Diffusion Source Detection in a Network using Partial Observations

Roxana Alexandru and Pier Luigi Dragotti

Department of Electrical and Electronic Engineering
Imperial College London

ABSTRACT

Diffusive phenomena are ubiquitous in nature and society, and have been extensively studied in various fields, such as natural sciences and engineering. Recently, however, the more challenging inverse problem of diffusion source detection in a network has started to receive a significant amount of attention. A lot of research has concentrated on finding origins in tree-like networks, however these approaches cannot be easily extended to generic networks. Furthermore, only some methods consider realistic temporal diffusion dynamics. We introduce a novel method to localise the source of multiple rumours in an arbitrary network of known topology, using partial observations of the network nodes. We first present two mathematical models of the discrete-time, susceptible-infected propagation dynamics, which accurately capture the diffusion process and have low computational complexity. The first one is a simplified likelihood of infection at a node, at a certain time after the rumour is initiated. The second is a formulation of the infection likelihood of a node, as a function of its shortest distance to the source. We then design an efficient single source detection algorithm, which leverages these mathematical models of diffusion, and the assumption that the start time of the propagation is known. Finally, we show how these methods can be extended to the case when the start time of the rumour is unknown, by taking advantage of the dissimilarity in dynamics of infection, of different nodes in the network. Simulation results show that a high source estimation probability is achieved using a small number of observations.

Keywords: social networks, diffusion of information, susceptible-infected epidemic model, rumour source identification, unknown start time of the epidemic.

1. INTRODUCTION

As social networks have developed and the spreading of information has greatly amplified, the dynamics of information dissemination within a network have attracted considerable attention. Recently, however, several authors have started to consider the more challenging inverse problem, of detecting the source responsible for the spreading of information.[1] This problem is motivated by interesting applications, such as: estimating the origin of rumours and finding influencers in social networks, determining the causes of cascading failures in large systems such as financial markets or sensor networks, and identifying the origin of infectious diseases or computer viruses.

Most state-of-the-art approaches focus on estimating diffusion sources in simple topologies such as trees or random geometric graphs,[2–6] with a few other methods tailored for generic topologies.[7–12] Many source identification methods are based on the assumption that there is access to information at all the nodes in the network, which is unrealistic for large graphs.[1, 13–16] On the other hand, only a few methods leverage partial observations of the network.[9, 17–22] Furthermore, regarding the epidemic model, most methods assume the susceptible-infected (SI) model, where the nodes can either be infected or susceptible and once a node has received the rumour, it cannot recover from it. Some other typical epidemic models include the susceptible-infected-recovered (SIR) model,[20] where nodes can transition from an infected to a recovered state (in which case they never become infected again), as well as the susceptible-infected-susceptible (SIS) model,[18, 23] where nodes

Further author information:
R.A.: E-mail: ria12@imperial.ac.uk
P.L.D.: E-mail: p.dragotti@imperial.ac.uk

can transition from a susceptible to an infected state and vice-versa. Most algorithms focus on the detection of a single source and there are a few methods that can be used to identify multiple sources, however these are typically more computationally expensive and challenging to implement.[3]

The objective of this work is to estimate the source of information on a general graph of known topology, using observations from a finite set of monitor nodes. Moreover, we assume that the source emits multiple rumours of information at the same time, which spread independently within the network, according to the susceptible-infected epidemic model. The assumption of multiple rumours increases the diversity of observations and makes the estimation of the source more reliable.

This paper is structured as follows. In Section 2, we introduce the problem of estimating the source of rumours in a network. Then, in Section 3, we describe two mathematical models of information propagation over graphs, previously introduced by Alexandru et al. in the context of rumour source detection with known activation time.[24] We also present an algorithm for estimating a single rumour source in Section 4.1, when we assume that the activation time of the rumour is known.[24] Furthermore, in Section 4.2 we develop an algorithm which achieves inference of a single rumour source, when the start time of the rumour is unknown. In Section 5 we assess the performance of the algorithms on synthetic as well as real-world graphs. Finally, we conclude in Section 6.

2. PROBLEM SETTING

2.1 Network Model

A graph is defined as a set of nodes (or vertices) connected by edges (or links), and is mathematically represented via an adjacency matrix G, where the entry $g_{i,j} = 1$ if nodes i and j are connected, and $g_{i,j} = 0$ otherwise.

Graphs can be classified according to the attributes they exhibit. Some of these attributes are: the node degree (the number of neighbours of a node), the average shortest path length (the average of all the shortest distances in the network), and the clustering coefficient (a measure of the tendency of nodes in a network to cluster together). Regular networks exhibit great clustering coefficient and long average shortest distance, whereas random networks typically have small clustering coefficient and short average distance. Real-world networks are neither completely regular nor completely random, and typically exhibit small-world and scale-free properties. In a small-world graph, most nodes are not neighbours of each other, but they can be reached from any other node in a small number of hops (edges). This leads to small average shortest distance and great clustering coefficient. Scale-free networks also have these attributes, and their node degree follows a power-law distribution, which means that the network will contain very high degree nodes (or *hubs*). The presence of hubs is very common in real-world networks, as highlighted in Fig. 1. Here, the small-world network was generated using the Watts-Strogatz model,[25] the scale-free graph with the Barabási-Albert model,[26] and the Facebook subgraphs are extracted from the SNAP dataset.[27]

Graphs can also be categorized according to the type of relationships between nodes. The first category is that of directed networks, where the links between nodes have a certain directionality. The second class is represented by the undirected graphs, where there is a two-way relationship between any connected nodes. Throughout this work, we primarily focus on undirected graphs, with small-world and scale-free properties, which model a wide range of social networks (e.g. Facebook or Linkedin), as well as biological networks (e.g. protein interaction networks).

2.2 Epidemic Model

We consider a discrete-time version of the susceptible-infected epidemic model. Initially all the nodes are in a susceptible state. Then, once the rumour starts propagating, each susceptible node can get infected from any of its infected neighbours at any discrete time step, with a certain probability μ. Moreover, once a node becomes infected, it cannot recover from this state. This epidemic model is depicted in Fig. 2, where the epidemic starts spreading at $t = 2$ in a small-world graph of 10 nodes. At subsequent discrete times, each infected node can transmit the infection to any of its susceptible neighbours, with probability $\mu = 0.5$. At time instance $t = 5$, all the network nodes are infected.

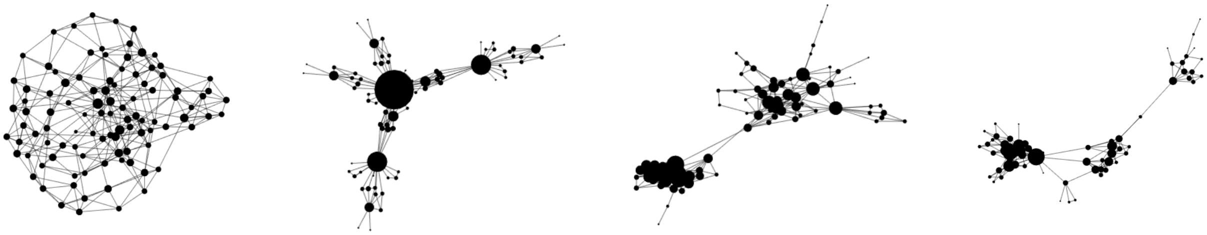

Figure 1. Topology of small-world network (left), scale-free network (second left), and two Facebook subgraphs (right).

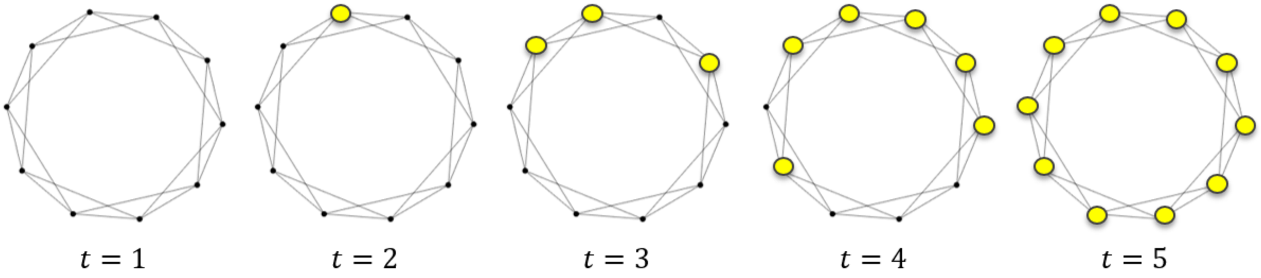

Figure 2. Spreading of information using the discrete-time susceptible-infected model, in a small-world graph of 10 nodes.

2.3 Rumour Source Localization

We assume that a single node in the network starts emitting multiple rumours of related information, at the same time instance. This is a realistic assumption, as influencers tend to disseminate multiple pieces of information in order to increase their presence in a social network. Then, we allow the rumours to spread independently of each other, according to the susceptible-infected epidemic model described in Section 2.2, and monitor a small set of nodes in the network throughout an observation window. We aim to localize the source of the rumours, using the observations at these monitors at different time instances. The problem statement is depicted in Fig. 3, where we monitor the state of three nodes (highlighted in red), at all the time instances in the observation window up to time $t = 5$.

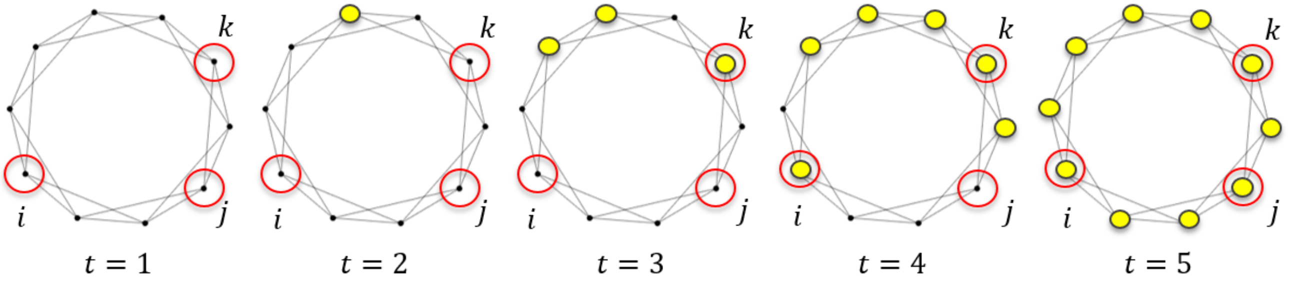

Figure 3. Observations at a small set of nodes (i, j and k), in a small-world network of 10 nodes.

We assume the source emits R rumours, and that we have knowledge of how many of these rumours reach each monitor node, at any time instance t. Then, we can define the observed probability of infection of a monitor i at time t as follows:

$$\tilde{F}_i(t) = \frac{R_i(t)}{R}, \tag{1}$$

where $R_i(t)$ is the number of rumours that reached node i by time t.

3. MATHEMATICAL MODEL OF DIFFUSION OVER NETWORKS

In this section we describe two mathematical models which accurately capture the diffusion process in a network, and which were previously introduced in the context of rumour source identification in social networks.[24]

3.1 Exact and Simplified Likelihood of Infection

Given a node s initiates the rumour at $t = 0$, the probability of first infection of node i at time t is given by:

$$f_{i|s}(t) = [1 - \prod_{j \in N_i} (1 - \mu F(x_j(t-1) = 1 | x_i(t-1) = 0))] \times \prod_{\tau=1}^{t-1} (1 - f(x_i(\tau) = 1)), \tag{2}$$

where the conditional probability can be further expanded using Bayes' rule and the law of total probability, as follows:

$$\begin{aligned} F(x_j(t-1) = 1 | x_i(t-1) = 0) &= \frac{F(x_j(t-1) = 1, x_i(t-1) = 0)}{F(x_i(t-1) = 0)} \\ &= \frac{\sum_{K_{t-1}} F(x_j(t-1) = 1, x_i(t-1) = 0, x_k(t-1), \forall k \in G \setminus \{i, j\})}{F(x_i(t-1) = 0)} \\ &= \frac{\sum_{Q_{t-2}} F(x_j(t-1) = 1, x_i(t-1) = 0, x_k(t-1), \forall k \in G \setminus \{i, j\} | x_q(t-2), \forall q \in G) \times F(x_q(t-2), \forall q \in G)}{F(x_i(t-1) = 0)}, \end{aligned}$$

where:

N_i = set of neighbours of node i,
μ = constant edge transmission likelihood in the network,
G = the set of all nodes in the graph,
K_{t-1} = set of possible states of nodes $k \in G \setminus \{i, j\}$, at time $t-1$,
Q_{t-2} = set of possible states of nodes $q \in G$, at time $t-2$,
$x_k(t)$ = state of node k at time t.

The above likelihood can be simplified, based on the observation that the state of node j is not significantly influenced by the state of a single node i.[24] The simplified likelihood is less computationally expensive, and is given by:

$$f_{i|s}(t) \approx [1 - \prod_{j \in N_i} (1 - \mu F(x_j(t-1) = 1))] \times \prod_{\tau=1}^{t-1} (1 - f(x_i(\tau) = 1)). \tag{3}$$

Finally, a node i is infected at time t if it *first* got infected at any time instance before. This means that the probability of a node i to have the infection at time t, given node s initiates the rumour at $t_0 = 0$, is given by:

$$F_{i|s}(t) = F(x_i(t) = 1 | x_s(0) = 1) = \sum_{\tau=1}^{t} f_{i|s}(\tau). \tag{4}$$

3.2 Distance-dependent Likelihood of Infection

Alexandru et al.[24] also propose a distance-dependent likelihood of infection, which gives the probability of infection of a node as a function of its shortest distance to the source of the rumours. This is computed as:

$$F_d(t) = \sum_{\tau=d}^{t} (\alpha_d \mu)^d (1 - \alpha_d \mu)^{\tau - d} \binom{\tau - 1}{d - 1}, \tag{5}$$

where:

d = shortest distance between the node and the source,
μ = constant edge transmission likelihood in the network,
α_d = parameter which captures the properties of the network.

The derivation of the distance-dependent likelihood in Eq. (5) is based on finding the *average* number of paths in which the rumour can reach a node at distance d from the source (the term $\binom{\tau-1}{d-1}$), and the probability of each of these paths (the term $(\alpha_d\mu)^d(1-\alpha_d\mu)^{\tau-d}$).[24] For example, in Fig. 4 we show that there are 4 different ways to reach node j located 3 hops away from the source s, in 3 time steps. At the same time there are only two ways to reach node i at distance $d = 3$ hops from the source, in exactly 3 time steps. As a result, the average number of paths from s to a node at distance $d = 3$ is equal to 3. Moreover, given the constant edge transmission rate of μ, the probability of each path is μ^3. Hence, the overall probability of a node located at distance $d = 3$ from the source, to get infected at $t = 3$ is $F_3(t = 3) = 3\mu^3$.

Furthermore, the parameter α_d captures the *avalanche* effect of the propagation of the rumour in the network. In other words, in a sufficiently dense network, the probability that a node at distance d from the source is infected at time $t \geq d$ increases for larger values of d, and this should be captured in a larger value of α_d.

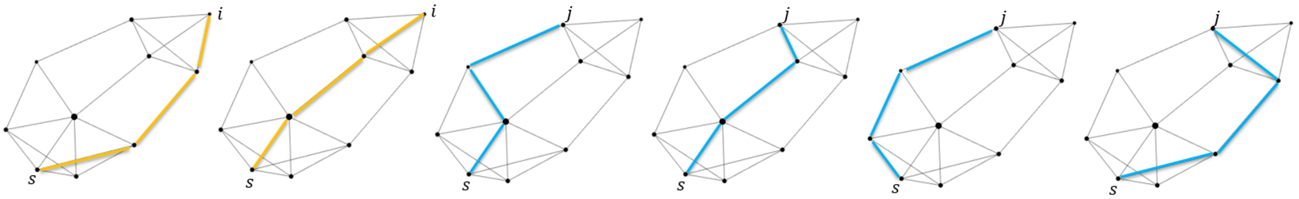

Figure 4. Different ways for node i (left subplots) and j (right subplots), located at shortest distance $d = 3$ hops from the source s, to get infected at time $t = 3$, given node s starts spreading the rumour at $t = 0$.

In Fig. 5, we show a comparison between the simplified likelihood of infection in Eq. (3) and (4) and the distance-dependent infection likelihood in Eq. (5). The results were obtained by simulating a spreading of 1000 rumours from a randomly selected node s, in a small-world network of 200 nodes, with average node degree 6. Even though the simplified and the distance-dependent likelihoods defined in Eq. (3) and Eq. (5) respectively, assume that the edge transmission rate μ is constant, in these results we allow μ to vary. In particular, we draw μ uniformly at random from $[0, 1]$. The average value of the edge transmission likelihood is $\mu = 0.5$, which we use for the computations of the likelihoods in Eq. (3) and Eq. (5).

We notice that the simplified infection likelihood follows more closely the infection pattern of individual nodes (top plots). On the other hand, the distance-dependent likelihood of infection has the same shape for any nodes i and j, since their shortest distances to the source are equal, $d_i = d_j = 2$.

4. RUMOUR SOURCE DETECTION

4.1 Rumour Source Detection with Known Activation Time

Let us consider the case when a single source emits R rumours of related information. Moreover, we assume that all these rumours are sent at the same time instance, which is known, and which we set as $t_0 = 0$ for simplicity. Our aim is to localize the source of the rumours, based on measurements at a small set of monitors S_M, througout an observation window $[0, T]$. The source detection algorithm we present in this section leverages the analytical infection likelihoods presented in Section 3.1 and 3.2, as well as the diversity of observations $\tilde{F}_i(t)$ in Eq. (1), created by multiple rumours being spread.

4.1.1 Computing the Distance-dependent Infection Likelihoods

In the first step of the algorithm, we compute the shortest distances between any two nodes in the network, using the Dijkstra algorithm.[28] Then, we learn the parameters α_d used in Eq. (5) as follows. We artificially simulate a spreading of rumours from a randomly selected node s in the network, and compute the observed

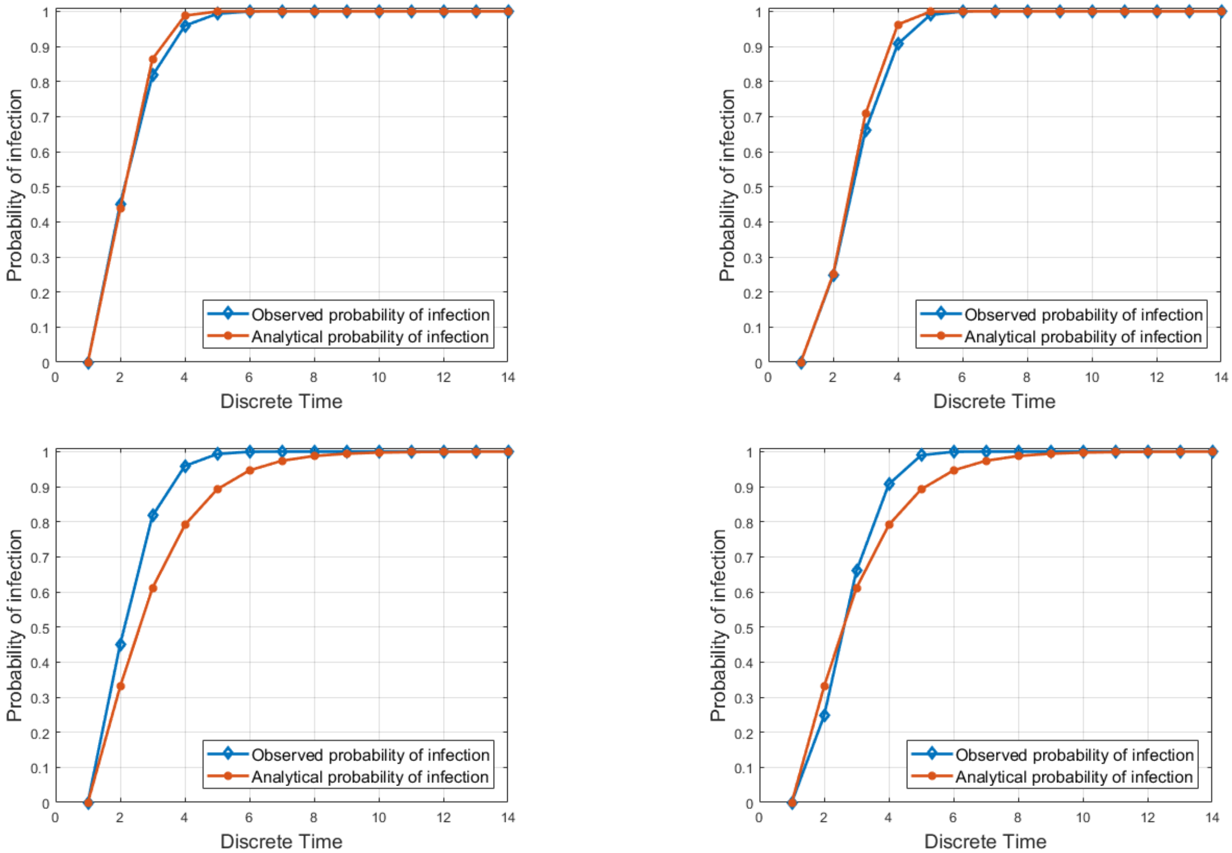

Figure 5. Comparison between the observed and simplified infection likelihoods of nodes i (top left) and j (top right). Comparison between the observed and distance-dependent likelihoods of infection of nodes i (bottom left) and j (bottom right). The distances between nodes i and j to the rumour source are equal, $d_i = d_j = 2$ hops.

likelihood of infection $\tilde{F}_i^{learning}(t)$ using Eq. (1), at all the nodes i in the network. Using the shortest distances we have calculated, we find the optimal parameter α_d for each shortest distance to the source s, by minimizing the divergence between the analytical infection likelihood $F_d(t)$ and the observations $\tilde{F}_i^{learning}(t)$:

$$\alpha_d^{opt} = \arg\min_{\alpha_d \in (1,\frac{1}{\mu})} [\sum_{i \in N_d} \sum_{t=0}^{T} \left\| F_d(t) - \tilde{F}_i^{learning}(t) \right\|^2], \tag{6}$$

where N_d is the set of nodes at shortest distance d from the source s, μ is the constant edge transmission rate and the upper bound $\frac{1}{\mu}$ of α_d ensures stability of Eq. (5).

4.1.2 Creating a Set of Potential Sources

By fitting the observations $\tilde{F}_i(t)$ at each monitor i to the distance-dependent infection likelihoods $F_d(t)$ in Eq. (5), we estimate the shortest distance between node i and the source. The optimal distance d_i^{OPT} for a monitor i is the one which minimises the divergence between $F_{d_i}(t)$ and $\tilde{F}_i(t)$:

$$d_i^{OPT} = \arg\min_{d_i \in [1,r]} [\sum_{t=0}^{T} \left\| F_{d_i}(t) - \tilde{F}_i(t) \right\|^2],, \tag{7}$$

where r is the network diameter (i.e. largest shortest distance between any two nodes in the network), and T is the length of the observation window.

Then, leveraging the estimated distances between all the monitors and the source, we build a set of potential sources using triangulation, as follows. For each monitor i whose estimated shortest distance to the source is d_i, we keep all the nodes at distance d_i from i as potential sources, and denote their set with G_i. We then repeat this for all the monitors in the set S_M. The final set of candidate sources P_S is the intersection of the sets of candidate sources found for all monitors: $P_s = \bigcap_{i \in S_M} G_i$.

For example, using the measurements at nodes i, j and q depicted in Fig. 6, we estimate the shortest distances between these nodes and the origin of the rumours as $d_i = 2$, $d_j = 2$ and $d_q = 1$. As highlighted in Fig. 7, we then use the estimated distances and triangulation to build a set of potential sources. In this example, the intersection of the three sets of potential sources contains node n only.

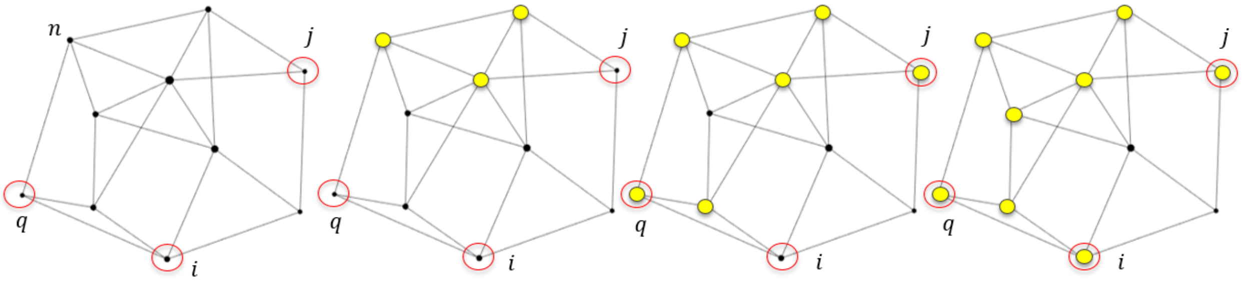

Figure 6. Observations at nodes i, j and q, when multiple rumours are initiated by node n, in a small-world network of 10 nodes.

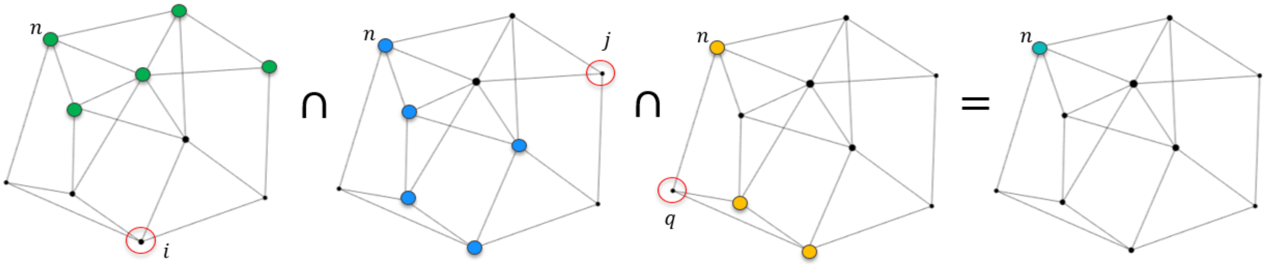

Figure 7. Using the estimated shortest distances between the source and monitors i, j and q, to build a set of potential sources. The final set of candidate sources (right subplot) contains node n only.

4.1.3 Finding a Unique Rumour Source

In some cases, we may only observe a very small set of monitors. Then, it may not be possible to find a unique source using triangulation based on the estimated shortest distances. This is illustrated in the example in Fig. 8, where we only monitor nodes i and j and estimate their distances to the source as $d_i = 2$ and $d_j = 2$ respectively. The final set of candidate sources contains both nodes n and k.

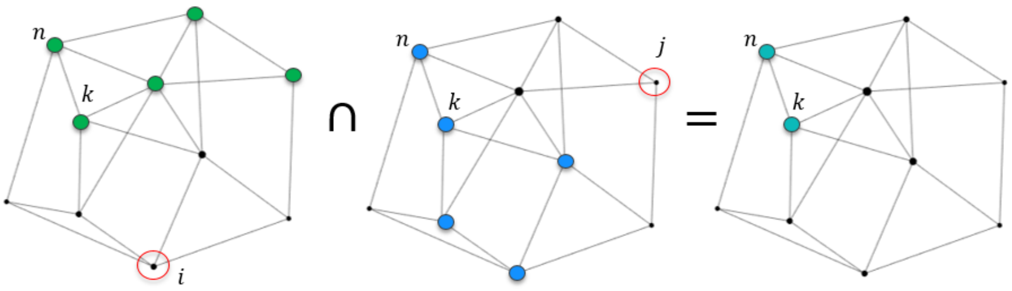

Figure 8. Using the estimated shortest distances between the source and monitors i and j, in order to build a set of potential sources. The final set of candidate sources (right subplot) contains nodes n and k.

The distance-dependent formulation described in Section 3.2 assigns the same infection likelihood to a node i at distance d_i from the source s, no matter where the source is localized. As a result, we cannot use this analytical likelihood to differentiate between the two potential sources n and k in Fig. 8. Nevertheless, the infection pattern of node i may be different, depending on the source node which started the spreading of the rumours. This is highlighted in Fig. 9, where we see that monitor i is more likely to get infected sooner if the rumour is initiated by node n, than in the case when the rumour is started by node k. An explanation for this infection pattern could be the fact that, even though the shortest distances between i and nodes k and n are equal, i.e. $d_{ik} = d_{in} = 2$, there are more paths of length equal to 2 between nodes k and i, than there are between n and i.

The analytical likelihood $F_{i|s}(t)$ defined in Eq. (4) captures the different infection patterns of the monitors, depending on which node initiates the rumours. Hence, we can fit this simplified analytical likelihood of infection $F_{i|s}(t)$ to the observations $\tilde{F}_i(t)$, in order to estimate the most likely origin of the rumours, from a set of candidate sources. The most likely rumour source s^{OPT} minimises the divergence between $F_{i|s}(t)$ and $\tilde{F}_i(t)$:

$$s^{OPT} = \arg\min_{s \in P_S} [\sum_{t=0}^{T} \left\| F_{i|s}(t) - \tilde{F}_i(t) \right\|^2], , \tag{8}$$

where P_S is the set of potential sources found using triangulation and T is the length of the observation window.

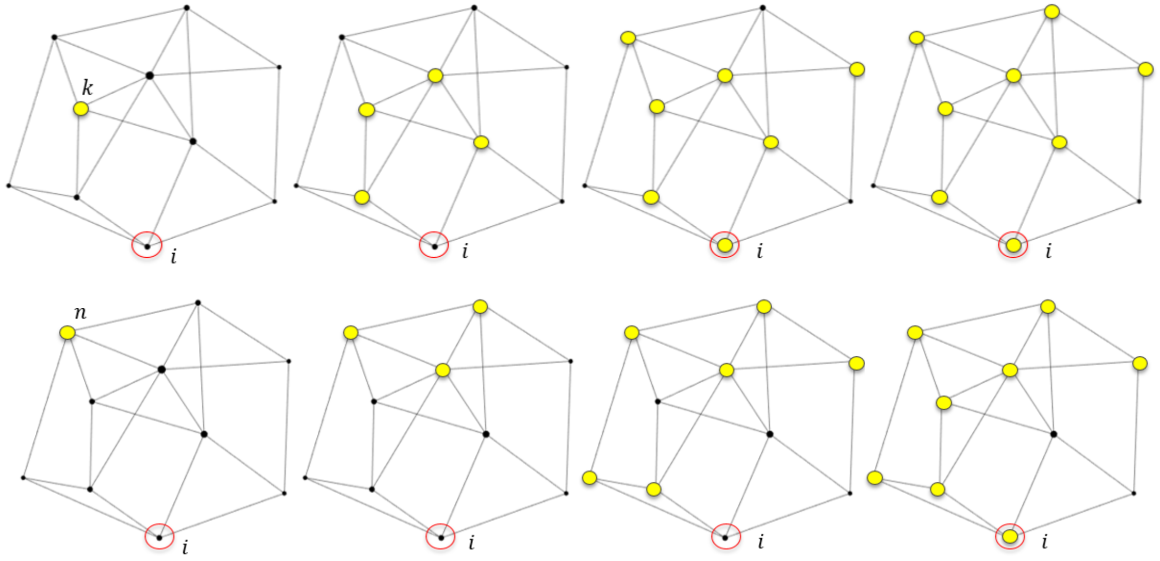

Figure 9. Spreading of multiple rumours initiated by node k (top), and node n (bottom), in a small-world network of 10 nodes. The infection pattern of node i is different depending on where the rumour starts.

4.1.4 Overall Algorithm

The overall source detection method is summarized in Algorithm 1. The required inputs are the network topology, the set of monitors S_M, the edge transmission likelihood μ and the observed likelihoods of infection $\tilde{F}_i(t)$ of each monitor node $i \in S_M$, computed using Eq. (1), at all the time instances within the observation window $[0, T]$.

Algorithm 1 Single Source Detection Algorithm, with Known Activation Time

```
Require: Network topology, measurements $\tilde{F}_i(t)$, for $t \in [0, T]$ and $i \in S_M$, edge transmission likelihood $\mu$.
1: Compute the shortest distances between any two nodes in the network, using Dijkstra algorithm,[28] and find the network diameter $r$ (i.e. the largest shortest distance between any two nodes).
2: Learn the optimal parameters $\alpha_d$ as in Eq. (6), using an artificial spreading of rumours from a random node in the network, which generates a set of measurements $\tilde{F}_i^{learning}(t)$. Hence, find the theoretic infection distributions $F_d(t)$ using Eq. (5), for all shortest distances $d \in \{1, 2, ..., r\}$ and for $t \in \{0, 1, ..., T\}$.
3: By fitting the observations $\tilde{F}_i(t)$ to $F_d(t)$ as in Eq. (7), estimate the shortest distances between any monitor node $i$ and the source.
4: Leveraging the estimated shortest distance between a monitor $i$ and the source, create a set of candidate sources. Repeat this for all the monitors in the set $S_M$ and find the intersection $P_S$ of all the sets of potential sources.
5: If the set $P_S$ contains more than one node, we find the most likely origin of the rumours as follows. We compare the simplified infection likelihoods $F_{i|s}(t)$ in Eq. (4) for each node $s$ in the set $P_S$, to the observed likelihoods $\tilde{F}_i(t)$, of all the monitor nodes. We then use Eq. (8) to select the most likely rumour source. This is node $s$ for which the cumulative divergence between $F_{i|s}(t)$ and $\tilde{F}_i(t)$ for $i \in S_M$, is minimised.
```

4.2 Rumour Source Detection with Unknown Activation Time

Let us now consider the case when a single source emits R rumours of related information, at the same time, which is unknown. For simplicity, let us assume that the observation starts at $t = 0$, and that the rumour is initiated at instance $t_0 \geq 0$. Our aim is to use the discrete observations of a small set of monitors S_M to estimate both the location of the source, as well as the activation time of the rumour. We first estimate the shortest distances between the monitors and the source, and then use a triangulation method to find a set of potential sources, based on these estimated distances.

4.2.1 Estimation of the Shortest Distances between the Monitors and the Source

We consider a K-medoids approach,[29] which uses the measurements $\tilde{F}_i(t)$ at all the monitor nodes $i \in S_M$, to estimate the shortest distances between the monitors and the source.

We define a *medoid* (or cluster) as a set of nodes C_d which are located at shortest distance d from the source. Moreover, each *medoid* C_d will be assigned a *prototype* funtion P_d, which is the distance-dependent likelihood of infection in Eq. (5), delayed by a certain time t_{start}:

$$P_d = F_d(t + t_{start}) = \sum_{\tau = t_{start}+d}^{t_{start}+t} (\alpha_d \mu)^d (1 - \alpha_d \mu)^{\tau - d} \binom{\tau - 1}{d - 1}, \tag{9}$$

where :

α_d = parameter which reflects the properties of the network, and which is specific to the cluster C_d,
t_{start} = the expected start time of the rumours.

We initialise the parameter α_d corresponding to each medoid C_d as follows:

$$\alpha_d = 1 + xd, \tag{10}$$

where the parameter x is chosen to guarantee that $\alpha \in (1, \frac{1}{\mu})$, in order to ensure stability of $F_d(t)$ in Eq. (5).

If we denote the network diameter with r (the largest shortest distance between any two nodes in the network), then setting $x < \frac{1-\mu}{\mu r}$ ensures that $\alpha_d < \alpha_r < 1 + \frac{1-\mu}{\mu r} r = \frac{1}{\mu}$.

In addition, we initialise the start time t_{start} as follows:

$$t_{start} = t_f - \frac{1}{\mu}, \tag{11}$$

where t_f is the infection time of the first infected monitor f in the set S_M.

This initialisation is based on the assumption that the delay between the time of infection of monitor f in the set S_M, and the rumour start time, is $\frac{1}{\mu}$. In other words, we assume monitor f is located at a small distance from the source of the rumour. In reality, this may not be true, especially if the set of observers S_M is sparse. Nevertheless, we shall see how to optimize t_{start} in the next part of the algorithm.

Once both parameters α_d and t_{start} have been initialised, we repeat the following two steps of the K-medoids algorithm until convergence to a local optimum. In the first step of the algorithm, each monitor in the set S_M is assigned to the *closest* medoid. For each monitor i, the closest medoid is the one for which the mean-squared error between the cluster's prototype defined in Eq. (9), and the observed infection probability $\tilde{F}_i(t)$, is minimised. This error is computed as follows:

$$\tilde{e}(i, C_d) = \sum_{t=0}^{T} \left\| F_d(t + t_{start}) - \tilde{F}_i(t) \right\|^2, \tag{12}$$

where T is length of observation window, $F_d(t + t_{start})$ is the cluster's prototype, and $\tilde{F}_i(t)$ is the observed probability of infection of monitor i at time t.

In the second step of the algorithm, we adjust α_d and t_{start} such that the *dissimilarity* of the medoid is minimised. The overall dissimilarity is the cumulative mean-squared error between all the data points belonging to the medoid C_d, and its prototype $P_d(t) = F_d(t + t_{start})$:

$$\tilde{E}(C_d) = \frac{1}{|C_d|} \sum_{i \in C_d} \sum_{t=0}^{T} \left\| F_d(t + t_{start}) - \tilde{F}_i(t) \right\|^2, \tag{13}$$

where $|C_d|$ is the number of data points in medoid C_d.

The two steps of the algorithm are repeated until convergence to a local optimum. When this is achieved, each monitor i will be assigned to the *closest* medoid C_d, which corresponds to the estimated shortest distance d between this monitor and the source.

In Algorithm 2 we formally present the K-medoids approach, which estimates the shortest distances between the monitors and the source. The inputs to the algorithm are the network topology, the constant edge transmission rate μ and the observations $\tilde{F}_i(t)$ computed using Eq. (1), of all the monitor nodes $i \in S_M$.

4.2.2 Estimation of a Set of Potential Sources

Once the shortest distances between the monitor nodes and the source have been estimated, we use the following triangulation method to build a set of potential sources. Suppose the estimated shortest distances between monitors i and j to the source are $\tilde{d}_i$ and $\tilde{d}_j$ respectively. Then, the set of potential sources will contain all the nodes located $\tilde{d}_i - \tilde{d}_j$ hops closer to node j, compared to node i. Since the start time of the rumours is unknown, the estimation of the absolute shortest distances $\tilde{d}_i$ and $\tilde{d}_j$ may not always be accurate. Nevertheless, the relative distance $\tilde{d}_i - \tilde{d}_j$ is likely to be correct.

Algorithm 2 K-medoids algorithm for estimation of the shortest distances between the monitors and the source, when the rumour start time is unknown.

```
Require: Network topology, measurements $\tilde{F}_i(t)$, for $t \in [0, T]$ and $i \in S_M$, edge transmission probability $\mu$.
1: Compute the pairwise shortest distances in the network, using Dijkstra algorithm,[28] and find the network diameter $r$.
2: Initialise the parameter $x$ in Eq. (10) with $x = \frac{1-\mu}{\mu(r+1)}$, which ensures $\alpha_d \in (1, \frac{1}{\mu})$, $\forall d$. Hence, initialise the parameters $\alpha_d$ corresponding to each medoid $C_d$ (the cluster of monitors at distance $d$ from the source).
3: Initialise the start time of the rumours as in Eq. (11), with $t_{start} = t_f - \frac{1}{\mu}$, where $t_f$ is the infection time of the first infected monitor in the set $S_M$.
4: Using the initial values of $\alpha_d$ and $t_{start}$, compute the prototype function $P_d$, of each medoid $C_d$.
5: Assign each monitor to the closest medoid, by minimising the error in Eq. (12).
6: While the dissimilarity of the medoid decreases, iterate:
    1. Adjust the parameters $\alpha_d \in (1, \frac{1}{\mu})$ and $t_{start}$ such that the cumulative divergence defined in Eq. (13), of all data points to the cluster prototype is minimised. Hence, re-compute the prototype $P_d$ of each medoid $C_d$ as in Eq. (9).
    2. Re-assign each monitor to the closest medoid, by minimising the error in Eq. (12).
```

5. EXPERIMENTAL RESULTS

5.1 Rumour Source Detection with Known Activation Time

We validate the proposed algorithms for three different network types: small-world graph, scale-free network and graphs extracted from Facebook. In all cases, a single source emits $R = 10$ rumours at time $t = 0$, and these propagate independently across different edges, with varying edge probability. In particular, the edge transmission likelihood is uniformly drawn from the interval $[0, 1]$. In Fig. 10, we highlight the high accuracy of estimation of a single rumour source, when the start time of the rumours is known. For example, when we observe 20% of the nodes in the small-world network, the probability of correctly estimating the rumour source is 1. When we observe at least 10% of the network nodes, the probability of correct source estimation is above 95%. The likelihood of correct estimation in scale-free networks and real-world graphs is slightly lower, and this may be due to the high node degree variability in these networks (see left subplots in Fig. 10 showing the degree of the nodes).

5.2 Rumour Source Detection with Unknown Activation Time

We first evaluate the efficiency of the K-medoids algorithm in estimating the parameters α_d used in Eq. (9), as well as the start time of the rumours t_{start}, as described in Section 4.2. For the results in Fig. 11, the observation begins at $t = 0$, and a random node in the network starts spreading $R = 1000$ rumours at time $t = 5$. Then, the prototype distance-dependent infection likelihoods $F_d(t + t_{start})$ are learnt using the method described in Section 4.2. From the results in Fig. 11, we notice that as expected, nodes at distance $d = 1, 2$ and 3 from the source, have a positive probability of infection at times $t = 6, 7$ and 8 respectively. Moreover, the shape of the analytical distance-dependent likelihoods follows closely the observations $\tilde{F}_i(t)$, which shows that the parameters α_d were correctly learnt.

The performance of the source detection algorithm is analysed in a small-world network of 200 nodes, when 10 rumours are initiated by the same source at time $t = 5$, following the start of the observation at $t = 0$. The results in Fig. 12 are averaged over 100 experiments. These show that the probability of correctly estimating the rumour source is above 90%, even when the fraction of observed nodes is small, in the case when we allow the edge probability μ to vary. In particular, the likelihood of transmission of each edge is uniformly drawn from the interval $[0, 1]$. When we fix $\mu = 0.5$, the probability of correct estimation increases above 95%.

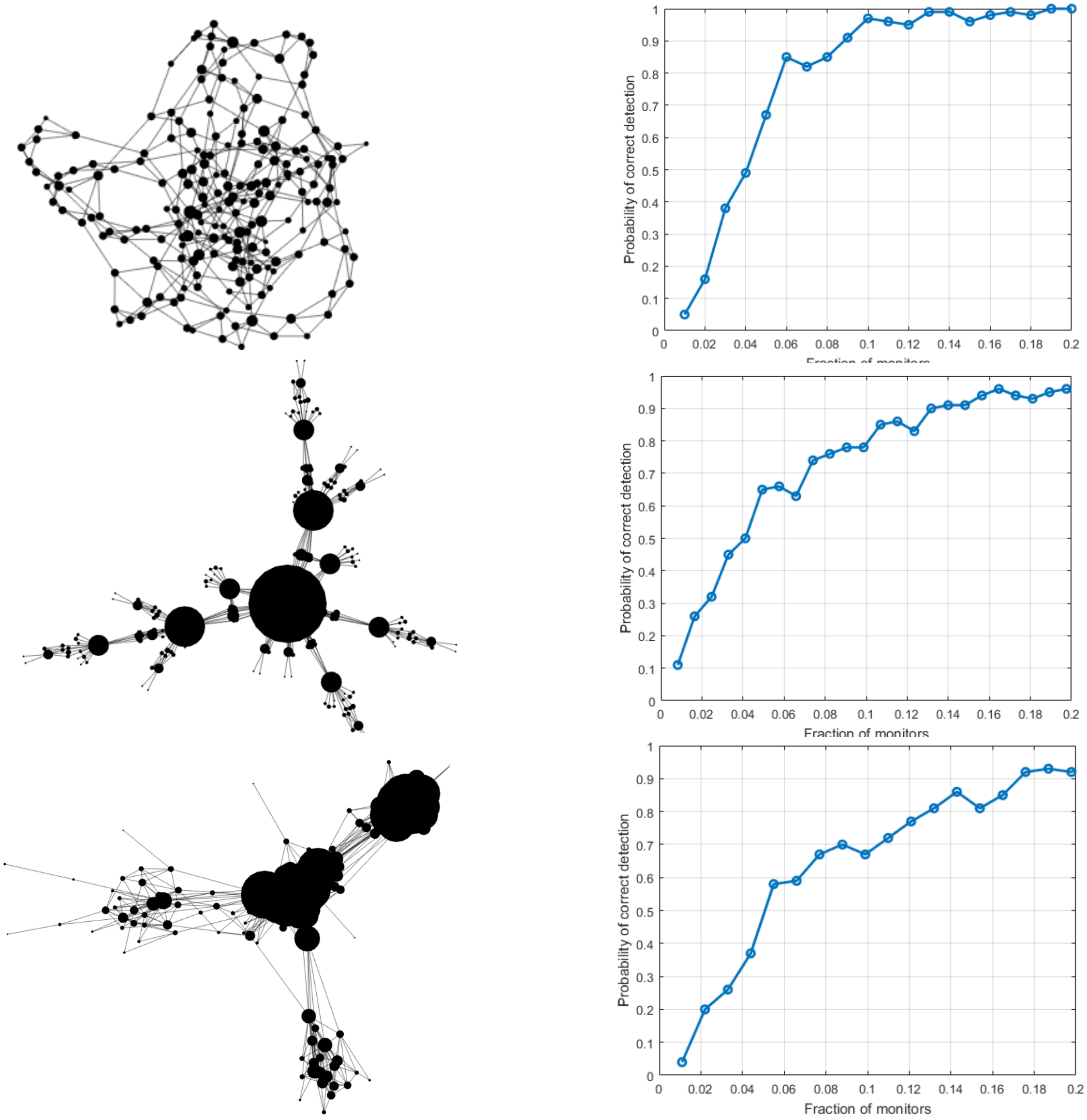

Figure 10. Probability of correctly estimating a single rumour source, when the activation time of the rumour is known, in a small-world network of 200 nodes (top), scale-free network of 243 nodes (middle), and Facebook subgraph of 182 nodes (bottom). In all cases, we show the network topology (left), and the probability of correctly estimating the rumour source for different fractions of observed nodes (right).

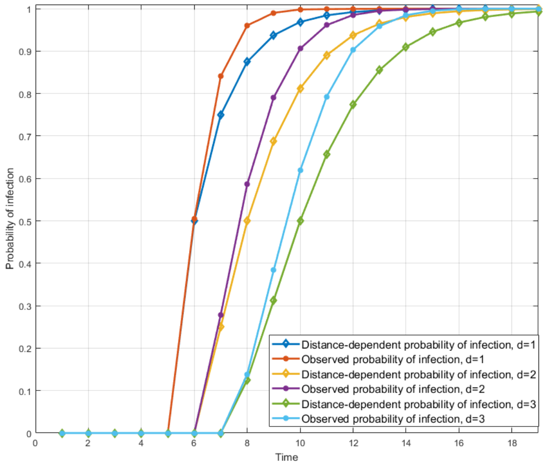

Figure 11. Comparison between the distance-dependent infection likelihoods $F_d(t + t_{start})$ and the average observations $\tilde{F}_i(t)$ at nodes at distance $d = 1$, $d = 2$ and $d = 3$ respectively, from the rumour source. The distance-dependent infection likelihoods are learnt using the K-medoids algorithm.

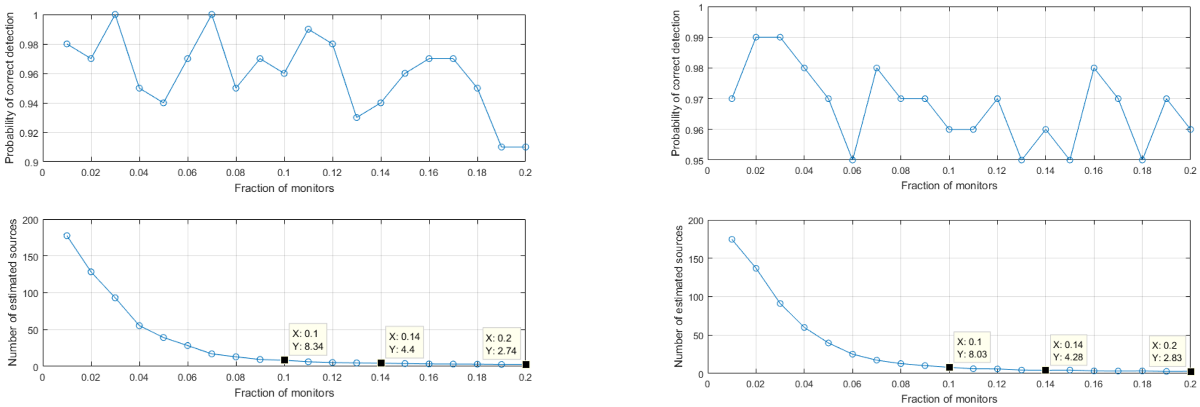

Figure 12. Probability that the true rumour source is in the set of estimated sources, when the rumours are spread in a small-world network, with varying edge transmission probability (left) and constant edge likelihood (right). The top plots show the probability of correct estimation of the rumour source, whereas the bottom ones show the number of nodes in the set of potential sources.

6. CONCLUSION

In this paper we described two mathematical models which accurately capture the diffusion process over complex networks. The first model gives the probability of infection of a node in the network, given a particular source initiates the rumour. The second formulates the probability of infection of a node, as a function of its shortest distance to the origin of the rumours. We then presented an algorithm for estimating a single rumour source from sparse observations, when the activation time of the rumours is known. This algorithm leverages the distance-dependent infection likelihood in order to estimate the distances between all monitor nodes and the source. Based on these estimated shortest distances, triangulation is then used to build a set of potential sources. The most likely rumour origin is found using the former mathematical model of infection, which gives the infection likelihood, given a particular node started the rumour. Furthermore, we have extended the single source estimation method to the case of unknown rumour start time. Finally, we evaluated the proposed algorithms in small-world and scale-free networks, as well as in real-world graphs, and results showed that the probability of correctly estimating the rumour source is high, even when the set of observations is small.

REFERENCES

[1] Jiang, J., Wen, S., Yu, S., Xiang, Y., and Zhou, W., "Identifying Propagation Sources in Networks: State-of-the-Art and Comparative Studies," *IEEE Communications Surveys and Tutorials* **19**(1), 465–481 (2017).

[2] Shah, D. and Zaman, T., "Rumor Centrality: A Universal Source Detector," *SIGMETRICS Performance Evaluation Review* **40**, 199–210 (June 2012).

[3] Luo, W., Tay, W. P., and Leng, M., "Identifying Infection Sources and Regions in Large Networks," *IEEE Transactions on Signal Processing* **61**, 2850–2865 (June 2013).

[4] Karamchandani, N. and Franceschetti, M., "Rumor source detection under probabilistic sampling," 2184–2188, IEEE International Symposium on Information Theory (ISIT), Istanbul, Turkey (2013).

[5] Nguyen, D. T., Nguyen, N. P., and Thai, M. T., "Sources of misinformation in Online Social Networks: Who to suspect?," 1–6, Proc. IEEE Military Communications Conference (MILCOM), Orlando, FL, USA (2012).

[6] Zheng, L. and Tan, C. W., "A probabilistic characterization of the rumor graph boundary in rumor source detection," in [*2015 IEEE International Conference on Digital Signal Processing (DSP)*], 765–769 (July 2015).

[7] Tang, W., Ji, F., and Tay, W. P., "Multiple sources identification in networks with partial timestamps," in [*2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*], 638–642 (Nov 2017).

[8] Zejnilovi, S., Gomes, J., and Sinopoli, B., "Sequential source localization on graphs: A case study of cholera outbreak," in [*2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*], 1010–1014 (Nov 2017).

[9] Agaskar, A. and Lu, Y. M., "A fast Monte Carlo algorithm for source localization on graphs," SPIE Optical Engineering and Applications, San Diego, CA, USA (2013).

[10] B. A. Prakash, J. V. and Faloutsos, C., "Spotting Culprits in Epidemics: How many and Which ones?," 11–20, Proc. IEEE 12th International Conference on Data Mining (ICDM), Brussels, Belgium (2012).

[11] Luo, W. and Tay, W. P., "Identifying multiple infection sources in a network," in [*2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*], 1483–1489 (Nov 2012).

[12] Fioriti, V. and Chinnici, M., "Predicting the sources of an outbreak with a spectral technique," *Applied Mathematical Sciences* **9**(135), 6775–6782 (2014).

[13] Shah, D. and Zaman, T., "Detecting Sources of Computer Viruses in Networks: Theory and Experiment," 203–214, Proc. ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, New York, NY, USA (Dec. 2010).

[14] Shah, D. and Zaman, T., "Rumors in a Network: Who's the Culprit?," *IEEE Transactions on Information Theory* **57**, 5163–5181 (Aug. 2011).

[15] Wang, Z., Dong, W., Zhang, W., and Tan, C. W., "Rumor source detection with multiple observations: Fundamental limits and algorithms," in [*The 2014 ACM International Conference on Measurement and Modeling of Computer Systems*], *SIGMETRICS '14*, 1–13, ACM, New York, NY, USA (2014).

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

[17] Zhu, K. and Ying, L., "Information source detection in the sir model: A sample-path-based approach," *IEEE/ACM Transactions on Networking* **24**, 408–421 (Feb 2016).

[18] Luo, W. and Tay, W. P., "Finding an infection source under the sis model," in [*2013 IEEE International Conference on Acoustics, Speech and Signal Processing*], 2930–2934 (May 2013).

[19] Luo, W., Tay, W. P., and Leng, M., "How to identify an infection source with limited observations," *IEEE Journal of Selected Topics in Signal Processing* **8**, 586–597 (Aug 2014).

[20] Zhu, K. and Ying, L., "A robust information source estimator with sparse observations," in [*IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*], 2211–2219 (April 2014).

[21] Lokhov, A. Y., Mézard, M., Ohta, H., and Zdeborová, L., "Inferring the origin of an epidemic with a dynamic message-passing algorithm," **90**, 012801 (Jul 2014).

[22] Brockmann, D. and Helbing, D., "The hidden geometry of complex, network-driven contagion phenomena," *Science* **342**, 1337–1342 (2013).

[23] Wang, Z., Zhang, W., and Tan, C. W., "On inferring rumor source for sis model under multiple observations," in [*2015 IEEE International Conference on Digital Signal Processing (DSP)*], 755–759 (July 2015).

[24] Alexandru, R. and Dragotti, P. L., "Rumour source detection in social networks using partial observations," in [*2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*], 730–734 (Nov 2018).

[25] Watts, D. J. and Strogatz, S. H., "Collective dynamics of small-world networks," *Nature* **393**, 440–442 (1998).

[26] Barabsi, A.-L., Ravasz, E., and Vicsek, T., "Deterministic scale-free networks," *Physica A: Statistical Mechanics and its Applications* **299**(3), 559 – 564 (2001).

[27] Leskovec, J. and Krevl, A., "SNAP Datasets: Stanford large network dataset collection." http://snap.stanford.edu/data (June 2014).

[28] Dijkstra, E. W., "A note on two problems in connexion with graphs," *Numerische Mathematik* **1**, 269–271 (Dec 1959).

[29] Bishop, C. M., [*Pattern Recognition and Machine Learning (Information Science and Statistics)*], Springer-Verlag, Berlin, Heidelberg (2006).