A General Framework for Sensor Placement in Source Localization

Brunella Spinelli, L. Elisa Celis, Patrick Thiran

Abstract—When an epidemic spreads in a given network of individuals or communities, can we detect its source using only the information provided by a small set of nodes? We propose a general framework that incorporates two dimensions. First, we can either rely exclusively on a set of selected nodes (i.e., *sensors*) which always reveal their state independently of any particular epidemic (these are called *static*), or we can add some sensors (called *dynamic*) as an epidemic spreads, depending on which additional information is required. Second, the method can either localizes the source after an epidemic has spread through the entire network (*offline*), or while the epidemic is ongoing (*online*).
We empirically study the performance of offline and online localization both with and without dynamic sensors. Our analysis shows that, by using dynamic sensors, the number of sensors necessary to localize the source is reduced by up to a factor of 10 and that, even with high-variance transmission delays, the source can be localized by using fewer than 5% of the nodes as sensors.

Index Terms—Epidemics; Source Localization; Sensor Placement.

1 Introduction

Computer worms, or rumors spreading on social networks, often trigger the question of how to identify the source of an epidemic. This question also arises in epidemiology, when the origin of a disease outbreak is investigated. For these reasons source localization has received considerable attention in the past few years. Because of its combinatorial nature, it is inherently difficult: the infection of a few nodes can be explained by multiple and possibly very different propagations. Researchers have considered various models and algorithms that differ in the epidemic model and in the information used for source localization. Such models are often not realistic, either because they rely on strong assumptions about the features of the epidemic (tree networks, deterministic transmission delays, etc.) or because they require an overwhelming amount of information. At the end of this section we position our work with respect to the assumptions that are most commonly made. In Section 8 we give a more general discussion of the state-of-the-art.

When studying source localization, the cost of collecting information cannot be disregarded. In fact, data collection is never inexpensive; moreover, due to privacy concerns, individuals are becoming aware of the value of their data, hence are resistant to share it for free [14]. In the case of infectious diseases, performing the necessary clinical tests and data analysis on many suspected households or communities can be very expensive, whereas the efficient allocation of resources can lead to enormous savings [47].

Another important concern is the timeliness of source localization: if an epidemic is detected while it is spreading, being able to promptly identify the source based on the incomplete information available can be essential for the activation of containment measures [1].

Driven by the demand for general models and by practical resource-allocation constraints, we make minimal assumptions on the epidemic spread and we design a flexible framework for information collection and source localization where the information can be either collected adaptively or at a fixed set of locations, and the source of an epidemic can be promptly localized.[1]

Model. We localize the source by using the information provided by a subset of nodes called sensors. When a node is chosen as a sensor, it can reveal its infection state and, if it is infected, its infection time. We have two possible types of sensors: *static* sensors and *dynamic* sensors. Static sensors are placed *a priori* in the network, independently of any particular epidemic instance. Dynamic sensors are placed *adaptively* while we perform source localization. Figure 1 depicts our approach to source localization.

Contributions. We propose a general framework for source localization that encompasses both *static* and *dynamic* sensor placement and that allows to localize the source both while the epidemic is still spreading (*online* localization) and after the epidemic has spread throughout the entire network (*offline* localization); see Table 1. This opens new possibilities, as to date most approaches assume that all sensors are static and that the source can be localized only after the epidemic spreads throughout the network. We show that when we can sequentially deploy dynamic sensors, the source is always correctly identified with only a few sensors even when the transmission delays are highly noisy. This result is very practical because it applies to general graphs.

We also propose several methods for choosing where to deploy the dynamic sensors and we compare them.

Because of its flexibility, the proposed framework can be used in a number of different applications, ranging from localizing the source of a belief that spread in the past to tracking the source of a disease outbreak in real time, and

• *All authors are with the School of Computer Science and Communication Systems, Ecole Polytechnique Fédérale de Lausanne, Switzerland*

1. A preliminary version of this work, focusing on online source localization using dynamic sensors (Section 6), was presented at the World Wide Web Conference in 2017 [49].

	Offline Localization	*Online Localization*
Static Sensor Placement	**S-OFF – Section 3** only static sensors, the source is localized after the epidemic, observations are always positive.	**S-ON – Section 4** only static sensors, the source is localized during the epidemic, observations can be positive or negative.
Dynamic Sensor Placement	**D-OFF – Section 5** static and dynamic sensors, the source is localized after the epidemic, observations are always positive.	**D-ON – Section 6** static and dynamic sensors, the source is localized during the epidemic, observations can be positive or negative.

TABLE 1: The source-localization methods and settings considered in this paper.

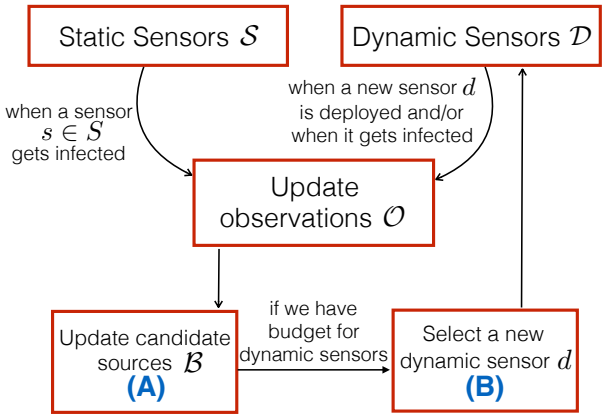


Fig. 1: Illustration of our approach to source localization. An epidemic is observed through the sensors (static $\mathcal{S}$ and dynamic $\mathcal{D}$). Based on the observations $\mathcal{O}$, we iteratively update a set $\mathcal{B}$ of candidate sources (step A) which is used, if the budget allows it, to guide the choice of an additional dynamic sensor (step B). The subroutines (A) and (B) depend on the setting used (among the four listed in Table 1).

from finding the source of a rumor in a network in which we constantly monitor a set of individuals, to detecting the patient-zero of an infection through ad-hoc interviews or clinical tests.

Experimental evaluation. Through extensive experiments on synthetic and real-world networks we evaluate our approach along two different axes: (1) Under budget-constraints for the number of sensors, we measure the uncertainty on the identity of the source (i.e., the number of nodes that have a positive probability of being the source given the available observations); and (2) when the budget for sensors is not limited, we assess the number of sensors needed to exactly identify the source.

Our analysis highlights that a strategy that uses dynamic sensors dramatically outperforms a static strategy with the same budget: By choosing fewer than 5% of the nodes as sensors we improve the success rate of finding the source from approximately 30% to approximately 92% (see Figure 9(a)). Moreover, when we do not have a limited budget on the number of dynamic sensors, we can localize the source with a small number of sensors: between 3% and 6% of the network nodes depending on the network topology (see Figures 7(a) and 7(b)). The reason for these improvements is that, using dynamic sensors, we can progressively reduce the network to a small sub-network whose nodes always include the source.

We also show that, given a set of constraints (how many sensors can be deployed, whether they can be deployed adaptively, ...), the choice of the sensors strongly affects the performance of source localization. In particular, we evaluate different choices of the static sensors (in Section 7.3) and different methods for choosing the dynamic sensors (in Section 7.6). We also study the effect of varying the proportion of static versus dynamic sensors, showing that we can save some resources by choosing a small budget for static sensors, but not too small as we might pay with a longer time for localizing the source (see Figure 9(b)).

Finally we demonstrate that, by using all the information available as soon as it becomes available, we can greatly enhance the timeliness of source localization and restrict the search to a small set of candidate sources when the number of infected nodes is still small (see Figures 6(a) and 6(b)).

What we assume.

(A.1) We assume that the contact network is known. This is a common assumption when studying source localization (see, e.g., [2], [36], [40], [41], [46]). In the contexts where the network topology is not known, or only partially known we should first estimate the network topology. This difficult and interesting task is out of the scope of this paper; recently, solutions were proposed, for example, by Farajtabar et al. [18], Fu et al. [19], [20] and Gomez-Rodriguez et al. [22].

(A.2) We assume that, when a node is a sensor, it reveals its *state* (healthy or infected). If it is infected, it also reveals the time at which it became infected. This is not a strong assumption because, by interviewing users of a social network or patients affected by a disease, a (possibly noisy) observation of the infection time might become available [55].

What we do *not* assume. Estimating the source of an epidemic is intrinsically difficult. For the sake of tractability, prior work often makes assumptions which are not always feasible in practice. We list some assumptions which are not needed for our work.

Notation

$\mathbb{N}$ / $\mathbb{N}^+$	positive integers including / excluding 0
$\mathcal{G}(E,V)$	contact network
$N = \lvert V \rvert$	network size
w_{uv}	weight of edge (u,v) ($\in \mathbb{R}^+$)
X_{uv}	infection delay on edge (u,v)
$d(x,y)$	weighted distance between x and y ($\in \mathbb{R}^+$)
$v^\star$	source
$t^\star$	starting time of the epidemic
t_u	infection time of node u
$T(v,u)$	$t_u - t^\star$, infection delay of node u when $v^\star = v$
$\mathcal{S}$	set of static sensors
$\mathcal{D}$	set of dynamic sensors
$\mathcal{U}$	$\mathcal{S} \cup \mathcal{D}$, set of all sensors
K_s	budget for static sensors
K_d	budget for dynamic sensors
K	$K_s + K_d$, total budget for sensors
$\tau^\star$	time at which source localization begins
θ	deployment delay
$\mathcal{O}$	set of observations
$\omega = (u, t_u)$	observation of node u: if u is not infected, $t_u = \emptyset$
$\mathcal{B}$	set of candidate sources

(B.1) Knowledge of the *state of every node* at a given point in time. It might be prohibitively expensive to maintain a very large number of monitoring systems [58]. Instead, we detect the source based on the infection time of a very small set of nodes.

(B.2) Knowledge of the *time at which the epidemic starts*. This information is, in most practical cases, not available [26], [40]. Hence we do not make assumptions about the starting time of the epidemic.

(B.3) Observation of *multiple epidemics*. Observing multiple epidemics started by the same source certainly helps in its localization [17], [40]. In this work, we consider a single epidemic, because we are interested in localizing the source *while* the epidemic spreads.

(B.4) A specific *class of network topologies*. Having a unique path between any two nodes makes source localization much easier [26], hence tree topologies are often assumed. Instead, our methods work on arbitrary graphs.

(B.5) *Deterministic or discretized transmission delays*. When the transmission delays are deterministic, the epidemic itself is deterministic given the position of the source. Therefore, if the source is unknown, tracking back its position becomes much easier [48]. Also, assuming that infection times are discrete-valued is limiting and can result in a loss of important information [7]. We derive our algorithm assuming transmission delays to be randomly drawn from bounded-support continuous distributions, which include deterministic delays as a particular case and can, in practice, approximate unimodal distributions with unbounded support (see Figures 10(c) and 10(d)).

2 PRELIMINARIES

2.1 Network Model

We model a set of contacts with a weighted graph $\mathcal{G}(V,E)$. For every $(u,v) \in E$, the weight $w_{uv} \in \mathbb{R}^+$ is equal to the average time it takes for an infection to spread from u to v. $\mathcal{G}$ is undirected, i.e., $w_{uv} = w_{vu}$ for every $(u,v) \in E$. The distance $d(u,v)$ between two nodes u and v is the minimal sum of edge weights along a path connecting u and v.

2.2 Epidemic Model

An epidemic starts from a single source at an unknown time $t^\star$. The identity of the source is an unobserved random variable $v^\star$ which takes values in the node set V.

At any time, every node is in one of two possible states: S (*Susceptible*) or I (*Infected*). For every edge $(u,v) \in E$, let X_{uv} be the time it takes for an infection to spread from u to v. X_{uv} is called the *infection delay* on edge (u,v) and is a positive random variable with mean $\mathbf{E}[X_{uv}] = w_{uv}$. Denote by t_u the infection time of a node u. Then a non-infected neighbor v of u gets infected at time $t_v = t_u + X_{uv}$. The variables $\{X_{uv}\}_{(u,v)\in E}$ are mutually independent. When $v^\star = v$ we denote by $T(v,u)$ the total time it takes for the infection to spread from v to a node $u \in V$. This model implies that all nodes eventually become infected.

We do not assume a precise distribution for the infection delays. Instead, we assume that there exists $\varepsilon \in [0,1)$ such that for every $(u,v) \in E$, the support of X_{uv} is contained in $[(1-\varepsilon)w_{uv}, (1+\varepsilon)w_{uv}]$. We call ε the *variance parameter* because it encodes how much the infection delays can deviate from their mean.

For $\varepsilon = 0$, $X_{uv} = w_{uv}$ for every $(u,v) \in E$, and we say that the epidemic is deterministic; for $\varepsilon > 0$, $w_{uv} \cdot \varepsilon$ gives an upper bound on the deviation of X_{uv} from its mean. By letting the maximum deviation be proportional to the edge weights we make sure that our transmission model is not trivial: If the maximum deviation was constant, the impact of the variance would depend on the scale of the edge weights and, for large w_{uv}, X_{uv} would be effectively deterministic.

In our experiments, we consider mostly the case in which X_{uv} is uniformly distributed on $[w_{uv}(1-\varepsilon), w_{uv}(1+\varepsilon)]$. In this case, the variance of X_{uv} is $\mathrm{Var}(X_{uv}) = w_{uv}^2\varepsilon^2/3$ (which is the maximum variance of unimodal distributions with support $[(1-\varepsilon)w_{uv}, (1+\varepsilon)w_{uv}]$ [43]). However we also test the performance of our methods when X_{uv} has unbounded support (see Section 7.8).

2.3 Online & Offline Source Localization

Depending on the application of interest, it can be desirable to localize the source *while* the epidemic is still spreading (e.g., for a disease spreading now) or only *after* it has propagated throughout the network (e.g., for an epidemic that happened in the past or when a timely investigation is not needed). In the former case, we speak about *online* source localization, and in the latter case about *offline* source localization. In contexts where the data becomes available while the epidemic spreads, online localization has the potential to identify the source (or a small set of candidate sources) before the epidemic propagates throughout the entire network.

Offline localization, instead, is the only possible approach to source localization when we study epidemics that occurred in the past. Moreover, it is the setting that is commonly used in the literature (see Section 8).

Source localization might not be instantaneous. Let $\tau^\star$ be the time at which we start investigating the identity of the source and, for every $v \in V$, let t_v denote the infection time of v. We can give the following definition.

Definition 1 (Online/offline method)**.** *A method for source localization is said to be an online (respectively, offline) method if* $\tau^\star < \max_{v\in V} t_v$ *(resp.,* $\tau^\star \geq \max_{v\in V} t_v$*).*

The main difference between online and offline source localization is that, when performing offline localization, the full picture of the process is already available at time $\tau^\star$.

We present a framework that naturally encompasses both the offline and the online regimes. We study offline source localization in Section 3 and 5, online source localization in Section 4 and 6.

2.4 Sensors

We use the information provided by a subset of nodes which we call sensors.

Definition 2 (Sensor)**.** *A node is a sensor if it can reveal its infection state (S or I) and, if it is infected, its infection time.*

Note that if $v^\star = v$ and v is a sensor, v provides information in the same way as any other sensor i.e., it only reveals, if infected, its infection time (which would be equal to $t^\star$), but it does not reveal itself to be the source.

Definition 3 (Static/dynamic sensor)**.** *A sensor is said to be static if it is chosen independently of any epidemic. In contrast, we say that a sensor is dynamic if it is chosen in order to localize the source of a particular epidemic.*

We assume that we have a budget K for the total number of sensors and we consider two regimes for sensor placement: static sensor placement (all K sensors are static) and dynamic sensor placement ($K_s > 0$ static sensors and $K_d = K - K_s$ dynamic sensors). Note that we never set $K_s = 0$ because otherwise no sensor would be deployed in the network when the epidemic starts spreading and the detection of the source would be trivially impossible.

The set of static (respectively, dynamic) sensors is denoted by $\mathcal{S}$ (resp., $\mathcal{D}$) and the set of all sensors is $\mathcal{U} = \mathcal{S} \cup \mathcal{D}$.

In online localization, the dynamic sensors are chosen while the epidemic spreads; in the offline regime, they are chosen while we perform source localization, i.e., by the time they are chosen the epidemic has already spread throughout the network.

As mentioned in Section 2.3, the localization process is generally not instantaneous because it can require a sequence of steps (e.g., updates of the estimated identity of the source when more information is available or when additional dynamic sensors are deployed). Let τ denote the time at which a step in the localization process is taken: At time τ a sensor u gives information in two possible ways: If it became infected at $t_u \leq \tau$, it reveals its infection time t_u; otherwise it informs about its susceptible state. In the first (respectively, second) case we say that the sensor gives a *positive* (resp., *negative*) *observation*.

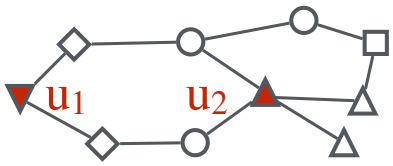


Fig. 2: An unweighted network with two sensors u_1 and u_2. Different shapes represent different equivalence classes, i.e., groups of nodes which are equivalent with respect to the sensors (red). In this example there are 5 classes.

In offline source localization, all observations are positive. Instead, in online source localization both static and dynamic sensors can give positive or negative observations and, as we will see in Section 4, both positive and negative observations contribute to the localization.

We represent each observation ω as a tuple $\omega \triangleq (u, t_u)$ where $u \in V$ denotes the sensor and $t_u \in \mathbb{R}$ is the infection time of u if the observation is positive, $t_u = \emptyset$ if the observation is negative.

Table 1 summarizes the source-localization settings that we consider and the relationships between the definitions of static/dynamic sensors, online/offline source-localization and positive/negative observations. Figure 1 illustrates our high-level approach to source localization, highlighting the different roles of static and dynamic sensors.

2.5 Localization Based on Relative Distances

Let $v^\star = v$. If $\varepsilon = 0$, $T(v,u) = t_u - t^\star = d(v,u)$; if $\varepsilon > 0$ and if the path connecting v and u is unique, then $\mathbf{E}[T(v,u)] = d(v,u)$.[2] Hence if $t^\star$ were known, $T(v,u)$ could be interpreted as a proxy for $d(v,u)$, hence the infection time t_u could be directly used to localize the source.

However, as we assume that $t^\star$ is unknown, we cannot use the infection time of a single sensor to infer the identity of the source. Instead, we use the *differences* between the infection times of pairs of sensors. If the sensor set is $\mathcal{U}$, we use the differences $\{t_u - t_z\}_{u,z\in\mathcal{U}}$. Borrowing the terminology used for the localization of transmitting devices, our work is a TDOA (Time Difference Of Arrivals) approach to source localization (in contrast with a TOA approach where the Time Of Arrivals - and the starting time $t^\star$ - are used) [31].

Consider now the case of a deterministic epidemic ($\varepsilon = 0$) and two possible sources v_1 and v_2. We can distinguish which of the nodes is the source based on the set $\{t_u - t_v\}_{u,v\in\mathcal{U}}$ if and only if there exist $u_1, u_2 \in \mathcal{U}$ such that

$$d(u_1, v_1) - d(u_1, v_2) \neq d(u_2, v_1) - d(u_2, v_2).$$

Definition 4 (Distinguished nodes)**.** *Let* $\mathcal{U} \subseteq V$ *with* $|\mathcal{U}| \geq 2$*. A node* v_1 *is distinguished from a node* v_2 *by* $\mathcal{U}$ *if and only if there exist* $u_1, u_2 \in \mathcal{U}$ *such that*

$$d(u_1, v_1) - d(u_1, v_2) \neq d(u_2, v_1) - d(u_2, v_2). \quad (1)$$

2. If the path connecting v and u is not unique, then $\mathbf{E}[T(v,u)] \leq d(v,u)$ and this bound becomes looser when there are many alternative paths connecting u and v whose length is similar to $d(u,v)$. In our setting, it does not hold, in general, that $\mathbf{E}[T(v,u)] = d(v,u)$ and none of our results relies on this identity. Instead, our results are based on bounds for the infection times that hold for general networks (see, e.g., Proposition 2).

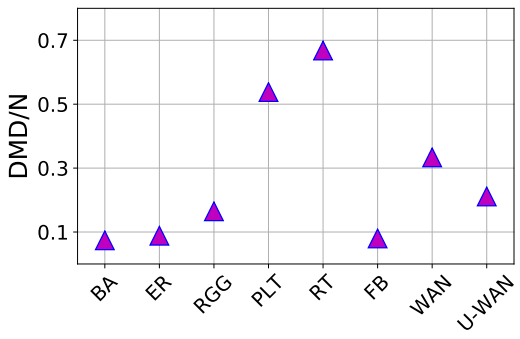


Fig. 3: Approximate DMD as a fraction of the network size. The approximation is computed with the $(1+o(1))\log(N)$-approximation algorithm of Chen et al. [9]. The networks considered are presented in Section 7.2.

In this case we say that v_1, v_2 are distinguished by the pair u_1, u_2.

Definition 5 (Equivalent nodes)**.** *Let $\mathcal{U} \subseteq V$ with $|\mathcal{U}| \geq 2$. A node v_1 is said to be equivalent to a node v_2 with respect to $\mathcal{U}$, (which we write $v_1 \sim v_2$) if and only if, for every $u_1, u_2 \in \mathcal{U}$*

$$d(u_1, v_1) - d(u_1, v_2) = d(u_2, v_1) - d(u_2, v_2). \tag{2}$$

The relation $\sim$ of Definition 5 is reflexive, symmetric, and transitive, hence it defines an *equivalence relation*. Therefore, a set of sensors $\mathcal{U}$ partitions V in *equivalence classes* (an example is given in Figure 2). We denote by $[v]_{\mathcal{U}}$ the class of v, i.e., the set of all nodes that are equivalent to v.

A set $\mathcal{Z}$ such that for every $v_1, v_2 \in V$, v_1 and v_2 are distinguished by $\mathcal{Z}$ is called a *Double Resolving Set* (DRS) of $\mathcal{G}$. The problem of finding the minimum-size DRS of a network is known as the *Minimum Double Resolving Set Problem* [6]. Finding a minimum-size DRS is NP-hard and a $(1+o(1))\log(N)$-approximation algorithm was proposed by Chen et al. [9].

In a setting where the starting time of the epidemic is unknown, yet some of the node infection-times can be used to localize the source, sensor placement is naturally related to the DRS problem. In fact, if the transmission delays are deterministic, choosing all nodes in a DRS as sensors guarantees that the source can always be localized. However, in addition to the hardness of finding a minimum-size DRS, there are two other drawbacks of this choice. First, choosing all the nodes in a minimum-size DRS as sensors is often not a feasible solution because the number of sensors required can be prohibitively large (see Figure 3). Second, even if we could choose all the nodes in a minimum-size DRS as sensors, we could guarantee that the source is correctly localized only when the transmission delays are deterministic.

For this reason, studying how to allocate a limited number of sensors in order to guarantee a good performance of source localization is a crucial aspect of the source localization problem [48].

The connection between source localization and the DRS problem is at the basis of the approach to static sensor placement proposed by Spinelli et al. [48]. In Section 7 we choose the static sensors based on the latter work.

We now define the *distance vector* of a candidate source.

Definition 6 (Distance vector)**.** *Let $\mathcal{U} \subseteq V$ with $|\mathcal{U}| = K \geq 2$ be a set of sensors and let $u_1 \in \mathcal{U}$. For each candidate source $v \in V$ the distance vector of v (with respect to u_1) is $\mathbf{d}_{v,u_1} \in \mathbb{R}^{K-1}$ with entries $d(v,u_i) - d(v,u_1)$ for $2 \leq i \leq K$.*

The following lemma, equivalent to Lemma 3.1 in [9], shows that the equality between distance vectors of different candidate sources does not depend on the choice of the reference sensor u_1 of Definition 6.

Lemma 1. *Let $\mathcal{U} \subseteq V$ with $|\mathcal{U}| = K \geq 2$, $u_1 \in \mathcal{U}$ and let $v_1, v_2 \in V$. Then, $[v_1]_{\mathcal{U}} = [v_2]_{\mathcal{U}}$ if and only if $\mathbf{d}_{v_1,u_1} = \mathbf{d}_{v_2,u_1}$, for any choice of the reference observer u_1.*

3 S-OFF: Offline Localization with Static Sensors

In this section and in the following ones (Sections 4-6), we present the four settings listed in Table 1. For the sake of readability, the technical details are presented in Appendices C-E.

We first describe the S-OFF algorithm with which we perform offline source localization using only static sensors. This is the setting most of the literature works with (see Section 8). In contrast to other approaches we are more interested in determining all nodes that are possible sources, to make sure that we did not miss the actual source, rather than in isolating only one node that would maximize the likelihood of being the source. This approach paves the way for the correctness results of D-OFF and D-ON in Sections 5 and 6.

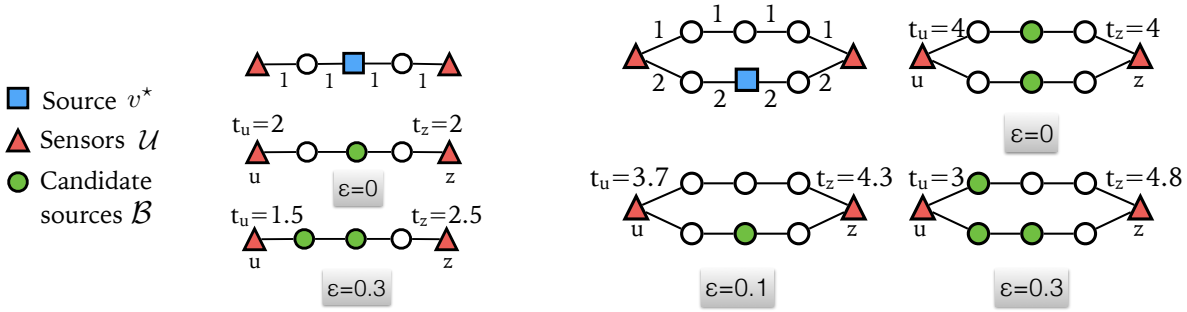
We do not use any dynamic sensor, hence $\mathcal{U} = \mathcal{S}$ and $|\mathcal{U}| = K = K_s$. As we localize the source offline, a set of positive observations of the form $\mathcal{O} = \{(u, t_u) : u \in \mathcal{U}\}$ is available at the beginning of the localization process.

Using $\mathcal{O}$, we want to determine the possible sources, i.e., the set of *candidate sources*

$$\mathcal{B} \triangleq \{v \in V : \mathbf{P}(\mathcal{O} | v^\star = v) > 0\}. \tag{3}$$

$\mathcal{B}$ depends not only on $v^\star$ but also on the particular realization of the infection times and on the variance parameter ε. In fact, when the variance parameter ε is larger, given a set of observations, the uncertainty on the identity of the source is higher because a larger set of nodes can initiate epidemics that result in the observed infection times.
In Figure 4, we display the set $\mathcal{B}$ for a few simple examples. Figure 4(a) illustrates that when the infection times are not deterministic, it is in general more difficult to localize the source, as we can expect, because the infection times can substantially deviate from their mean value. More surprisingly, the reverse can also occur, as shown in the example of Figure 4(b): For specific network topologies and moderate values of ε, non-deterministic infection delays can make source-localization easier than in the deterministic case. This is due to the maximum deviation of the infection delay X_{uv} from w_{uv} being proportional to w_{uv} itself (see Section 2.2): Observing a very extreme value of the difference $t_u - t_v$ between the infection time of two sensors u and v can give information about the path along which the epidemic spread, hence about the identity of $v^\star$.

Sections 3.1 and 3.2 explain how $\mathcal{B}$ can be computed in practice.



(a) (b)

Fig. 4: Examples of sets of candidate sources $\mathcal{B}$: The set $\mathcal{B}$ depends on the (unknown) source $v^\star$, on the variance parameter ε, and on the realization of the random infection times. For each of the two setups (a) and (b), the first graph shows the actual source and the network topology, and the following graphs show different realizations of the observations $\mathcal{O}$ for different noise parameters ε. **(a)**: For the weighted graph at the top, when ε is large (bottom), $\mathcal{B}$ can be larger than for $\varepsilon = 0$ (middle). **(b)**: For the weighted graph at the top-left, when ε is positive but small (bottom-left), $\mathcal{B}$ can be smaller than when $\varepsilon = 0$ (top-right); when ε is large (bottom-right) $\mathcal{B}$ can be larger than in the two previous cases.

3.1 Deterministic Epidemics

We explained in Section 2.5 that, when the starting time $t^\star$ of the epidemic is unknown, no single observation taken in isolation is informative about the identity of the source. Instead, a set of two (or more) observations gives information on the identity of the source. Therefore we start defining, for two observations ω_1, ω_2, the event of observing ω_1 and ω_2 jointly.

Definition 7 (Event A_{ω_1,ω_2})**.** *Let $\omega_1 \triangleq (u_1, t_{u_1})$, and $\omega_2 \triangleq (u_2, t_{u_2})$, $\omega_1 \neq \omega_2$, be two observations. We define the event A_{ω_1,ω_2} as $A_{\omega_1,\omega_2} \triangleq \{T(v^\star, u_1) - T(v^\star, u_2) = t_{u_1} - t_{u_2}\}$.*

For every pair of observations ω_1, ω_2, $\omega_1 \neq \omega_2$, we define

$$\mathcal{B}_{\omega_1,\omega_2} \triangleq \{v \in V : \mathbf{P}(A_{\omega_1,\omega_2} | v^\star = v) > 0\}. \tag{4}$$

When epidemics are deterministic ($\varepsilon = 0$), $\mathcal{B}$ can be easily computed using the result of the following proposition (see Appendix C for proof and complementary lemmas).

Proposition 1. *Let $\mathcal{O}$ be a set of observations and let $\varepsilon = 0$. Let $\omega_1 \triangleq (u_1, t_{u_1}) \in \mathcal{O}$ be a fixed observation, which we call the reference observation. Then, the set of candidate sources $\mathcal{B}$ is*

$$\mathcal{B} = \bigcap_{\omega \in \mathcal{O} \setminus \{\omega_1\}} \mathcal{B}_{\omega_1,\omega}.$$

From Proposition 1, we can compute the candidate set $\mathcal{B}$ with Algorithm 1.

Algorithm 1 S-OFF - deterministic epidemic

Require: $\mathcal{O}$ set of observations
  $\mathcal{B} \leftarrow V$
  $\omega_1 \triangleq (u_1, t_{u_1}) \leftarrow$ Sample($\mathcal{O}$)
  for $(u, t_u) \in \mathcal{O} \setminus \{\omega_1\}$ **do**
    for $v \in \mathcal{B}$ **do**
      if $d(v, u) - d(v, u_1) \neq t_u - t_{u_1}$ **then**
        remove v from $\mathcal{B}$
  return $\mathcal{B}$

The running time of Algorithm 1 is $O(K_s N)$. When $v^\star = v$, since $T(v, u_i) - T(v, u_1) = t_i - t_1$ is deterministic and equal to $d(v, u_i) - d(v, u_1)$ for any $u_1, u_i \in \mathcal{U}$, the set $\mathcal{B}$ of candidate sources returned by Algorithm 1 is equal, because of Lemma 1, to $[v]_{\mathcal{U}}$. Hence, in the inner **for** of Algorithm 1 it would be enough to loop over a set of representatives v of the equivalence classes in $\mathcal{B}$. However, for consistency with the algorithms presented in the following sections, we keep the version of the algorithm given in Algorithm 1, where the loop is over all $v \in \mathcal{B}$.

3.2 Non-deterministic Epidemics

When the infection delays are not deterministic ($0 < \varepsilon < 1$), verifying analytically if $\mathbf{P}(\mathcal{O} | v^\star = v) > 0$ is computationally intense for two reasons: the interdependence of the events $\{A_{\omega_i,\omega_j}\}_{\omega_i \neq \omega_j}$ and, in meshed networks, the multiplicity of possible propagation paths. To overcome these difficulties, we compute a superset $\widetilde{\mathcal{B}} \supseteq \mathcal{B}$ of the set of candidate sources using the result of the following proposition.

Proposition 2. *Let $0 < \varepsilon < 1$, let $\omega_1 \triangleq (u_1, t_{u_1})$, $\omega_2 \triangleq (u_2, t_{u_2}) \in \mathcal{O}$, $\omega_1 \neq \omega_2$, and let $v \in \mathcal{B}$. Then*

$$|d(v, u_1) - d(v, u_2) - t_{u_1} + t_{u_2}| \leq \varepsilon(d(v, u_1) + d(v, u_2)). \tag{5}$$

Based on Proposition 2, we define

$$\widetilde{\mathcal{B}} \triangleq \{v \in V : |d(v, u_1) - d(v, u_2) - t_{u_1} + t_{u_2}| \leq \varepsilon(d(v, u_1) + d(v, u_2)) \quad \forall (u_1, t_{u_1}), (u_2, t_{u_2}) \in \mathcal{O}\}. \tag{6}$$

Combining (6) and Proposition 2 we have that $v^\star \in \mathcal{B}$.

Remark 1. *The running time of computing $\widetilde{\mathcal{B}}$ is $O(K_s^2 N)$. In fact, in contrast to Algorithm 1, we need to loop over all pairs $(\omega_1, \omega_2) \in \mathcal{O} \times \mathcal{O}$ with $\omega_1 \neq \omega_2$. If we fix a reference observation ω_1 as in Proposition 1 and, for a candidate source v, we verify (5) only for the pairs $(\omega_1, \omega_i), \omega_i \in \mathcal{O}$, we would obtain a larger set of candidate sources that might also includes some nodes v such that $\mathbf{P}(\mathcal{O} | v^\star = v) = 0$.*

In fact, given $v \in V$ and $\omega_1 \triangleq (u_1, t_1), \omega_i \triangleq (u_i, t_i), \omega_j \triangleq (u_j, t_j) \in \mathcal{O} \backslash \{\omega_1\}$, $\omega_i \neq \omega_j$, it is possible that

$$\begin{cases} |d(u_1, v) - d(u_i, v) - t_1 + t_i| \leq \varepsilon(d(u_1, v) + d(u_i, v)) \\ |d(u_1, v) - d(u_j, v) - t_1 + t_j| \leq \varepsilon(d(u_1, v) + d(u_j, v)) \end{cases}$$

and yet $|d(u_i, v) - d(u_j, v) - t_i + t_j| > \varepsilon(d(u_1, v) + d(u_j, v))$, whence $v \notin \mathcal{B}$ (see Appendix F for an example).

Remark 2. *The set $\widetilde{\mathcal{B}}$ defined in* (6) *is, in general, strictly larger than $\mathcal{B}$. When computing $\widetilde{\mathcal{B}}$ we consider only two observations at a time; however, if ω_1, ω_2 and ω_3 are three distinct observations, it is possible that $\mathbf{P}(v^\star = v | \omega_i, \omega_j) > 0$ for every $i, j \in \{1, 2, 3\}$ but $\mathbf{P}(v^\star = v | \omega_1, \omega_2, \omega_3) = 0$ (similar situations can arise for larger sets of observations). When epidemics are non-deterministic, verifying if $\mathbf{P}(v^\star = v | \mathcal{O}) > 0$ for a set $\mathcal{O}$ of arbitrary cardinality would be computationally intractable, roughly exponential in the cardinality of the set. Thus, we approximate $\mathcal{B}$ with $\widetilde{\mathcal{B}}$. Section 7 demonstrates empirically that the size of $\widetilde{\mathcal{B}}$ decreases very fast over the iterations of the algorithm, indicating that our approximation is not too loose.*

We conclude this section with a proposition stating that, for low ε and $v^\star = v$, $\widetilde{\mathcal{B}} \subseteq [v]_{\mathcal{U}}$. This guarantees that when ε is sufficiently small, identifying the source is at least as easy as in the deterministic case.

Proposition 3. *Let $\mathcal{U}$ be the sensor set. Let*

$$\Delta(\mathcal{U}) \triangleq \max_{u \in \mathcal{U}, v \in V} d(v, u)$$

and

$$\delta(\mathcal{U}) \triangleq \min_{[v_1]_{\mathcal{U}} \neq [v_2]_{\mathcal{U}}} \max_{u_1, u_2 \in \mathcal{U}} |d(v_1, u_1) - d(v_1, u_2) - d(v_2, u_1) + d(v_2, u_2)|. \quad (7)$$

If $\varepsilon < \varepsilon_0 \triangleq \delta(\mathcal{U})/4\Delta(\mathcal{U})$ and $v^\star = v$, then $\widetilde{\mathcal{B}} \subseteq [v]_{\mathcal{U}}$.

If additional conditions on the edge weights or on the network topology are given, more refined bounds ε_0 in Proposition 3 can be derived. For example, in a *tree* with weights $w_{uv} > C \in \mathbb{R}^+$, the uniqueness of the path between two any nodes yields that $\delta(\mathcal{U}) \geq 2C$ for every $\mathcal{U}$. Hence, in this case, the statement holds for $\varepsilon < C/2\Delta(\mathcal{U})$.

4 S-ON: Online Localization with Static Sensors

In online localization with static sensors (S-ON), the localization of the source can be seen as a process in which we iteratively refine the set $\mathcal{B}$ of candidate sources while we gather more and more information about the epidemic.

Given a static sensor set $\mathcal{U}$, the final outcome of S-ON and S-OFF is identical in terms of the nodes that are identified as candidate sources. The difference is the ability of S-ON to restrict the search for the source to a very small subset of nodes when the epidemic has not yet propagated throughout the network.

In contrast to Section 3, where the observation set $\mathcal{O}$ contained all the observations available *at the end* of the epidemic, which could therefore only be positive, here $\mathcal{O}$ changes *while* the epidemic progresses. From a technical point of view, the difference between S-ON and S-OFF is that in S-ON some observations are negative. Hence the results of Proposition 1 (respectively, 2) must be refined to include in the computation of $\mathcal{B}$ (resp., $\widetilde{\mathcal{B}}$) the information given by negative observations. We start the localization process as soon as a sensor gets infected, i.e., at time $\tau^\star \triangleq \min_{u \in \mathcal{U}} t_u$. We denote by $\mathcal{O}_t$ the observation set at time t. At every time $t \geq \tau^\star$, $\mathcal{O}_t$ is the union of a set of positive observations and of a set of negative observations (see Section 2.4). We denote by $\mathcal{O}_t^+$ (respectively, $\mathcal{O}_t^-$) the set of positive (respectively, negative) observations at time t. This means that, for every $t \geq \tau^\star$ we have

$$\mathcal{O}_t^+ = \{(u, t_u) : u \in \mathcal{U}, t_u \leq t\},$$

$$\mathcal{O}_t^- = \{(u, \emptyset) : u \in \mathcal{U}, t_u > t\},$$

and

$$\mathcal{O}_t = \mathcal{O}_t^+ \cup \mathcal{O}_t^-.$$

Definition 8 (Event $A^t_{\omega_1, \omega_2}$)**.** *Let $t \in \mathbb{R}$ and let $\omega_1 \triangleq (u, t_u) \in \mathcal{O}_t^+, \omega_2 \triangleq (w, \emptyset) \in \mathcal{O}_t^-$ be two observations, one positive and one negative. We define the event $A^t_{\omega_1, \omega_2}$ as $A^t_{\omega_1, \omega_2} \triangleq \{T(v^\star, u) - T(v^\star, w) < t_u - t\}$.*

Note that, if $\omega_1 \neq \omega_2$ are two negative observations in $\mathcal{O}_t^-$, for every possible source there exists a starting time $t^\star \in \mathbb{R}$ such that both ω_1 and ω_2 hold at time t (and this remark generalizes to larger sets of negative observations). For this reason, sets of only negative observations are not useful to localize the source.

If only negative observations are available, we do not even know if there is an ongoing epidemic. However, combining negative observations and positive observations we can design an algorithm, S-ON, which, at any time t during the localization process, computes the smallest possible set of candidate sources: Given the information contained in $\mathcal{O}_t$, the set of candidates $\mathcal{B}_t$ computed by S-ON contains all and only the nodes that have a positive probability of being the source.

More formally, we define the candidate sources set at time t as

$$\mathcal{B}_t \triangleq \{v \in V : \mathbf{P}(\mathcal{O}_t | v^\star = v) > 0\}. \quad (8)$$

For every pair of positive observations ω_1, ω_2, $\omega_1 \neq \omega_2$, let $\mathcal{B}_{\omega_1, \omega_2}$ be defined as in (4). At time t, for every positive observation ω_1 and negative observation ω_2 we also define

$$\mathcal{B}_{\omega_1, \omega_2, t} \triangleq \{v \in V : \mathbf{P}(A^t_{\omega_1, \omega_2} | v^\star = v) > 0\}.$$

Then Proposition 1 can be extended as follows (see Appendix D for proof and complementary lemmas).

Proposition 4. *Let $t \in \mathbb{R}$, $\mathcal{O}_t$ be the set of observations at time t and $\varepsilon = 0$. Let $\omega_1 \triangleq (u_1, t_{u_1}) \in \mathcal{O}^+_{\tau^\star}$ be the first positive observation that we call the reference observation. Then, the set of candidate sources $\mathcal{B}_t$ is*

$$\mathcal{B}_t = \left(\bigcap_{\omega \in \mathcal{O}_t^+ \backslash \{\omega_1\}} \mathcal{B}_{\omega_1, \omega} \right) \bigcap \left(\bigcap_{\omega \in \mathcal{O}_t^-} \mathcal{B}_{\omega_1, \omega, t} \right).$$

Moreover, if $t, t' \in \mathbb{R}$, $t' > t$, $\mathcal{B}_{t'} \subseteq \mathcal{B}_t$.

Call $\tau^\star = t_1 < t_2 < .. < t_F = \tau_F$ the times at which the observation set changes. Denoting by t_u the infection time

of sensor u we have $\tau_F \triangleq \max_{u\in\mathcal{U}} t_u$. Let us also denote $\mathcal{O}_i \triangleq \mathcal{O}_{t_i}$.

Using the result of Proposition 4 we compute and update the set of candidate sources with Algorithm S-ON (see Algorithm 2). S-ON updates the set of candidate sources $\mathcal{B}$ at every time t_i, $1 \le i \le F$, producing a set that we call $\mathcal{B}_i$.

Algorithm 2 S-ON - deterministic epidemic

Require: Observation sets $\{\mathcal{O}_i^+\}_{i=1}^F$, $\{\mathcal{O}_i^-\}_{i=1}^F$
  $\mathcal{B}_0 \leftarrow V$
  $\omega_1 \triangleq (u_1, t_{u_1}) \leftarrow$ Sample($\mathcal{O}_1^+$)
  $\mathcal{O}_0^+ \leftarrow \{\omega_1\}$
  $i \leftarrow 1$
  while $i \le F$ and $|\mathcal{B}_{i-1}| > 1$ **do**
    $i \leftarrow i + 1$
    $\mathcal{B}_i \leftarrow \mathcal{B}_{i-1}$
    for $(u, t_u) \in \mathcal{O}_i^+ \backslash \mathcal{O}_{i-1}^+$ **do**
      for $v \in \mathcal{B}_i$ **do**
        if $d(u, v) - d(u_1, v) \neq t_u - t_{u_1}$ **then**
          remove v from $\mathcal{B}_i$
    for $(u, \emptyset) \in \mathcal{O}_i^-$ **do**
      for $v \in \mathcal{B}_i$ **do**
        if $d(u, v) - d(u_1, v) < t_i - t_{u_1}$ **then**
          remove v from $\mathcal{B}_i$
  return $\mathcal{B}_i$

At every time step t_i S-ON produces the smallest possible set of candidate sources that always contain the source $v^\star$. More formally, we have the following.

Proposition 5. *For every $1 \le i \le F$ there is no algorithm $\mathcal{A}$ different from S-ON which, given $\mathcal{O}_i$, produces a set of candidate sources $\mathcal{B}_i(\mathcal{A}) \subsetneq \mathcal{B}_i$ and $\mathbf{P}(v^\star \in \mathcal{B}_i(\mathcal{A})) = 1$.*

Proof. By Proposition 4, the set $\mathcal{B}_i$ produced by S-ON is equal to $\mathcal{B}_{t_i}$ for all $1 \le i \le F$. Hence, by (8), for every $v \in \mathcal{B}_i$, $\mathbf{P}(v^\star = v|\mathcal{O}_i) > 0$. If an algorithm $\mathcal{A}$ produces $\mathcal{B}_i(\mathcal{A}) \subsetneq \mathcal{B}_i$ there exists $v \in \mathcal{B}_i \backslash \mathcal{B}_i(\mathcal{A})$ such that $\mathbf{P}(v^\star = v|\mathcal{O}_i) > 0$. Therefore $\mathbf{P}(v^\star \in \mathcal{B}_i(\mathcal{A})) < 1$. □

Like for Algorithm 1, Algorithm 2 runs in time $O(K_s N)$.

The extension of S-ON to $\varepsilon > 0$ follows similarly to Section 3.2 and is presented in Appendix D.

5 D-OFF: Offline Localization with Static and Dynamic Sensors

We now study D-OFF: offline source localization with static and dynamic sensors. After computing the set of candidate sources $\mathcal{B}$ by using the observations gathered by the static sensors, we use dynamic sensors to refine $\mathcal{B}$, i.e., to remove as many nodes as possible from $\mathcal{B}$. Hence our D-OFF algorithm is, in its first part, identical to Algorithm 1 for S-OFF whereas, in its second part, it consists of an iterative refinement of $\mathcal{B}$ based on the observations obtained through the newly-deployed dynamic sensors. If we obtain a candidate sources set $\mathcal{B}$ such that $|\mathcal{B}| = 1$ before deploying the entire budget K_d, we stop deploying dynamic sensors.

Clearly, not all possible dynamic sensors are equally informative about the identity of the source. Our strategy is to iteratively choose where to place the dynamic sensors in order to maximize the progress in the localization of the source, which we refer to as Gain. In Section 5.2 we compare three possible notions of Gain.

In Algorithm 3 we present the pseudo-code for deterministic epidemics. The extension to non-deterministic epidemics follows directly from the results of Section 3.2.

Algorithm 3 D-OFF - deterministic epidemic

Require: $\mathcal{O}$ set of observations, K_d budget for dynamic sensors
  $\mathcal{B} \leftarrow V$
  $\omega_1 \triangleq (u_1, t_1) \leftarrow$ Sample($\mathcal{O}$)
  for $(u, t) \in \mathcal{O} \backslash \omega_1$ **do**
    for $v \in \mathcal{B}$ **do**
      if $d(u, v) - d(u_1, v) \neq t - t_1$ **then**
        remove v from $\mathcal{B}$
  $\mathcal{B}_0 \leftarrow \mathcal{B}$
  $i \leftarrow 0$
  while $|\mathcal{B}_i| > 1$ and $i < K_d$ **do**
    $i \leftarrow i + 1$
    $d_i \leftarrow \text{argmax}_{d\in V\backslash\mathcal{U}} \textsc{Gain}_{\mathcal{U}}(d)$
    $\mathcal{U} \leftarrow \mathcal{U} \cup \{d_i\}$
    $t_{d_i} \leftarrow$ infection time of d_i
    $\mathcal{B}_i \leftarrow \mathcal{B}_{i-1}$
    for $v \in \mathcal{B}_i$ **do**
      if $d(d_i, v) - d(u_1, v) \neq t_{d_i} - t_1$ **then**
        remove v from $\mathcal{B}_i$
  return $\mathcal{B}_i$

Let $\chi_{\textsc{Gain}}$ denote the time required to compute Gain. The running time of Algorithm 3 is $O(N(K_s + \chi_{\textsc{Gain}} K_d))$.

5.1 Correctness

If we could observe the infection time of all nodes in V, we could identify the source trivially by looking at the node with the smallest infection time. We now prove that when the budget for dynamic sensors is unrestricted, Algorithm 3 converges to the set containing only the source $v^\star$ independently of the variance parameter ε. In other words, Algorithm 3 never misses the source.

This result is tight in the budget $K_s + K_d$ of dynamic sensors (for some network topologies the number of sensors needed to localize the source can go up to $N-1$ [9].) Proving tighter results for particular classes of network topologies is an interesting direction for future work.

The correctness of Algorithm 3 does not depend on the definition of Gain: As we will see in Section 7, Gain has an effect on the number of sensors required in Algorithm 3 but not on its correctness.

Theorem 1. *Let $\varepsilon \in [0, 1)$ and X_{uv} be a random variable with support $[(1-\varepsilon)w_{uv}, (1+\varepsilon)w_{uv}]$ for every $uv \in E$. Moreover let the budget for dynamic sensors be unrestricted ($K_s + K_d = N$). Algorithm 3 always returns $\{v^\star\}$.*

Proof. We prove the statement for $\varepsilon > 0$, the proof for $\varepsilon = 0$ can be derived in a similar way. First, note that nodes are removed from the set of candidate sources if and only if they do not satisfy, for some u_1, u_2, the necessary conditions expressed by (5). Hence, due to Proposition 2, the source $v^\star$ is never removed from the set of candidates. Next, we want

to prove that, for every node $w \neq v^\star$, there exist $z, y \in V$ such that, when the infection times of z, y are observed, w is removed from the set of candidate sources. Suppose that $v^\star = v$ and that its infection time t_v is observed. Let $w \neq v$ be another node for which the infection time t_w is also observed. As $v^\star = v$, we have $t_w > t_v$. Note that (5) cannot hold for w with $u_1 = w$ and $u_2 = v$: Indeed, we would have $0 < t_w - t_v \leq (\varepsilon - 1)d(v, w) < 0$, which gives a contradiction. Let $i' \in \mathbb{N}^+$ such that, when $i = i'$ in Algorithm 3, both $v \in \mathcal{U}$ and $w \in \mathcal{U}$. Then, $w \notin \mathcal{B}_{i'}$. □

5.2 Gain Functions

We consider three possible GAIN functions to be used for the selection of the dynamic sensors.

SIZE-GAIN. Perhaps the most natural GAIN function is the one that computes the expected reduction in the number of candidate sources. Let $\mathcal{B}_\mathcal{U}$ denote the set of candidate sources computed based on the information given by the sensors in $\mathcal{U}$ and $\mathcal{B}^c_\mathcal{U} \subseteq \mathcal{B}_\mathcal{U}$ the set of candidate sources after adding $c \in V \backslash \mathcal{U}$ as a dynamic sensor. We define the SIZE-GAIN of choosing c as a dynamic sensor as $g^{\text{SIZE}}_\mathcal{U}(c) \triangleq \mathbf{E}[|\mathcal{B}_\mathcal{U}| - |\mathcal{B}^c_\mathcal{U}|]$. Hence, maximizing $g^{\text{SIZE}}_\mathcal{U}$ is equivalent to minimizing the size of $\mathcal{B}^{(c)}_\mathcal{U}$ and maximizing $g^{\text{SIZE}}_\mathcal{U}$ gives, at any step, a sensor choice that is locally optimal.

For deterministic epidemics, $g^{\text{SIZE}}_\mathcal{U}(c)$ can be easily computed by summing over the set $\mathcal{T}^c_\mathcal{U}$ of the possible infection times for c (see Definition 9 below). For $\varepsilon \in (0, 1)$ we propose an approximation of $g^{\text{SIZE}}_\mathcal{U}(c)$ in Appendix G.

Definition 9 (Possible infection times)**.** *Let $\mathcal{U}$ be a set of sensors, $c \in V \backslash \mathcal{U}$, $\mathcal{O}_\mathcal{U} \triangleq \{(u, t_u), u \in \mathcal{U}\}$ and fix $(u_1, t_1) \in \mathcal{O}_\mathcal{U}$ arbitrarily. Let $\mathcal{B}_\mathcal{U}$ be the set of candidate sources after observing the infection times of the nodes in $\mathcal{U}$, i.e., $\mathcal{B}_\mathcal{U} = \{v \in V : \mathbf{P}(v = v^\star | \mathcal{O}_\mathcal{U}) > 0\}$. Then*

$$\mathcal{T}^c_\mathcal{U} \triangleq \{h \in \mathbb{R} : h = d(v, c) - d(v, u_1) - t_1 \quad \text{for some } v \in \mathcal{B}_\mathcal{U}\} \tag{9}$$

is the set of possible infection times of c.

Note that when $\varepsilon = 0$, the cardinality of $\mathcal{T}^c_\mathcal{U}$ is always finite and equal to the number of equivalence classes in which $\mathcal{U} \cup \{c\}$ partitions $\mathcal{U}$ (see Definition 5). With techniques similar to those of the proof of Proposition 1, it is easy to prove that Definition 9 does not depend on the choice of $(u_1, t_1) \in \mathcal{O}_\mathcal{U}$. The next proposition shows how $g^{\text{SIZE}}_\mathcal{U}$ can be computed in practice.

Proposition 6. *Let $\mathcal{U}$ be a set of sensors, $c \in V \backslash \mathcal{U}$, $\mathcal{O}_\mathcal{U}, \mathcal{B}_\mathcal{U}$ as in Definition 9 and fix $(u_1, t_1) \in \mathcal{O}_\mathcal{U}$ arbitrarily. Call t_c the infection time of c and define*

$$\begin{aligned} b_\mathcal{U}(c, h) &\triangleq \{v \in \mathcal{B}_\mathcal{U} : \mathbf{P}(v = v^\star | t_c = h) > 0\} \\ &= \{v \in \mathcal{B}_\mathcal{U} : h = d(v, c) - d(v, u_1) + t_1\}. \end{aligned}$$

Then,

$$g^{\text{SIZE}}_\mathcal{U}(c) = \sum_{h \in \mathcal{T}_c} \mathbf{P}(v^\star \in b_\mathcal{U}(c, h)) \cdot (|\mathcal{B}_\mathcal{U}| - |b_\mathcal{U}(c, h)|). \tag{10}$$

Proof. Follows from the definition of $g^{\text{SIZE}}_\mathcal{U}$, $\mathcal{T}_c$ and $b_\mathcal{U}(\cdot, \cdot)$. □

DRS-GAIN. The definition of this second GAIN function is inspired by the notion of DRS (see Section 2.5). When epidemics spread deterministically, observing the infection times of a DRS of the candidate sources set $\mathcal{B}$ removes all ambiguities about the source identity. With DRS-GAIN, we iteratively choose the sensor that, added to the current sensor set $\mathcal{U}$, gives the most progress in forming a DRS of $\mathcal{B}$.

Let $c \in V \backslash \mathcal{U}$ and $\mathcal{T}^c_\mathcal{U}$ as in Definition 9. Then, the DRS-GAIN of adding c to $\mathcal{U}$ is

$$g^{\text{DRS}}_\mathcal{U}(c) \triangleq |\mathcal{T}^c_\mathcal{U}|. \tag{11}$$

Since there is no direct extension of $g^{\text{DRS}}_\mathcal{U}$ to the non-deterministic case, we use the above definition of $g^{\text{DRS}}_\mathcal{U}$ independently of the variance parameter ε.

RC-GAIN. RC-GAIN (random candidate GAIN) assigns gain 1 to all candidates sources and gain 0 to all nodes that are not candidate sources, i.e., when the sensor set is $\mathcal{U}$ and $\mathcal{B}_\mathcal{U}$ is the set of candidate sources, for $c \in V \backslash \mathcal{U}$ we set $g^{\text{RC}}_\mathcal{U}(c) = 1$ if $c \in \mathcal{B}_\mathcal{U}$, $g^{\text{RC}}_\mathcal{U}(c) = 0$ otherwise. In other words, we randomly choose the dynamic sensors among the candidate sources. Note that if the infection time of at least one node in $\mathcal{B}_\mathcal{U}$ is already observed, adding a sensor in any other node in $\mathcal{B}_\mathcal{U}$ implies $|\mathcal{B}_{\mathcal{U} \cup \{c\}}| < |\mathcal{B}_\mathcal{U}|$, independently of the variance parameter ε.[3] Hence, this very simple GAIN ensures that the localization of the source makes progress whenever a new dynamic sensor is chosen.

For any of the proposed GAIN functions, the computation time χ_{GAIN} is $O(|\mathcal{B}|) \subseteq O(N)$.

As it is not *a priori* clear which version of GAIN leads to a better performance of Algorithm 3, in Section 7 we experiment with all of them.

6 D-ON: ONLINE LOCALIZATION WITH STATIC AND DYNAMIC SENSORS

We now turn to the online version of D-OFF: D-ON.

As in Section 4, we set the time $\tau^\star$ at which the localization starts to the earliest time at which a static sensor gets infected, i.e., $\tau^\star = \min_{s \in \mathcal{S}} t_s$. Starting from $\tau^\star$, we run online source localization as per S-ON and, in addition, we deploy dynamic sensors one after the other to refine the localization of the source till the budget K_d for dynamic sensors is exhausted. Specifically, a new dynamic sensor is deployed at times $\tau^\star + j\theta$, $j \in \{1, .., K_d\}$, where $\theta \in \mathbb{R}^+$ is fixed. We call θ the *deployment delay*. The choice of θ, which will be discussed in Section 7, requires the evaluation of the trade-off between timely localization and resource-savings: With a large θ we are likely to have less negative observations, hence to reach the localization with few dynamic sensors, but a long time after the beginning of the epidemic. Viceversa, with a small θ, we are likely to reach the localization earlier but by deploying more dynamic sensors.

At time t, the candidate set $\mathcal{B}$, the sensors set $\mathcal{U}$, and the observation set $\mathcal{O} = \mathcal{O}^+ \cup \mathcal{O}^-$ are updated in two cases:

I) if $t = \tau^\star + j\theta$, $j \in \mathbb{N}$, i.e., at time t a new dynamic sensor is added;
II) if $t = t_u > \tau^\star$, i.e., t is the infection time of a static sensor or of a node that was chosen as dynamic sensor before time t but was not infected before time t.

3. This can be proven with an argument analogous to the one used in the proof of Theorem 1: If the infection time of two nodes is observed, only one of the two (the one with smaller infection time) can belong to the set of candidate sources $\mathcal{B}$ (or, for $\varepsilon \in (0, 1)$, to the superset $\widetilde{\mathcal{B}}$).

For both cases, technical details are given in Appendix E. Note that D-ON includes D-OFF as a special case. If we run D-ON starting at $\tau^\star > \max_{v\in V} t_v$, the initial observations set is $\mathcal{O}_1 = \{(s, t_s) : s \in \mathcal{S}\}$. Moreover, throughout the process all observations are positive and we recover D-OFF.

6.1 Correctness

The correctness result of Theorem 1 holds also for D-ON.

Theorem 2. *Let $\varepsilon \in [0, 1)$ and X_{uv} be a random variable with support $[(1-\varepsilon)w_{uv}, (1+\varepsilon)w_{uv}]$ for every $uv \in E$. Moreover let the budget for dynamic sensors be unrestricted ($K_s + K_d = N$). D-ON always returns $\{v^\star\}$.*

Proof. The proof is almost identical to that of Theorem 1, the only necessary change is in the last step. With the notations of the proof of Theorem 1, at the minimum time t such that $(v, t_v), (w, t_w) \in \mathcal{O}_t^+$, it is guaranteed that $w \notin \mathcal{B}$. □

Finally, online localization needs, on average, more resources to reach convergence to the source with respect to offline localization. This is due to the fact that, when running offline localization, every deployed sensor can directly reveal its infection time (i.e., there are no negative observations). As in the choice of the deployment delay θ, in order to choose between D-ON and D-OFF, the trade-off between resource-savings and timeliness must be evaluated.

For the extension of the gain functions of Section 5.2 to the online setting, i.e., to the case in which we can have negative observations, we refer the reader to Appendix E.1.

7 Experiments

7.1 Experimental Setup

Transmission delays. In our experiments, the *transmission delays* are *uniformly distributed* (except in Section 7.8). The uniform distribution is, among the unimodal distributions on a bounded support, the one that maximizes the variance [23], which makes it a very challenging setting for source localization.

Static sensors. We choose the K_s static sensors with one of the following two rules:

- ◇ KDRS (K-nodes approximation of a Double Resolving Set): This rule computes the set of K_s sensors that maximize the number of equivalence classes (see Section 2.5). KDRS was shown to outperform several common heuristics for sensor placement in the case of deterministic or low-variance epidemics [48].
- ◇ KMED (K-Medians): This rule computes the optimal placement of K_s sensors for the closely-related problem of maximizing the detectability of a flow [4]. The KMED placement is the set of K_s nodes $\mathcal{S}$ such that

$$\mathcal{S} = \text{argmin}_{|\mathcal{S}|=K_s} \sum_{v\in V} (\min_{s\in\mathcal{S}} d(v, s)).$$

  Determining the K-Medians of a network is NP-hard [28], hence we approximate KMED with a greedy heuristic. Contrary to KDRS sensors, which are generally placed in the periphery of the network, KMED sensors are more uniformly spread [48].

Spinelli et al. [48] recently showed that the optimal placement of static sensors depends on the variance parameter ε and that, given a value of ε, a suitable sensor placement can be found interpolating between KDRS and KMED. For this reason, we limit ourselves to considering these two alternative static-sensor placements, and we evaluate their respective benefits for the different localization settings.

Default parameters. If it is not differently specified, we set the budget for static sensors to $K/N = 2\%$, the gain function for the choice of the dynamic sensors to SizeGain and the deployment delay to $\theta = 0.5$. The reasons for these choices will be clear in the following discussion.

All results are averaged over at least 100 simulations in which the source is chosen uniformly at random.

For readability, throughout this section the set of candidate sources is always denoted by $\mathcal{B}$ even when $\varepsilon > 0$ and we are actually computing the approximate set $\widetilde{\mathcal{B}}$.

7.2 Network Topologies

We experiment with both synthetic and real-world networks; the network properties and statistics are reported in Table 2.

Synthetic networks. We generated synthetic networks from the following classes: Erdös-Rényi networks (ER) [15], Barabási-Albert networks (BA) [3], random geometric networks on the sphere (RGG) [39], regular trees of degree 3 (RT) and trees with power-law distributed node degree (PLT). For each network class, 10 connected instances of size 250 were generated.

Real-world networks. *Facebook Egonets (FB)*. This dataset is a subset of the Facebook network, consisting of 3732 nodes. It was obtained from the union of 10 Facebook egonet networks [37], after removing the ego nodes[4] and taking the largest connected component.
World Airline Network (WAN). This network is obtained from a publicly available dataset [38] that provides the aircraft type used for every daily connection between over three thousands airports. Using these data we can derive the number of seats available on each route daily. We preprocess the network by removing the connections on which fewer than 20 seats per day are available and by assigning to each connection uv the average between the number of seats available from u to v and from v to u. Also, we iteratively remove leaf nodes (for which we believe connections are not well represented in the dataset), and we obtain a network of 2258 nodes.

Edge weights. All our results are valid for arbitrary edge weights $w_{uv} \in \mathbb{R}^+$. For our experiments we consider integer edge weights for several reasons. When $\varepsilon = 0$, integer weights actually make the problem more challenging because it is more difficult to distinguish among nodes based on their distances to the sensors; when $\varepsilon > 0$ instead, taking integer weights does not affect the difficulty of the problem because graph distances cannot anyway be precisely recovered from the observations. Moreover, in practice, distances are always known up to some degree of precision and, up to

4. The ego nodes were removed in order to ensure that the sampling of contacts across the nodes in the network is uniform.

	ER (p=0.016)	BA (m=2)	RGG (R=0.3)	RT	PLT	FB	U-WAN	WAN
$\|V\|$	250	250	250	250	250	3732	2258	2258
$\|E\|$	511	496	696	249	249	82305	17695	17695
avg degree	4.09	3.96	5.6	1.99	1.99	44.1	15.67	15.67
avg shortest path	4.09	3.47	9.68	7.45	37.8	5.34	6.94	3.56
avg clustering	0.02	0.06	0.56	0	0	0.54	0.65	0.65

TABLE 2: Statistics for the networks used in the experiments.

a multiplication factor, edge-weights can always be assumed to be integer.

All synthetic networks and the FB network are given unit edge weights as there is not a straightforward method for deriving realistic edge weights for these networks. For WAN the definition of the edge weights is inspired by a work by Colizza et al [11]. An edge uv is weighted with an approximation of the expected time between the infection of city u and the arrival of an infected individual at city v (see Appendix H for details). This gives a very skewed weight distribution. Our experiments show that the variability of the edge weights brings an additional challenge to source localization. In order to evaluate the impact of non-uniform weights, we also run our experiments on an unweighted version (U-WAN) of this network (in which all weights are set to 1).

7.3 Choice of the Static Sensors

When we adopt a static approach, the only degree of freedom we have is the choice of the static sensors. It is known that this choice has an impact on the performance of source localization [44], [48], [53] and that the optimal choice depends on the variance parameter [48]. Figure 5(a) compares the performance of KDRS and KMED sensors in terms of the final size of the set of candidate sources $\mathcal{B}$. As in [48], we observe that for moderate variance, KDRS sensors are better than KMED sensors; for larger ε we have the inverse result. In fact, for large ε, the more uniform placement achieved by KMED can better deal with the noisy observations. Instead KDRS enforces the possibility of distinguishing among possible sources in the deterministic case.

We compare KDRS and KMED also for the case where we use dynamic sensors, and we do not restrict the budget for dynamic sensors ($K_d = N - K_s$). As the final set of candidate sources has always cardinality 1, we look instead at the total number of sensors $|\mathcal{U}|$ needed to localize the source. We could think that when we use dynamic sensors and place only a few static sensors ($K_s = 0.02 \cdot N$), the choice of the static sensors has a much smaller effect. Instead, we observe the same result of the static approach. Figures 5(b) and 5(c) show that for both online and offline localization, KDRS emerges as the best choice for static sensors when ε is small but it is outperformed by KMED for larger values of ε.

7.4 Online vs Offline Localization

Static approach. When taking an online approach (S-ON), the computation of $\mathcal{B}$ occurs while the epidemic spreads, hence we can localize the source when many nodes are still not infected. Figure 6(a) (respectively, 6(b)) show the fraction μ of infected nodes when $\mathcal{B}$ contains fewer than 5% of the nodes with KDRS (resp., KMED) sensors. With KDRS sensors, μ is smaller than 60% for all synthetic topologies and smaller than 35% for the real-world networks. As we could expect, with KMED sensors, μ is even smaller (less than 20%), giving an argument for the choice of KMED sensors rather than KDRS sensors also for deterministic epidemics. However, for the tree topologies, the final size of $\mathcal{B}$ can heavily deteriorate when using KMED sensors: For RT, for example, the probability of obtaining $|\mathcal{B}| < 5\% \cdot N$ is lower than 0.05.

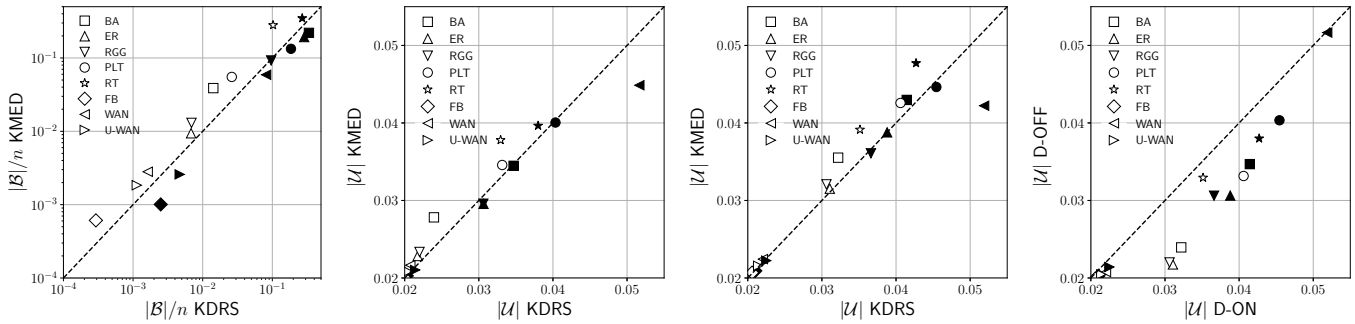
Dynamic approach. Also D-ON dramatically reduces, with respect to D-OFF, the fraction of infected nodes at the time when our algorithm terminates. However, in this case, there is a trade-off between μ and the cost in number of sensors $|\mathcal{U}|$ used for the localization. Figure 5(d) compares the average $|\mathcal{U}|$ for online and offline localization. We see that, independently of the variance parameter and of the network topology, $|\mathcal{U}|$ is smaller for offline localization than for online localization. In fact, in D-OFF, every sensor is already infected by the time it is deployed, hence, when localizing the source offline, we have access to more information.

7.5 Static vs Dynamic

Cost of source localization. As recalled in Section 2, with a static sensor-placement (i.e, $K_d = 0$), the minimum number of sensors required to localize the source when the transmission delays are deterministic is the DMD of the network [9]. Hence, DMD is a natural benchmark for the cost in terms of number of sensors $|\mathcal{U}|$ of our dynamic approach.

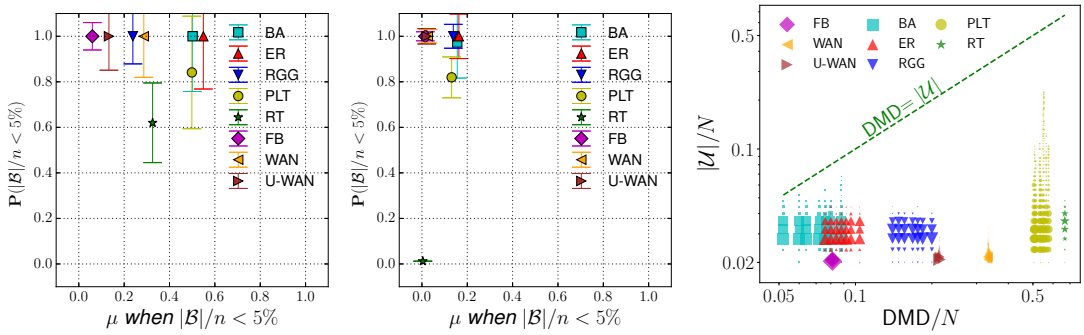
We focus on the deterministic case ($\varepsilon = 0$) with no constraints on the budget for dynamic sensors ($K_d = N - K_s$). We run D-ON and we compare $|\mathcal{U}|/N$ with the (approximate) DMD. The results are depicted in Figure 6(c). For all topologies, $|\mathcal{U}|/N$ is much smaller than DMD/N. The improvement is particularly significant for trees whose DMD is very large (equal to the number of leaves [9]) but where the topology makes it easy for our algorithm to rapidly narrow the search for the source to a small set of candidates. Moreover, we note that $|\mathcal{U}|/N$ is smaller for the real-world topologies than for the synthetic ones and, across all topologies, never exceeds an average of $0.03 \cdot N$, whereas DMD goes up to $0.7 \cdot N$.

Performance with limited budget. We compare the success rate in localizing the source (i.e., the probability $\mathbf{P}(|\mathcal{B}| = 1)$) of a purely-static approach with that of a dynamic approach in both its online and offline variants. We fix a total budget of $K/N = 5\%$ sensors. For the purely-static approach, we have $K_s = K$ and $K_d = 0$; for the dynamic approach we have $K_s = 0.02 \cdot N$ and $K_d = K - K_s = 0.03 \cdot N$. For this experiment, we set $\varepsilon = 0.2$ and we choose KMED



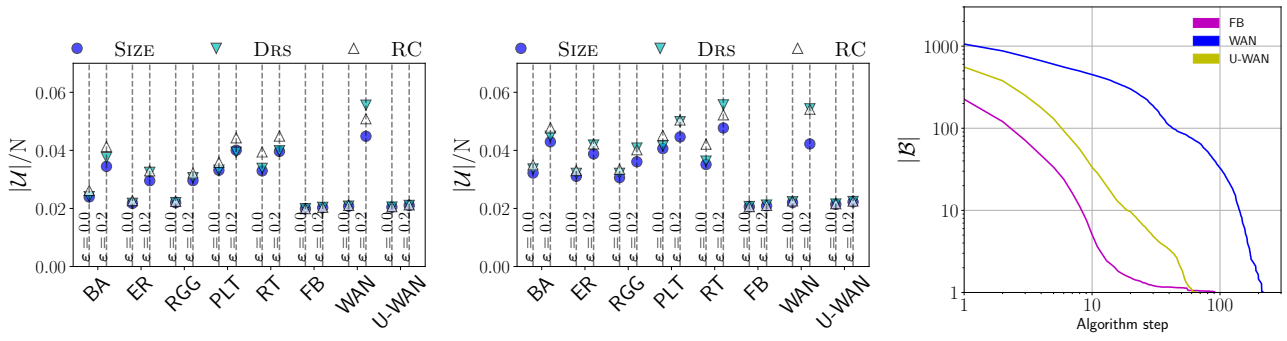
(a) S-OFF & S-ON (b) $Ks + K_d = N$, D-OFF (c) $K_s + K_d = N$, D-ON (d) D-ON vs D-OFF, $\mathcal{S}$=KDRS

Fig. 5: **(a-c)**: Comparison of KDRS and KMED for the choice of the static sensors. **(d)**: Number of sensors needed to localize the source by D-ON and D-OFF. In all plots we use white markers for $\varepsilon = 0$, black for $\varepsilon = 0.3$).



(a) S-ON, $\mathcal{S}$=KDRS (b) S-ON, $\mathcal{S}$=KMED (c) D-ON

Fig. 6: **(a-b)**: S-ON, Fraction μ of infected nodes at the time when $\mathcal{B}$ contains fewer than 5% of the nodes using KDRS sensors (a) and KMED sensors (b). The variance parameter is $\varepsilon = 0$. **(c)**: Fraction of sensors needed by D-ON to localize the source compared with the number needed by an optimal offline placement (DMD). Larger markers represent higher concentrations of data points.



(a) D-OFF (b) D-ON (c) D-ON

Fig. 7: **(a-b)**: Comparison of different GAIN functions. **(c)**: Cardinality of the set $\mathcal{B}$ of candidate sources at successive steps for D-ON (KMED static sensors, $\varepsilon = 0.2$).

static sensors. Figure 9(a) shows that a dynamic approach outperforms a static approach on all topologies. Moreover, in the dynamic case, in line with the results displayed in Figure 5(d), D-OFF achieves a better success rate than D-ON. The only exception is represented by the WAN network: This is probably due to the high variability of the edge weights, which makes it such that sometimes choosing sensors while the epidemic is still local avoids dealing with very noisy transmission times. However, even for this network, D-ON outperforms D-OFF only at an intermediate step in the localization process (i.e., if we stop when $^K/_N = 0.05$): Looking at Figure 5(d), we see that also for WAN, the total number of sensors needed to localize the source is smaller with D-OFF.

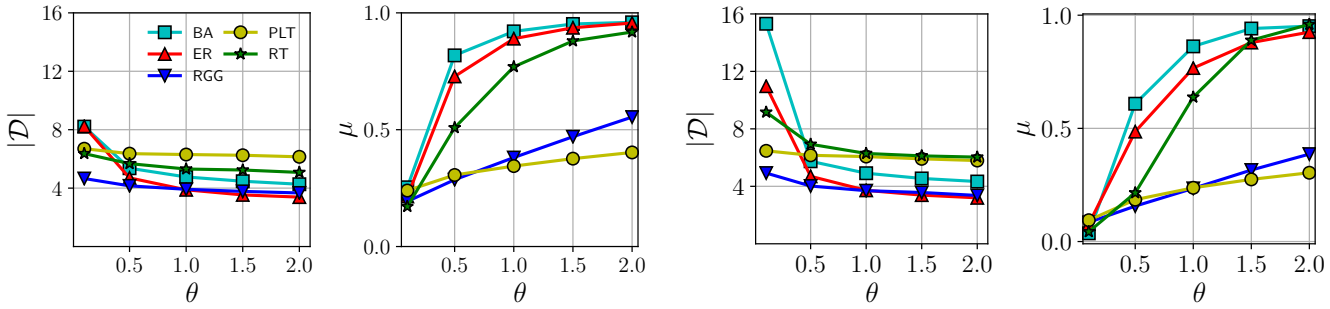
7.6 Source Localization with Dynamic Sensors

Different GAIN functions. We study the effect of GAIN on the performance of our dynamic algorithm. For each variant, i.e., SIZE-GAIN, DRS-GAIN, RC-GAIN, we report the relative cost in terms of number of sensors $^{|\mathcal{U}|}/_N$ when $K_s + K_d = N$. We experiment with both a deterministic setting ($\varepsilon = 0$) with KDRS static sensors and a non-deterministic setting ($\varepsilon = 0.2$) with KMED static sensors (see Section 7.3). The results are depicted in Figures 7(a) and 7(b). We observe that for the real-world networks and $\varepsilon = 0$, all proposed GAIN functions have similar performances. For FB and U-WAN, this is true also when $\varepsilon > 0$. These are the cases in which source localization is achieved with the smallest number of sensors. We conclude that, when source localization is less challenging, GAIN does not have a strong effect. In all other cases, SIZE-GAIN consistently gives the best performance. The improvement, with respect to DRS-GAIN, is most noticeable when $\varepsilon > 0$; indeed, in this setting, and in particular for online localization, DRS-GAIN is outperformed by the simple RC-GAIN. We attribute this to the fact that, when there is high variance in the transmission delays, splitting the candidate sources into subsets of nodes that have different average infection times (see the definition of DRS-GAIN in Eq. (11)), does not guarantee that we are able to distinguish them based on the observed infection times [48]. Instead, as mentioned in Section 5, RC-GAIN enforces a continuous progress in shrinking the set of candidate sources.

Deployment delay. When applying D-ON, an important parameter is the deployment delay θ, i.e., the time between two consecutive placements of a dynamic sensor. On the one hand, the larger θ is, the smaller we expect the cost in terms of number of sensors $|\mathcal{U}|$ to be; on the other hand, the smaller θ is, the less time we expect to need for localizing the source, hence the fewer individuals are infected before we do so. To choose θ, we must also account for the scale of edge weights because, when transmission delays have larger (respectively, smaller) mean, the optimal θ is also likely to be larger (resp., smaller). Here, for simplicity of exposition, we ignore this aspect and experiment only with networks in which all weights are equal to 1. We fix $\varepsilon = 0.2$. We vary θ and look at the number $|\mathcal{D}|$ of dynamic sensors used to localize the source and at the fraction μ of infected individuals at the time of localization. Figures 8(a) and 8(b) display the results for KDRS and KMED sensors, respectively. In both cases, we observe a trade-off between $|\mathcal{D}|$ and μ. When using KDRS sensors, $|\mathcal{D}|$ is smaller, especially for small θ. However, in line with the results of Figure 6(a), KMED sensors guarantee a smaller fraction of infected μ.

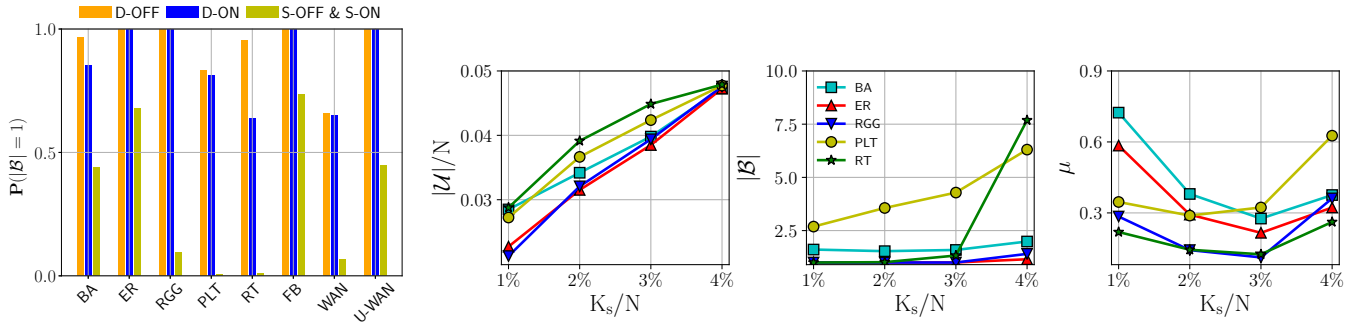
Budget allocation Given a total budget K, how much of it should we allocate for static and how much for dynamic sensors? We choose KMED static sensors, set $\varepsilon = 0.2$ and run D-ON. We take a total budget $^K/_N = 5\%$ and we look at the impact of different values of K_s and $K_d = K - K_s$ on the total number of sensors needed to localize the source (which can be smaller than K if $|\mathcal{B}| = 1$ is reached before K_d is exhaust). We also evaluate the cardinality of the final set $\mathcal{B}$ and the fraction of infected nodes μ at the final stage of the algorithm. The results are displayed in Figure 9(b). We observe that both $|\mathcal{U}|$ and $|\mathcal{B}|$ increase with K_s, which indicates that a small budget for static sensors guarantees both a lower cost in terms of number of sensors $|\mathcal{U}|$ and a higher precision. However μ is minimized when we use around half of our budget for static sensors. In fact, when K_s is small, we need to deploy many dynamic sensors in order to localize the source; instead, when K_s is large and K_d is small, after placing a few dynamic sensors, we have to wait for the static sensors to get infected in order to use this information to refine $\mathcal{B}$.

Size of $|\mathcal{B}|$ at successive iterations. We evaluate the running time of D-ON and D-OFF. The running time of each iteration of our algorithms is linear in $|\mathcal{B}|$. Hence to estimate the running time of source localization we look, for the real-world topologies, at how many iterations are needed to localize the source and at how $|\mathcal{B}|$ decreases along the successive iterations of the algorithm. The running time of D-OFF is smaller than that of D-ON, so we focus on this last case. As we can see in Figure 6(c), the approximate DMD is 303 (around $0.08 \cdot N$) for the FB network, 751 (around $0.3 \cdot N$) for WAN and 484 for U-WAN (around $0.2 \cdot N$). Hence, source localization is more challenging on the WAN network. This is confirmed by the results shown in Figure 7(c). In the FB network, with variance parameter $\varepsilon = 0.2$, the source is localized with in average 15 iterations of the localization algorithm. For the U-WAN network, the average number of iterations needed is larger (around 37). We attribute this effect to the presence of *bottleneck* edges, i.e., edges that appear on many different shortest paths and make it difficult to estimate the source based on its distance to the sensors. This effect becomes even stronger with the weighted version of the WAN network (where the number of iterations needed is in average 98) and it is reflected in the average total number of sensors $\mathcal{U}$ used for localization: $0.021 \cdot N$ for FB, $0.022 \cdot N$ for U-WAN and $0.042 \cdot N$ for WAN (see Figures 7(a) and 7(b)). This result highlights that the high variability among the edge-weights makes source localization substantially more difficult, especially for $\varepsilon > 0$ (see Figures 7(a) and 7(b) for a comparison of the cost in terms of number of sensors $|\mathcal{U}|$ between deterministic and non-deterministic delays). Also the cardinality of $|\mathcal{B}|$ decreases more slowly for the weighted network WAN than for FB and U-WAN. However, for all three topologies, $|\mathcal{B}|$ decreases faster than linearly (note the logarithmic scale in



(a) D-ON, $\mathcal{S}$=KDRS (b) D-ON, $\mathcal{S}$=KMED

Fig. 8: D-ON with varying deployment delay θ: Number of dynamic sensors $|\mathcal{D}|$ needed for source localization and fraction of infected nodes μ at the final stage of the algorithm for varying K_s.



(a) (b)

Fig. 9: **(a)**: Success rate of source localization when $K = K_s + K_d = 0.05 \cdot N$ and $\varepsilon = 0.2$. **(b)**: D-ON with constant total budget $K = K_s + K_d = 0.05 \cdot N$: Number of sensors $|\mathcal{U}|$ used by the algorithm, cardinality of the final set of candidate sources $\mathcal{B}$ and fraction of infected nodes μ at the final stage of the algorithm for varying K_s.

Figure 7(c)), confirming the feasibility of our approach to source localization for a variety of real-world networks.

7.7 Comparison with existing methods

We compare our algorithms for source localization with the two following existing methods:

- ◇ GAU: this method estimates the source through maximum-likelihood assuming Gaussian transmission delays. It was initially proposed by Pinto et al. [40] under the assumption that each sensor reveals its infection time and from which node it received the infection. It was then generalized to the setting where the latter information is not available [48]. In this last work, it was also shown that the method can be applied with different infection-delays distributions (such as truncated-Gaussian and uniform). Here, we consider the improved estimator used in Spinelli et al. [48].
- ◇ EIF: this method estimates the source by computing, for every candidate source v, an infection tree $\mathcal{T}_v$ rooted at v that is compatible with the observed infection times. Every spreading tree $\mathcal{T}_v$ is given a cost that quantifies how much the observed infection times deviate from the expected infection times when the infection spreads along $\mathcal{T}_v$. The estimated source is the root v of the spreading tree $\mathcal{T}_v$ with minimal cost. The method was proposed by Zhu et al. [55] and it is independent from the transmission-delay distribution, only the average delays are needed for the estimation.

We take $K = 5\%N$. For GAU, EIF and S-ON/-OFF, all sensors are static; for D-ON $K_s = 2\%N$ sensors are static and $K_d = K - K_s = 3\%N$ sensors are dynamic. We look at $\mathbf{P}(\mathcal{B} = \{s^*\})$, i.e., the probability that the source is correctly identified without ties, for $\varepsilon = 0$ (Figure 10(a)) and for $\varepsilon = 0.2$ (Figure 10(b)). For $\varepsilon = 0$, the performance of GAU and S-ON/-OFF is identical: They both identify the correct equivalence class of the source (see Definition 5) but cannot distinguish among nodes in the same equivalence class. For tree networks (PLT and RT) the performance of EIF is also identical to that of S-ON/-OFF and GAU: On trees, all the nodes in the equivalence class of the source, and only these nodes, are the roots of the minimal cost spreading trees. Instead, on non-tree networks, the presence of loops makes it more challenging to correctly identify the spreading tree and the performance of EIF is poorer than that of GAU and S-ON/-OFF. For all network topologies, D-ON beats all other methods by a large margin. For $\varepsilon \neq 0$, the performance of S-ON/-OFF is lower than that of GAU. In fact, our methods are designed to have maximum recall ($\mathbf{P}(s^\star \in \mathcal{B}) = 1$) and, for both S-ON/-OFF and D-ON, all nodes that have a positive probability of being the source

are contained in $\mathcal{B}$. When moving from a deterministic to a non-deterministic setting, the various methods suffer a performance drop in different ways: S-ON/-OFF and D-ON have a drop in precision; GAU and EIF have a drop in recall. In a non-deterministic setting, D-ON still detects the source with no ambiguity with a high probability, again strongly outperforming the alternative methods.

7.8 Resistance to unbounded delay distributions

Our theoretical results are derived under the hypothesis of bounded-support for the distribution of the transmission delays (see Section 3-6). However in some applications we would like to work with transmission delays that are not upper bounded by a constant. We test our D-ON method when each transmission delay X_{uv} is drawn from a Gamma distribution $\Gamma(k, 1/k)$ (hence $\mathbb{E}[X_{uv}] = 1$ and $Var(X_{uv}) = 1/k$) which include the case of exponential transmission delays ($k = 1$). In this setting, D-ON is not guaranteed to always detect the source (i.e., Theorem 2 does not hold) and it can happen that $v^\star$ is removed from the set of candidates $\mathcal{B}$.

Let ε_0 be the minimum value such that $\mathbf{P}(X_{uv} \in [w_{uv}(1-\varepsilon_0), w_{uv}(1+\varepsilon_0)]) = 0.75$. In order to account for the variance of X_{uv} but still enforce the removal of nodes from $\mathcal{B}$ we run D-ON with $\varepsilon = \min(\varepsilon_0(k), 0.6)$. Figure 10(c) shows the final size of the set $\mathcal{B}$ while Figure 10(d) depicts the average distance $\bar{d}$ from $s^\star$ to the nodes in $\mathcal{B}$. For moderate variance, $\bar{d}$ is very small for all the topologies considered; for PLT, RGG and RT this distance is very small even for large values of the variance, indicating a good performance of our methods, especially for tree networks, even with unbounded and highly variable delays.

8 Related work

Adaptive resource allocation. Two-stage resource allocation is studied in several contexts, including information diffusion [12], curing policies for epidemics [13], [42] and more general *robust-optimization* problems where, to reach some objective, we allocate a-priori only a part of the resources and we deploy the rest, at a higher cost, when more information is available [24]. Another related line of work in the field of artificial intelligence is that of *active learning*: It studies how to adaptively take a sequence of decisions, based on sparse data, in order to optimize a given objective [21], [45].

We briefly review some important contributions to source localization (see [26] for an in-depth discussion).

Complete observation. The first source-estimator was proposed by Shah and Zaman [46] in 2009. This work, and many others that followed, rely on what is often called a *complete observation* of the epidemic (see Assumption (B.1) in Section 1) [41], [54]. In these models, the source is estimated by maximum likelihood estimation (MLE).

The results of [46] have been extended in many ways, e.g., to the case of multiple sources [10], [25], [35] or to a setting where (B.2) is replaced with an assumption similar to (A.2) [29]. An alternate line of work that also uses Assumption (B.1), allows the observed states to be *noisy*, i.e., potentially inaccurate. For example, a model in which it is not possible to distinguish between susceptible and recovered nodes was studied by Zhu et al. [57].

Partial observation. Follow-up work considers a *partial observation* setting where a randomly-selected fraction of nodes reveal their state [32], [36], [50], [56], [58]. These works do not assume that the infection times are known (see Assumption (A.2)), hence they need a large fraction of the nodes to be sensors (typically more than 30%).

Static sensor placement. Other works address the problem of strategically selecting sensor nodes *a-priori*, i.e., finding a *static* sensor placement. In the deterministic setting (see Assumption (B.5)) some works considered the problem of *minimizing* the budget required for detecting the source. This question is similar to the one we address, except that we allow random transmission delays and, most importantly, we propose an online solution. On trees, under (B.2) and (B.5), the minimization of the number of sensors has been studied [51]. Without (B.2) and (B.4), but with (B.5), approximation algorithms were developed by Chen et al. [9].

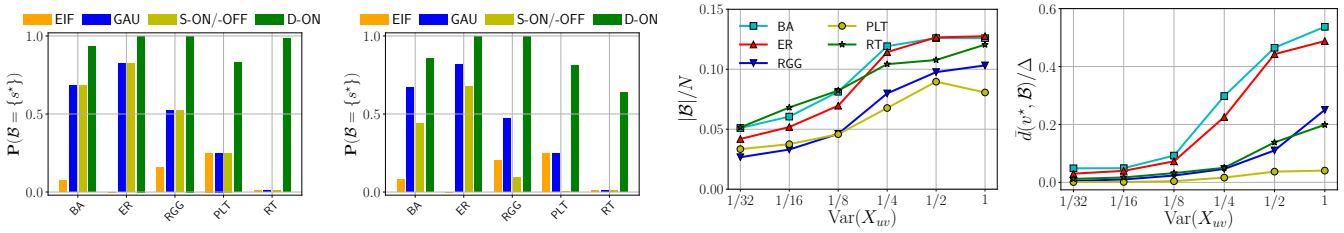
Budgeted sensor placement. In a network of N nodes, the minimal budget required for source-localization can go up to $N-1$, in which case the result of Chen et al. is not practical. Hence, researchers have looked into a *budgeted* version of the problem, i.e., how to place sensors given that only a limited number of them is available. In this direction, "common sense" approaches, e.g., using high-degree vertices, or centrality measures were first evaluated [34], [40]. Later, the budgeted optimization problem was solved on trees [8] (B.4). Without (B.4), a heuristic approach, based on the definition of a double resolving set of a graph (see Section 2), has been shown to outperform all previous heuristics [48]. Due to budget restrictions, none of the works mentioned above can guarantee exact source localization.

Sequential sensor placement. Working under (B.5) and (B.2), Zejnilovic et al. [52], propose an algorithm that sequentially places sensors in order to localize the source *after* the epidemic has spread through the entire network. Adopting very different techniques, we propose a solution that selects the sensors *while* the epidemic evolves, enhancing both cost- and time-efficiency. Moreover, our approach works without (B.5) and (B.2).

Transmission delays. Several models of how the epidemic spreads have been studied [30]. Discrete-time transmission delays were initially very common (see Assumption (B.5)) [2], [36], [41]. Then, to better approximate realistic settings, continuous-time transmission models with varying distributions for the transmission delays have been adopted; e.g., exponential [35], [46], Gaussian [33], [34], [40], [53] or truncated Gaussians [48]. We mainly consider continuous bounded-support distributions that are tractable yet versatile.

9 Future Work

Several research directions can be investigated using the framework and the formalism introduced in this work. First, a natural and realistic extension would attribute a different cost to static and dynamic sensors or would give the sensors a cost that depends on the time at which they are deployed. Second, in order to further decrease the number of sensors needed, we could approximate the set of candidate sources



(a) $\varepsilon = 0$ (b) $\varepsilon = 0.2$ (c) (d)

Fig. 10: **(a-b)**: Comparison of D-ON and S-ON/-OFF with the baseline methods GAU and EIF. **(c-d)**: Performance of D-ON when the transmission delays are Gamma random variables with mean 1 in terms of (a) $\mathbf{P}(\mathcal{B} = \{v^\star\})$ and (b) average distance $\bar{d}(v^\star, \mathcal{B})$ from the nodes in $\mathcal{B}$ to $v^\star$ (rescaled with the diameter Δ). $K_d = 10\%$ and ε is chosen as described in Section 7.8.

$\mathcal{B}$ with a smaller set $\overline{\mathcal{B}} \subseteq \mathcal{B}$ excluding the nodes that have a small probability to be the source. Clearly this operation leads to possible errors (i.e., cases in which the source $v^\star$ is removed from the set of candidates) but, depending on the budget available, a favourable trade-off between cost and precision could be achieved.

A more theoretical and very interesting direction is the investigation of upper bounds for the number of dynamic sensors needed to reach $\mathcal{B} = \{v^\star\}$ for special classes of networks. This would lead to a deeper understanding of the inherent difficulties of source localization.

An interesting and closely related line of work would investigate source localization and sensor placement in adversarial settings, e.g., where the epidemic spread is designed to obfuscate the position of the source [16] or where an adversary knows the position of the static sensors and chooses the source in order to maximize the difficulty of source-localization. We believe that these settings would require different assumptions about the sensors: for example, sensors that can be iteratively moved in the network or that can reveal information about the infection provenance could be considered.

References

[1] M. Al Qathrady, A. Helmy, and K. Almuzaini. Infection tracing in smart hospitals. In *Wireless and Mobile Computing, Networking and Communications (WiMob), 2016 IEEE 12th International Conference on*, pages 1–8. IEEE, 2016.

[2] 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), 2014.

[3] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286, 1999.

[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] D. Bradley and R. Gupta. On the distribution of the sum of n non-identically distributed uniform random variables. *Annals of the Institute of Statistical Mathematics*, 54(3), 2002.

[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] S. Cauchemez and N. Ferguson. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in london. *Journal of the Royal Society Interface*, 5(25), 2008.

[8] L. Celis, F. Pavetić, B. Spinelli, and P. Thiran. Budgeted sensor placement for source localization on trees. In *LAGOS*, 2015.

[9] X. Chen, X. Hu, and C. Wang. Approximability of the minimum weighted doubly resolving set problem. In *COCOON*, 2014.

[10] Z. Chen, K. Zhu, and L. Ying. Detecting multiple information sources in networks under the sir model. *Transactions on Network Science and Engineering*, 2016.

[11] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc. of the National Academy of Sciences of the USA*, 103(7), 2006.

[12] S. Dhamal, K. Prabuchandran, and Y. Narahari. Information diffusion in social networks in two phases. *Transactions on Network Science and Engineering*, 2016.

[13] K. Drakopoulos, A. Ozdaglar, and J. N. Tsitsiklis. An efficient curing policy for epidemics on graphs. *IEEE Transactions on Network Science and Engineering*, 2014.

[14] B. Ehrenberg. How much is your personal data worth? https://www.theguardian.com/news/datablog/2014/apr/22/how-much-is-personal-data-worth, 2014.

[15] P. Erdös and A. Rényi. On random graphs. *I. Publ. Math. Debrecen*, 6, 1959.

[16] G. C. Fanti, P. Kairouz, S. Oh, and P. Viswanath. Spy vs. spy: rumor source obfuscation. In *SIGMETRICS*, 2015.

[17] M. Farajtabar, M. Gomez-Rodriguez, M. Zamani, N. Du, H. Zha, and L. Song. Back to the past: Source identification in diffusion networks from partially observed cascades. In *AISTATS*, 2015.

[18] M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha, and L. Song. Coevolve: A joint point process model for information diffusion and network co-evolution. In *Advances in Neural Information Processing Systems*, 2015.

[19] L. Fu, X. Fu, Z. Xu, Q. Peng, X. Wang, and S. Lu. Determining source-destination connectivity in uncertain networks: Modeling and solutions. *IEEE/ACM Transactions on Networking*, 2017.

[20] L. Fu, X. Wang, and K. P.R. Are we connected? optimal determination of source-destination connectivity in random graphs. *IEEE/ACM Transactions on Networking*, 2016.

[21] D. Golovin and A. Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42, 2011.

[22] M. Gomez-Rodriguez, J. Leskovec, D. Balduzzi, and B. Schölkopf. Uncovering the structure and temporal dynamics of information propagation. *Network Science*, 2(1), 2014.

[23] H. Gray and P. Odell. On least favorable density functions. *SIAM Review*, 9, 1967.

[24] A. Gupta, V. Nagarajan, and R. Ravi. Thresholded covering algorithms for robust and max-min optimization. In *International Colloquium on Automata, Languages, and Programming*, 2010.

[25] F. Ji and W. P. Tay. An algorithmic framework for estimating rumor sources with different start times. *Transactions on Signal Processing*, 2017.

[26] 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.

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

[28] O. Kariv and S. Hakimi. An algorithmic approach to network

location problems. ii: The p-medians. *SIAM journal of Applied Mathematics*, 37, 1979.
[29] A. Kumar, V. Borkar, and N. Karamchandani. Temporally agnostic rumor source detection. *Transactions on Signal and Information Processing over Networks*, 2017.
[30] M. Lelarge. Efficient control of epidemics over random networks. In *SIGMETRICS/Performance*, 2009.
[31] X. Li, Z. D. Deng, L. T. Rauchenstein, and T. J. Carlson. Contributed review: Source-localization algorithms and applications using time of arrival and time difference of arrival measurements. *Review of Scientific Instruments*, 87(4):041502, 2016.
[32] A. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Ph. Review E*, 90(1), 2014.
[33] A. Louni, A. Santhanakrishnan, and K. Subbalakshmi. Identification of source of rumors in social networks with incomplete information. *ASE SocialCom)*, 2015.
[34] 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.
[35] W. Luo and W. Tay. Identifying infection sources in large tree networks. In *IEEE SECON*, 2012.
[36] W. Luo, W. Tay, and M. Leng. How to identify an infection source with limited observations. *IEEE Journal of Sel. Topics in Signal Processing*, 8(4), 2014.
[37] J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In *NIPS*, 2012.
[38] OpenFlights. Route dataset. http://openflights.org/data.html\#route.
[39] M. Penrose. *Random Geometric Graphs*. Oxford Studies in Probability, 2003.
[40] P. Pinto, P. Thiran, and M. Vetterli. Locating the source of diffusion in large-scale networks. *Physical Review Letters*, 109, 2012.
[41] B. Prakash, J. Vreeken, and C. Faloutsos. Spotting culprits in epidemics: How many and which ones? *IEEE ICDM*, 2012.
[42] K. Scaman, A. Kalogeratos, and N. Vayatis. Suppressing epidemics in networks using priority planning. *Transactions on Network Science and Engineering*, 2016.
[43] J. W. Seaman, P. S. Odell, and D. M. Young. Maximum variance unimodal distributions. *Statistics & Probability Letters*, 3(5):255 – 260, 1985.
[44] 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.
[45] B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1), 2012.
[46] D. Shah and T. Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on information theory*, 57, 2011.
[47] Z. Somda, M. Meltzer, H. Perry, N. Messonnier, U. Abdulmumini, G. Mebrahtu, M. Sacko, K. Touré, S. O. Ki, T. Okorosobo, W. Alemu, and I. Sow. Cost analysis of an integrated disease surveillance and response system: case of burkina faso, eritrea, and mali. *Cost Effectiveness and Resource Allocation*, 7(1), 2009.
[48] B. Spinelli, L. Celis, and P. Thiran. Observer placement for source localization: The effect of budgets and transmission variance. In *Allerton Conference*, 2016.
[49] B. Spinelli, L. Celis, and P. Thiran. Back to the source: an online approach to sensor placement and source localization. In *World Wide Web Conference*, 2017.
[50] H. Wang, P. Zhang, L. Chen, H. Liu, and C. Zhang. Online diffusion source detection in social networks. In *IEEE Neural Networks (IJCNN)*, 2015.
[51] S. Zejnilovic, J. Gomes, and B. Sinopoli. Network observability and localization of the source of diffusion based on a subset of vertices. In *Allerton Conf.*, 2013.
[52] S. Zejnilović, J. Gomes, and B. Sinopoli. Sequential observer selection for source localization. In *IEEE GlobalSIP*, pages 1220–1224, 2015.
[53] 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, 2016.
[54] L. Zheng and C. Tan. A probabilistic characterization of the rumor graph boundary in rumor source detection. In *IEEE DSP*, 2015.
[55] K. Zhu, Z. Chen, and L. Ying. Locating the contagion source in networks with partial timestamps. *Data Mining and Knowledge Discovery*, 2015.
[56] K. Zhu, Z. Chen, and L. Ying. Catch'em all: Locating multiple diffusion sources in networks with partial observations. *arXiv preprint arXiv:1611.06963*, 2016.
[57] K. Zhu and L. Ying. Information source detection in the SIR model: A sample path based approach. In *IEEE ITA*, 2013.
[58] K. Zhu and L. Ying. A robust information source estimator with sparse observations. *Computational Social Networks*, 1(1), 2014.

Brunella Spinelli earned a B.Sci. and M.Sci. degrees in Mathematics at the State University of Milan in 2010 and 2012 respectively. She is currently a PhD candidate at the School of Computer and Communication Sciences at EPFL. Her advisors are Dr. L.E. Celis and Prof. P. Thiran. Her research focuses on algorithms for source-localization and epidemic-containment on networks.

L.Elisa Celis earned a B.Sci. degree in Computer Science and Mathematics at Harvey Mudd College in 2006, and a M.Sci. in Mathematics, M.Sci. in Computer Science and Ph.D. in Computer Science at the University of Washington in 2008, 2009 and 2012 respectively. She is currently a senior research scientist at the School of Computer and Communication Sciences at EPFL. Prior to joining EPFL, she worked as a Research Scientist at Xerox Research where she was the worldwide head of the Crowdsourcing and Human Computation thrust. Her research expertise include networks, online learning, algorithms, and game theory. She primarily works on novel optimization problems which have arisen in the highly interconnected modern world, including crowdsourcing, online advertising, ranking and recommendation systems, and their implications to fairness and accountability in algorithms. She is the recipient of a Yahoo! Key Scientific Challenges Award and the China Theory Week Prize.

Patrick Thiran (S'89 - M'96 - SM'12 - F'14) earned his electrical engineering degree from the Université Catholique de Louvain, Louvain-la-Neuve, Belgium, in 1989, and his M.S. degree in electrical engineering from the University of California at Berkeley, USA, in 1990, and his Ph.D. degree from EPFL, in 1996. He is a Full Professor at EPFL. He became an Adjunct Professor in 1998, an Assistant Professor in 2002, an Associate Professor in 2006 and a Full Professor in 2011. From 2000 to 2001, he was with Sprint Advanced Technology Labs, Burlingame, CA. His research interests include networks, performance analysis and stochastic models. He is currently active in the analysis and design of wireless and PLC networks, in network measurements and inference, and in dynamic processes on graphs. Dr. Thiran served as an Associate Editor for the IEEE Transactions on Circuits and Systems in 1997-99, and for the IEEE/ACM Transactions on Networking in 2006-10. He is currently serving on the editorial board of the IEEE Journal on Selected Areas in Communication. He was the recipient of the 1996 EPFL Ph.D. award and of the 2008 Crédit Suisse Teaching Award.

A General Framework for Sensor Placement in Source Localization

B. Spinelli, L.E. Celis, P. Thiran

SUPPLEMENTARY MATERIAL

APPENDIX A
HARDNESS OF KDRS

We approximate the k-DRS set following the approach of Spinelli et al. [48]. The underlying idea to this approach is that any set $W \subseteq V$ partitions V in a set of equivalence classes in the following way: any two nodes $u, v \in V$ are equivalent if for all $w_1, w_2 \in W$, $d(u, w_1) - d(u, w_2) = d(v, w_1) - d(v, w_2)$. Clearly, if W is a DRS, we have n equivalence classes, each consisting of only one node. A k-DRS is a set that maximizes the number of equivalence classes among the sets of cardinality smaller or equal than k. Computing a k-DRS is NP-hard, hence we use a greedy approximation. For every $v \in V$ we initialize $W_v = \{v\}$ and add for $k-1$ times the node that maximize the number of equivalence classes in which V is partitioned. We then choose the set W_v that maximizes the number of equivalence classes as approximation of k-DRS.

The number of equivalence classes can be seen, in a natural way, as a measure of the success of source localization for deterministic epidemics using static sensors. Given a set of static sensors $\mathcal{U}$, we define the average error $\mathcal{E}$

$$\mathcal{E}(\mathcal{U}) = \frac{1}{N}\sum_{v\in V}\mathcal{E}(\mathcal{U}|v = v^\star) = \frac{1}{N}\sum_{v\in V}\frac{|\mathcal{B}_{v=v^\star}(\mathcal{U})| - 1}{|\mathcal{B}_{v=v^\star}(\mathcal{U})|},$$

where $\mathcal{B}_{v=v^\star}(\mathcal{U}) = [v]_{\mathcal{U}}$ is the set of candidate sources when $v = v^\star$.

In this way, $\mathcal{E} = 0$ if and only if $|\mathcal{B}| = 1$. Moreover, we always have $\mathcal{E} < 1$ and $\mathcal{E}$ grows with the number of candidate sources. In fact, $\mathcal{E}$ is equal to the average probability that, choosing a node uniformly at random from $\mathcal{B}$, we do *not* pick $v^\star$.

Hence,

$$\mathcal{E}(\mathcal{U}) = \frac{1}{N}\sum_{v\in V}\frac{|[v]_{\mathcal{U}}| - 1}{|[v]_{\mathcal{U}}|} = 1 - \frac{q}{N}$$

and, maximizing the number of equivalence classes is actually equivalent to minimize $\mathcal{E}$.

APPENDIX B
APPROXIMATION ALGORITHM FOR DMD

The problem of *minimizing* the required number of sensors in order to identify the source in the zero-variance setting has first been studied in relation to the DRS problem by Chen et al. [9]; in fact, a sensor set $\mathcal{U}$ such that the number of equivalence classes is $q = N$ (and hence the source can always be identified) is nothing but a DRS.

Finding a Doubly Resolving Set of minimum size is known to be NP-hard [27]. Chen et al. proposed an approximation algorithm based on a greedy minimization of an *entropy* function [9]. Note that this has no connection to true information-theoretic entropy.

Definition 10. *Let $\mathcal{G}$ a network, $\mathcal{U} \subseteq V$, $|\mathcal{U}| = k$. The entropy of $\mathcal{U}$ is*

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

Note that $H_{\mathcal{U}}$ is minimized if and only if each equivalence class consists of only one node and hence if and only if the average error $\mathcal{E}$ defined in Section A is equal to 0. However, despite the fact that $H_{\mathcal{U}}$ is minimized when $\mathcal{E}$ is minimized and that both are computed based the same set of equivalence classes for a given $\mathcal{U}$, the greedy processes that minimize $H_{\mathcal{U}}$ and $\mathcal{E}$ are not the same. This can be seen by rewriting both objective functions in the following way. Let $[c_1, \dots, c_q]$ be the sequence of equivalence class sizes. Then $H_{\mathcal{U}}$ can be written as

$$H_{\mathcal{U}}([c_1, .., c_q]) = \sum_{i=1}^{l}\sum_{j=2}^{c_i}\log(j) = \sum_{i=2}^{\max c_j}\log(i)\#\{c_j \geq i\}.$$

Analogously we have the following equality for the error $\mathcal{E}([c_1, \dots, c_q])$:

$$n\mathcal{E}([c_1, \dots, c_q])) = n - q = \sum_{i=2}^{\max c_j}\#\{c_j \geq i\}.$$

Hence, though similar in spirit, a greedy minimization of $H_{\mathcal{U}}$ is not related to a greedy optimization of $\mathcal{E}$ and the greedy algorithm of Chen et al. [9] is effectively different from the one of Spinelli et al. [48].

APPENDIX C
S-OFF: TECHNICAL DETAILS

We give the technical details regarding the computation of the set of candidate sources $\mathcal{B}$, starting from the case of deterministic epidemics and later extending our results to the more general case when $\varepsilon > 0$.

We first note that, using Definition 7 we can rewrite $\mathbf{P}(\mathcal{O}|v^\star = v)$ as

$$\mathbf{P}(\mathcal{O}|v^\star = v) = \mathbf{P}\Big(\bigcap_{\omega_i\neq\omega_j\in\mathcal{O}} A_{\omega_i,\omega_j}\Big| v^\star = v\Big). \tag{12}$$

The next lemma formalizes that, when epidemics spread deterministically, the only source of randomness in the epidemic is the value of $v^\star$.

Lemma 2. *Let $\mathcal{O}$ be a set of observations and let $\varepsilon = 0$. Then, for all $v \in V$, $\mathbf{P}(\mathcal{O}|v^\star = v) \in \{0, 1\}$.*

Proof. Let us pick $v \in V$ such that $\mathbf{P}(\mathcal{O}|v^\star = v) > 0$. We want to prove that $\mathbf{P}(\mathcal{O}|v^\star = v) = 1$. We have

$$\begin{aligned}
&\mathbf{P}(\mathcal{O}|v^\star = v) > 0 \\
&\Leftrightarrow \mathbf{P}\Big(\bigcap_{\omega_i \neq \omega_j \in \mathcal{O}} A_{\omega_i,\omega_j}|v^\star = v\Big) > 0 \\
&\Rightarrow \mathbf{P}(A_{\omega_i,\omega_j}|v^\star = v) > 0 \ \ \forall \ \ \omega_i \neq \omega_j \in \mathcal{O} \\
&\Leftrightarrow \mathbf{P}(T(v,u_i) - T(v,u_j) = t_i - t_j) > 0 \ \ \forall \ \ \omega_i \neq \omega_j \in \mathcal{O} \\
&\overset{(a)}{\Leftrightarrow} \mathbf{P}(A_{\omega_i,\omega_j}|v^\star = v) = 1 \ \ \forall \ \ \omega_i \neq \omega_j \in \mathcal{O} \\
&\Leftrightarrow \mathbf{P}(\mathcal{O}|v^\star = v) = 1,
\end{aligned}$$

where (a) holds because $T(v,u_i) - T(v,u_j)$ is deterministic and equal to $d(v,u_i) - d(v,u_j)$. □

A particular case of Lemma 2 is the following.

Lemma 3. *Let $\varepsilon = 0$ and let $\omega_1 \triangleq (u_1, t_1)$ and $\omega_2 \triangleq (u_2, t_2)$ be two observations. Then $\mathbf{P}(A_{\omega_1,\omega_2}|v^\star = v) > 0$ if and only if $\mathbf{P}(A_{\omega_1,\omega_2}|v^\star = v) = 1$. Moreover, $\mathbf{P}(A_{\omega_1,\omega_2}|v^\star = v) > 0$ if and only if $d(v,u_1) - d(v,u_2) = t_1 - t_2$.*

Proof. As in the proof of Lemma 2, $\mathbf{P}(A_{\omega_1,\omega_2}|v^\star = v) > 0 \Leftrightarrow \mathbf{P}(T(v,u_1) - T(v,u_2) = t_1 - t_2) > 0 \Leftrightarrow \mathbf{P}(T(v,u_1) - T(v,u_2) = t_1 - t_2) = 1 \Leftrightarrow d(v,u_1) - d(v,u_2) = t_1 - t_2$. □

We are now ready to prove Proposition 1 which gives a practical way of computing $\mathcal{B}$.

Proposition 1. *Let $\mathcal{O}$ be a set of observations and let $\varepsilon = 0$. Let $\omega_1 \triangleq (u_1, t_{u_1}) \in \mathcal{O}$ be a fixed observation, which we call the reference observation. Then, the set of candidate sources $\mathcal{B}$ is*

$$\mathcal{B} = \bigcap_{\omega \in \mathcal{O}\setminus\{\omega_1\}} \mathcal{B}_{\omega_1,\omega}.$$

Proof. We have

$$\begin{aligned}
v \in \mathcal{B} &\overset{(a)}{\Leftrightarrow} \mathbf{P}(\mathcal{O}|v^\star = v) = 1 \\
&\overset{(b)}{\Leftrightarrow} \mathbf{P}(A_{\omega_i,\omega_j}|v^\star = v) = 1 \ \ \forall \omega_i \neq \omega_j \in \mathcal{O} \\
&\overset{(c)}{\Leftrightarrow} v \in \mathcal{B}_{\omega_i,\omega_j} \ \ \forall \omega_i \neq \omega_j \ \ \in \mathcal{O} \\
&\Leftrightarrow v \in \bigcap_{\omega_i \neq \omega_j \in \mathcal{O}} \mathcal{B}_{\omega_i,\omega_j},
\end{aligned}$$

where (a) holds by Lemma 2, (b) follows from (12) and (c) holds by Lemma 3.

To prove $\mathcal{B} \subseteq \bigcap_{\omega \in \mathcal{O}\setminus\{\omega_1\}} \mathcal{B}_{\omega_1,\omega}$ it is enough to note that $\bigcap_{\omega_i \neq \omega_j \in \mathcal{O}} \mathcal{B}_{\omega_i,\omega_j} \subseteq \bigcap_{\omega \in \mathcal{O}\setminus\{\omega_1\}} \mathcal{B}_{\omega_1,\omega}$. For the reverse inclusion, take $v \in \bigcap_{\omega \in \mathcal{O}\setminus\{\omega_1\}} \mathcal{B}_{\omega_1,\omega}$ and $\omega_i \triangleq (u_i, t_{u_i}), \omega_j \triangleq (u_j, t_{u_j}) \in \mathcal{O}\setminus\{\omega_1\}$, $\omega_i \neq \omega_j$. Since $v \in \mathcal{B}_{\omega_1,\omega_i} \cap \mathcal{B}_{\omega_1,\omega_j}$, by Lemma 3 we have

$$d(v,u_i) - d(v,u_1) = t_i - t_1, \tag{13}$$
$$d(v,u_j) - d(v,u_1) = t_j - t_1. \tag{14}$$

By subtracting (13) from (14), we get $d(v,u_i) - d(v,u_j) = t_i - t_j$, which, again by Lemma 3 implies $v \in \mathcal{B}_{\omega_i,\omega_j}$. Hence we can conclude that $v \in \bigcap_{\omega_i \neq \omega_j \in \mathcal{O}} \mathcal{B}_{\omega_i,\omega_j}$. □

We now turn to non-deterministic epidemics and give a proof of Proposition 2 which is at the basis of the definition of the superset $\widetilde{\mathcal{B}}$ of candidate sources (see (6)).

Proposition 2. *Let $0 < \varepsilon < 1$, let $\omega_1 \triangleq (u_1, t_{u_1})$, $\omega_2 \triangleq (u_2, t_{u_2}) \in \mathcal{O}$, $\omega_1 \neq \omega_2$, and let $v \in \mathcal{B}$. Then*

$$|d(v,u_1) - d(v,u_2) - t_{u_1} + t_{u_2}| \leq \varepsilon(d(v,u_1) + d(v,u_2)). \tag{5}$$

Proof. Since $v \in \mathcal{B}$, $\mathbf{P}(A_{\omega_1,\omega_2}|v^\star = v) > 0$. We prove that, if $v^\star = v$, (5) holds. From this we can conclude that (5) holds for every $v \in \mathcal{B}$, because if there were $v \in \mathcal{B}$ such that (5) does not hold, we would have $\mathbf{P}(A_{\omega_1,\omega_2}|v^\star = v) = 0$, giving a contradiction with $v \in \mathcal{B}$.

Recall that, if $v^\star = v$, the infection time of u is $t_u = t^\star + T(v,u)$. Since the infection delay along edge (x,y) has range $[(1-\varepsilon)w_{xy}, (1+\varepsilon)w_{xy}]$, we have

$$T(v,u) \leq (1+\varepsilon)d(v,u). \tag{15}$$

If $\mathcal{Q}$ is the collection of all paths connecting v and u and, for $p \in \mathcal{Q}$, $d_p(v,u)$ is the weighted length of path p we have

$$T(v,u) \geq (1-\varepsilon)\min_{p \in \mathcal{Q}} d_p(v,u) = (1-\varepsilon)d(v,u). \tag{16}$$

Combining inequalities (15) and (16) we obtain

$$|T(v,u_1) - d(v,u_1)| \leq \varepsilon d(v,u_1), \tag{17}$$
$$|T(v,u_2) - d(v,u_2)| \leq \varepsilon d(v,u_2). \tag{18}$$

From (17) and (18) and using the relation $T(v,u_1) - T(v,u_2) = t_{u_1} - t_{u_2}$ we obtain (5). □

Algoritm 4 gives the pseudo-code for computing of $\widetilde{\mathcal{B}}$.

Algorithm 4 S-OFF - non-deterministic epidemic

Require: $\mathcal{O}$ set of observations
  $\widetilde{\mathcal{B}} \leftarrow V$
  for $(u,t_u), (z,t_z) \in \mathcal{O}$, $u \neq z$ **do**
    for $v \in \widetilde{\mathcal{B}}$ **do**
      $D \leftarrow |d(v,u) - d(v,z) - t_u + t_z|$
      $E \leftarrow \varepsilon(d(v,u) + d(v,z))$
      if $D > E$ **then**
        remove v from $\widetilde{\mathcal{B}}$
  return $\widetilde{\mathcal{B}}$

Finally, we give a proof of Proposition 3.

Proposition 3. *Let $\mathcal{U}$ be the sensor set. Let*

$$\Delta(\mathcal{U}) \triangleq \max_{u \in \mathcal{U}, v \in V} d(v,u)$$

and

$$\delta(\mathcal{U}) \triangleq \min_{[v_1]_\mathcal{U} \neq [v_2]_\mathcal{U}} \max_{u_1,u_2 \in \mathcal{U}} |d(v_1,u_1) - d(v_1,u_2) - d(v_2,u_1) + d(v_2,u_2)|. \tag{7}$$

If $\varepsilon < \varepsilon_0 \triangleq \delta(\mathcal{U})/4\Delta(\mathcal{U})$ and $v^\star = v$, then $\widetilde{\mathcal{B}} \subseteq [v]_\mathcal{U}$.

Proof. Let $v^\star = v$ and $w \notin [v]_\mathcal{U}$. We want to prove that $w \notin \widetilde{\mathcal{B}}$. By hypothesis, there exist $u_1, u_2 \in \mathcal{U}$ such that

$$|d(v,u_1) - d(v,u_2) - d(w,u_1) + d(w,u_2)| \geq \delta(\mathcal{U}). \tag{19}$$

For every $z \in V$, let $\mu_z(u_1,u_2) \triangleq d(z,u_2) - d(z,u_1)$. By Equation (5), the deviation of $t_{u_2} - t_{u_1}$ from $\mu_v(u_1,u_2)$ is upper bounded by

$$|t_{u_2} - t_{u_1} - \mu_v(u_1,u_2)| \leq \varepsilon(d(v,u_2) + d(v,u_1)) \leq 2\varepsilon\Delta(\mathcal{U}).$$

Moreover, by definition of $\widetilde{\mathcal{B}}$, a similar bound holds for every $z \in \widetilde{\mathcal{B}}$:

$$|t_{u_2} - t_{u_1} - \mu_z(u_1, u_2)| \leq \varepsilon(d(z, u_2) + d(z, u_1)) \leq 2\varepsilon\Delta(\mathcal{U}).$$

Assume by contradiction that $w \in \widetilde{\mathcal{B}}$. Then, by applying the triangle inequality and the hypothesis $\varepsilon < {}^{\delta(\mathcal{U})}/_{4\Delta(\mathcal{U})}$ we have

$$\begin{aligned} |\mu_v(u_1, u_2) - \mu_w(u_1, u_2)| &\leq |t_{u_2} - t_{u_1} - \mu_v(u_1, u_2)| \\ &\quad + |t_{u_2} - t_{u_1} - \mu_w(u_1, u_2)| \\ &\leq 4\varepsilon\Delta(\mathcal{U}) < \delta(\mathcal{U}) \end{aligned} \tag{20}$$

which contradicts (19). Hence, $\widetilde{\mathcal{B}} \subseteq [v]_{\mathcal{U}}$. □

APPENDIX D
S-ON: TECHNICAL DETAILS

We show how $\mathcal{B}$ and $\widetilde{\mathcal{B}}$ can be updated when we have negative observations.

As in Appendix C, we start with the case of deterministic epidemics. The next lemma extends Lemma 2 and Lemma 3 to the case in which $\mathcal{O}$ contains negative observations.

Similarly to Section C, using Definitions 7 and 8 we can rewrite $\mathbf{P}(\mathcal{O}_t | v^\star = v)$ as

$$\mathbf{P}\Bigg(\Big(\bigcap_{\omega_i \neq \omega_j \in \mathcal{O}_t^+} A_{\omega_i,\omega_j}\Big) \bigcap \Big(\bigcap_{\substack{\omega_i \in \mathcal{O}_t^+, \\ \omega_j \in \mathcal{O}_t^-}} A^t_{\omega_i,\omega_j}\Big) \Bigg| v^\star = v\Bigg).$$

Lemma 4. *Let $t \in \mathbb{R}$, $\omega_1 \triangleq (u_1, t_{u_1}) \in \mathcal{O}_t^+$ and $\omega_2 \triangleq (u_2, \emptyset) \in \mathcal{O}_t^-$. Then $\mathbf{P}(A^t_{\omega_1,\omega_2} | v^\star = v) > 0$ if and only if $\mathbf{P}(A^t_{\omega_1,\omega_2} | v^\star = v) = 1$ and $d(v, u_1) - d(v, u_2) < t_{u_1} - t$.*

Proof. We have the following sequence of equivalences.

$$\begin{aligned} &\mathbf{P}(A^t_{\omega_1,\omega_2} | v^\star = v) > 0 \\ &\Leftrightarrow \mathbf{P}(T(v^\star, u_1) - T(v^\star, u_2) < t_{u_1} - t) > 0 \\ &\overset{(a)}{\Leftrightarrow} \mathbf{P}(T(v^\star, u_1) - T(v^\star, u_2) < t_{u_1} - t) = 1 \\ &\Leftrightarrow d(v^\star, u_1) - d(v^\star, u_2) < t_{u_1} - t \end{aligned} \tag{21}$$

where (a) holds because, given the value of $v^\star$, $T(v^\star, u_1) - T(v^\star, u_2)$ is deterministic and equal to $d(v^\star, u_1) - d(v^\star, u_2)$. □

Proposition 4. *Let $t \in \mathbb{R}$, $\mathcal{O}_t$ be the set of observations at time t and $\varepsilon = 0$. Let $\omega_1 \triangleq (u_1, t_{u_1}) \in \mathcal{O}^+_{\tau^\star}$ be the first positive observation that we call the reference observation. Then, the set of candidate sources $\mathcal{B}_t$ is*

$$\mathcal{B}_t = \Bigg(\bigcap_{\omega \in \mathcal{O}_t^+ \setminus \{\omega_1\}} \mathcal{B}_{\omega_1,\omega}\Bigg) \bigcap \Bigg(\bigcap_{\omega \in \mathcal{O}_t^-} \mathcal{B}_{\omega_1,\omega,t}\Bigg).$$

Moreover, if $t, t' \in \mathbb{R}$, $t' > t$, $\mathcal{B}_{t'} \subseteq \mathcal{B}_t$.

Proof. The fact that $\mathcal{B}_t$ contains all and only the nodes that have a positive probability to be the source given the information available at time t is a direct consequence of the definition of $\mathcal{B}_t$ and $\mathcal{O}_t$.

Similarly to the proof of Proposition 1 we have

$$\begin{aligned} &v \in \mathcal{B}_t \\ &\Leftrightarrow v \in \Bigg(\bigcap_{\omega_i \neq \omega_j \in \mathcal{O}_t^+} \mathcal{B}_{\omega_i,\omega_j}\Bigg) \bigcap \Bigg(\bigcap_{\omega_i \in \mathcal{O}_t^+, \omega_j \in \mathcal{O}_t^-} \mathcal{B}_{\omega_i,\omega_j,t}\Bigg) \end{aligned}$$

and hence

$$\mathcal{B}_t \subseteq \Bigg(\bigcap_{\omega \in \mathcal{O}_t^+ \setminus \{\omega_1\}} \mathcal{B}_{\omega_1,\omega}\Bigg) \bigcap \Bigg(\bigcap_{\omega \in \mathcal{O}_t^-} \mathcal{B}_{\omega_1,\omega,t}\Bigg).$$

If $\omega_i, \omega_j \in \mathcal{O}_t^+ \setminus \{\omega_1\}$, $\omega_i \neq \omega_j$, as in the proof of Proposition 1 we have that $v \in \mathcal{B}_{\omega_1,\omega_i} \cap \in \mathcal{B}_{\omega_1,\omega_j}$ implies $v \in \mathcal{B}_{\omega_i,\omega_j}$. Let now $\omega_i \triangleq (u_i, t_{u_i}) \in \mathcal{O}_t^+ \setminus \{\omega_1\}$ and $\omega_j \triangleq (u_j, \emptyset) \in \mathcal{O}_t^-$ and take $v \in \mathcal{B}_{\omega_1,\omega_i} \cap \mathcal{B}_{\omega_j,\omega_1,t}$. By Lemma 3 and Lemma 4 we have

$$d(u_i, v) - d(u_1, v) = t_{u_i} - t_{u_1} \tag{22}$$
$$d(u_1, v) - d(u_j, v) < t_{u_1} - t. \tag{23}$$

Combining (22) and (23), we have $d(u_i, v) - d(u_j, v) < t_{u_i} - t$ and, by Lemma 4, $v \in \mathcal{B}_{\omega_i,\omega_j,t}$. Hence we proved

$$\mathcal{B}_t \supseteq \Big(\bigcap_{\omega \in \mathcal{O}_t^+ \setminus \{\omega_1\}} \mathcal{B}_{\omega_1,\omega}\Big) \bigcap \Big(\bigcap_{\omega \in \mathcal{O}_t^-} \mathcal{B}_{\omega_1,\omega,t}\Big).$$

Let now $t, t' \in \mathbb{R}$ and $t' > t$ and take $v \in \mathcal{B}_{t'}$. As $\mathcal{O}^+_{t'} \supseteq \mathcal{O}^+_t$,

$$v \in \bigcap_{\omega_i \in \mathcal{O}_t^+ \setminus \{\omega_1\}} \mathcal{B}_{\omega_1,\omega}.$$

If $\omega_j \triangleq (u_j, \emptyset) \in \mathcal{O}^-_{t'}$, by Lemma 4, $d(u_1, v) - d(u_j, v) < t_{u_1} - t'$ and, since $t' > t$, $d(u_1, v) - d(u_j, v) < t_{u_1} - t$. Hence, again by Lemma 4, $v \in \mathcal{B}_{\omega_1,\omega_j,t}$ and $\mathcal{B}_{t'} \subseteq \mathcal{B}_t$. □

We now turn to non-deterministic epidemics. As in S-OFF, when $0 < \varepsilon < 1$, we compute a superset of the candidate set $\widetilde{\mathcal{B}}_t \supseteq \mathcal{B}_t$. To do this, we extend Proposition 2 to account for negative observations.

Proposition 7. *Let $t \geq \tau^\star$, $\omega_1 \triangleq (u_1, t_{u_1}) \in \mathcal{O}_t^+$, $\omega_2 \triangleq (u_2, \emptyset) \in \mathcal{O}_t^-$ and let $v \in \mathcal{B}_t$. Then,*

$$d(u_1, v) - d(u_2, v) - t_{u_1} + t < \varepsilon(d(u_1, v) + d(u_2, v)). \tag{24}$$

Proof. The proof of the result follows closely that of Proposition 2. We limit ourselves to highlighting the differences. If $v^\star = v$, we have

$$T(v, u_1) \geq d(v, u_1) - \varepsilon d(v, u_1), \tag{25}$$
$$T(v, u_2) \leq d(v, u_2) + \varepsilon d(v, u_2). \tag{26}$$

Combining (25) and (26) and using the relation $T(v, u_1) - T(v, u_2) < t_{u_1} - t$, we obtain (24). □

In view of Proposition 7 and Remark 1, we can compute and update $\widetilde{\mathcal{B}}$ with Algorithm 5. Like for Algorithm 4, the running time of Algorithm 5 is $O(K_s^2 N)$.

Algorithm 5 S-ON - non-deterministic epidemic

Require: Observation sets $\{\mathcal{O}_i^+\}_{i=1}^F$, $\{\mathcal{O}_i^-\}_{i=1}^F$
  $\widetilde{\mathcal{B}}_0 \leftarrow V$
  $i \leftarrow 1$
  while $i \leq F$ and $|\widetilde{\mathcal{B}}_{i-1}| > 1$ **do**
    $i \leftarrow i+1$
    $\widetilde{\mathcal{B}}_i \leftarrow \widetilde{\mathcal{B}}_{i-1}$
    for $(u,t_u) \in \mathcal{O}_i^+ \backslash \mathcal{O}_{i-1}^+, (z,t_z) \in \mathcal{O}_i^+$ **do**
      for $v \in \widetilde{\mathcal{B}}_i$ **do**
        $D \leftarrow |d(u,v) - d(z,v) - t_u + t_z|$
        $E \leftarrow \varepsilon(d(u,v) + d(u,v))$
        if $D > E$ **then**
          remove v from $\widetilde{\mathcal{B}}_i$
    for $(u,\emptyset) \in \mathcal{O}_i^-, (z,t_z) \in \mathcal{O}_i^+$ **do**
      for $v \in \widetilde{\mathcal{B}}_i$ **do**
        $D \leftarrow d(z,v) - d(u,v) - t_z + t_i$
        $E \leftarrow \varepsilon(d(u,v) + d(z,v))$
        if $D \geq E$ **then**
          remove v from $\widetilde{\mathcal{B}}_i$
  return $\mathcal{B}_i$

APPENDIX E
D-ON: TECHNICAL DETAILS

We give here some details concerning our D-ON algorithm.

As a pseudo-code for the complete algorithm would be quite involved (hence not very helpful for the reader), we limit ourselves to giving the pseudo-code for the subroutines with which, at a time t, we update the candidate set $\mathcal{B}$, the sensors set $\mathcal{U}$, and the observation set $\mathcal{O} = \mathcal{O}^+ \cup \mathcal{O}^-$. The initialization, i.e., the first computation of $\mathcal{B}$ at time $\tau^\star$ is done in D-ON as for S-ON (i.e., as in the first iteration of the **while** loop of Algorithm 2).

At time t, the candidate set $\mathcal{B}$, the sensors set $\mathcal{U}$, and the observation set $\mathcal{O} = \mathcal{O}^+ \cup \mathcal{O}^-$ are updated in two cases:

I) if $t = \tau^\star + \theta j$, $j \in \mathbb{N}$, i.e., at time t a new dynamic sensor is added. In this case $\mathcal{B}, \mathcal{U}$ and $\mathcal{O} = \mathcal{O}^+ \cup \mathcal{O}^-$ are updated with the subroutine presented in Algorithm 6.

II) if $t = t_u > \tau^\star$, i.e., t is the infection time of a static sensor or of a node that was chosen as dynamic sensor before time t but was not yet infected at time t. In this case $\mathcal{B}$, $\mathcal{U}$ and $\mathcal{O} = \mathcal{O}^+ \cup \mathcal{O}^-$ are updated with the subroutine presented in Algorithm 7.

In Algorithm 6 and 7, $\bar{t}$ denotes the time at which $\mathcal{B}, \mathcal{U}$ and $\mathcal{O}$ where last updated before time t. If t is the time of the first update, $\bar{t} = \tau^\star$. To simplify the notation, the time index for the set $\mathcal{B}$ is omitted.

The extensions of Algorithm 6 and 7 to non-deterministic epidemics follow from Proposition 7 and Algorithm 5.

E.1 Extending the Gain Functions to Negative Observations

In online source localization, dynamic sensors can yield negative observations. For this reason, the computation of $g_{\mathcal{U}}^{\mathrm{SIZE}}$ and $g_{\mathcal{U}}^{\mathrm{DRS}}$ given in Section 5.2 should slightly change to account for the case in which a dynamic sensor is not infected by the time at which it is deployed.

Algorithm 6 D-ON - Update I - deterministic epidemic

Require: $\mathcal{B}, \mathcal{U}, \mathcal{O}_{\bar{t}}, \omega_1 \triangleq (u_1,t_1) \in \mathcal{O}_{\tau^\star}^+$
  $d' \leftarrow \mathrm{argmax}_{d \in V\backslash\mathcal{U}} \textsc{Gain}_{\mathcal{U}}(d)$
  $\mathcal{U} \leftarrow \mathcal{U} \cup \{d'\}$
  if d' is infected **then**
    $t_{d'} \leftarrow$ infection time of d'
    $\mathcal{O}_t^+ \leftarrow \mathcal{O}_{\bar{t}}^+ \cup (d', t_{d'})$
    $\mathcal{O}_t^- \leftarrow \mathcal{O}_{\bar{t}}^-$
    for $v \in \mathcal{B}$ **do**
      if $d(d',v) - d(u_1,v) \neq t_{d'} - t_1$ **then**
        remove v from $\mathcal{B}$
    for $\omega \triangleq (u,\emptyset) \in \mathcal{O}_t^-$ **do**
      for $v \in \mathcal{B}$ **do**
        if $d(u,v) - d(u_1,v) < t - t_1$ **then**
          remove v from $\mathcal{B}$
  else
    $\mathcal{O}_t^+ \leftarrow \mathcal{O}_{\bar{t}}^+$
    $\mathcal{O}_t^- \leftarrow \mathcal{O}_{\bar{t}}^- \cup (d',\emptyset)$
    for $\omega \triangleq (u,\emptyset) \in \mathcal{O}_t^-$ **do**
      for $v \in \mathcal{B}$ **do**
        if $d(u,v) - d(u_1,v) < t - t_1$ **then**
          remove v from $\mathcal{B}$

Algorithm 7 D-ON - Update II - deterministic epidemic

Require: $\mathcal{B}, \mathcal{O}_{\bar{t}}, \omega_1 \triangleq (u_1,t_1) \in \mathcal{O}_{\tau^\star}^+$,
    (u,t_u) new positive observation
  $\mathcal{O}_t^+ \leftarrow \mathcal{O}_{\bar{t}}^+ \cup (u,t_u)$
  $\mathcal{O}_t^- \leftarrow \mathcal{O}_{\bar{t}}^-$
  for $v \in \mathcal{B}$ **do**
    if $d(u,v) - d(u_1,v) \neq t_u - t_1$ **then**
      remove v from $\mathcal{B}$
  for $\omega \triangleq (w,\emptyset) \in \mathcal{O}_t^-$ **do**
    for $v \in \mathcal{B}$ **do**
      if $d(w,v) - d(u_1,v) < t - t_1$ **then**
        remove v from $\mathcal{B}$

Definition 11 (Possible infection times)**.** *Let $\mathcal{U}$ be a set of sensors, $c \in V\backslash\mathcal{U}$, $\mathcal{O}_{\mathcal{U}} \triangleq \{(u,t_u), u \in \mathcal{U}\}$ and fix $(u_1,t_1) \in \mathcal{O}_{\mathcal{U}}$ arbitrarily. Let $\mathcal{B}_{\mathcal{U}}$ be the set of candidate sources after observing the infection times of the nodes in $\mathcal{U}$, i.e., $\mathcal{B}_{\mathcal{U}} = \{v \in V : \mathbf{P}(v = v^\star | \mathcal{O}_{\mathcal{U}}) > 0\}$. Then*

$$\mathcal{T}_{\mathcal{U},t}^c \triangleq \{h \in (-\infty, t] : h = d(v,c) - d(v,u_1) - t_1 \text{ for some } v \in \mathcal{B}_{\mathcal{U}}\} \quad (27)$$

is the set of possible infection times of c that are smaller than t.

Again, Definition 11 does not depend on the choice of $(u_1,t_1) \in \mathcal{O}_{\mathcal{U}}$. The next proposition extends Proposition 6 to online localization.

Proposition 8. *Let $\mathcal{U}$ be a set of sensors, $c \in V\backslash\mathcal{U}$, $\mathcal{O}_{\mathcal{U}}, \mathcal{B}_{\mathcal{U}}$ as in and Definition 9 and fix $(u_1,t_1) \in \mathcal{O}_{\mathcal{U}}$ arbitrarily. Call t_c the infection time of c and define*

$$\begin{aligned} b_{\mathcal{U}}(c,h) &\triangleq \{v \in \mathcal{B}_{i-1} : \mathbf{P}(v = v^\star | t_c = h) > 0\} \\ &= \{v \in \mathcal{B}_{\mathcal{U}} : h = d(v,c) - d(v,u_1) + t_1\}, \end{aligned}$$

$$\begin{aligned} \tilde{b}_{\mathcal{U}}(c) &\triangleq \{v \in B_{i-1} : \mathbf{P}(v = v^\star | t_c > t) > 0\} \\ &= \{v \in B_{i-1} : t < d(v,c) - d(v,u_1) + t_1\}. \end{aligned}$$

Then at time t, g^{SIZE} can be computed as,

$$g^{\mathrm{SIZE}}_{\mathcal{U}}(c) = \sum_{h\in\mathcal{T}^c_{\mathcal{U},t}} \mathbf{P}(v^\star \in b_{\mathcal{U}}(c,h)) \cdot (|\mathcal{B}_{\mathcal{U}}| - |b_{\mathcal{U}}(c,h)|) + \mathbf{P}(v^\star \in \tilde{b}_{\mathcal{U}}(c)) \cdot (|B_{\mathcal{U}}| - |\tilde{b}_{\mathcal{U}}(c)|). \quad (28)$$

Proof. Follows from the definition of $g^{\mathrm{SIZE}}_{\mathcal{U}}$, $\mathcal{T}_c$ and $b_{\mathcal{U}}(\cdot,\cdot)$. □

For g^{DRS}, let $X_c = 1$ if there exists $v \in \mathcal{B}_{\mathcal{U}}$ such that the infection time t_c of c is larger than t (i.e., such that $d(v,c) - d(v,u_1) - t_1 > t$), $X_c = 0$ otherwise. Then, the value of DRS-GAIN at time t is defined as

$$g^{\mathrm{DRS}}_{\mathcal{U}}(c) \triangleq |\mathcal{T}^c_{\mathcal{U},t}| + X_c. \quad (29)$$

As in Section 5.2, we use the same definition of g^{DRS} for both deterministic and non-deterministic epidemics. Instead, an approximation of g^{SIZE} is given in Appendix G.

APPENDIX F
EXAMPLE TO COMPLEMENT REMARK 1

Figure 11 complements Remark 1 showing why, when $\varepsilon > 0$, in order to obtain a smaller set of candidate sources, we do not use a single sensor as reference point. In fact, if we take $\varepsilon = 0.25$ and the infection times as in Figure 11, for node $v \neq$ we have

$$|d(v,u) - d(v,z) - t_u + t_z| \leq \varepsilon(d(v,u) + d(v,z))$$

and

$$|d(v,u) - d(v,w) - t_u + t_w| \leq \varepsilon(d(v,u) + d(v,w))$$

but

$$|d(v,w) - d(v,z) - t_w + t_z| > \varepsilon(d(v,w) + d(v,z)).$$

Hence, taking u as reference sensor and comparing t_u with t_z and t_w we would not remove v from the set of candidate sources, which instead we do if we further compare t_w and t_z.

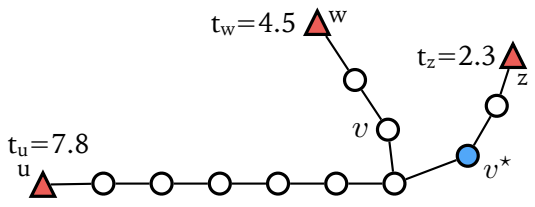


Fig. 11: When $\varepsilon > 0$, taking a single reference sensor, we get a larger set of candidate sources. In this example, taking u as reference sensor and comparing t_u with t_z and t_w we would not remove v from the set of candidate sources, which instead we do if we further compare t_w and t_z.

APPENDIX G
APPROXIMATE SIZE-GAIN FOR THE NON-DETERMINISTIC CASE

When epidemics spread deterministically, Proposition 6 and 8 show that, for any candidate sensor c, the probability of it being infected at time h, can be computed summing over the possible the nodes $v = v^\star$ such that c is infected at time h. We limit ourselves to give a generalization of Proposition 8 to non-deterministic epidemic for the online localization setting For offline localization, Proposition 6 can be generalised to non-deterministic epidemic analogously. We adopt the notations of Section 6.

Proposition 9. *Let t_c be the infection time of $c \in V\backslash\mathcal{U}$ and t'_c, t''_c the minimum and maximum values for t_c given $\mathcal{O}_t$, then*

$$t'_c \geq \min_{v\in\mathcal{B}} \Big(\max_{(u,t_u)\in\mathcal{O}_t, t_u\neq\emptyset} \Big\{ d(c,v) - d(u,v) + t_u - \varepsilon(d(c,v) + d(u,v)) \Big\} \Big),$$

$$t''_c \leq \max_{v\in\mathcal{B}} \Big(\min_{(u,t_u)\in\mathcal{O}_t, t_u\neq\emptyset} \Big\{ d(c,v) - d(u,v) + t_\omega + \varepsilon(d(c,v) + d(u,v)) \Big\} \Big)$$

Proof. We prove the bound for t'_c, the one for t''_c is analogous. Take $v \in \mathcal{B}$. If $v = v^\star$, then for every $(u,t_u) \in \mathcal{O}_t$

$$t'_c \geq d(c,v) - d(u,v) + t_u - \varepsilon(d(c,v) + d(u,v)),$$

hence

$$t'_c \geq \max_{(u,t_u)\in\mathcal{O}_t} \Big\{ d(c,v) - d(u,v) + t_u - \varepsilon(d(c,v) + d(u,v)) \Big\}.$$

The bound follows then from the fact that $v^\star$ can be any node in $\mathcal{B}$. □

For $h \in [t'_c, t''_c]$, let $a(c,h)$ be the set of nodes v that satisfy (5) and (24) with $v = v^\star$ for all observations in $\mathcal{O}_t \cup \{(c,h)\}$, and let $\tilde{a}(c)$ be the set of nodes v that satisfy (24) at time t for all observations in $\mathcal{O}_t \cup \{(c,\emptyset)\}$. Then we define

$$g^{\mathrm{SIZE}}_{\mathcal{U}}(c) = \int_{\min(t'_c,t)}^{\min(t''_c,t)} (|\mathcal{B}| - |a(c,h)|) f_{t_c}(h)dh + (|\mathcal{B}| - |\tilde{a}(c)|)(1 - F_{t_c}(t)), \quad (30)$$

where $f_{t_c}(\cdot)$ denotes the density of the infection time t_c of c conditioned on $\mathcal{O}_t$ and F_{t_c} is its cumulative function.

Let $(u_0, t_{u_0}) \in \mathcal{O}_{\tau^\star}$ and, for $h \in \mathbb{R}$, let us denote by J_h the interval $[h - \frac{1}{2}, h + \frac{1}{2}]$, by J'_h the interval $[h - \frac{1}{2} - t_{u_0}, h + \frac{1}{2} - t_{u_0}]$. In order to compute (30), we make the following approximations:

1) we approximate the integrand with a stepwise constant function with steps of unity length centered around the integer values in $[t'_c, t''_c]$, i.e.

$$\mathrm{E}[g^{\mathrm{SIZE}}_{\mathcal{U}}(c)] \approx \sum_{h\in\mathbb{Z}, h\in[t'_c,t''_c], h\leq t} (|\mathcal{B}| - |a(c,h)|)\mathbf{P}(t_c \in J_h|\mathcal{O}_t) + (|\mathcal{B}| - |\tilde{a}(c)|)\mathbf{P}(t_c > t|\mathcal{O}_t);$$

2) we compute $\mathbf{P}(t_c \in J_h|\mathcal{O}_t)$ by summing over $\mathcal{B}$:

$$\mathbf{P}(t_c \in J_h|\mathcal{O}_{i-1}) = \sum_{v\in\mathcal{B}} \mathbf{P}(t_c \in J_h|v = v^\star, \mathcal{O}_t)\mathbf{P}(v = v^\star|\mathcal{O}_t).$$

In order to further limit the computational costs, if $\mathbf{P}(v = v^\star|\mathcal{O}_{i-1}) > 0$, we approximate

$$\mathbf{P}(v = v^\star|\mathcal{O}_t) \approx \frac{\mathbf{P}(v = v^\star)}{\mathbf{P}(v^\star \in \mathcal{B})},$$

i.e., we ignore the fact that, conditioned on the observations in $\mathcal{O}_t$ the probability of a node being the source can differ from the (rescaled) prior. Moreover, we approximate $\mathbf{P}(t_c \in J_h|\mathcal{O}_t)$ as follows. We take (u_0, t_{u_0}) as reference observation[5] and we approximate $\mathbf{P}(t_c \in J_h|\mathcal{O}_t) \approx \mathbf{P}(t_c - t_{u_0} \in J'_h)$.[6]

An important side-effect of the approximation of $\mathbf{P}(t_c \in J_h)$ is that the event $g^{\text{SIZE}}_{\mathcal{U}}(c) = |\mathcal{B}|$, i.e., no node is a valid candidate source after adding c, might have a positive *weight* in the computation of $\mathrm{E}[g^{\text{SIZE}}_{\mathcal{U}}]$. Specifically, there might be a value of h such that $\mathbf{P}(t_c - t_{u_0} \in J'_h) > 0$ but $|a_{c,h}| = 0$. This can lead our algorithm to slow down by choosing sensors that do not reduce the number of candidate sources. We address this problem applying the following heuristic: Whenever the number of candidate sources does not decrease in two consecutive steps we restrict the choice of the new sensor to the set of candidate sources $\mathcal{B}$. In fact, if the infection time of at least one node in $\mathcal{B}$ is already observed, adding a sensor in any other node in $\mathcal{B}$ implies that the cardinality of $\mathcal{B}$ decreases at the next step.

APPENDIX H
WEIGHTS FOR THE WAN NETWORK

Our definition of the edge weights for the WAN network is inspired by the work of Colizza et al. [11].

Let s_{ij} be the number of seats available on a flight from airport i to airport j. The number of seats can be inferred by the aircraft with which the flight is operated [38]. Moreover, let $\alpha = 0.7$ denote the average occupancy rate on a flight [11] and N_i denote the population of city i. We approximate the probability that an individual flies from i to j as $\alpha s_{ij}/N_i$.

Let θ be the probability that an individual is infected when the infection reached the city where he leaves. Then the probability that a sick individual travels from i to j is $1-(1-\alpha s_{ij}/N_i)^{\theta N_i}$. Hence the average delay for the infection to spread from city i to city j can be estimated to be

$$w_{ij} = [1 - (1 - \alpha s_{ij}/N_i)^{\theta N_i}]^{-1} \approx [1 - \exp^{-\alpha s_{ij}\theta}]^{-1}.$$

We assume $\theta = 0.05$ and we round all weights w_{ij} to the closest integer. Figure 12 shows the resulting weight distribution (note the log-scale of the y-axis, hence the skewness of the distribution).

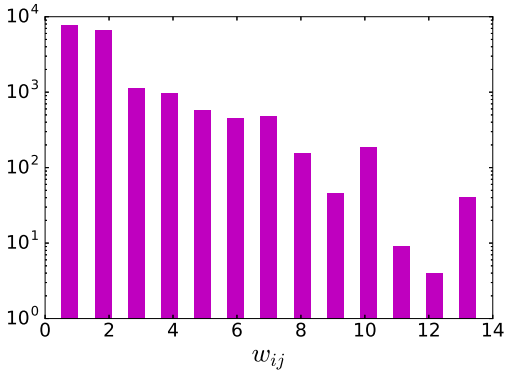


Fig. 12: Histogram of edge weights for the WAN network.

5. In case of a large diameter network, this choice could be optimized taking as reference the sensor u (static or dynamic) which is closer to the candidate source v; for a small-diameter network this would not yield a substantial improvement.

6. If the time delays are all uniformly distributed with equal expected values, we can normalize $t_c - t_{u_0}$ to obtain a sum of uniform $U([0, 1])$ variables, i.e., an Irwin-Hall random variable and the latter probability can be computed exactly. If time delays are uniformly distributed but with different expected values, the probability $\mathbf{P}(t_c - t_{u_0} \in J'_h)$ is not easily computable [5], hence we approximate the distribution of $t_c - t_{u_0}$ with a Gaussian distribution with mean and variance equal to the mean and variance of $t_c - t_{u_0}$. The latter Gaussian approximation can be used for generally distributed transmission delays.