - |V_u^v| > 0$, since it would imply $d_\tau(v) - d_G(v) > d_\tau(u) - d_G(u)$. Towards this goal, let us partition the set of nodes V as $V = V_v^u \cup V_u^v \cup \left((V_v^u)^\complement \cap (V_u^v)^\complement\right)$. There are three possible scenarios:

1) $(V_v^u)^\complement \cap (V_u^v)^\complement \neq \emptyset$;
2) $(V_v^u)^\complement \cap (V_u^v)^\complement = \emptyset$ and $|V_v^u| > 1$;
3) $(V_v^u)^\complement \cap (V_u^v)^\complement = \emptyset$ and $|V_v^u| = 1$.

In 1), we necessarily have $|V_u^v| \leq |V| - 2$ (since all three sets are non-empty), which implies $|\tau_u^v| - |V_u^v| > 0$. In 2), since $|V_v^u| > 1$ we obtain $|V_u^v| \leq |V| - 2$ and thus $|\tau_u^v| - |V_u^v| > 0$. Lastly, in 3), we have $|V_v^u| = 1$ and $(V_v^u)^\complement \cap (V_u^v)^\complement = \emptyset$. Note that, in this latter case v must necessarily be a leaf in G. The proof of the second claim follows from noticing that if v is a leaf in G, then 3) holds and, given that v must necessarily also be a leaf in τ, from Corollary 4.8, we obtain $d_\tau(v) - d_G(v) = d_\tau(u) - d_G(u)$. □

Lemma 4.7. Given $\{u,v\} \in E$ the set of nodes V can be partitioned as $V = V_v^u \cup V_u^v \cup \left((V_v^u)^{\complement} \cap (V_u^v)^{\complement}\right)$. Moreover,

$$d_G(v,w) = \begin{cases} d_G(u,w) - 1 \,, & \text{if } w \in V_v^u \,, \\ d_G(u,w) + 1 \,, & \text{if } w \in V_u^v \,, \\ d_G(u,w) \,, & \text{if } w \in (V_v^u \cup V_u^v)^{\complement} \,. \end{cases}$$

Therefore, we obtain that

$$\begin{aligned} d_G(v) &= \sum_{w \in V} d_\tau(v,w) = \sum_{w \in V_v^u} d_G(v,w) + \sum_{w \in V_u^v} d_G(v,w) \\ &+ \sum_{w \in (V_v^u \cup V_u^v)^{\complement}} d_G(v,w) = \sum_{w \in V_v^u} d_G(u,w) - |V_v^u| \\ &+ \sum_{w \in V_u^v} d_G(u,w) + |V_u^v| + \sum_{w \in (V_u^v \cup V_u^v)^{\complement}} d_G(u,w) \\ &= d_G(u) - |V_v^u| + |V_u^v| \,. \end{aligned}$$

□

5. Epicenter of spanning trees in vertex-transitive graphs

In this section, we show that whenever the underlying graph G is vertex-transitive, the epicenter of any spanning tree of G can be computed more efficiently. Notable examples of such graphs are: complete, complete bipartite balanced, cycle, hypercubes amongst others.

Let us begin observing that if $f : V \to V$ is an automorphism of $G = (V,E)$, then for every node $u \in V$, it holds that $d_G(u) = d_G(f(u))$. Hence, for a vertex-transitive[2] G, it holds that $d_G(v) = d_G(u)$, $\forall u,v \in V$. Thus, the epicenter of a spanning tree τ of a vertex-transitive G is equivalent to

$$r^*_{\mathrm{EPI}}(G,\tau) = \arg\min_{v \in V} \sum_{u \in V} d_\tau(v,u) \,. \tag{5.1}$$

In this case the epicenter only depends on τ, and reduces to a well-known notion for the center of a tree τ called *distance center*, which is defined as $r^*_{\mathrm{DC}}(\tau) := \arg\min_{v \in V} \sum_{u \in V} d_\tau(v,u)$ [16]. In general, however, the epicenter cannot be easily compared to the distance center because it depends on G (network) and τ.

Remark 5.1. The distance center of trees is related to another network centrality concept called *rumor center*, defined for a tree τ as[3] [24].

$$r^*_{\mathrm{RC}}(\tau) := \arg\max_{v \in V} \frac{|\tau|!}{\prod_{u \in V} |\tau_u^v|} \,.$$

In particular, it can be shown that, ruling out possible tie breaking, $r^*_{\mathrm{DC}}(\tau) = r^*_{\mathrm{RC}}(\tau)$. However, for arbitrary graphs it is often the case that $r^*_{\mathrm{DC}}(G) \neq r^*_{\mathrm{RC}}(G)$ (see, Prop. 2 in [23]).

[2] A graph $G = (V,E)$ is vertex-transitive if for every two nodes $u,v \in V$ there exists an automorphism f mapping u to v.

[3] The rumor center can be defined for arbitrary graphs by executing a BFS from each node v to generate a spanning tree τ^v rooted at v that is used to determine $|\tau_u^v|$.

Before stating the main result of this section, which provides a linear time algorithm to compute the epicenter of trees in vertex-transitive graphs, let us introduce some notation. Given a tree τ and $v \in V(\tau)$ a root, let $S_v := \{u \in V(\tau) : |\tau_u^v| \geq |\tau|/2\}$ be the set of vertices such that the tree dangling from them with respect to v has size at least half of $|\tau|$; note that, $S_v \neq \emptyset$, since $v \in S_v$. Let us define,

$$u^*(\tau, v) := \arg\max_{u \in S_v} d_\tau(u, v) \ ,$$

i.e., the node in S_v with the maximal distance from the chosen root v. Note that for a fixed τ by varying v, the node $u^*(\tau, v)$ may change. However, for a fixed τ and v the node $u^*(\tau, v)$ is unique; as a matter of fact, if the maximal distance is zero, then the only possible node is v itself. Assuming that the maximal distance is $k \geq 1$, then if there were another node with the same distance from v in S_v, it would necessarily imply that the tree τ has size at least $|\tau|/2 + |\tau|/2 + 1$, which is clearly a contradiction.

In the following proposition we show that:

i) the node $u^*(\tau, v)$ is such that after removing any edge incident to it, u^* always belongs to a subtree of size at least half of τ, and this holds regardless the specific v chosen; in formula, u^* is such that $|\tau_{u^*}^{w'}| \geq n/2$, for all w' which are neighbors of u^* in τ.

ii) the node $u^*(\tau, v)$ is an epicenter (ruling out tie breaking), regardless the specific root v chosen. Since $u^*(\tau, v)$ can be easily computed in linear time (see, Algor. 2), this provides an efficient way to find the epicenter of a spanning tree of a transitive graph.

Proposition 5.2. *Let $G = (V, E)$ be a vertex-transitive graph and τ a spanning tree of G. Then, it holds that:*

i) $\{u^*(\tau, v) : v \in V\} = \{w \in V : |\tau_w^v| \geq n/2 \ , \forall v\}$.
ii) $\{u^*(\tau, v) : v \in V\} = \{w \in V : w = r^*_{\mathrm{EPI}}(G, \tau)\}$.

Remark 5.3. The set $\{w \in V : |\tau_w^v| \geq n/2 \ , \forall v\}$ is equivalent to

$$\{w \in V : |\tau_w^v| \geq n/2 \ , \text{ for all } v \text{ neighbors of } w \text{ in } \tau\} \ , \tag{5.2}$$

and, for every tree τ, is non-empty. Moreover, if the tree τ has a bisection (*i.e.*, there exists an edge whose removal partitions the tree into two equal size subtrees), then the set in (5.2) contains two elements, namely the nodes corresponding to the edge whose removal halves the tree. If the tree does not admit a bisection, the set in (5.2) contains a unique element. In passing, note that if the size of the tree n is odd, then the epicenter is uniquely determined.

Proposition 5.2. We begin proving *i*), and specifically that $\{w \in V : |\tau_w^v| \geq n/2 \ , \forall v\} \subseteq \{u^*(\tau, v) : v \in V\}$. For that, it is enough to show that if $w' \in \{w \in V : |\tau_w^v| \geq |\tau|/2 \ , \forall v \in V\}$, then there exists a v such that $w' = u^*(\tau, v)$. Let us assume, towards a contradiction, that for all v, $u^*(\tau, v) \neq w'$. Let v be arbitrary; by definition of w', we know that $|\tau_{w'}^v| \geq n/2$, *i.e.*, $w' \in S_v$. Therefore, we must have $d_\tau(u^*, v) > d_\tau(w', v)$. Since $u^* \in S_v$, we know that $|\tau_{u^*}^v| \geq n/2$. Note that, if $|\tau_{w'}^v| \geq n/2$ and $|\tau_{u^*}^v| \geq n/2$, necessarily w' and u^* must be neighbors in τ, and $|\tau_{u^*}^{w'}| = |\tau_{w'}^{u^*}| = n/2$, *i.e.*, the removal of the edge $\{u^*, w'\} \in E(\tau)$ bisects τ. Thus, if τ does not have a bisection, we obtain a contradiction. If τ admits a bisection and we denote by $\{u', v'\}$ the edge whose removal halves the tree, then it is not difficult to see that $\{u^*(\tau, v) : v \in V\} = \{w \in V : |\tau_w^v| \geq n/2 \ , \forall v\} = \{u', v'\}$.

We now show that $\{u^*(\tau, v) : v \in V\} \subseteq \{w \in V : |\tau_w^v| \geq n/2 \ , \forall v \in V\}$. Let $v \in V$ be arbitrary; given that $u^* = u^*(\tau, v) \in S_v$, we have that $|\tau_{u^*}^v| \geq n/2$. Also, since for all $v' \in V \setminus V(\tau_{u^*}^v)$, it holds that $|\tau_{u^*}^{v'}| = |\tau_{u^*}^v|$, we also have that $\tau_{u^*}^{v'} \geq n/2$, $\forall v' \in V \setminus V(\tau_{u^*}^v)$. Thus, it remains to consider the nodes $v' \in V(\tau_{u^*}^v) \setminus \{u^*\}$. Note that all nodes in the latter set have distance from v strictly bigger than the distance $d_\tau(u^*, v)$. Consequently, these nodes cannot belong to the set S_v (otherwise it would contradict the definition of u^*), and therefore we have that $|\tau_{v'}^v| < n/2$ for all $v' \in V(\tau_{u^*}^v) \setminus \{u^*\}$. Also, for all these nodes it holds that $|\tau_{v'}^v| = |\tau_{v'}^{u^*}| < n/2$. Let $\widehat{w}$ be a node in $V(\tau_{u^*}^v) \setminus \{u^*\}$ having distance one from u^*; then, $|\tau_{\widehat{w}}^{u^*}| < n/2$. However, since $\widehat{w}$ and u^* are neighbors,

it holds that $|\tau^{u^*}_{\widehat{w}}| + |\tau^{\widehat{w}}_{u^*}| = |\tau| = n$, which implies $|\tau^{\widehat{w}}_{u^*}| \geq n/2$, for all $\widehat{w} \in V(\tau^v_{u^*}) \setminus \{u^*\}$, which are neighbors of u^*. Noticing that, for all $v' \in V(\tau^v_{u^*}) \setminus \{u^*\}$ it exists a $\widehat{w} \in V(\tau^v_{u^*}) \setminus \{u^*\}$ neighbor of u^* such that $|\tau^{v'}_{u^*}| \geq |\tau^{\widehat{w}}_{u^*}|$, claim i) follows.

We now proceed proving ii). For a spanning tree τ of a vertex-transitive G, we know that, ruling out tie breaking, $r^*_{\text{EPI}}(G,\tau) = \arg\min_{v\in V} d_\tau(v)$. Thus, it is enough to show that $\{u^*(\tau,v) : v \in V\} = \{w \in V : w = \arg\min_{v\in V} d_\tau(v)\}$. We first show that $\{u^*(\tau,v) : v \in V\} \subseteq \{w \in V : w = \arg\min_{v\in V} d_\tau(v)\}$. Given $u^* \in \{u^*(\tau,v) : v \in V\}$, let us assume towards a contradiction, that $u^* \neq \arg\min_{v\in V} d_\tau(v)$, which is equivalent to say that $\exists r \in V$ such that, $d_\tau(r) < d_\tau(u^*)$. Let p denote the unique path in τ connecting u^* and r, and without loss of generality we assume $p = w_0, w_1 \dots w_{k-1}, w_k$, with $w_0 = u^*$ and $w_k = r$. Let $j := \min\{i \geq 0 : d_\tau(w_i) > d_\tau(w_{i+1})\}$, and note that, by the hypothesis $d_\tau(r) < d_\tau(u^*)$, j is always less or equal than $k-1$. Applying Lemma 4.7 we have that $d_\tau(w_j) + |\tau^{j+1}_j| = d_\tau(w_{j+1}) + |\tau^j_{j+1}|$. Note that $|\tau^{j+1}_j| + |\tau^j_{j+1}| = n$ (because j and $j+1$ are at distance one in τ), and that $|\tau^{j+1}_j| \geq |\tau^{j+1}_{u^*}|$. By point i) above, u^* satisfies $|\tau^w_{u^*}| \geq \frac{n}{2}\ \forall w \in V$, which implies $|\tau^{j+1}_j| \geq \frac{n}{2}$, and thus $d_\tau(w_j) \leq d_\tau(w_{j+1})$, which is a contradiction.

Let us now show that $\{u^*(\tau,v) : v \in V\} \supseteq \{w \in V : w = r^*_{\text{EPI}}(G,\tau)\}$. Let $u' = \arg\min_{v\in V} d_\tau(v)$ and assume, towards a contradiction, that there exists an $r \in V$ such that $|\tau^r_{u'}| < n/2$ (we are using i)). Note that, without loss of generality, we may assume r is a neighbor of u' in τ. Given that $|\tau^r_{u'}| < n/2$ it is also the case that $|\tau^{u'}_r| \geq n/2$, since r and u' are neighbors. Applying Lemma 4.7 to τ, we have that $d_\tau(r) + |\tau^{u'}_r| = d_\tau(u') + |\tau^r_{u'}|$, which implies $d_\tau(r) < d_\tau(u')$, and thus a contradiction. □

5.1. Computing the epicenter in vertex-transitive graphs

The special structure of vertex-transitive graphs provided theoretical results that allow for the design of an efficient algorithm to compute its epicenter. In particular, Proposition 5.2 establishes that the epicenter of τ in G is given by $u^*(\tau,v)$. The goal of the algorithm is to compute $u^*(\tau,v)$ efficiently, and its pseudo-code is shown in Algorithm 2.

The main idea of the algorithm is to create an orientation for τ using an arbitrary node as root (*e.g.*, the first node of V), and then compute the subtree sizes from the leaves towards the root: leaves have subtree size equal to one, a parent has subtree size that is one plus the subtree size of its children. The algorithm stops when it reaches a node that has subtree size of at least $\lceil n/2 \rceil$, and returns this node as the epicenter.

The algorithm is iterative and prunes the leaves of the rooted tree τ_{v_1} which in turn may create new leaves. When a leaf is pruned, the subtree size of its parent is updated, as well as the number of children of its parent (lines 15−16).

The algorithm stops when it encounters a node (*i.e.*, a leaf in the pruned tree) that has subtree size of at least $\lceil n/2 \rceil$ (line 12−13). Since the algorithm iterates from the leaves towards the root, the stopping condition is satisfied by a node at the largest possible distance from v_1. In light of Proposition 5.2, this is the epicenter of the tree τ (and does not depend on the choice of v_1).

Algorithm 2 has computational complexity $\Theta(n)$ where n is the number of nodes in τ. Note that the BFS in line 1 runs on the tree τ which has $n-1$ edges. Moreover, a node enters the leaf set Leafs only once, and thus the main loop requires at most n iterations. Finally, all computations with in the main loop (lines 10−19) require constant time.

The running time complexity of Algorithm 2, $\Theta(n)$, is in sharp contrast with the complexity of the general Algorithm 1, $O(nm)$. Indeed, finding the epicenter of vertex-transitive graphs requires significantly less effort.

5.2. Epicenter in complete graphs

The random network epidemic model under study (described in Sect. 3) has a distinctive feature on complete graphs, henceforth denoted as K: the edge cut size depends only on time t. Specifically, it holds that $|C(t)| =$

Algorithm 2: *Epicenter of a spanning tree of a vertex-transitive graph*

Input: τ `/* a tree with node set` $V = \{v_1, \dots, v_n\}$ `*/`
Output: $r \in V$ `/* epicenter of` τ `*/`

`/* initialization */`

Parent, Children = `bfs`(τ, v_1)
`/* Parent[`i`] is the parent node of` i `in` τ_{v_1} `*/`
`/* Children[`i`] is the number of children of` i `*/`
Leafs = $\emptyset$
TreeSize[ ] = 0
for $i \in V$ **do**
    if Children[i] == 0 **then**
        Leafs.push(i)
    end
end

`/* main body */`

while Leafs $\neq \emptyset$ **do**
    i = Leafs.pop()
    TreeSize[i] += 1
    if TreeSize[i] $\geq \lceil n/2 \rceil$ **then**
        if n % 2 == 0 & TreeSize[i] == $n/2$ **then**
            $C = \{i\} \cup \{\text{Parent}[i]\}$
            return `random.uniform` (C)
        else
            return i
        end
    end
    TreeSize[Parent[i]] += TreeSize[i]
    Children[Parent[i]] -= 1
    if Children[Parent[i]] == 0 **then**
        Leafs.push(Parent[i])
    end
end

$t(n-t)$, for any $t = 1, \dots, n-1$, regardless of the nodes that have been infected up to time t, *i.e.*, $|C_{b^v}(t)| = t(n-t)$, for any $v \in V$ and any rooted sequence b^v starting at v. Due to this inherent symmetry of the complete graph, equation (4.1) can be greatly simplified, leading to:

$$P_K(\tau|\text{root} = v) = \sum_{b^v \in B(\tau_v)} P_K\left((e_1, \dots, e_{n-1}) = b^v\right) = \frac{|B(\tau_v)|}{(n-1)!^2}\,, \tag{5.3}$$

where $|B(\tau_v)|$ denotes the number of edge sequences rooted at v which generates the rooted tree τ_v. In essence, when the underlying graph is complete, the probability of an edge sequence does not depend on the specific sequence. This implies that the ML estimator r^*_{ML} for the epidemic source is given by the node v which maximizes $|B(\tau_v)|$.

Let us point out that in a vertex-transitive graphs, equation (5.3) will not hold in general. For example, in the $3d$-hypercube, there exist two different rooted sequences which generate the same rooted tree but with different probabilities. Similar counter-example can be found in the complete (balanced) bipartite graph $K_{3,3}$.

Theorem 5.4. *If $G = K$, then for every spanning tree τ of K, it holds that*

$$r^*_{\mathrm{EPI}}(K, \tau) = r^*_{\mathrm{ML}}(K, \tau) \, ,$$

ruling out possible tie breaking.

Thus, on a complete graph, the two estimators for the epidemic source, ML e EPI, coincide. Note that, however, $r^*_{\mathrm{EPI}}(K, \tau)$ depends only on τ (not on any probability model), whereas $r^*_{\mathrm{ML}}(K, \tau)$ depends on the probability model that generates τ (see, Eq. (4.3)). Moreover, due to Theorem 5.4, for a complete graph Algorithm 2 also computes the MLE for the epidemic source (and equivalently, the rumor and distance center).

Theorem 5.4. In light of equation (5.3) and Proposition 4.3, the ML estimator for the epidemic source of a spanning tree of a complete graph will be the node v which minimize $\prod_{u \neq v} |\tau^v_u|$. Thus, it holds that $r^*_{\mathrm{ML}}(K, \tau) = r^*_{\mathrm{RC}}(\tau)$ (see, the first remark in Sect. 5). Furthermore, for any tree it holds $r^*_{\mathrm{RC}}(\tau) = r^*_{\mathrm{DC}}(\tau)$. Finally, given that $r^*_{\mathrm{EPI}}(K, \tau) = r^*_{\mathrm{DC}}(\tau)$ concludes the proof. □

Remark 5.5. Another example of a vertex-transitive graph for which Theorem 5.4 holds, is the cycle graph $\mathcal{C}$. Indeed, in a cycle graph the size of the edge cut induced by the epidemic model is always equal to 2, for every time t. Thus, equation (4.1) reduces to

$$P_{\mathcal{C}}(\tau | \mathrm{root} = v) = \sum_{b^v \in B(\tau_v)} P_K\left((e_1, \ldots, e_{n-1}) = b^v\right) = \frac{|B(\tau_v)|}{2^{n-1}} \, .$$

Note that the cycle graph yields a relatively simple scenario because all spanning trees of $\mathcal{C}$ are isomorphic paths, contrary to the spanning trees of the complete graph.

6. Numerical evaluation

Is the epicenter the epidemic source? What is the distance between the epicenter and the epidemic source? Is the epicenter significantly more accurate than the distance center? Clearly, the answers to such questions depend on the underlying graph structure, and this section provides empirical evidence to address them.

6.1. Graph models and performance metrics

The following graph models are considered in the characterization and evaluation of the epicenter [16]:

- Complete Graph (**CO**): The complete graph has no structure since all possible edges are present.
- 2-dimensional Torus (**TO**): Nodes are arranged in a 2-dimensional square lattice with size length of $\sqrt{n}$, yielding a total of n nodes where every node has degree 4.
- Erdös-Réyni random graph (**ER**): This classic random graph model has n nodes and every node pair is connected by an edge with probability p, independently.
- Watts-Strogatz random graph (**WS**): Also known as *Small World*, this random graph model starts from n nodes arranged in a ring and for each node add edges to nodes at distance k or less on the ring. Every edge is then rewired with probability p, choosing one of its endpoint uniformly at random among the nodes. This sparse random graph model yields high clustering and short distances.
- Barabási-Albert random graph (**BA**): This generative random graph model follows the *preferential attachment* principle, adding k edges per node. The model generates graphs with power-law degree distribution and short distances (but low clustering).

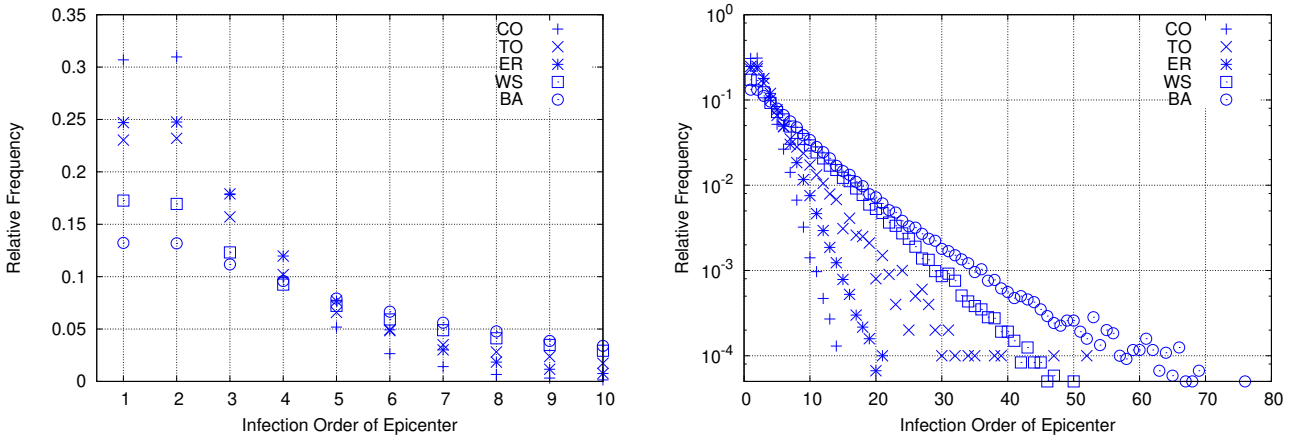


FIGURE 3. Infection order of epicenter for different graph models (1 corresponds to epidemic source). Left plot shows first 10 infections; right plot is in semi-log scale.

Given a random graph instance G, the epidemic source is a node chosen uniformly at random and the epidemic process is simulated on G to generate the epidemic tree. Let r denote the epidemic source and τ the tree generated. Given G and τ, let node $s = \text{epi}(G, \tau)$ denote its epicenter. Let $o(v)$ denote the time at which node v became infected; in particular $o(v) = t$ if, and only if, $I(t) \setminus I(t-1) = \{v\}$. Note that $o(r) = 1$ as r is the actual epidemic source. The effectiveness of the epicenter in identifying the epidemic source is captured by $o(s)$ as this indicates the time at which the epicenter became infected. Note that if $o(s) = 1$ then the epicenter is the epidemic source.

Another important characteristic is the distance between the epicenter and the epidemic source on τ. Intuitively, the epicenter should be close to the epidemic source even if this node was infected rather late in the epidemic. In particular, the epicenter should be much closer to the epidemic source than most nodes on the tree. Let $d(s)$ denote the hop distance on the tree τ between node $s = \text{epi}(G, \tau)$ and node r, the epidemic source. Note that when $d(s) = 0$ the epicenter is the epidemic source.

The following methodology is adopted to characterize $o(s)$ and $d(s)$. Consider R independent runs of the simulator, each generating a random graph instance and a random epidemic tree τ_j, for $j = 1, \ldots, R$. Let f_i denote the fraction of runs that the node returned by Algorithm 1 or 2 was the i-th infected node in the simulation, *i.e.*, $f_i = 1/R \sum_{j=1}^{R} \mathbb{1}(o(\text{epi}(G, \tau_j)) = i)$, where $\mathbb{1}$ is the indicator function. Note that f_1 is the fraction of runs where the algorithm identified the epidemic source. Similarly, let g_i denote the fraction of runs that the node returned by Algorithm 1 or 2 was at distance i from the epidemic source on tree τ_j, *i.e.*, $g_i = 1/R \sum_{j=1}^{R} \mathbb{1}(d(\text{epi}(G, \tau_j)) = i)$. Note that g_0 is the fraction of runs where the algorithm identified the epidemic source, and thus, $f_1 = g_0$.

In what follows, the graphs have size $n = 1000$ (with the exception of TO which has $n = 1024$ nodes) and for all random graphs the the average degree $\overline{d} \in \{6, 12\}$. Note that $\overline{d}$ determines the parameters for each model accordingly: in ER, $p = \overline{d}/1000$, in WS, $k = \overline{d}/2$ and $p = 0.05$, and in BA, $k = \overline{d}/2$. Algorithm 2 was used on CO and TO (since they are vertex-transitive) and Algorithm 1 was used on the random graph models. In all scenarios, $R = 1.2 \times 10^5$ runs (or higher). Thus, the standard error for the average accuracy reported in any scenario is always less than 0.0014, giving rise to very small confidence intervals that have been omitted from the results.

6.2. Infection order of the epicenter

Figure 3 shows the fraction of time that the epicenter is the i-th infected node where the left plot shows a restricted range. Interestingly, the trend is similar for all graph models: the values for f_i decrease monotonically

and fast with i and f_1 (correct identification of the epidemic source) shows the highest value. However, f_1 greatly depends on the graph model being highest for CO (at 31%) and lowest for BA (at 13%). Clearly, the power-law degree distribution of BA poses a challenge in identifying the epidemic source. Recall that for CO the epicenter coincides with the MLE so their accuracy is just 31%. For ER graphs, the accuracy of the epicenter is 24%, a value relatively high given its sparseness and structure in light of the complete graph.

It is curious that f_1 and f_2 are practically identical for all graph models. Indeed, for vertex-transitive graphs (such as CO and TO), given just the first edge of the epidemic tree, there is no information on which of the two nodes is the epidemic source. The epidemic tree that will be generated and hung on each of these two nodes are statistically equivalent. Thus, the algorithm returns each one of them with the same frequency. While this is not the case for general graphs, such information depends on the degrees of the nodes, and its connection to the epicenter should be further studied.

The right plot of Figure 3 is also revealing that nodes infected late in the epidemic process are never identified as the epicenter. Again, this clearly depends on the graph model: for CO, only the first 15 infected nodes were ever identified as the epicenter, while for BA this number is around 75 (still very small if compared to $n = 1000$). Moreover, the apparent straight line for each model shown in the semi-log plot indicates that f_i has an exponential decay. Again, the decay rate (*i.e.*, slope) depends on the graph model with BA being the slowest.

6.3. Distance to the epicenter

Figure 4 shows the relative frequency of distances between the epidemic source and the epicenter (left) and other nodes (right). Recall that $g_0 = f_1$ and this indicates that the epicenter corresponds to the epidemic source. Interestingly, all graphs models show a very similar trend: g_1 is much larger than g_0 and then it decreases monotonically and fast (with the exception of BA, where the peak is in g_2). Thus, the epicenter is more likely to be a neighbor of the epidemic source (on the epidemic tree) than the epidemic source itself. This results follows from the fact that the epidemic source has at least one neighbor in the epidemic tree (and possibly more) and the second infected node is in this neighborhood.

The peak in g_2 for BA in Figure 4 is a consequence of its power-law degrees and the uniform choice for the epidemic source. In particular, the source is likely to be a small degree node that has as neighbor a high degree node which in turn has many other neighbors. Thus, the epicenter is often a node that has a common neighbor with the epidemic source, and thus is at distance two.

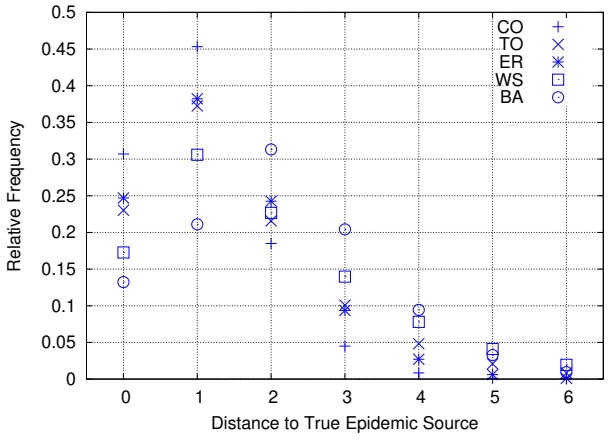
The right plot of Figure 4 shows the distance between the epidemic source and all other nodes on the epidemic tree (empirical distribution). Interestingly, distances follow a bell-shaped curve independent of the graph model. However, the mode and variance of the distribution strongly depend on the graph model. In particular, note that CO and BA are quite similar and generate trees where nodes are closest to the epidemic source. In sharp contrast, distances are much larger for WS and even more so for TO (distribution for TO is not shown completely).

6.4. Comparison with distance center

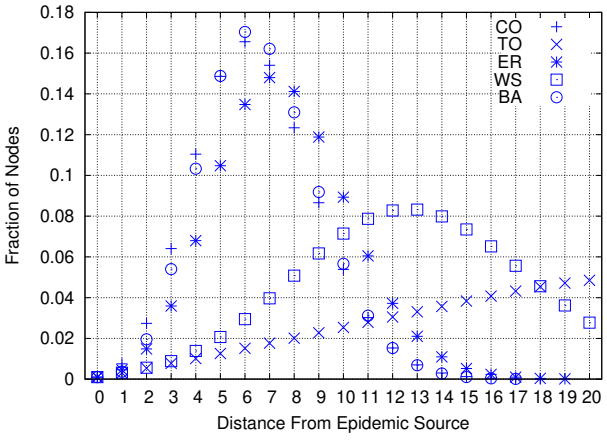
While the epicenter of a spanning tree τ of vertex-transitive graph G is equivalent to the distance center of τ, this is clearly not the case for arbitrary graphs. Figure 5 shows a comparison between the accuracy of the epicenter (given by f_1) and the distance center in identifying the epidemic source (computed on the exact same spanning trees generated for each run).[4]

Note that for all random graph models, the accuracy of the epicenter is significantly higher than that of the distance center. The relative improvement depends on the graph model and average degree: for ER the epicenter is around 26% more accurate than the distance center, but for BA this superiority is around 226% (3.26 times more accurate), for $\overline{d} = 6$. Note that with an average degree $\overline{d} = 12$ both estimators improve their accuracy (with respect to $\overline{d} = 6$) and their relative difference becomes smaller. Indeed, as the density of the random graphs increases, the diversity of paths of short length also increases and the underlying network becomes less

[4]Note that the distance center is identical to the rumor center since the input is the epidemic tree τ. Moreover, G is arbitrary and has no information concerning the epidemic and thus cannot be used to estimate the epidemic source.



(A) Distance between epicenter and epidemic source.



(B) Distance from epidemic source on epidemic tree.

FIGURE 4. Distribution of distances between epicenter and epidemic source (*left*), and distances from epidemic source for the epidemic tree (*right*).

TABLE 1. Average distance from epicenter (EPI column) and distance center (DC column) to the epidemic source given that the source was not identified (*i.e.*, distance is greater than zero). Average distance from source to all nodes on epidemic tree (All column). For all graphs $n = 1000$ (except TO that has $n = 1024$), and for all random graphs $\overline{d} = 6$.

Graph	EPI	DC	All
CO	1.441	1.441	6.496
TO	1.921	1.921	19.56
ER	1.710	1.709	7.557
WS	2.367	2.751	13.04
BA	2.366	2.322	6.639
SB	1.722	1.742	7.573

informative (recall the complete graph). In any case, the epicenter leverages the underlying graph in combination with τ to provide a better estimator for the epidemic source.

Beyond accuracy, it is interesting to compare the distances between the epicenter and the distance center to the epidemic source given that they have not identified the epidemic source. Table 1 shows the average distance for the epicenter and distance center (given they have not identified the source) as well as the average distance between the source and all nodes on the tree. The average results for the epicenter are supported by Figure 4. Interestingly, the average distance of the distance center are very similar to that of the epicenter, while much smaller than the average for the entire tree. Thus, given that the epicenter and distance center failed to identify the source, the nodes identified are at a similar distance to the epidemic source (on average).

6.5. Comparison on real networks

While random network models allows for a more principled evaluation of epidemic source estimators, real networks often exhibit structural features that are not captured by such models. Thus, two real social contact networks are considered, both publicly available. The Haslemere dataset is the result of a three-day experiment with 469 volunteers in the town of Haslemere that were tracked continuously using a mobile phone app [12]. The dataset has also been used to evaluate localized COVID-19 control strategies [7]. In the network generated from the raw dataset, an edge between two individuals is present if they were closer than 12m in any time step (this

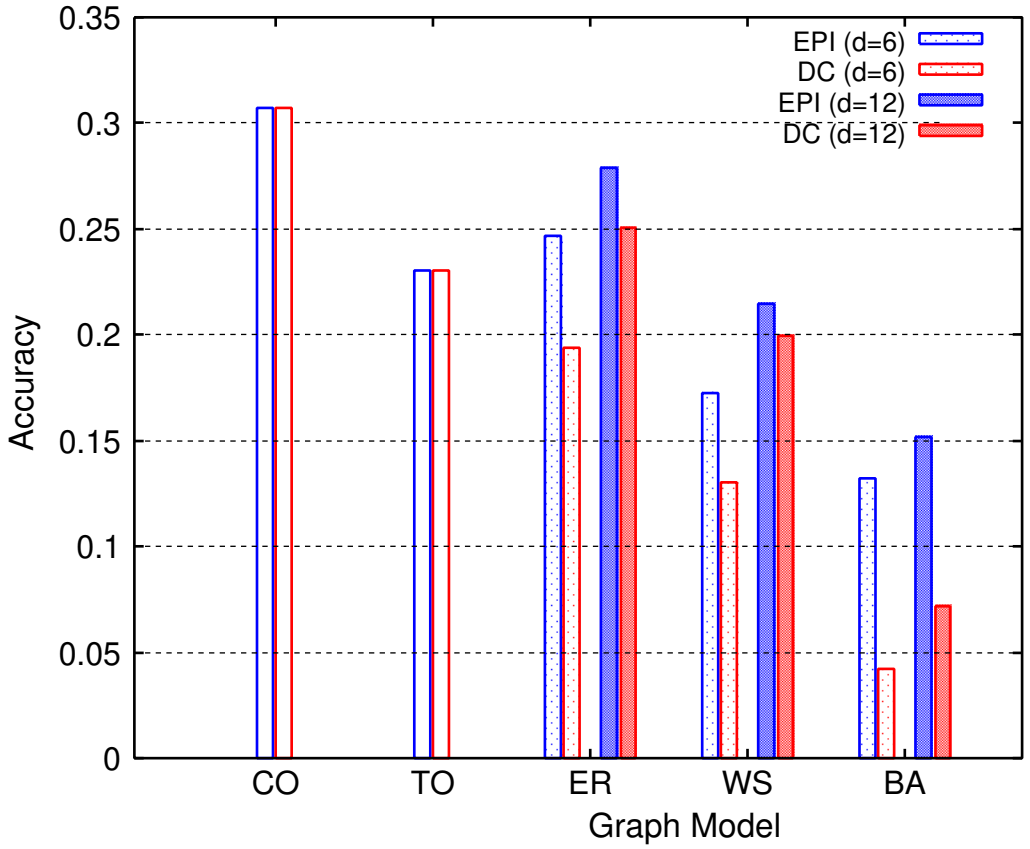


FIGURE 5. Average accuracy of epicenter and distance center for different graph models. For each random graph model, the average degree $\overline{d} \in \{6, 12\}$.

threshold was required to obtain a connected graph). This network has 449 nodes and average degree of 9.3. The US High School dataset is a high-resolution (in time and space) recording of contacts between 788 individuals during a typical day at an American high school [22]. In the network generated from the raw dataset, an edge between two individuals is present if they were together (less than 3 m apart) for more than 5 minutes (this time threshold was required to obtain a connected graph). This network has 786 nodes and average degree of 51.8.

Figure 6 shows a comparison between the accuracy of the epicenter (as predicted by Algorithm 1) and the distance center in identifying the epidemic source (computed on the exact same spanning trees generated for each run, using f_1 and $R = 10^3$ runs). For the Haslemere network, the epicenter is 54% more accurate than the distance center (17.8% versus 11.5% accuracy, respectively). For the US High School network, the epicenter is only 13% more accurate. While the US High School network is larger, it is also significantly denser (the average degree is five times larger). Interestingly, as with the random graph models (Fig. 5), the relative superiority of the epicenter over the distance center seems to decrease with the increase of the average degree.

6.6. Convergence of epicenter accuracy

While the previous results characterize the epicenter for a fixed graph size of $n = 1000$, it is interesting to consider its dependence on n. Growing n in random graph models generally requires scaling the average degree and such choice has a fundamental impact on the graph structure. For example, if the average degree is kept

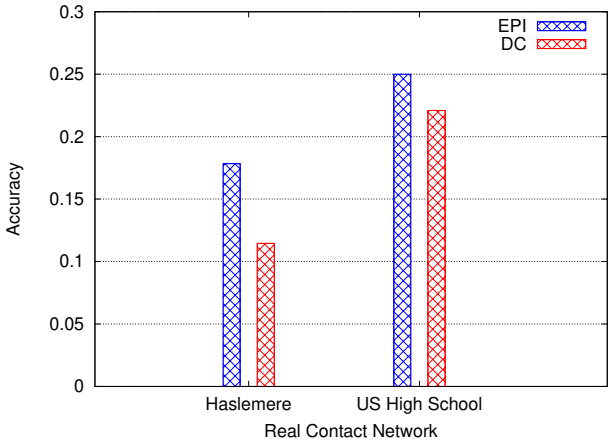


FIGURE 6. Average accuracy of epicenter and distance center for two real social contact networks (Haslemere with $n = 449$ and $\overline{d} = 9.3$, and US High School with $n = 786$ and $\overline{d} = 51.8$).

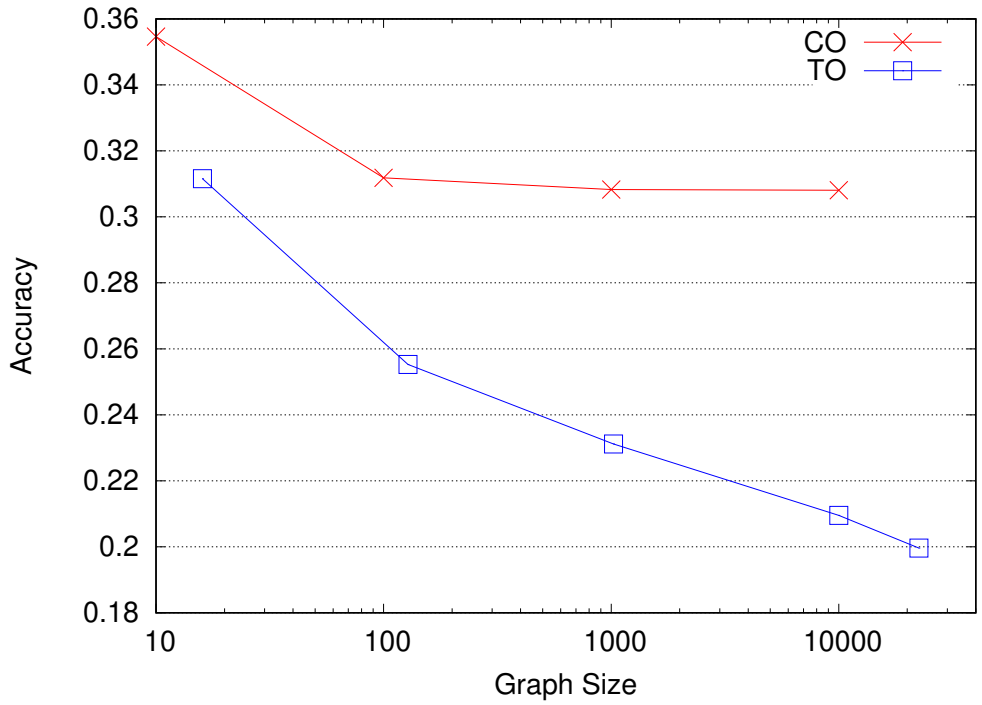


FIGURE 7. Accuracy of epicenter for CO and TO as a function of the graph size (n).

constant as n grows in ER model, then the graph is surely not connected as n grows. However, if the average degree grows as $\Omega(\log n)$ then the graph is surely connected. To avoid this difficulty, only CO and TO are considered in the following evaluation since their degrees is a fixed function of n (*i.e.*, $n-1$ for CO and 4 for TO).

Figure 7 shows the accuracy of the epicenter (f_1) as a function of n for CO and TO (x-axis in log-scale). Interestingly, the accuracy of CO is larger for smaller graphs but converges relatively quickly to about 0.308.

Thus, larger graphs and epidemic trees do not provide additional structural information to increase the accuracy of the epicenter. On the positive side, it also does not diminish the accuracy of the epicenter!

However, the story is quite different for TO (also in Figure 7, since the accuracy decreases monotonically with n. In this case, larger graphs and epidemic trees reduce the accuracy of the epicenter. A key distinctive feature with respect to CO are the distances both on the graph and on the epidemic tree. While all distances on CO are 1, in TO it scales with $\sqrt{n}$ (recall that the graph is a square lattice with n nodes). Moreover, Figure 4 (left plot) and Table 1 indicate that distances from the source on the epidemic tree are significantly smaller for CO, and in particular the average distance for TO is three times larger (see Tab. 1). Thus, since the epicenter is determined by the sum of distance differences and such distances increase with n, the estimator seems to become less accurate with the increase of distances.

The dependence of epicenter accuracy on the underlying graph size and structure is an interesting and challenging question. On a vertex-transitive graph of size n, in light of Proposition 5.2, the probability of correct detection is equal to the probability that the size of the tree dangling from all the neighbors of the epidemic source is less than $n/2$. Specifically, for a complete graph of size n, the epicenter accuracy corresponds to the probability that in the Chinese restaurant process at time n all occupied tables have less than $n/2$ people (each table corresponds to a subtree hanging from the epidemic source). In any case, it can be argued that the epicenter accuracy converges, as n tends to infinity, to $1 - \ln 2$, in accordance with simulation results shown in Figure 7.

The asymptotic result above is related to the accuracy of the rumor center on infinite d-regular trees which also converges to $1 - \ln 2$, as d tends to infinity [24]. Recall that in this model, for any fixed d, the probability of correct detection is computed when the time (in our case n) goes to infinity. This is possible since the d-regular tree is infinite. In our case, since the graph is finite, the epidemics stops at time n (size of the graph). Despite a technical issue (related to the double limit in [24], first letting time go to infinity and then d, while in our case there is only one limit in n) there is an intuitive explanation why these two results are related. In particular, for any given n, one can find a $d_0 = d_0(n)$ sufficiently large, such that the epidemic process up to time n on a complete graph of size n or on a d_0-regular tree, are both essentially evolving according to recursive random uniform tree; *i.e.*, for every $t \leq n$ the node infected at time t connects uniformly at random to one of the already infected nodes in the tree.

For a Torus of size n, the probability of correct detection also corresponds to the probability that the size of the tree dangling from all the neighbors of the epidemic source is less than $n/2$. However, in this case, the comparison with the Chinese restaurant process does not hold. On the one hand, the epidemic source has at most four neighbors (thus the number of possible tables is bounded) and, on the other hand, the probability of a node infected at time $t \leq n$ to be connected to one of these subtrees need not be proportional to the subtree size. In particular, this probability might even be zero, since every node has at most four neighbors and all neighboring nodes of one subtree might have already been infected by others. The behavior observed for the Torus (TO) in Figure 7 showing that the accuracy decreases monotonically as n goes to infinity might be explained by the fact that one of the four possible dangling subtrees of the epidemic source will eventually dominate the others, and therefore for n sufficiently large its size will be larger than $n/2$. Establishing the limiting accuracy of the epicenter on the torus and other graphs is an open question that subject of future investigation.

7. Conclusion

The problem of identifying the epidemic source by partially observing a network epidemic is quite fundamental and has been explored over the last decade both from a practical and theoretical perspective. While most works assume that infected nodes are observed (or partially observed) at a given point in time during the epidemic, this work observes the epidemic after it terminates. Note that in this scenario, observing the set of infected nodes provides no information, as this is simply the set of nodes of the graph. Thus, the observation here is the undirected and unrooted tree (a spanning tree of the graph) that encodes the edges responsible for infections, but not their direction (who infected whom).

TABLE 2. Notation.

$G = (V, E)$	Graph
τ	Tree
$V(\tau), E(\tau), \lvert\tau\rvert$	Set of nodes, set of edges and size (number of nodes) of the tree τ
τ_v	Tree rooted at v
$B(\tau_v)$	Set of edge sequences rooted at v that generate the rooted tree τ_v
τ_u^v	Subtree of τ_v rooted at u obtained by removing the neighbor of u in the unique path between u and v in τ_v (if $u = v$, $\tau_u^v = \tau_v$)
$d_G(u, v)$	Graph distance between node u and v in G
$d_G(v)$	Distance centrality of v in G ($d_G(v) := \sum_{u \in V} d_G(v, u)$)
$\mathcal{P}_G(u, v)$	Set of shortest paths in G between u and v
$V_v^u(G)$ or V_v^u	Set of nodes w for which there exists a shortest path between w and u crossing v in G
	$V_v^u(G) := \{w \in V : \exists p \in \mathcal{P}_G(w, u) \text{ such that } p \ni v\}$
$r^*_{\mathrm{EPI}}(G, \tau)$	Epicenter of τ in G
	$r^*_{\mathrm{EPI}}(G, \tau) := \arg\min_{v \in V} \sum_{u \in V} \Big(d_\tau(v, u) - d_G(v, u)\Big)$

The proposed epicenter is an estimator that leverages both the epidemic tree and the graph to estimate the epidemic source (and thus the direction of all edges) by identifying the node that better aligns the graph and the tree in terms of their distances to corresponding nodes. The epicenter can be computed in time $O(nm)$ in general, and $\Theta(n)$ when the graph is vertex-transitive. Moreover, the epicenter and distance center (of the epidemic tree) coincide for vertex-transitive graphs. However, numerical simulations indicated that the epicenter is always more accurate than the distance center (for all random graph models considered), with larger relative improvements for sparse and power-law graphs.

Last, numerical results strongly suggest that when the epicenter (and distance center) makes a wrong prediction, this is often close (in the tree) to the epidemic source. Can a new estimator that leverages the epicenter be designed to correct for such mistakes? This seems possible, opening the doors to further explorations.

APPENDIX A. PROOF SKETCH OF LEMMA 4.2

Given the rooted tree τ_v, we identify $d_v = |N_v|$ rooted subtrees τ_u^v, with $u \in N_v$. Consider a rooted sequence which generates τ_v. Assume for the moment that such an edge sequence corresponds to building one of the d_v branches of τ_v entirely before moving to a different branch and so on until all branches have been generated. The number of possible ways to do that is accounted in the first term of equation (4.2), *i.e.*, $\prod_{u \in N_v} |B(\tau_u^v)|$ (the subtrees τ_u^v, with $u \in N_v$ are disjoint, thus the product). Note that, for every $u \in N_v$, the size of every rooted edge sequence which generates τ_u^v is $|\tau_u^v| - 1$, and $\sum_{u \in N_v} |\tau_u^v| = |\tau_v| - 1$. Thus, when concatenating a rooted edge sequence for each τ_u^v we obtain an edge sequence of size $|\tau_v| - 1 - d_v$, despite a rooted sequence which generates τ_v having size $|\tau_v| - 1$. However, given that for every $u \in N_v$ there is only one edge connecting v to u, a concatenation of the d_v sequences can be modified in a unique manner to give rise to a rooted edge sequence which generates τ_v, namely by inserting the edge $\{v, u\}$ right before the corresponding edge sequence generating τ_u^v.

Note that a rooted edge sequence which generates τ_v, does not necessarily correspond to a concatenation of d_v rooted sequences corresponding to the d_v different branches. In particular, a rooted edge sequence which generates τ_v may correspond to alternating between the edges of the rooted sequences generating the different branches. In order to account for the latter, we need to compute the number of different ways a given concatenation of the rooted sequences generating the subtrees τ_u^v, with $u \in N_v$, can be rearranged to give rise to different rooted edge sequence which generates τ_v. The second factor in equation (4.2), *i.e.*, $\frac{(\sum_{u\in N_v} |\tau_u^v|)!}{\prod_{u\in N_v} |\tau_u^v|!}$ accounts for this number.

References

[1] S. Bubeck, L. Devroye and G. Lugosi, Finding Adam in random growing trees. *Random Struct. Algor.* **50** (2017) 158–172.

[2] K. Cai, H. Xie and J.C.S. Lui, Information spreading forensics via sequential dependent snapshots. *IEEE/ACM Trans. Netw.* **26** (2018) 478–491.

[3] W. Dong, W. Zhang and C.W. Tan, Rooting out the rumor culprit from suspects, in IEEE International Symposium on Information Theory (2013) 2671–2675.

[4] M. Draief and L. Massouli, Epidemics and rumours in complex networks. Cambridge University Press (2010).

[5] M. Drmota, Random trees: an interplay between combinatorics and probability. Springer Science & Business Media (2009).

[6] S. Feizi, M. Médard, G. Quon, M. Kellis and K. Duffy, Network infusion to infer information sources in networks. *IEEE Trans. Netw. Sci. Eng.* **6** (2019) 402–417.

[7] J.A. Firth, J. Hellewell, P. Klepac, S. Kissler, A.J. Kucharski and L.G. Spurgin, Using a real-world network to model localized COVID-19 control strategies. *Nat. Med.* **26** (2020) 1616–1622.

[8] A. Frieze and W. Pegden, Looking for vertex number one. *Ann. Appl. Probab.* **27** (2017) 582–630.

[9] J. Jiang, S. Wen, S. Yu, Y. Xiang and W. Zhou, K-center: an approach on the multi-source identification of information diffusion. *IEEE Trans. Inf. Forensics Secur.* **10** (2015) 2616–2626.

[10] J. Jiang, S. Wen, S. Yu, Y. Xiang and W. Zhou, Identifying propagation sources in networks: state-of-the-art and comparative studies. *IEEE Commun. Surv. Tutor.* **19** (2017) 465–481.

[11] N. Kamiyama, Arborescence problems in directed graphs: theorems and algorithms. *Interdiscip. Inf. Sci.* **20** (2014) 51–70.

[12] P. Klepac, S. Kissler and J. Gog, Contagion! The BBC four pandemic – the model behind the documentary. *Epidemics* **24** (2018) 49–59.

[13] A. Kumar, V.S. Borkar and N. Karamchandani, Temporally agnostic rumor-source detection. *IEEE Trans. Signal Inf. Process. Netw.* **3** (2017) 316–329.

[14] G. Lugosi and A.S. Pereira, Finding the seed of uniform attachment trees. *Electr. J. Probab.* **24** (2019) 15 pp.

[15] W. Luo, W.P. Tay and M. Leng, How to identify an infection source with limited observations. *IEEE J. Selected Topics Signal Process.* **8** (2014) 586–597.

[16] M. Newman, Networks. Oxford University Press (2018).

[17] C. Nowzari, V.M. Preciado and G.J. Pappas, Analysis and control of epidemics: a survey of spreading processes on complex networks. *IEEE Control Syst. Mag.* **36** (2016) 26–46.

[18] R. Paluch, X. Lu, K. Suchecki, B.K. Szymański and J.A. Hołyst, Fast and accurate detection of spread source in large complex networks. *Sci. Rep.* **8** (2018) 1–10.

[19] R. Pastor-Satorras, C. Castellano, P. Van Mieghem and A. Vespignani, Epidemic processes in complex networks. *Rev. Mod. Phys.* **87** (2015) 925.

[20] R. Pastor-Satorras and A. Vespignani, Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86** (2001) 3200.

[21] P.C. Pinto, P. Thiran and M. Vetterli, Locating the source of diffusion in large-scale networks. *Phys. Rev. Lett.* **109** (2012) 068702.

[22] M. Salathé, M. Kazandjieva, J.W. Lee, P. Levis, M.W. Feldman and J.H. Jones, A high-resolution human contact network for infectious disease transmission. *Proc. Natl. Acad. Sci.* **107** (2010) 22020–22025.

[23] D. Shah and T. Zaman, Detecting sources of computer viruses in networks: theory and experiment, in ACM SIGMETRICS Perf Eval Rev, vol. 38 (2010) 203–214.

[24] D. Shah and T. Zaman, Rumor centrality: a universal source detector, in ACM SIGMETRICS Perf Eval Rev, vol. 40 (2012) 199–210.

[25] S. Shelke and V. Attar, Source detection of rumor in social network – a review. *Online Social Netw. Media* **9** (2019) 30–42.

[26] W. Tang, F. Ji and W.P. Tay, Estimating infection sources in networks using partial timestamps. *IEEE Trans. Inf. Forens. Secur.* **13** (2018) 3035–3049.

[27] P.-D. Yu, C.W. Tan and H.-L. Fu, Epidemic source detection in contact tracing networks: epidemic centrality in graphs and message-passing algorithms. *IEEE J. Selected Top. Signal Process.* **16** (2022) 234–249.

[28] K. Zhu, Z. Chen and L. Ying, Catch'em all: Locating multiple diffusion sources in networks with partial observations, in AAAI Conference on Artificial Intelligence (2017).