Locating the Source of Diffusion in Large-Scale Networks

Supplemental Material

Pedro C. Pinto, Patrick Thiran, Martin Vetterli

Contents

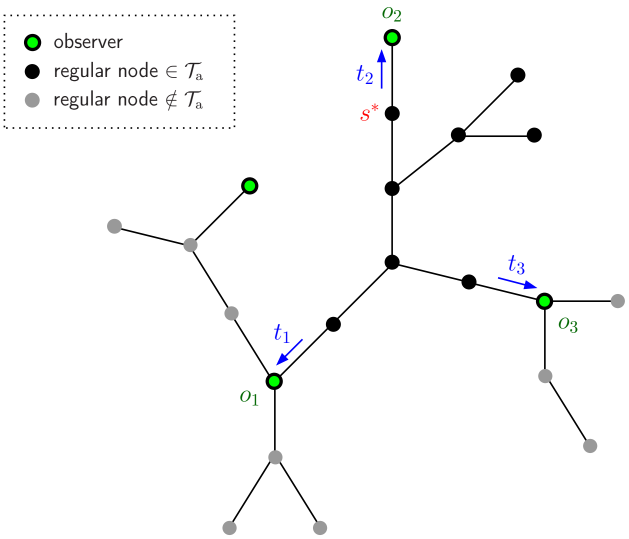

observer
regular node $\in \mathcal{T}_a$
regular node $\notin \mathcal{T}_a$
o_2
t_2
s^*
t_3
o_3
t_1
o_1

Figure S1: Source estimation on an arbitrary tree $\mathcal{T}$.

S1. Detailed of Proof of Proposition 1

Let $\mathcal{T}$ denote the underlying tree on which information is diffused (Fig. S1). Because a tree does not contain cycles, in general only a subset $O_a \subseteq O$ of the observers will receive information emitted by the unknown source. We call $O_a = \{o_k\}_{k=1}^{K_a}$ the set of K_a *active observers*. The observations made by the nodes in O_a provide two types of information:

1. *Direction:* The direction in which information arrives to the active observers uniquely determines a subset $\mathcal{T}_a \subseteq \mathcal{T}$ of regular nodes (called *active subtree*), which is guaranteed to contain the unknown source s^*.

2. *Timing:* The times at which the information arrives to the active observers, denoted by $\{t_k\}_{k=1}^{K_a}$, are used to localize the source within the set $\mathcal{T}_a$.

The arrival times $\{t_k\}$ can be written as a function of the unknown source location $s^* \in \mathcal{T}_a$. It is first convenient to label the edges of $\mathcal{T}_a$ as $E(\mathcal{T}_a) = \{1, 2, \ldots, E_a\}$, so that the propagation delay associated with edge $i \in E$ is denoted by the RV θ_i (Fig. 2a in main paper). Let $\mathcal{P}(u, v) \subseteq E(\mathcal{T}_a)$ denote the set of edges (*path*) connecting vertices u and v in the underlying graph $\mathcal{G}$. Then,

$$t_k = t^* + \sum_{i \in \mathcal{P}(s^*, o_k)} \theta_i, \tag{1}$$

for $k = 1, \ldots, K_a$. Since all the arrival times $\{t_k\}$ are shifted by the unknown parameter t^*, an equivalent sufficient statistic for estimating s^* is the set $\{d_k\}$ of time difference of arrivals (TDOA) where

$$d_k \triangleq t_{k+1} - t_1 = \sum_{i \in \mathcal{P}(s^*, o_{k+1})} \theta_i - \sum_{i \in \mathcal{P}(s^*, o_1)} \theta_i, \tag{2}$$

for $k = 1, \ldots, K_a - 1$. Collecting the $\{d_k\}$ measurements into the *observed delay vector* $\mathbf{d} \triangleq [d_1, \ldots, d_{K_a-1}]^T$, and the $\{\theta_i\}$ into the *propagation delay vector* $\boldsymbol{\theta} \triangleq [\theta_1, \ldots, \theta_{E_a}]^T$, we rewrite (2) in the matrix form as

$$\mathbf{d} = \mathbf{C}_s \boldsymbol{\theta}, \tag{3}$$

where $\mathbf{C}_s$ is a $(K_\mathrm{a}-1)\times E_\mathrm{a}$ matrix whose (k,i)-th element is given by

$$[\mathbf{C}_s]_{k,i} = \begin{cases} 1, & \text{if edge } i \in \mathcal{P}(s^*, o_{k+1}), \\ -1, & \text{if edge } i \in \mathcal{P}(s^*, o_1), \\ 0, & \text{otherwise.} \end{cases}$$

If the observers are sparse and the RVs $\{\theta_i\}$ have finite moments, then the observer delay vector $\mathbf{d}$ can be closely approximated by a Gaussian random vector, due to the central limit theorem (see Section S3 for details on the validity of this approximation). Therefore, we consider that the propagation delays associated with the edges of $\mathcal{T}$ are *jointly Gaussian* with known means and covariances, i.e., $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Lambda}_{\boldsymbol{\theta}})$.[1] This implies that the observed delay $\mathbf{d}$ is also Gaussian, resulting in the following Gaussian source estimator.

Source estimator for general trees (jointly-Gaussian diffusion)

$$\hat{s} = \underset{s\in\mathcal{T}_\mathrm{a}}{\operatorname{argmax}} \frac{\exp\left(-\frac{1}{2}(\mathbf{d}-\boldsymbol{\mu}_s)^\mathrm{T}\boldsymbol{\Lambda}_s^{-1}(\mathbf{d}-\boldsymbol{\mu}_s)\right)}{|\boldsymbol{\Lambda}_s|^{1/2}}, \tag{4}$$

where the *deterministic delay* $\boldsymbol{\mu}_s$ and the *delay covariance* $\boldsymbol{\Lambda}_s$ follow directly from (3) as

$$\boldsymbol{\mu}_s = \mathbf{C}_s\boldsymbol{\mu}_{\boldsymbol{\theta}}, \tag{5}$$

$$\boldsymbol{\Lambda}_s = \mathbf{C}_s\boldsymbol{\Lambda}_{\boldsymbol{\theta}}\mathbf{C}_s^\mathrm{T}. \tag{6}$$

A scenario of special interest is when all the propagation delays are independent identically distributed (IID) Gaussian $\mathcal{N}(\mu, \sigma^2)$, with a common mean μ and variance σ^2. In this case, $\boldsymbol{\Lambda}_{\boldsymbol{\theta}} = \sigma^2\mathbf{I}$, and $\boldsymbol{\Lambda}_s = \boldsymbol{\Lambda}$ (independent of s), so the optimal estimator in (4) can be further simplified as follows, leading to **Proposition 1** in the main paper.

Source estimator for general trees (Gaussian-IID diffusion)

$$\hat{s} = \underset{s\in\mathcal{T}_\mathrm{a}}{\operatorname{argmax}}\ \boldsymbol{\mu}_s^\mathrm{T}\boldsymbol{\Lambda}^{-1}\left(\mathbf{d}-\frac{1}{2}\boldsymbol{\mu}_s\right) \tag{7}$$

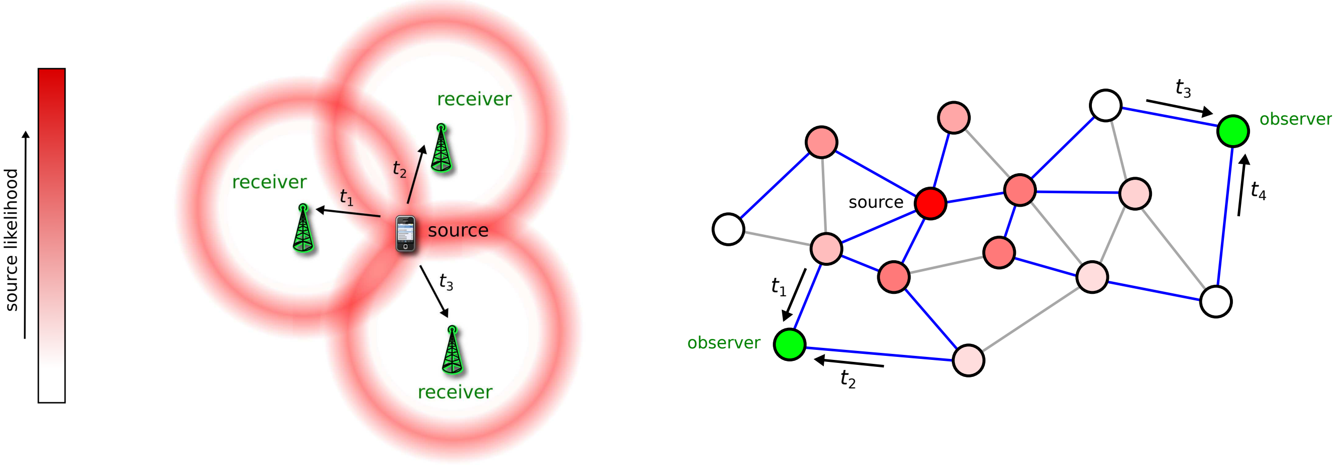
where (5)-(6) simplify to

$$[\boldsymbol{\mu}_s]_k = \mu\cdot\left(|\mathcal{P}(s,o_{k+1})| - |\mathcal{P}(s,o_1)|\right), \tag{8}$$

$$[\boldsymbol{\Lambda}]_{k,i} = \sigma^2\cdot\begin{cases} |\mathcal{P}(o_1,o_{k+1})|, & k=i, \\ |\mathcal{P}(o_1,o_{k+1})\cap\mathcal{P}(o_1,o_{i+1})|, & k\neq i, \end{cases} \tag{9}$$

for $k,i = 1,\ldots,K_\mathrm{a}-1$, and $|\mathcal{P}(u,v)|$ denoting the number of edges (*length*) of the path connecting vertices u and v.

[1] We use $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ to denote that a jointly Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Lambda}$. Since delays are nonnegative, the mean of each delay must be much larger than its standard deviation, so that the model has practical significance.



(a) Locating a cellphone on a wireless network. The cellphone transmits a signal in all directions of space. Each receiver measures the travel time of such a signal. The measurements are then combined to generate a *likelihood* for each point in space, which measures how strongly we believe that it coincides with the source location.

(b) Locating the source of information on a graphical network. A *source* initiates the diffusion of *information* over the network (e.g., a rumor, a computer virus, or a biological virus). The two observers measure the arrival time of such information. The measurements are then combined to generate a *likelihood* for each candidate source in the graph, which measures how strongly we believe that it coincides with the source location.

Figure S2: Conceptual comparison between source localization on a wireless network and on a graph network.

Since $\boldsymbol{\mu}_s$ is proportional to μ and $\boldsymbol{\Lambda}$ is proportional to σ^2, it is useful to define their normalized versions, $\tilde{\boldsymbol{\mu}}_s \triangleq \frac{\boldsymbol{\mu}_s}{\mu}$ and $\tilde{\boldsymbol{\Lambda}} \triangleq \frac{\boldsymbol{\Lambda}}{\sigma^2}$. Figure S2 provides a comparison between source localization on a wireless network and on a graph network.

S2. Algorithms and Complexity

The procedure for source localization is summarized in Algorithms 1 and 2. If $N \triangleq |\mathcal{T}|$ denotes the total number of vertices of the tree $\mathcal{T}$, then the overall complexity of Algorithm 1 is simply proportional to the number of nodes, or $O(N)$. On the other hand, If $N \triangleq |\mathcal{G}|$ denotes the total number of vertices of the graph $\mathcal{G}$, then the worst-case complexity of Algorithm 2 is $O(N^3)$, which is smaller than exponential, as desired.[2]

S3. Accuracy of Gaussian Approximation

We illustrate the accuracy of the Gaussian approximation using the line network in Fig. S3(a). The propagation delays $\{\theta_i\}$ are IID exponential RVs with mean 1, rather than Gaussian RVs. According to (2), the observed delay $d \triangleq t_2 - t_1$ is a linear combination of six RVs, $\{\theta_i\}_{i=1}^{6}$, with weights ± 1 that depend on the source location. Fig. S3(c) shows that the PDF of d indeed resembles that of a Gaussian RV, when conditioned on $s^* = s$. Table S3(b) further illustrates the validity of the Gaussian approximation in terms

[2] The BFS tree $\mathcal{T}_{\text{bfs},s}$ can be computed with worst-case complexity $O(N^2)$ per node $s \in \mathcal{G}_{\text{a}}$, resulting in an overall worst-case complexity of $O(N^3)$.

Algorithm 1 Proposed algorithm for source localization in general trees.

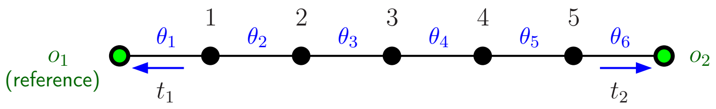
```
select one observer from $O_\mathrm{a}$ as reference, and label it $o_1$
compute the delay vector $\mathbf{d}$ relative to $o_1$
compute the matrix $\mathbf{\Lambda}$ for subtree $\mathcal{T}_\mathrm{a}$
for $k = 1$ to $K_\mathrm{a} - 1$ do
    for every $s \in \mathcal{P}(o_1, o_{k+1})$ do
        if $s = o_1$ then
            send message $[\tilde{\mathbf{\Lambda}}]_{k,k}$ to next neighbour of $s$ in $\mathcal{P}(o_1, o_{k+1})$, in the direction of $o_{k+1}$
        else
            set $[\tilde{\boldsymbol{\mu}}_s]_k$ = received message $- 2$
            send message $[\tilde{\boldsymbol{\mu}}_s]_k$ to next neighbour of $s$ in $\mathcal{P}(o_1, o_{k+1})$, in the direction of $o_{k+1}$
            broadcast $[\tilde{\boldsymbol{\mu}}_s]_k$ on all subtrees rooted at $s$, disjoint with $\mathcal{P}(o_1, o_{k+1})$
        end if
    end for
end for
pick $\hat{s}$ according to the maximization in (7)
```

Algorithm 2 Proposed algorithm for source localization in general graphs.

```
select one arrival time as reference, and label it $t_1$
compute the delay vector $\mathbf{d}$ relative to $t_1$
for every $s \in \mathcal{G}_\mathrm{a}$ do
    compute the spanning tree $\mathcal{T}_{\mathrm{bfs},s}$ rooted at $s$
    compute $\boldsymbol{\mu}_s$ and $\mathbf{\Lambda}_s$ with respect to tree $\mathcal{T}_{\mathrm{bfs},s}$
    compute the source likelihood in (7) for node $s$
end for
pick $\hat{s}$ according to the maximization in (7)
```

of the resulting probability of localization. In particular, if we use an estimator that perfectly matches the exponential statistics of propagation, we achieve an optimal performance of $P_\mathrm{loc} = 0.6208$. In general, this approach presents several drawbacks: i) the estimator must be developed in a case-by-base basis (using eq. 1 in the main paper) to optimally match the propagation statistics, which must be exactly known; and ii) the resulting estimator can exhibit large complexity, possibly exponential in the size of the network. On the other hand, if we use an estimator that is optimal for Gaussian statistics, although not perfectly matched to the underlying exponential statistics, it achieves a slightly worse performance of $P_\mathrm{loc} = 0.6196$. The advantage is that the complexity of the Gaussian estimator scales linearly with the size $N \triangleq |\mathcal{T}|$ of the network. In addition, we do not need to know the full distribution of the propagation delays, but only its second order moments, to which the Gaussian estimator must be matched to.

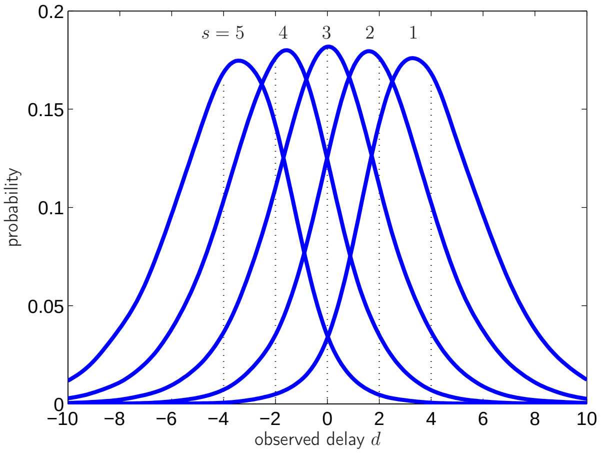
However, there are some special cases where the proposed Gaussian estimator is not applicable. If the delays $\{\theta_i\}$ have a "heavy-tail" distribution with infinite mean or variance (e.g., alpha-stable and Cauchy distributions), then the regular central limit theorem does not apply. For example, if $\{\theta_i\}$ are IID alpha-stable RVs, then their sum tends to another stable distribution, rather than a Gaussian. In such cases, the moments of the underlying distribution cannot be matched, so the exact estimator must be used, despite the previously mentioned drawbacks.



(a) A line graph with two observers at the extremities, which measure the delay $d \triangleq t_2 - t_1$. The propagation delays $\{\theta_i\}$ are independent identically distributed (IID) exponential RVs with mean 1.

Estimator	$P_{\mathrm{loc}} = \mathbf{P}(\hat{s} = s^*)$
Exact (exponential)	0.6208
Approximate (Gaussian)	0.6196

(b) Comparison between the exact and approximated estimators in terms of the resulting localization probability P_{loc}.



(c) Probability density function $\mathbf{P}(d|s^* = s)$ of the observed delay d, for all $s \in \{1, 2, 3, 4, 5\}$. Even though the individual propagation delays are exponentially distributed, the PDF of the overall delay d can be closely approximated by a Gaussian distribution.

Figure S3: Accuracy of the Gaussian estimator in a scenario with exponential propagation delays.

S4. Detailed Proof of Proposition 2

This section presents a detailed proof of Proposition 2, first by deriving the localization probability P_{loc} for a single cascade, and then generalizing to an arbitrary number of cascades.

Performance of the Optimal Estimator with Single Cascade

We start by partitioning the underlying tree $\mathcal{T}$ in a way that will facilitate the determination of P_{loc}. Recall Fig. S1, where the set O of observers is scattered across $\mathcal{T}$. These observers break the tree $\mathcal{T}$ into smaller subtrees, denoted by $\{\mathcal{T}_i\}$, so that the vertices of $\mathcal{T}$ can be partitioned as

$$\mathcal{T} = \bigcup_i \mathcal{T}_i \cup O. \tag{10}$$

We further classify each subtree $\mathcal{T}_i$ according to different cases.

1. *The subtree $\mathcal{T}_i$ is adjacent to one observer only.* In this case, we denote the subtree by $\mathcal{L}_i$.

2. *The subtree $\mathcal{T}_i$ is adjacent to at least two observers.* In this case, we partition the subtree as $\mathcal{T}_i = \mathcal{R}_i \cup \bigcup_j \mathcal{U}_j$, where $\mathcal{R}_i$ is the union of paths between every pair of observers adjacent to $\mathcal{T}_i$, and $\{\mathcal{U}_j\}$ are the subtrees that are created by removal of $\mathcal{R}_i$ from $\mathcal{T}_i$.

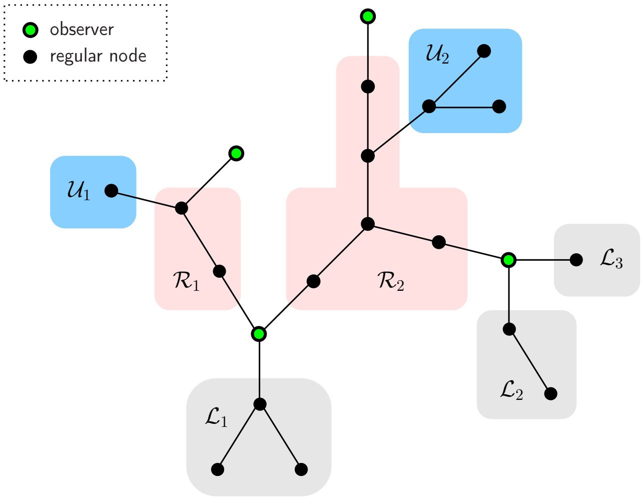


Figure S4: Partition of an arbitrary tree into observers O, resolvable sets $\{\mathcal{R}_i\}$, and unresolvable sets $\{\mathcal{L}_i\}$ and $\{\mathcal{U}_i\}$.

With this notation, the partition in (10) can be further expanded as

$$\mathcal{T} = \bigcup_i \mathcal{R}_i \cup \bigcup_i \mathcal{U}_i \cup \bigcup_i \mathcal{L}_i \cup O. \tag{11}$$

Figure S4 shows an example of such partition.

We call the sets $\{\mathcal{L}_i\}$ and $\{\mathcal{U}_i\}$ *unresolvable*, since it is not possible to distinguish between different source locations inside these sets. Specifically, if $s^* \in \mathcal{L}_i$, there is only one active observer, and only one arrival time. Since in (1) the start time of the diffusion, t^*, is an unknown parameter, the single arrival time does not contain any information about the location of the source in $\mathcal{L}_i$, so all nodes in $\mathcal{L}_i$ are equally likely candidates for the source location. On the other hand, if $s^* \in \mathcal{U}_i$, then there are at least two observers, and TDOAs can be computed. However, the deterministic delay $\boldsymbol{\mu}_s$ in (8) is the same for all $s \in \mathcal{U}_i$, since the propagation delays occurred in the edges of $\mathcal{U}_i$ cancel out in the calculation of the TDOAs. As a result, all nodes in $\mathcal{U}_i$ have the same source similarity, and are indistinguishable in terms of source estimation.

Based on these considerations, and assuming IID propagation delays with distribution $\mathcal{N}(\mu, \sigma^2)$, we can use Boole's inequality to write the following lower-bound

$$P_{\text{loc}} \geq \frac{1}{|\mathcal{T}|} \left(\#\text{sets } \mathcal{L} + K + \sum_i |\mathcal{R}_i| \cdot \left(1 - (|\mathcal{R}_i| - 1) \cdot Q\left(\frac{\mu}{2\sigma} d_{\min,i}\right)\right)\right), \tag{12}$$

where

$$d_{\min,i} \triangleq \min_{m \neq n} \left\| \tilde{\boldsymbol{\mu}}_m^{(i)} - \tilde{\boldsymbol{\mu}}_n^{(i)} \right\|_{\tilde{\boldsymbol{\Lambda}}^{(i)}}; \tag{13}$$

$\tilde{\boldsymbol{\mu}}_m^{(i)}$ and $\tilde{\boldsymbol{\Lambda}}^{(i)}$ are, respectively, the (normalized) deterministic delay and delay covariance associated with subtree $\mathcal{T}_i$; $|\mathcal{T}|$ is the number of vertices of the underlying tree $\mathcal{T}$; #sets $\mathcal{L}$ is the number of non-empty sets of type $\mathcal{L}$; K is the total number of observers; and $Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du$ is the Gaussian Q-function. A fact that will later be useful is that this bound holds with equality in the limit of deterministic propagation

$(\mu/\sigma \to \infty)$, i.e.,

$$P_{\text{loc}} = \frac{1}{|\mathcal{T}|}\left(\#\text{sets } \mathcal{L} + K + \sum_i |\mathcal{R}_i|\right) \tag{14}$$

$$= 1 - \frac{\sum_i |\mathcal{U}_i|}{|\mathcal{T}|} - \frac{\sum_i |\mathcal{L}_i|}{|\mathcal{T}|} + \frac{\#\text{sets } \mathcal{L}}{|\mathcal{T}|}, \tag{15}$$

where we used the fact that $|\mathcal{T}| = \sum_i |\mathcal{R}_i| + \sum_i |\mathcal{U}_i| + \sum_i |\mathcal{L}_i| + K$.

Performance of the Optimal Estimator with Multiple Cascades

We now generalize (12) to account for the scenario where the information source s^* transmits C different cascades of information. We assume that the random propagation delay $\theta_{uv}^{(c)}$ associated with cascade c and edge uv is independent for different cascades c. As a result, the observed delays vectors $\{\mathbf{d}^{(c)}\}$ are also independent for different cascades c, conditional on the source location s^*. By jointly considering the observations $\{\mathbf{d}^{(c)}\}$ for all cascades, the optimal estimator in (7) can be written more generally as

$$\hat{s} = \underset{s\in\mathcal{T}_{\text{a}}}{\text{argmax}}\; \boldsymbol{\mu}_s^{\text{T}} \boldsymbol{\Lambda}^{-1}\left(\sum_{c=1}^{C} \mathbf{d}^{(c)} - \frac{C}{2}\boldsymbol{\mu}_s\right). \tag{16}$$

In this case, the lower-bound in (12) generalizes to

$$P_{\text{loc}} \geq \frac{1}{|\mathcal{T}|}\left(\#\text{sets } \mathcal{L} + K + \sum_i |\mathcal{R}_i| \cdot \left(1 - (|\mathcal{R}_i| - 1)\cdot Q\left(\frac{\mu\sqrt{C}}{2\sigma} d_{\min,i}\right)\right)\right). \tag{17}$$

Since $Q(x) \leq \frac{1}{2}e^{-x^2/2}$ for $x > 0$, we can write $P_{\text{loc}} = P_{\max} - O\left(e^{-aC}\right)$ as $C \to \infty$, where $P_{\max}$ is the probability of localization under deterministic propagation given in (14), and a is a constant. This is the result of **Proposition 2** in the main paper.

S5. Details of Case Study: *Cholera Outbreak*

The dataset was compiled by the KwaZulu-Natal Health Department, and kindly provided to us by the authors of ref. [10]. It consists of a record of each single cholera case since August 2000, specified by the date and health subdistrict where it occurred. These reports were mapped onto a graphical model of the Thukela river basin—a tree $\mathcal{T}$ composed of $N = 287$ nodes. All the channels of perennial rivers are considered edges, and all the endpoints of these channels are considered as nodes. The cholera bacteria is diffused across this network by a multitude of mechanisms, including downstream hydrological transport, and mobility of infected individuals. The forward analysis in ref. [10] shows that the overall drift of the bacteria is only 8% downstream. Therefore, we ignore this bias and consider the tree $\mathcal{T}$ representing the basin to be *undirected.*

The system parameters are summarized in Table 1. The spatial drift v of cholera was estimated at approximately 3 km/day [17]. We consider that a given population is *infected* whenever the cumulative number of cholera cases since August 2000 exceeds the threshold Θ of 50 cases. The propagation delay θ_{uv} between communities u and v is modeled according to a Gaussian RV $\mathcal{N}(\mu_{uv}, \sigma_{uv}^2)$. The mean μ_{uv} is approximated by $r_{uv}\Theta/v$, where r_{uv} is the physical distance between communities u and v. The standard deviation σ_{uv} is considered proportional to the mean μ_{uv}, with a fixed propagation ratio $\beta \triangleq \sigma_{uv}/\mu_{uv} = 0.5$.

Parameter	Value	Unit	Description
b	≈ 0	–	Transport bias
v	3.0	km/day	Spatial drift of *Vibrio cholerae*
Θ	50	cases	Infection threshold at each node
β	0.5	–	Ratio between standard deviation and mean of propagation delay
σ_m	1.0	days	Standard deviation of the measurement delay
K_max	3	observers	Maximum number of observers used for source localization
d_max	2	hops	Maximum search distance to first infected observer

Table 1: System parameters of cholera outbreak case study.

The general source estimator in (4) cannot be directly applied here because of two particularities of the cholera diffusion process: i) the *direction of arrival* of the vibrios cannot be observed, only its *timing*, and ii) there is a non-negligible *measurement delay* between infection by the vibrios and reporting to local health authorities. It is straightforward to show that the estimator in (4) can be extended in order to accommodate these differences, as follows.

Source estimator for general trees (jointly-Gaussian diffusion, Gaussian IID measurement delays)

$$\hat{s} = \underset{s \in \mathcal{T}}{\operatorname{argmax}} \frac{\exp\left(-\frac{1}{2}(\mathbf{d}-\boldsymbol{\mu}_s)^\mathrm{T}\boldsymbol{\Lambda}_s^{-1}(\mathbf{d}-\boldsymbol{\mu}_s)\right)}{|\boldsymbol{\Lambda}_s|^{1/2}}, \tag{18}$$

with

$$\boldsymbol{\mu}_s = \mathbf{C}_s\boldsymbol{\mu}_{\boldsymbol{\theta}}, \tag{19}$$

$$\boldsymbol{\Lambda}_s = \mathbf{C}_s\boldsymbol{\Lambda}_{\boldsymbol{\theta}}\mathbf{C}_s^\mathrm{T} + (\mathbf{1}_{K-1} + \mathbf{I}_{K-1}) \cdot \sigma_\mathrm{m}^2, \tag{20}$$

where $\mathbf{1}_n$ is the $n \times n$ matrix of ones, $\mathbf{I}_n$ is the $n \times n$ identity matrix, and σ_m is the standard deviation of the measurement delay.

We performed two additional optimizations. First, since it is often desirable to localize and limit the outbreak as soon as possible, we do not wait until all K observers get infected in order to estimate the source location. Instead, we perform the estimation as soon as the first $K_\mathrm{max} = 3$ observers get infected. Second, since it is likely that the actual source is in the neighbourhood of the *first infected observer,* we limit the maximization in (18) to all nodes within $d_\mathrm{max} = 2$ hops of the first infected observer.

Note that the resulting *error distance* is a random variable, since it depends on the (random) location of the observers. Figure S5 plots the distribution of the error distance, for various observer densities.

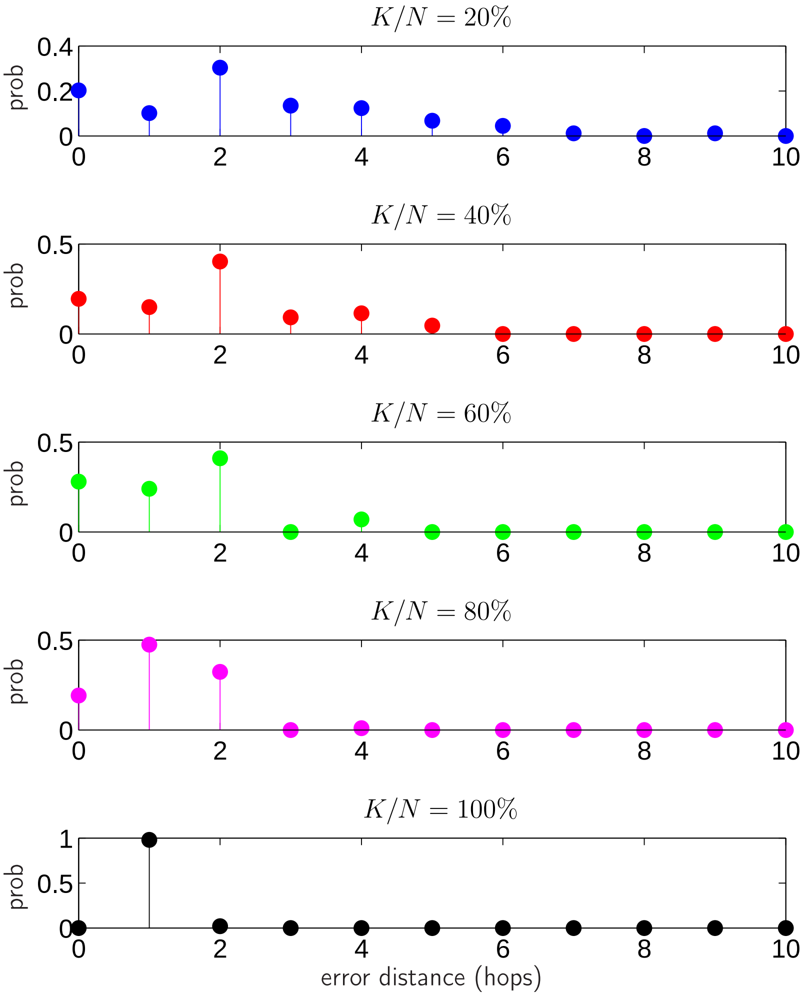


Figure S5: Probability mass function of the error distance (in hops), for various observer densities K/N.