Observer Placement for Source Localization: The Effect of Budgets and Transmission Variance

Brunella Spinelli, L. Elisa Celis, Patrick Thiran

Abstract*— When an epidemic spreads in a network, a key question is where was its *source*, i.e., the node that started the epidemic. If we know the time at which various nodes were infected, we can attempt to use this information in order to identify the source. However, maintaining *observer* nodes that can provide their infection time may be costly, and we may have a *budget* k on the number of observer nodes we can maintain. Moreover, some nodes are more informative than others due to their location in the network. Hence, a pertinent question arises: *Which nodes should we select as observers in order to maximize the probability that we can accurately identify the source?

Inspired by the simple setting in which the node-to-node delays in the transmission of the epidemic are deterministic, we develop a principled approach for addressing the problem even when transmission delays are random. We show that the optimal observer-placement differs depending on the *variance* of the transmission delays and propose approaches in both low- and high-variance settings. We validate our methods by comparing them against state-of-the-art observer-placements and show that, in both settings, our approach identifies the source with higher accuracy.

I. Introduction

Regardless of whether a network comprises computers, individuals or cities, in many applications we want to detect whenever any anomalous or malicious activity spreads across the network and, in particular, where the activity originated.[1] We call the spread of such activity an *epidemic* and the originator the *source*.

Clearly, monitoring all nodes is not feasible due to cost and overhead constraints: The number of nodes in the network may be prohibitively large and some of them may be unable or unwilling to provide information about their state. Thus, studies have focused on how to estimate the source based on information from a few nodes (called *observers*). Given a set of observers, many models and estimators for source localization have been developed [25], [20], [31]. However, the *selection* of observers has not yet received a satisfactory answer: Most of state-of-the-art methods are based on common centrality heuristics (e.g., degree- or betweenness-centrality) or on more advanced heuristic approaches that do not directly optimize source localization (see [31] for a survey) or are limited to simple networks such as trees (e.g., [15]). Moreover, such methods consider only the structure of the network when placing observers. However, depending on the particular epidemic, the expected transmission delay between two nodes, and its variance, can differ widely. We show that different transmission models require different observer placements: This is illustrated in Figure 1: As the variance of the transmission delays changes, the optimal set of observers also changes (see also Figure 2 for a concrete example).

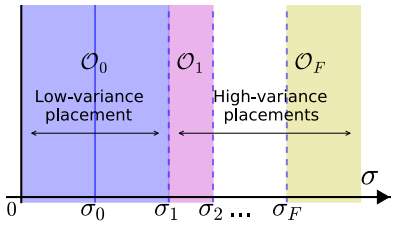


Fig. 1. Transmission variance σ and optimal observer placement. For $\sigma \in (0, \sigma_0)$ the transmission delays are effectively deterministic. For $\sigma \in (\sigma_0, \sigma_1)$ the variance σ affects the accuracy of source localization but the optimal observer placement is still $\mathcal{O}_0$. For larger σ, the optimal observer placement may change, possibly multiple times ($\mathcal{O}_k$ denotes the optimal placement for $\sigma \in (\sigma_k, \sigma_{k+1})$) up to $\sigma = \sigma_F$. For $\sigma > \sigma_F$ the optimal placement remains the same ($\mathcal{O}_F$).

The difficulties faced in finding the optimal observers are two-fold. First, computing the likelihood of a node being the source conditional on the available observations can be computationally prohibitive [28], [25]; evaluating the probability of detection given a set of observers is, in general, even harder. Second, the optimal selection of a limited number of observers is NP-hard, even when the transmission times are deterministic. We take a principled approach that begins with considering deterministic transition delays, and build on this intuition in order to develop heuristics for both the low-variance and high-variance regimes.

A. Model and Problem Statement

Our Transmission Model. We assume that the contact network $\mathcal{G} = (V, E, w)$ is known and is *weighted*. The weight $w_{uv} \in \mathbb{R}_+$ of edge $uv \in E$ is the mean of the *transmission delay* encoded by the random variable X_{uv}; this is the time it would take for u to infect v.[2] This transmission model is both natural and versatile as it comprises deterministic transmissions (i.e., if $X_{uv} = w_{uv} \in \mathbb{R}_+$ a.s. for all edges $uv \in E$), which we call *zero-variance*, and arbitrary *random* independent transmission models. It naturally captures the SI epidemic model adopted, e.g., in [25], [22] and related SIR/SIS/SEIR models (see [14] and the discussion in [31]). We study, in particular, how the *amount* of randomness (i.e.,

[1] In effect, we wish to answer questions such as *what was the origin of a worm in a computer network?*, *who was the instigator of a false rumor in a social network?* and *can we identify patient zero of a virulent disease?*

[2] For ease of presentation we assume the graph is undirected and $w_{uv} = w_{vu}$; however our definitions and approach extend straightforwardly to the directed case.

the variance of X_{uv}) in the transmission delays affects the choice of observers for source localization. Towards this, we are the first to separately analyze two different regimes for the amount of randomness in transmission delays: *low-variance* and *high-variance*. A dichotomy exists between the two, and our approach for observer placement differs.

Our Source Estimation. We assume that there is a single source that initiates the epidemic[3] and let $\mathcal{O} \subseteq V$ (which we will select) be the set of observer nodes. We assume we know the time at which each observer is infected, and refer to this vector of infection times as $T_{\mathcal{O}}$. This is a standard (see, e.g., [23]) and realistic assumption (for example, clinics keep records of patients and carefully record outbreaks so can provide such information). To identify the source, we use this (and only this) information.

We use maximum likelihood estimation (MLE) to produce an estimate $\widehat{s}$ of the true unknown source s^* as in [25],[4]i.e.,

$$\widehat{s} \in \text{argmax}_{s \in V} \mathrm{P}(T_{\mathcal{O}} | s^* = s) \mathrm{P}(s^* = s).$$

We assume the prior on s^* is uniform unless otherwise specified (i.e., $\mathrm{P}(s^* = s) = 1/n$ for all nodes $s \in V$ where $n = |V|$).

Our Observer Placement. We assume that we are given a *budget* k on the number of observers we can use, and that we must select our observers *once and for all*. In order to select the *best set of observers* $\mathcal{O}$ *of size* k we must first define our metric of interest. We consider the two metrics proposed by [15], although variations (including worst-case versions) exist [15]:

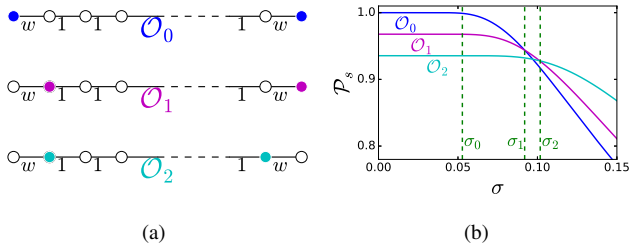
1) the *success probability* $\mathcal{P}_s = \mathrm{P}(\widehat{s} = s^*)$, and
2) the *expected distance* between estimated source and real source, i.e., $\mathrm{E}[d(s^*, \widehat{s})]$ with d denoting the distance between two nodes in the network.

The two metrics might require different sets of observers [15], however we show experimentally that maximizing $\mathcal{P}_s$ is a good proxy for minimizing $\mathrm{E}[d(s^*, \widehat{s})]$ (see Section III). Hence, due to space constraints, we focus on the minimization of the former.

B. Main Contributions

Low-Variance Regime. When the variance in the transmission delays is *low* (see Section III), we prove that the set of optimal observers is equal to the optimal set for the zero-variance regime. In the zero- and low- variance regime, both the probability of success $\mathcal{P}_s$ and the expected distance $\mathrm{E}[d(s^*, \widehat{s})]$ can be explicitly computed. Despite this seeming simplicity, the problem remains NP-hard. We tackle the problem by using its connection with the well-studied related Double Resolving Set (*DRS*) problem [6] that minimizes the number of observers for perfect detection.[5] From this connection we find inspiration for our algorithm that, by selecting one observer at a time until the budget is exhausted in order to reach a *DRS* set, greedily improves $\mathcal{P}_s$.

High-Variance Regime. When the noise in the transmission delays is *high* (see Section IV), it is no longer negligible and it poses an additional challenge to source localization; in effect, the accumulation of noise from node to node as the epidemic spreads might no longer enable us to distinguish between two potential sources, especially when they are both *far* from all observers. Hence, we must *strengthen* the requirements for observer placement in order to ensure that the nodes can be distinguished by observers that are *near* to them; this nearness is a function of the noise, the budget k, and the network topology. We define a novel objective function that both maximizes the success probability and imposes a *uniform* spread of observers in the network. Taking inspiration from the low-variance regime, we design an algorithm that greedily maximizes this new objective (see Section IV).



(a) (b)

Fig. 2. Optimal observers for Gaussian transmission delays with variance σ^2. (a): different observer placements; (b): their performance in terms of probability of success ($\mathcal{P}_s$) for $w = 20$ and 30 edges.

Empirical Results. Our methods perform favourably against state-of-the-art approaches in both the low- and high-variance regimes (see Section V-B). In Appendix III, for the low-variance regime, we further compare them against two other natural objective functions; we show that our approach is the best. Moreover, in the empirical results the dichotomy between the low- and high-variance regimes becomes apparent.

II. Related Work

The problem of source localization has been widely studied in recent years, we survey the works that are more relevant to ours and refer the reader to the survey by Jiang et. al. [10] for a more complete review of the different approaches.

Transmission delays. Many transmission models for epidemics have been studied [16] and considered for source localization. Although discrete-time transmission delays are common [21], [26], [3], in order to better approximate realistic settings, much work (including ours) adopt continuous-time models with varying distributions for the transmission delays; e.g., exponential [28], [22] or Gaussian [25], [20], [19], [31]. In the same line of the latter class of works, we use *truncated* Gaussian variables, which gives us the advantage of ensuring that infection delays are strictly positive.

[3]Our results can be extended to the case of multiple sources following the recent work by [32] on a related problem.

[4]This approach is common (see e.g., [28], [8]), although the exact form of the estimator depends on the model and assumptions.

[5]This minimum number is, in many cases, still prohibitively large, and can be as much as $n - 1$, hence we cannot use this approach directly.

Source localization. Many approaches [33], [26], [29], beginning with the seminal work by Shah and Zaman [28], rely on knowing the state of the *entire* network at a fixed point in time t; this is often called a *complete observation* of the epidemic. These models use maximum likelihood estimation (MLE) to estimate the source. The results of [28] have been extended in many ways, for example in the case of multiple sources [22] or to obtain a *local* source estimator [8]. An alternate line of work considers a complete observation of the epidemic, except that the observed states are *noisy*, i.e., potentially inaccurate [34], [29]. As assuming the knowledge of the state of all the nodes is often not realistic, *partial observation* settings have also been studied. In such a setting, only a subset of nodes $\mathcal{O}$ reveal their state. In this line of work, the observers are mainly *given*, either arbitrarily or via a random process, and the problem of *selecting* observers is not addressed. For example, when a fraction x of nodes are randomly selected, Lokhov et al. [18] propose an algorithm that relies on the knowledge of the state (S, I or R) of a fraction of the nodes in the graph at a given moment in time. This approach, however, crucially relies on the assumption that the starting time of the epidemic is known, which is often not realistic [10], [25]. When the nodes are independently selected to be observers, an approach to source estimation based on the notion of *Jordan center* was proposed [21] and has since been used in other work for source estimation, especially with regard to a *game theoretic* version of epidemics [9]. This line of work does not assume infection times are known, which we believe is, in many cases, an unnecessary limitation. Indeed by using infection times we can achieve exact source localization in the zero-variance setting with sufficiently many observers [7], whereas this is not true otherwise.

Observer placement. Natural heuristics for observer placement (e.g., using high-degree vertices or optimizing for distance centrality) were first evaluated under the additional assumption that infected nodes know which neighbor infected them [25]. Later, Luoni et al. [20] proposed, for a similar model, to place the observers using a Betweenness-Centrality criterion (which we use as a benchmark, see Section V-B), and extended it to noisy observations [19]. These and other heuristic approaches for observer placement are evaluated empirically by Seo et al. [27]; they reach the conclusion that, among the placements they evaluate, the Betweenness-Centrality criterion performs the best. In their work the source is estimated by ranking candidates according to their distance to the set of observers, without using the time at which the observers became infected. Once again, this approach is inherently limited by the fact that it does not make use of the time of infection. The problem of *minimizing* the number of observers required to detect the precise source (as opposed to *maximizing* the performance given a *budget* of observers) has been considered in the zero-variance setting. For trees, given the time at which the epidemic starts, the minimization problem was solved by Zejnilovic et al. [30]. Without assuming a tree topology and a known starting time, approximation algorithms have been developed towards this end [7] (still in a zero-variance setting). However, in a network of size n, the number of observers required, even if minimized, can be up to $n-1$, hence, a budgeted setting is practically more interesting. For trees, the budgeted placement of observers was solved by using techniques different from ours [15]. However these techniques heavily rely on the tree structure of the network and do not seem to be extendible to other topologies. In a very recent work, Zhang et al. [31] consider selecting a fixed number of observers using several heuristics such as Betweenness-Centrality, Degree-Centrality and Closeness-Centrality and they show that none of these methods are satisfactory. They introduce a new heuristic for the choice of observers, called *Coverage-Rate*, which is linked to the total number of nodes neighboring observers, and show that an approximated optimization of this metric yields better performance. Connecting the budgeted placement problem to the un-budgeted minimization problem, we provably outperform their approach in low-variance settings.[6] Moreover, the effect of the variance in the transmission delays is neglected by Zhang et al., leaving open the question of whether their approach works in general. However, we consider Coverage-Rate as one of our baselines.

III. The Low-Variance Regime

In this section, we focus on the low-variance regime. We start by introducing the setting and the definitions we adopt.

A. Preliminaries

Let $\mathcal{G} = (V, E)$ be an undirected weighted network. Assuming u is infected, the weight w_{uv} of edge $uv \in E$ represents the expected time it takes for u to infect v. As the network is undirected, we assume $w_{uv} = w_{vu}$ for all $uv \in E$.

We assume that the epidemic is initiated by a single unknown source s^* at an unknown time t^*. If a node u gets infected at time t_u, a non-infected neighbor v of u will become infected at time $t_v = t_u + X_{uv}$ where X_{uv} is a random variable with $\mathrm{E}[X_{uv}] = w_{uv}$.

The *time* t^* at which an epidemic starts is unknown. This adds a significant difficulty to the problem because a *single* observation is not *per se* informative. Instead, we must use the collection of *differences* between observed infection times. If the variance is zero or if it is low compared to edge weights, network distances are a good proxy for time delays (see Proposition 1). We refer to this setting as a *low-variance* regime, as opposed to the *high-variance* regime in which time delays are highly noisy and network distances no longer work as a proxy for time delays.

Distance vectors and equivalence between nodes. We start with a few definitions. Our setting is similar to [15].

Definition 1 (Equivalence): Let $\mathcal{G} = (V, E)$ and $\mathcal{O} \subseteq V$ with $|\mathcal{O}| = k \geq 2$ be a set of observers on $\mathcal{G}$. A node u is

[6]For example, on cycles of odd-length d with a budget $k = 2$ in the low-variance setting, any two nodes at distance more than 2 are equivalent with respect to the coverage rate, but only optimal if the observers are at distance $(d-1)/2$; our approach selects this optimal placement.

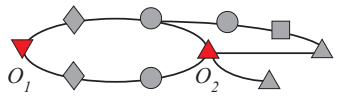


Fig. 3. An unweighted network with two observer nodes o_1 and o_2. Different shapes represent different equivalence classes, i.e., groups of nodes which are not distinguishable from the point of view of the observers (solid red). In this example there are $q = 5$ equivalence classes.

said to be equivalent to a node v (which we write $u \sim v$) if and only if, for every $o_i, o_j \in \mathcal{O}$

$$d(u, o_i) - d(u, o_j) = d(v, o_i) - d(v, o_j). \tag{1}$$

where $d(x, y)$ is the (weighted) distance between x and y. The relation $\sim$ is reflexive, symmetric, and transitive, hence it defines an *equivalence relation*. Therefore, a set of observers $\mathcal{O}$ partitions V in *equivalence classes* (an example is given in Figure 3). We denote by q the number of equivalence classes and we let $[u]_{\mathcal{O}}$ be the class of u, i.e., the set of all nodes that are equivalent to u.

When the variance is zero, given an observer set, we can *distinguish* u from v if Equation (1) does *not* hold for u, v and a pair of observers o_i, o_j, i.e., if $[u]_{\mathcal{O}} \neq [v]_{\mathcal{O}}$.

The problem of finding the minimum-size set of nodes S, such that for every u, v there exist $s_i, s_j \in S$ for which $d(u, s_i) - d(u, s_j) \neq d(v, s_i) - d(v, s_j)$ is known as the *Double Resolving Set (DRS) Problem* [6]. Our problem differs from *DRS* because we focus on the more realistic context in which, due to limited resources, we want to allocate a *finite budget* in order to maximize the detection probability (as opposed to minimizing the number of observers for perfect detection, which is, in many cases, still prohibitively large). However, the connection between our problem and *DRS* paves the way for a principled approach.

We now define a *distance vector* associated with a candidate source, which, as we will prove in Lemma 1, mathematically captures equivalence in a manner that is easy to work with.

Definition 2 (Distance Vector): Let $\mathcal{G} = (V, E)$ and $\mathcal{O} \subseteq V$ with $|\mathcal{O}| = k \geq 2$ a set of observers on $\mathcal{G}$. For each candidate source s the distance vector is $\mathbf{d}_{s,\mathcal{O}} \in \mathbb{R}^{k-1}$ with entries $d(s, o_{i+1}) - d(s, o_1)$ for $1 \leq i \leq k - 1$.

The following lemma, similar in spirit to Lemma 3.1 in [7], shows that, the equality between distance vectors of different candidate sources does not depend on the choice of the *reference observer* o_1.

Lemma 1: Let $\mathcal{G} = (V, E)$ and $\mathcal{O} \subseteq V$ with $|\mathcal{O}| = k \geq 2$ and let $u, v \in V$. Then, $[u]_{\mathcal{O}}$=$[v]_{\mathcal{O}}$ if and only if $\mathbf{d}_{u,\mathcal{O}} = \mathbf{d}_{v,\mathcal{O}}$, independently of the choice of the reference observer o_1 in Definition 2.

Estimating the source in the low-variance setting. We are now ready to describe how we can estimate the source, and define the probability of correct detection in the zero- and low- variance setting, i.e., when $X_{uv} = w_{uv}$ a.s. for every edge (u, v).

For every observer $o_i \in \mathcal{O}$, denote by t_i the time at which o_i gets infected. In the zero-variance setting, the observed infection times of nodes $o_2, .., o_K$ with respect to observer o_1, i.e., the vector $\tau \stackrel{\text{def}}{=} t_2 - t_1, \ldots, t_k - t_1$, is exactly the distance vector of the *unknown* source s^*. Then, if for every $u, v \in V$, $[u]_{\mathcal{O}} \neq [v]_{\mathcal{O}}$, the source can be always correctly identified. We will see in Proposition 1 that this is true also in a more general *low-variance* framework where we are always able to identify the equivalence class to which the real source belongs.

We assume a prior probability distribution on the location of the source to be given, i.e., $Q(u) \stackrel{\text{def}}{=} \mathrm{P}(s^* = u)$. As we cannot distinguish between vertices inside $[s^*]_{\mathcal{O}}$ (otherwise they would not be in the same equivalence class), we let our estimated source $\widehat{s}$ be chosen at random from the conditional probability $Q|_{E^*}(u) \stackrel{\text{def}}{=} \mathrm{P}(s^* = u | u \in E^*)$. Hence the success probability is

$$\begin{aligned} \mathcal{P}_s(\mathcal{O}) &\stackrel{\text{def}}{=} \sum_{s \in V} \mathrm{P}(\widehat{s} = s | s^* = s) \mathrm{P}(s^* = s) \\ &= \sum_{s \in V} Q|_{[s]_{\mathcal{O}}}(s) Q(s) = \sum_{s \in V} \frac{Q(s)}{Q([s]_{\mathcal{O}})} Q(s), \end{aligned} \tag{2}$$

and is 1 if all equivalence classes are singletons.

In the experimental results in Section V we also look at another relevant metric for the source localization problem, the *expected distance* (weighted or in hops) between the true and estimated source:

$$\begin{aligned} \mathrm{E}[d(s^*, \widehat{s})] &\stackrel{\text{def}}{=} \sum_{s \in V} \mathrm{P}(s^* = s) \sum_{u \in [s]_{\mathcal{O}}} \mathrm{P}(\widehat{s} = u | s^* = s) d(s, u) \\ &= \sum_{s \in V} \sum_{u \in [s]_{\mathcal{O}}} \frac{Q(s) Q(u)}{Q([s]_{\mathcal{O}})} \cdot d(s, u). \end{aligned} \tag{3}$$

Alternative metrics, including worst-case metrics, also exist [15] (see Appendix I for some examples).

B. Setting

For ease of exposition, we focus on the case in which the prior distribution on the position of the source is uniform, hence $Q(u) = 1/n$ for all $u \in V$.[7]

Proposition 1: Let $\mathcal{G} = (V, E)$ be a network of size n and $\mathcal{O} \subseteq V$. Call $\delta = \min_{u,v:\mathbf{d}_{u,\mathcal{O}} \neq \mathbf{d}_{v,\mathcal{O}}} \|\mathbf{d}_{u,\mathcal{O}} - \mathbf{d}_{v,\mathcal{O}}\|_\infty$. Assume a uniform prior $Q(u) = 1/n$ for all $u \in V$ and call D the maximum distance in hops in any shortest path between any node and any observer.

1) In the zero-variance case, then $\mathcal{P}_s(\mathcal{O}) = q/n$, where q is the number of equivalence classes for $\mathcal{O}$;
2) If the transmissions are such that for each $uv \in E$, $X_{uv} \in [w_{uv} - \varepsilon, w_{uv} + \varepsilon]$, we denote as $\mathcal{P}_s^\varepsilon(\mathcal{O})$ the probability of success and we define $\varepsilon_0 = \sup\{\varepsilon > 0 : \mathcal{P}_s^\varepsilon(\mathcal{O}) = \mathcal{P}_s^0(\mathcal{O})\}$, we have $\varepsilon_0 > \delta/2D$.

Proof:

1) By definition,

$$\mathcal{P}_s(\mathcal{O}) = \sum_{[u]_{\mathcal{O}}} \mathrm{P}(\widehat{s} = s^* | s^* \in [u]) \mathrm{P}(s^* \in [u]).$$

[7]Our algorithms and observations can be generalized using Equation (2) instead of the simpler formula that we now derive for the uniform case.

Hence,

$$\mathcal{P}_s(\mathcal{O}) = \sum_{[u]_\mathcal{O}} \frac{1}{|[u]_\mathcal{O}|} \cdot \frac{|[u]_\mathcal{O}|}{n} = \frac{1}{n} \sum_{[u]_\mathcal{O}} 1 = \frac{q}{n}.$$

2) Recall that, for $u, v \in V$, $[u]_\mathcal{O} \neq [v]_\mathcal{O}$ if and only if $\mathbf{d}_{u,\mathcal{O}} \neq \mathbf{d}_{v,\mathcal{O}}$. Since $\mathbf{d}_{u,\mathcal{O}} \neq \mathbf{d}_{v,\mathcal{O}}$ implies $\|\mathbf{d}_{u,\mathcal{O}} - \mathbf{d}_{v,\mathcal{O}}\|_\infty \geq \delta$, if $\varepsilon < \delta/2D$, no estimation error is possible between $u, v \in V$ such that $\mathbf{d}_{u,\mathcal{O}} \neq \mathbf{d}_{v,\mathcal{O}}$. Hence $\varepsilon_0 > \delta/2D$.

■

Note that here ε_0 plays the role of σ_0 in Figure 1. Indeed, for $\varepsilon < \varepsilon_0$ the variance of the transmission delays does not affect the accuracy of source localization.

If additional conditions on the weights or on the network topology are made, more refined bounds for ε_0 can be derived. For example, in a *tree* with integer weights, due the uniqueness of the path between two any vertices, the minimum distance (in the infinity norm) between two distance vectors is 2. Hence, in this case, an accumulated variance of less than 1 can be tolerated and we have $\varepsilon_0 > 1/D$.

For the remainder of this section, we will assume $\varepsilon < \delta/2D$, which we call the low-variance case. Independently of the topology of the network $\mathcal{G}$, the success probability $\mathcal{P}_s$, as well as other possible metrics of interest, can be computed exactly in polynomial time (see e.g., Equation (2) and (3)). In fact, due to Lemma 1, it is enough to compute the distance vector of Definition 1 for all the nodes. Nonetheless, if we have a budget $k \geq 2$ of nodes that we can choose as observers, finding the configuration that maximizes $\mathcal{P}_s$ is an NP-hard problem. This is a direct consequence of the hardness result of Chen et al. [7].

Theorem 1: Let $k \geq 2$ be the budget on the number of nodes we can select as observers. Finding $\mathcal{O} \subseteq V$ such that $\mathcal{O} \in \text{argmax}_{|\mathcal{O}|=k} \mathcal{P}_s(\mathcal{O})$ is NP-hard.
The proof follows straightforwardly with a reduction from the *DRS* problem (see Appendix IV).

C. Observer Placement

Our first main contribution in this paper is a solution to the budgeted observer-placement problem. Our approach, presented in Algorithm 1, is specifically designed for the source localization problem and has a simple greedy structure: for every node $v \in V$, initialize $\mathcal{O} \leftarrow \{v\}$ and iteratively add to $\mathcal{O}$ the node u that maximizes the gain with respect to the success probability until we either run out of budget or $\mathcal{P}_s = 1$. Proposition 1 ensures that greedily maximizing the success probability is equivalent to greedily maximizing the number q of equivalence classes. When adding an element to the observer set, the partition in equivalence classes can be updated in linear time, hence the total running time of our algorithm is $O(kn^3)$. Despite bypassing the NP-hardness of the problem, this might not be sufficiently fast for very large graphs. However, the procedure is extremely parallelizable and well suited, e.g., for Map-Reduce (see, for example, the main for loop and the **argmax** in the **while** loop).

The observer placement obtained through Algorithm 1 will be denoted LV-OBS to emphasize the fact that it has been designed for the case in which the variance is absent or very small (LV stands for *low-variance* regime).

Algorithm 1 (LV-OBS): Observer placement for the low-variance setting.

Require: Network G, budget k
  for $v \in V$ **do**
    $\mathcal{O}_v \leftarrow v$
    while $\mathcal{P}_s(\mathcal{O}_v) \neq 1$ **and** $\mathcal{O}_v < k$ **do**
      $u \leftarrow \text{argmax}_{z \in V \setminus \mathcal{O}_v} [\mathcal{P}_s(\mathcal{O}_v \cup \{z\}) - \mathcal{P}_s(\mathcal{O}_v)]$
      $\mathcal{O}_v \leftarrow \mathcal{O}_v \cup \{u\}$.
  return $\text{argmax}_{v \in V} \mathcal{P}_s(\mathcal{O}_v)$

D. Performance

As budgeted observer placement (even in the zero-variance setting) is NP-hard, there is no optimal algorithm to compare against. Instead, we evaluate the performance of our algorithm against a set of natural benchmarks that have shown to have good performance in other works [27], [4], [31] (see Section V-B for a discussion of these benchmarks, Figure 7 for the results).

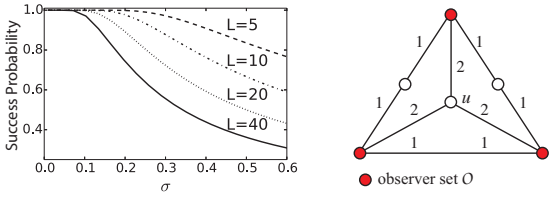
We further compare against two other natural heuristics that also optimize an objective function greedily. The first is an adapted version of the approximation algorithm for the *DRS* problem proposed by Chen at al. [7] and described in Appendix II. By stopping the greedy process after it selects k nodes, we can adapt in a natural way this approximation algorithm and create a heuristic for the budgeted version. The second is a direct minimization of the expected error distance obtained by Equation (3) with $Q(u) = 1/n$ for all $u \in V$. Comparing all three approaches, our algorithm outperforms the other two (see Appendix III for details).

IV. The High-Variance Regime

When the variance is not guaranteed to be low, as defined in Section III, computing analytically the success probability - or other metrics of interest - is unfortunately not possible (except for very simple graphs, like the path in the example of Figure 2). Moreover, the estimation of the source is more challenging because the observed infection delay $t_i - t_j$ can be misleading, especially if the corresponding observers o_i and o_j are *far* from the source. Take, for example, a path of length L where the two leaves are the only two observers and all edges have weight 1. Figure IV shows how the success probability $\mathcal{P}_s$ decays faster for increasing values of L. Building on this observation, we propose a strategy for observer placement that enforces a controlled distance from a general source node to the observer set.

A. Diffusion Model and Source Estimation

For every edge (u, v) the infection delay X_{uv} is distributed as a truncated Gaussian random variable with parameters $(w_{uv}, \sigma w_{uv}, [w_{uv}/2, 3w_{uv}/2])$. More precisely, if $Y_{uv} \sim \mathcal{N}(w_{uv}, \sigma w_{uv})$ is a Gaussian random variable, X_{uv} is obtained by conditioning Y_{uv} with $Y_{uv} \in [w_{uv}/2, 3w_{uv}/2]$. This delay distribution has two advantages with respect to



(a) (b)

Fig. 4. (a): Success probability $\mathcal{P}_s$ on a path of length L for increasing variance σ. (b): Counterexample for the converse of Lemma 2; for each pair of observers in $\mathcal{O}$, u is not contained in the shortest path between them, yet $\mathcal{O}$ is a *DRS*.

the one of [25], i.e., that $X_{(u,v)} \sim \mathcal{N}(w_{uv}, \sigma w_{uv})$. First, the model admits only strictly positive infection delays. Second, different values of the standard deviation σ result in different regimes for the propagation, making our model very versatile. When $\sigma = 0$, X_{uv} boils down to a deterministic value equal to the edge weight w_{uv}; when σ is large, the distribution of X_{uv} becomes closer to uniform $U([w_{uv}/2, 3w_{uv}/2])$. Finally, when σ is strictly positive but small, $X_{uv} \approx \mathcal{N}(w_{uv}, (\sigma w_{uv})^2)$. In Appendix V, we explain how an approximated maximum likelihood estimator for the source can be derived in this setting.

B. Observer Placement

First, we formalize why distances between observers are important: If o_i, o_j are two observers and the source is $s^* \in \mathcal{P}(o_i, o_j)$, then

$$\text{var}(t_i - t_j) \approx \sigma^2 \left[\sum_{(uv) \in \mathcal{P}(o_i, o_j)} w_{uv}^2 \right] \tag{4}$$

where $\mathcal{P}(x, y)$ denotes the shortest path from x to y, written as a sequence of edges. Although we cannot control σ, we can control the *path length* between observers.[8] We make use of the following sufficient condition for a set to be a *DRS*, i.e., for an observer set to guarantee optimal source detection.

Lemma 2: Let $\mathcal{G} = (V, E)$ be a network, $\mathcal{O} \subseteq V$. If for every $u \in V$ there exist $o_1, o_2 \in \mathcal{O}$ such that there is a unique shortest path $\mathcal{P}(o_1, o_2)$ between o_1 and o_2 and $u \in \mathcal{P}(o_1, o_2)$, then $\mathcal{O}$ is a *DRS* for G.

Proof: Let $u, v \in V \backslash \mathcal{O}$. We will prove that there exist $o_1, o_2 \in \mathcal{O}$ such that the pair (u, v) is resolved by (o_1, o_2), i.e., $d(v, o_1) - d(u, o_1) \neq d(v, o_2) - d(u, o_2)$. Let $o_1, o_2 \in \mathcal{O}$ such that u appears in the unique shortest path $\mathcal{P}(o_1, o_2)$ and $o_3, o_4 \in S$ such that v appears in the unique shortest path $\mathcal{P}(o_3, o_4)$. If $v \in \mathcal{P}(o_1, o_2)$ or $u \in \mathcal{P}(o_3, o_4)$ than u and v are resolved by, respectively, (o_1, o_2) or (o_3, o_4). Take $v \notin \mathcal{P}(o_1, o_2)$ and $u \notin \mathcal{P}(o_3, o_4)$. In this case, $\{o_1, o_2\} \neq \{o_3, o_4\}$. Let us suppose without loss of generality that $o_1 \notin \{o_3, o_4\}$. We look only at the case where (o_1, o_2) does not resolve (u, v) and prove that the pair is indeed resolved by two vertices in $\mathcal{O}$. Since (o_1, o_2) does not resolve (u, v), there exists $c \in \mathbb{R}$ such that $d(v, o_1) - d(u, o_1) = c = d(v, o_2) - d(u, o_2)$. Since the unique shortest path between o_1 and o_2 goes through u we have that $c > 0$. We prove that either (o_1, o_3) or (o_1, o_4) resolves (u, v). If this was not the case, we would have the following equalities:

$$c = d(v, o_1) - d(u, o_1) = d(v, o_3) - d(u, o_3)$$
$$c = d(v, o_1) - d(u, o_1) = d(v, o_4) - d(u, o_4).$$

Since $c > 0$, $d(v, o_3) > d(u, o_3)$ and $d(v, o_4) > d(u, o_4)$ giving a contradiction with v (and not u) being on the shortest path $\mathcal{P}(o_3, o_4)$. We conclude that (u, v) are resolved by either (o_1, o_3) or (o_1, o_4). ■

The converse of this lemma is not true: If $\mathcal{O}$ double resolves $\mathcal{G}$, it is not even true that for every node u there must exist $o_1, o_2 \in \mathcal{O}$ such that u is contained in *some* shortest path between o_1 and o_2 of (see the Example in Figure IV).

Path covering strategy. We take Lemma 2 as a basis for deriving a *path covering* strategy for observer placement. In practice, the condition about the *uniqueness* of the shortest path is too strong and excludes many potentially useful observer nodes[9]. This is why we relax the condition of Lemma 2 and we prefer, when the shortest path is not unique, to select one arbitrarily. Let $S \subseteq V$ be a set of observers and L a positive integer: We call $P_L(S)$ the set of nodes that lie on a shortest path of length at most L between any two observers in the set S. Given a budget k, and a positive integer L, we denote by $S^*_{k,L}$ the set of k vertices that maximize the cardinality of $P_L(S)$. We call L the *length constraint* for the observer placement because we consider an observer to be *useful* for source localization only if it is within distance L from another observer. $S^*_{k,L}$ can be approximated greedily as in Algorithm 2.[10]

Algorithm 2 (HV-OBS): Observer placement for the high-variance setting.

Require: Network $G(V, E)$, budget k, length constraint L
  $n \leftarrow |G|$
  for $v \in V$ **do**
    $\mathcal{O}_v \leftarrow v$
    while $|P_L(\mathcal{O}_v)| \neq n$ **and** $\mathcal{O}_v < k$ **do**
      $u \leftarrow \text{argmax}_{z \in V \backslash \mathcal{O}_v} [|P_L(\mathcal{O}_v \cup \{z\})| - |P_L(\mathcal{O}_v)|]$
      $\mathcal{O}_v \leftarrow \mathcal{O}_v \cup \{u\}$.
  return $\text{argmax}_{v \in V} |P_L(\mathcal{O}_v)|$

We will refer to the observer placement produced by Algorithm 2 as HV-OBS(L) to emphasize that it is designed for the high-variance case.

Comparison with Algorithm 1. Note that taking L equal to the maximum weighted distance Δ does not make

[8] A relevant but orthogonal line of work would study how to control the parameter σ by, e.g., immunizations, quarantines, or other preventative measures and is outside the scope of our work.

[9] Experimentally we see that in many practical situations two shortest paths differ only by a few nodes and the majority of nodes on the path are resolved by the two extreme nodes.

[10] The running time of Algorithm 2 is $O(n^2 k^2)$, however, as with the low-variance case, this is highly parallelizable and hence tractable even for large networks.

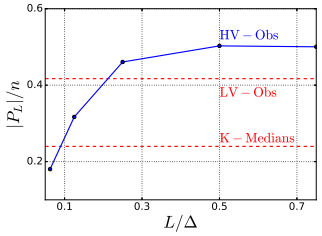


Fig. 5. Fraction of nodes in $P_L(\cdot)$ for the California dataset with 2% of observers.

Algorithm 2 equivalent to Algorithm 1, i.e., we do not obtain LV-OBS. To see how the two algorithms could give different results, take a cycle of odd length d with a leaf node ℓ added as a neighbor to an arbitrary node v and assume to start the algorithm with initial set $\{v\}$. At the first step, the two algorithms will make the same choice, choosing one of the two nodes that is at distance $(d-1)/2$ from v. At the second step however, LV-OBS will add ℓ (a DRS contains all leaves [7]), whereas Algorithm 2 will add a node on the cycle. This observation is key to our results because it explains why Algorithm 2 results in a more uniform (and hence *variance-resistant*) observer placement with respect to LV-OBS. HV-OBS operates a trade-off between the average distance to the observers and the maximization of $\mathcal{P}_s$.

Choice of the L parameter. How could one optimally set L? Needless to say, the optimal L depends on the network topology and on the available budget: Clearly, for a larger budget a smaller L is preferred.

The cardinality of $P_L(\mathcal{O})$ is a good proxy for the performance of $\mathcal{O}$. The value $|P_L|$ is increasing in L and reaches its maximum for L equal to the maximum weighted distance ($L = \Delta$). For small L, $|P_L(\text{HV-OBS})| < |P_\Delta(\text{LV-OBS})|$ but for L large enough this is no more the case. See Figure 5 for an example. Our empirical results suggest that L should be chosen as the maximum for which $|P_L(\text{HV-OBS})| \leq |P_\Delta(\text{LV-OBS})|$. The key property of HV-OBS with respect to LV-OBS is that observers are spread more *uniformly* without *losing* too much in terms of success probability $\mathcal{P}_s$: Figure 6 shows $|P_L(\text{HV-OBS})|$ and $\mathcal{P}_s$ as a function of L.

LV-OBS and HV-OBS can give drastically different observers (see Appendix VI for an example).

V. EMPIRICAL RESULTS

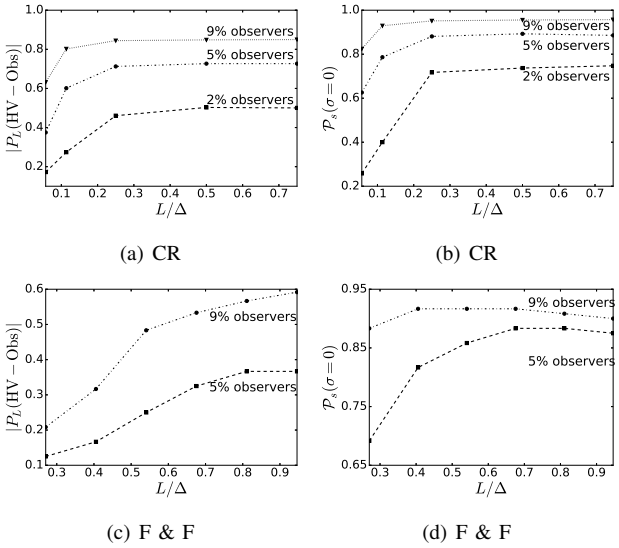
We purposely run our experiments on three very different real-world networks that, in addition to being relevant examples of networks for epidemic spread, display different characteristics in terms of size, diameter, clustering coefficient and average degree (see Table I), enabling us to test the performance of our methods on various topologies.

A. Datasets

The three networks we consider are:

- Friend & Families (F & F). This is a dataset containing phone calls, SMS exchanges and bluetooth proximity, among a community living in the proximity of a university campus [2]. We select the largest connected component of individuals who took part in the experiment during its whole duration. The edges are weighted, according to the number of phone calls, SMSs, and bluetooth contacts.
- Facebook-like Message Exchange (FB) [24]. As the individuals included in this dataset were living on the same university-campus, the number of messages exchanged is likely to be a good measure of in-person interaction. We selected links on which at least one message was sent in both directions and individuals that had more than 1 contact.
- California Road Network (CR) [1]. In order to obtain a single connected component and remove points that effectively represent the same location, we collapsed the points falling within a distance of 2 km. Moreover we iteratively deleted all leaves.[11] The diameter of this network is very large compared with that of the other two networks. The edges are weighted according to a rescaled version of the real distance (measured in km).

In all three networks, edges are given (non-unit) integer *weights*, which is realistic in many applications as the expected transmission delays are known only up to some level of precision. Integer weights do *not* simplify the estimation of the source; in fact, this makes it *more* difficult to distinguish between vertices. For example, if the edges of the CR network were weighted according to the Euclidean distance between the two endpoints, LV-OBS would use only a very small portion of the budget and the comparison would not be meaningful.



(a) CR (b) CR

(c) F & F (d) F & F

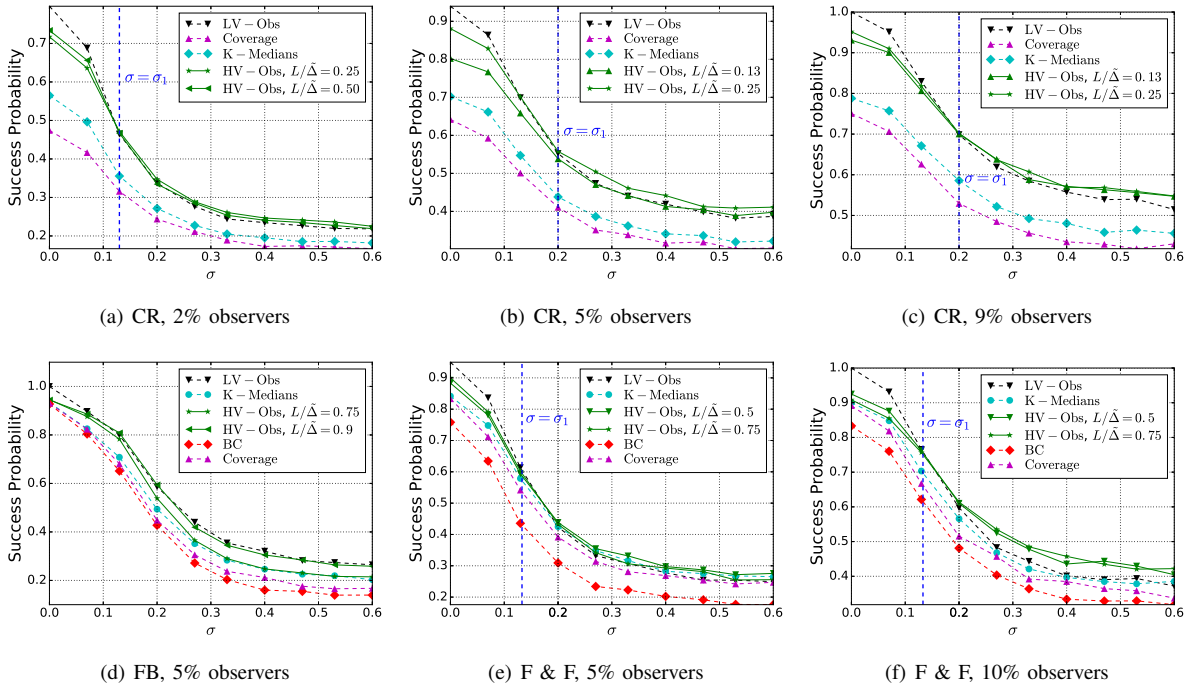
Fig. 6. Fraction of nodes in $P_L(\text{HV-OBS})$ and success probability as a function of L/Δ for the CR and the F & F datasets comparing with the zero-variance setting.

[11]The roads that cross the state border are not completely tracked in this dataset and terminate with a leaf. Some other leaves might represent remote locations, not necessarily close to the borders, but their influence on the epidemic should anyway be very low.

	$\|V\|$	$\|E\|$	$\min(w_{uv})$	$\text{avg}(w_{uv})$	$\max(w_{uv})$	Avg Degree	Diameter	Avg Dist	Avg Clust.
Friends & Families	120	563	4	5.58	7	9.38	6	17.5	0.67
Facebook Messages	1020	6205	1	2.97	5	12.16	5	6.69	0.09
California Roads	1259	1801	1	1.71	9	2.86	66	55.3	0.2

TABLE I

DISPLAYS STATISTICS FOR NETWORKS EXAMINED



(a) CR, 2% observers (b) CR, 5% observers (c) CR, 9% observers

(d) FB, 5% observers (e) F & F, 5% observers (f) F & F, 10% observers

Fig. 7. Success probability $\mathcal{P}_s$ as variance σ is increased.

B. Comparison against Benchmarks

We compare HV-OBS against the following benchmarks:

1) LV-OBS: this is our solution for the low-variance case (see Section III).
2) BC (Betweenness Centrality): This is a popular method for placing observers for source-localization (see, e.g., [20] and [27], where it emerges as the best heuristic for observer placement among those tested). It consists of the k nodes having the largest BC, which is defined, for all $u \in V$ as
$$BC(u) = \sum_{x,y \in V, x \neq y} \frac{\sigma_{x,y}(u)}{\sigma_{x,y}}$$
where $\sigma_{x,y}$ is the number of shortest paths between x and y and $\sigma_{x,y}(u)$ is the number of those paths that passes through u.
3) Coverage-rate (COVERAGE) [31]: This approach maximizes the number of nodes that have an observer as a neighbor, i.e.,
$$\mathcal{C}(\mathcal{O}) = |\cup_{o \in \mathcal{O}} N_o|/n,$$
where N_o denotes the set of neighbors of o and $n = |V|$. It has been shown to outperform several heuristics with a diffusion model and an estimation setting that are very similar to ours.
4) K-MEDIAN: this is the optimal placement for the closely-related problem of maximizing the detectability of a flow [4]. The K-MEDIAN placement is the set of k nodes $\mathcal{O}$ such that
$$\mathcal{O} = \text{argmin}_{|\mathcal{O}|=k} \sum_{s \in V} (\min_{o \in \mathcal{O}} d(s,o)).$$
Determining the K-MEDIANS of a network is NP-hard [12]; we use a greedy heuristic for K-MEDIANS.

C. Experimental Results

We estimate $\mathcal{P}_s$ and $\text{E}[d(s^*, \widehat{s})]$ for different values of the variance σ. We generate epidemics by using each node in turn as the source. For the FB and CR datasets, we run 5 simulations per node and variance level; and for the F & F dataset, as the network is smaller, we run 20 simulations per node and variance level. For the FB and CR datasets, we estimate the source based on the first 20 observations only: Given the large size of the network, it would be unrealistic to wait for all the network to get infected before running the algorithm. The results for $\mathcal{P}_s$ are displayed in Figure 7. An approximation of the value σ_1, above which HV-OBS outperforms LV-OBS, is marked with a vertical line. For the expected distance (weighted and in hops), see Appendix VI.

We first take as budget for the observers the minimum budget for which $\mathcal{P}_s(\text{LV-OBS}) = 1$. This corresponds to

$k \sim 9\%$ for the F & F dataset, $k \sim 9\%$ for the CR network and $k \sim 5\%$ for the FB dataset. This is the setting in which we expect the improvement of HV-OBS over LV-OBS to be especially strong: For smaller values of k we expect LV-OBS to be nearly optimal even in the high-variance regime because we do not have enough budget to contrast both the topological *undistinguishability* among nodes (what LV-OBS is designed for) and the accumulation of variance (what HV-OBS is designed for). For the F & F and the CR networks, we also experiment with smaller percentages of observers and consistently find an improvement of HV-OBS over LV-OBS in the high-variance regime: Below a certain amount of variance σ_1 LV-OBS performs better than HV-OBS for any choice of the parameter L, whereas above σ_1 a calibrated choice of L leads to a significant improvement. Such L stays constant for all $\sigma > \sigma_1$, i.e., with the notation of Figure 1 we have $\sigma_1 = \sigma_F$. For the FB dataset instead, probably due to the low diameter with respect to the number of nodes, we observe that HV-OBS does not improve on LV-OBS for any value of L. Both LV-OBS and HV-OBS systematically outperform the baseline heuristics for observer placement that we described in Section V-B. For the CR dataset the performance of Betweenness Centrality is particularly poor and the results are not shown. The Coverage Rate heuristic outperforms Betweenness Centrality on all three networks (confirming what found by by Zhang et al. [31]) but is consistently less effective than K-Medians and our methods.

D. Robustness

To measure the robustness of our approach, we consider an alternate transmission model, and we measure whether, without making any changes, our observer placement still performs well. For every edge $uv \in E$ with weight w_{uv}, we take $X_{uv} \sim \text{Unif}([(1-\varepsilon)w_{uv}, (1+\varepsilon)w_{uv}])$. We find comparable results (see Appendix VI); they suggest that our observer placement is not dependant on the exact transmission model and that the variance of the transmission delays is really a key factor for a good observer placement.

VI. Conclusion & Future Work

In this work, we have taken a principled approach towards budgeted observer placement for source localization. We are the first to have observed a dichotomy between the low and high-variance regimes, and we developed complementary approaches for both. We have evaluated our approaches against state-of-the-art and alternative heuristics and find that the performance of our algorithms is favourable.

One natural extension would account for two stages of observation; in the first stage, as in this work, we select a set of observers to monitor the network. In the next stage, once an epidemic begins, we deploy additional observers in the relevant region of the network. This would pave the way for other types of *adaptive* models, including ones where we not only *observe* a node but can act to *immunize* it or in which we can *move* the observers as required.

Acknowledgements

B.Spinelli was partially supported by the Bill & Melinda Gates Foundation, under Grant No. OPP1070273

References

[1] California road network. `http://www.census.gov/geography.html`.

[2] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland. Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659, 2011.

[3] F. Altarelli, A. Braunstein, L. Dall'Asta, A. Lage-Castellanos, and R. Zecchina. Bayesian inference of epidemics on networks via belief propagation. *Physical review letters*, 112(11):118701, 2014.

[4] J. Berry, W. Hart, C. Phillips, J. Uber, and J. Watson. Sensor placement in municipal water networks with temporal integer programming models. *Journal of Water Resources Planning and Management*, 132(4), 2006.

[5] P. Billingsley. *Probability and Measure*. John Wiley & Sons, 1995.

[6] J. Cáceres, M. Hernando, M. Mora, I. Pelayo, M. Puertas, C. Seara, and D. Wood. On the metric dimension of cartesian products of graphs. *SIAM J. Discrete Mathematics*, 21(2):423–441, 2007.

[7] X. Chen, X. Hu, and C. Wang. Approximability of the minimum weighted doubly resolving set problem. In *20th Annual Int. Computing and Combinatorics Conf. (COCOON)*, pages 357–368. LNCS, 2014.

[8] W. Dong, W. Zhang, and C. W. Tan. Rooting out the rumor culprit from suspects. In *IEEE Int. Symposium on Information Theory (ISIT)*, pages 2671–2675. IEEE, 2013.

[9] G. C. Fanti, P. Kairouz, S. Oh, and P. Viswanath. Spy vs. spy: Rumor source obfuscation. In *SIGMETRICS*, pages 271–284. ACM, 2015.

[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 Communication Survey Tutorials*, 2014.

[11] J.Kratica, M. Čangalović, and V. Kovačević-Vujčić. Computing minimal doubly resolving sets of graphs. *Computers & Operations Research*, 36(7):2149–2159, 2009.

[12] O. Kariv and S. Hakimi. An algorithmic approach to network location problems. ii: The p-medians. *SIAM journal of Applied Mathematics*, 37, 1979.

[13] A. Krause, J. Leskovec, C. Guestrin, J. Vanbriesen, and C. Faloutsos. Efficient sensor placement optimization for securing large water distribution networks. *Journal of Water Resources Planning and Management*, 134(6), 2008.

[14] S. Krishnasamy, S. Banerjee, and S. Shakkottai. The behavior of epidemics under bounded susceptibility. *SIGMETRICS Performormance Evaluation Review*, 42(1):263–275, June 2014.

[15] L.E.Celis, F.Pavetić, B.Spinelli, and P.Thiran. Budgeted sensor placement for source localization on trees. In *Latin-American Algorithms, Graphs and Optimization Symposium (LAGOS)*, 2015.

[16] M. Lelarge. Efficient control of epidemics over random networks. In *11th Int. Joint Conf. on Measurement and Modeling of Computer Systems, SIGMETRICS/Performance*, pages 1–12, 2009.

[17] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *13th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pages 420–429. ACM, 2007.

[18] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Physical Review E*, 90(1):012801, 2014.

[19] A. Louni, A. Santhanakrishnan, and K. Subbalakshmi. Identification of source of rumors in social networks with incomplete information. *ASE Eighth Int. Conf. on Social Computing (SocialCom)*, 2015.

[20] A. Louni and K. Subbalakshmi. A two-stage algorithm to estimate the source of information diffusion in social media networks. *IEEE INFOCOM Workshop on Dynamic Social Networks*, 2014.

[21] W. Luo, W. Tay, and M. Leng. How to identify an infection source with limited observations. *IEEE Journal of Selelcted Topics in Signal Processing*, 8(4):586–597, 2014.

[22] W. Luo and W. P. Tay. Identifying infection sources in large tree networks. In *9th Annual IEEE Communications Society Conf. on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, pages 281–289. IEEE, 2012.

[23] P. Netrapalli and S. Sanghavi. Learning the graph of epidemic cascades. *SIGMETRICS Performance Evalulation Review*, 40(1):211–222, June 2012.

[24] T. Opsahl and P. Panzarasa. Clustering in weighted networks. *Social networks*, 31(2):155–163, 2009.

[25] P. Pinto, P. Thiran, and M. Vetterli. Locating the source of diffusion in large-scale networks. *Physical Review Letters*, 109, 2012.

[26] B. A. Prakash, J. Vreeken, and C. Faloutsos. Spotting culprits in epidemics: How many and which ones? *IEEE 12th Int. Conf. on Data Mining (ICDM)*, 12:11–20, 2012.

[27] E. Seo, P. Mohapatra, and T. Abdelzaher. Identifying rumors and their sources in social networks. In *SPIE Defense, Security, and Sensing*, pages 83891I–83891I. Int. Society for Optics and Photonics, 2012.

[28] D. Shah and T. Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on information theory*, 57, 2011.

[29] S. Sundareisan, J. Vreeken, and B. A. Prakash. Hidden hazards: Finding missing nodes in large graph epidemics. In *SIAM Int. Conf. on Data Mining (SDM)*. SIAM, 2015.

[30] S. Zejnilovic, J. Gomes, and B. Sinopoli. Network observability and localization of the source of diffusion based on a subset of vertices. In 51^{st} *Annual Allerton Conf. on Communication, Control & Computing*, pages 847–852, 2013.

[31] X. Zhang, Y. Zhang, T. Lv, and Y. Yin. Identification of efficient observers for locating spreading source in complex networks. *Physica A: Statistical Mechanics and its Applications*, 442:100–109, 2016.

[32] Z. Zhang, W. Xu, W. Wu, and D.-Z. Du. A novel approach for detecting multiple rumor sources in networks with partial observations. *Journal of Combinatorial Optimization*, pages 1–15, 2015.

[33] L. Zheng and C. W. Tan. A probabilistic characterization of the rumor graph boundary in rumor source detection. In *IEEE Int. Conf. on Digital Signal Processing (DSP)*, pages 765–769. IEEE, 2015.

[34] K. Zhu and L. Ying. Information source detection in the SIR model: A sample path based approach. In *Information Theory and Applications Workshop (ITA), 2013*, pages 1–9. IEEE, 2013.

APPENDIX I
OTHER METRICS

In this section we define other metrics of interest. In particular, we consider an adversarial setting (e.g., in the case of bio-warfare) where if our observers are known, the adversary would select the worst location for the source.

First, we consider minimum success probability, which is

$$\widehat{\mathcal{P}_s}(\mathcal{O}) \quad \stackrel{\text{def}}{=} \quad 1 - \max_{E_i} \left(\mathrm{P}(\widehat{s} \neq s^* | s^* \in E_i)\right)$$

where $\{E_i\}$ are the equivalence classes with respect to $\mathcal{O}$. Note that in an adversarial setting, we would not consider any prior, rather would select $\widehat{s} \in \mathrm{argmax}_{E_i} \mathrm{P}(\widehat{s} \neq s^* | s^* \in E_i)$ uniformly at random; given any non-uniform distribution, the adversary could place the source at the location with lowest probability.

For the same reasons, we may also wish to consider the *maximum distance* between the true and the estimated source as a metric.

$$\max(d(s^*, \widehat{s})) \quad \stackrel{\text{def}}{=} \quad \max_{s \in V} \Delta_s = \max_i \Delta_i,$$

where Δ_s (similarly Δ_i) denotes the diameter of equivalence class $[s]_{\mathcal{O}}$ (similarly E_i). Note that, in particular, this is independent of any prior.

Another natural consideration which interpolates between expected and worst-case metrics is the *expected maximum distance* between the true and the estimated source. This captures the case where there is a prior Q on the source, and we are able to identify the equivalence class of s^*, but make the *worst-case* estimation $\widehat{s}$ within that class.

$$\begin{aligned} \mathrm{E}[\max(d(s^*, \widehat{s}))] \quad &\stackrel{\text{def}}{=} \quad \sum_{s \in V} \mathrm{P}(s^* = s)\big(\max_{u \in [s]_{\mathcal{O}}} d(s, u)\big) \\ &= \quad \sum_{s \in V} Q(s)\Delta_s = \sum_i Q(E_i)\Delta_i. \end{aligned}$$

APPENDIX II
DOUBLE RESOLVING SETS

The problem of *minimizing* the required number of observers in order to perfectly identify the source in the zero-variance setting has been studied [7]; an observer set $\mathcal{O}$ such that $\mathcal{P}_s(\mathcal{O}) = 1$ is called a Doubly Resolving Set (DRS). While the original formulation of the DRS problem is slightly different, this version follows straightforwardly from our observations in Section III.

Definition 3 (Double Resolving Set): Given a network $\mathcal{G}$, $S \subseteq V$ is said to be a Double Resolving Set of $\mathcal{G}$ if for any $x, y \in V$ there exist $u, v \in S$ s.t. $d(x, u) - d(x, v) \neq d(y, u) - d(y, v)$.

Finding a Doubly Resolving Set of minimum size is known to be NP-hard [11]. An approximation algorithm, based on a greedy minimization of an *entropy* function, has been studied. Note that this has no connection to true information-theoretic entropy.

Definition 4 (Entropy [7]): Let $\mathcal{G}$ a network, $\mathcal{O} \subseteq V$, $|\mathcal{O}| = k$ a set of observers. The entropy of $\mathcal{O}$ is

$$H_{\mathcal{O}} = \log_2\Big(\prod_{[u]_{\mathcal{O}} \subseteq V} |[u]_{\mathcal{O}}|!\Big).$$

Note that $H_{\mathcal{O}}$ is minimized if and only if each equivalence class consists of only one node and hence if and only if $\mathcal{P}_s = 1$. However, despite the fact that $H_{\mathcal{O}}$ is minimized when $\mathcal{P}_s$ is maximized and that both act on the same set of equivalence classes for a given $\mathcal{O}$, the greedy processes that minimize $H_{\mathcal{O}}$ and maximize $\mathcal{P}_s$ are not the same. This can be seen by rewriting both objective functions in the following way. Let $[c_1, \ldots, c_q]$ be the sequence of equivalence class sizes. Then $H_{\mathcal{O}}$ can be written as $H_{\mathcal{O}}([c_1, .., c_q]) = \sum_{i=1}^{l} \sum_{j=2}^{c_i} \log(j) = \sum_{i=2}^{\max c_j} \log(i) \#\{c_j \geq i\}$. Analogously we have the following equality for the success probability $\mathcal{P}_s([c_1, \ldots, c_q])$: $n(1 - \mathcal{P}_s([c_1, \ldots, c_q])) = n - q = \sum_{i=2}^{\max c_j} \#\{c_j \geq i\}$ Hence, though similar in spirit, a greedy minimization of $H_{\mathcal{O}}$ is not related to a greedy optimization of $\mathcal{P}_s$ (or $\mathrm{E}[d(s^*, \widehat{s})]$).

APPENDIX III
ALTERNATE OBJECTIVE FUNCTIONS

Random Geometric Graph, $N = 100$, $r = 0.2$			
	$\frac{\mathcal{P}_s(\Phi_{dist}) - \mathcal{P}_s(\Phi)}{\mathcal{P}_s(\Phi)}$	$\frac{\mathrm{E}_d(\Phi_{dist}) - \mathrm{E}_d(\Phi))}{\mathrm{E}_d(\Phi_{dist})+1}$	$\frac{\mathcal{P}_s(\Phi_{ent}) - \mathcal{P}_s(\Phi)}{\mathcal{P}_s(\Phi)}$
$k = 2$	-0.205	-0.101	-0.033
$k = 4$	-0.014	0.003	-0.007
$k = 8$	-0.003	0.002	-0.003
Barabàsi Albert Graph, $N = 100$, $m = 3$			
	$\frac{\mathcal{P}_s(\Phi_{dist}) - \mathcal{P}_s(\Phi)}{\mathcal{P}_s(\Phi)}$	$\frac{\mathrm{E}_d(\Phi_{dist}) - \mathrm{E}_d(\Phi))}{\mathrm{E}_d(\Phi_{dist})+1}$	$\frac{\mathcal{P}_s(\Phi_{ent}) - \mathcal{P}_s(\Phi)}{\mathcal{P}_s(\Phi)}$
$k = 2$	-0.168	-0.023	-0.037
$k = 4$	-0.039	-0.025	-0.028
$k = 8$	-0.004	0.003	0.005

TABLE II

COMPARISON OF LV-OBS (Φ) WITH THE GREEDY ALGORITHMS THAT MINIMIZE THE ENTROPY FUNCTION OF [7] (Φ_{ent}) AND THE EXPECTED DISTANCE (Φ_{dist})

Here we compare Algorithm 1, denoted in this section as Φ, with two other greedy algorithms that allocate the budget for observers according to different objective functions:

1) Φ_{ent} minimizes the entropy function $H_{\mathcal{O}}$ [7] (see Section II);
2) Φ_{dist} minimizes the expected distance (see Equation (3)).

We considered different topologies and different budgets k for the observers. The results are given in the form of (averaged) relative differences in Table II. The standard error of measurement is not reported for the sake of readability but it was checked to be small: approximately 10^{-2} for $k = 2$ and $(\mathcal{P}_s(\Phi_{dist}) - \mathcal{P}_s(\Phi))/\mathcal{P}_s(\Phi)$; on the order of 10^{-3} or smaller in all the other cases. Note that, since the expected distance can be 0 we add 1 in the denominator when comparing $\mathrm{E}_d(\Phi_{dist})$ and $\mathrm{E}_d(\Phi)$. The results achieved by these algorithms are, on average, worse than those of Algorithm 1 (Φ) independently of the graph topology. The only exception is the minimization of the expected distance when k is very small.

APPENDIX IV
HARDNESS OF BUDGETED OBSERVER PLACEMENT

Theorem 2: Given a network $\mathcal{G} = (V, E)$ and a budget k, finding an observer set $\mathcal{O}$ which maximizes $\mathcal{P}_s$ is NP-hard.

Proof: We will prove that the budgeted observer placement is NP-hard with a reduction from the DRS problem (see Section II), i.e., given a polynomial-time algorithm for the budgeted observer placement problem, we will prove that we can solve the DRS problem in polynomial time.

Assume that we have a polynomial-time algorithm $\mathcal{A}$ that takes as input a network $\mathcal{G} = (V, E)$ and a budget k, and outputs a set $\mathcal{O} \subseteq V$ of size k such that $\mathcal{P}_s$ is maximized. Recall from Section III that given a network $\mathcal{G}$ and a set $\mathcal{O}$, the probability $\mathcal{P}_s$ can be calculated in time $O(n)$ where $n = |V|$ (it is enough to compute the n distances vector with respect to $\mathcal{O}$ and any reference observer $o_1 \in \mathcal{O}$). Hence, we will construct an algorithm for the DRS problem.

Algorithm 3 Finds the minimum cardinality DRS given an algorithm to compute the k-nodes set that maximizes $\mathcal{P}_s$.

Require: Network $\mathcal{G} = (V, E)$
  for $k = 1, \dots, |V|$ **do**
    $\mathcal{O} := \mathcal{A}(G, k)$
    $P := \mathcal{P}_s(\mathcal{O})$
    if $P = 1$ **then**
      return k

Since the full set V always resolves the network, the program is well defined (i.e., it always returns *some* k). Moreover, it returns precisely the minimum budget k required in order to attain $\mathcal{P}_s = 1$. Lastly, it is clear that the runtime is at most $O(n(p_{\mathcal{A}}(n) + n))$ where $p_{\mathcal{A}}(n)$ is the running time of algorithm $\mathcal{A}$. Hence, we have a polynomial-time algorithm for the DRS problem. ∎

APPENDIX V
HIGH-VARIANCE SOURCE ESTIMATION

Denote by $T_{\mathcal{O}}$ the observed infection process. If the infection delays are Gaussian, $\mathcal{G}$ is a tree and no prior information about the source position is available, the maximum likelihood (ML) estimator is defined as $\widehat{s} \in \arg\max_{s \in V} \mathrm{P}(s|T_{\mathcal{O}})$, which has a tractable closed form [25].[12] In particular, given a set of observers $\mathcal{O} = \{o_1, o_2, \dots, o_k\} \subseteq V$, the vector of observed infection delays $\tau = [t_2 - t_1, \dots, t_k - t_1] \in \mathbb{R}^{k-1}$ is distributed as $\mathcal{N}(d_{s,\mathcal{O}}, \mathbf{\Lambda}_{\mathcal{O}})$ where $d_{s,\mathcal{O}}$ is the distance vector of Definition 2 and the covariance matrix $\mathbf{\Lambda}_{\mathcal{O}}$ is

$$\mathbf{\Lambda}_{\mathcal{O},(k,i)} = \sigma^2 \begin{cases} \sum_{(u,v)\in\mathcal{P}(o_1,o_{k+1})} w_{uv}^2 & k = i \\ \sum_{(u,v)\in\mathcal{P}(o_1,o_{k+1})\cap\mathcal{P}(o_1,o_{i+1})} w_{uv}^2 & k \neq i, \end{cases} \quad (5)$$

with $\mathcal{P}(x, y)$ denoting the set of edges in the unique path between node x and node y. Hence the ML estimator is

$$\begin{aligned} \widehat{s} &\in \arg\max_{s\in V} \frac{\exp\Big(-\frac{1}{2}(\tau - \mathbf{d}_{s,\mathcal{O}})^\top \mathbf{\Lambda}_{\mathcal{O}}{}^{-1}(\tau - \mathbf{d}_{s,\mathcal{O}})\Big)}{|\mathbf{\Lambda}_{\mathcal{O}}|^{1/2}} \\ &= \arg\max_{s\in V} \Big[\mathbf{d}_s^\top \mathbf{\Lambda}_{\mathcal{O}}{}^{-1}(\tau - \frac{1}{2}\mathbf{d}_{s,\mathcal{O}})\Big]. \end{aligned} \quad (6)$$

On non-tree networks, the multiplicity of paths linking any two nodes makes source estimation more challenging. As claimed in [25], the same estimator can be used as an approximation of the ML estimator for a non-tree network by assuming that the diffusion happens only through a BFS (*Breadth-First-Search*) tree rooted at the (unknown) source. In this case the paths which appear in the definition of the covariance matrix $\mathbf{\Lambda}_{\mathcal{O}}$ are computed on the BFS tree rooted at the candidate source considered. Hence $\mathbf{\Lambda}_{\mathcal{O}}$ depends on the candidate source and the ML estimator is

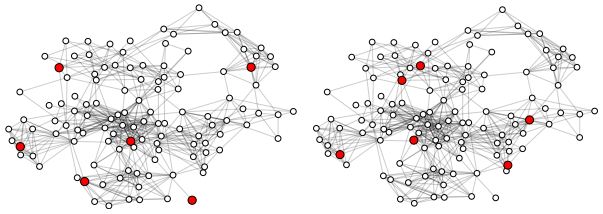
$$\widehat{s}_{\mathrm{bfs}} \in \arg\max_{s\in V} \frac{\exp\Big(-\frac{1}{2}(\tau - \mathbf{d}_{s,\mathcal{O}})^\top \mathbf{\Lambda}^s_{\mathcal{O}}{}^{-1}(\tau - \mathbf{d}_{s,\mathcal{O}})\Big)}{|\mathbf{\Lambda}^s_{\mathcal{O}}|^{1/2}}. \quad (7)$$

In this work, we adopt (7) as the source estimator in the noisy case. In fact, even if our edge delays are truncated Gaussians, under the hypothesis of sparse observations, we can apply the Central Limit Theorem (CLT) to approximate the sum of the edge delays with Gaussian random variables: if all edges have the same weight we can apply the CLT for i.i.d. random variables; if this is not the case, we can apply Lyapunov's version of CLT.[13]

[12]Note that the model of [25] additionally assumed infected observers knew the neighbor that infected them; this assumption is not required for our work.

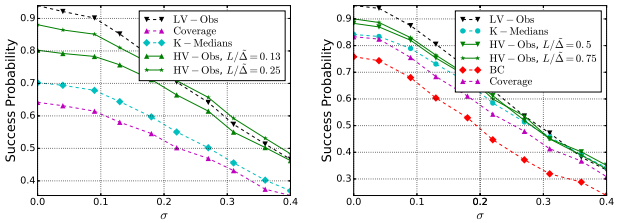
[13]Lyapunov condition with $\delta = 1$ is easily verified for a sequence of independent and uniformly bounded random variables (see Example 27.4 in [5] for more details).

APPENDIX VI
ADDITIONAL FIGURES

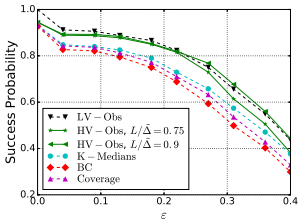


(a) LV-OBS (b) HV-OBS, $L/\widetilde{\Delta} = 0.5$

Fig. 8. The oberver placements of LV-OBS and HV-OBS with $L/\widetilde{\Delta} = 0.5$ and $k = 5\%$ on the F & F network are very different; LV-OBS contains leafs while HV-OBS has shorter spacing.

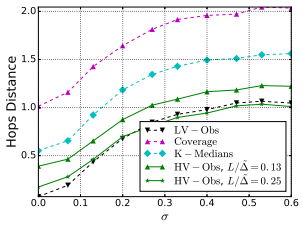


(a) CR, 5% observers (b) F & F, 5% observers

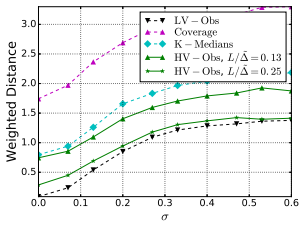


(c) FB, 5% observers

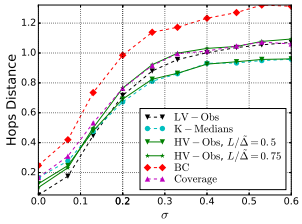
Fig. 9. Success probability $\mathcal{P}_s$ as variance is increased on a uniform transmission model (Section V-D).



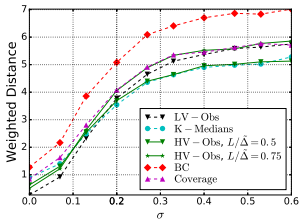
(a) California, 5% observers



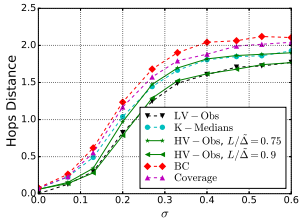
(b) California, 5% observers



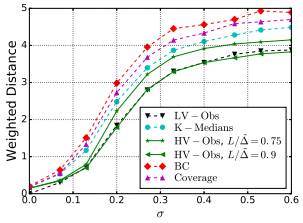
(c) Friends & Families, 5% observers



(d) Friends & Families, 5% observers



(e) Facebook, 5% observers



(f) Facebook, 5% observers

Fig. 10. Expected distance in number of edges (left column) and in weighted path length (right column) for the datasets and setting of Section IV