Statistical inference framework for source detection of contagion processes on arbitrary network structures

Nino Antulov-Fantulin[a,*], Alen Lančić[b], Hrvoje Štefančić[c,d], Mile Šikić[e,f], Tomislav Šmuc[a]

[a]Computational Biology and Bioinformatics Group, Division of Electronics,
Rudjer Bošković Institute, Zagreb, Croatia
[b]Faculty of Science, Department of Mathematics, University of Zagreb, Zagreb, Croatia
[c]Theoretical Physics Division, Rudjer Bošković Institute, Zagreb, Croatia
[d]Catholic University of Croatia, Zagreb, Croatia
[e]Faculty of Electrical Engineering and Computing, Department of Electronic Systems and Information Processing,
University of Zagreb, Croatia
*[f]Bioinformatics Institute, A*STAR, Singapore, Republic of Singapore*

Abstract

In this paper we introduce a statistical inference framework for estimating the contagion source from a partially observed contagion spreading process on an arbitrary network structure. The framework is based on a maximum likelihood estimation of a partial epidemic realization and involves large scale simulation of contagion spreading processes from the set of potential source locations. We present a number of different likelihood estimators that are used to determine the conditional probabilities associated to observing partial epidemic realization with particular source location candidates. This statistical inference framework is also applicable for arbitrary compartment contagion spreading processes on networks. We compare estimation accuracy of these approaches in a number of computational experiments performed with the SIR (susceptible-infected-recovered), SI (susceptible-infected) and ISS (ignorant-spreading-stifler) contagion spreading models on synthetic and real-world complex networks.

The structure of vast majority of biological networks (biochemical, ecological), technological networks (internet, transportation, power grids), social networks and information networks (citation, WWW) can be represented by complex networks [16], [7], [3]. Epidemic or contagion processes are amongst the most prevalent type of dynamic processes of interest characteristic for these real-life complex networks and they include disease epidemics, computer virus spreading, information and rumor propagation [23]. Different mathematical frameworks have been used to study epidemic spreading. We can divide them into two major categories based upon assumptions they make: the homogeneous mixing framework and the heterogeneous mixing framework. The homogeneous mixing framework assumes that all individuals in a population have an equal probability of contact. This is a traditional mathematical framework [12], [10] in which differential equations are used to model epidemic dynamics. The heterogeneous mixing framework assumes

[*]Corresponding author. Adress: Rudjer Bošković Institute, Bijenička cesta 54, 10000 Zagreb

Email addresses: `nino.antulov@irb.hr` (Nino Antulov-Fantulin), `alen@student.math.hr` (Alen Lančić), `shrvoje@thphys.irb.hr` (Hrvoje Štefančić), `mile.sikic@fer.hr` (Mile Šikić), `tomislav.smuc@irb.hr` (Tomislav Šmuc)

April 2, 2013

that properties of contact interactions among individuals are defined via some underlying network structure. The small world network property [24] and the scale-free network property [2] [8] have a great impact on the outcome of an epidemic spreading. We can further divide the heterogeneous mixing framework by other assumptions: the bond percolation, the mean-field and the particle network frameworks. The bond percolation approach applies the percolation theory to describe epidemic processes on networks [14], [11]. The mean-field approach assumes that all nodes having the same degree with respect to an epidemic process are statistically equivalent [4], [17]. The particle network approach assumes that spreading process is characterized by particles which diffuse along edges on a transportation network and each node contains some non-negative integer number of particles (reaction-diffusion processes).

The main question we address in this work is: Is it possible to detect location of the initial source from partial information on the contagion spread over a network structure ? This research question is useful for many realistic scenarios in which we observe epidemic spread at certain temporal moment and would like to infer the source location (patient-zero). Our statistical inference framework is applicable for arbitrary compartment contagion spreading model on arbitrary network structure. We have based our main case study on the SIR (susceptible-infected-recovered) model but we have demonstrated the applicability of inference framework on other contagion processes like the SI (susceptible-infected) and the ISS (ignorant-spreading-stifler) model. The SIR model [12] is an adequate model for many contagious processes like disease modelling, virus propagation [20] or rumour propagation [15]. We base our inference study on rather general assumptions which can be relaxed: (i) that observed partial epidemic realization is defined by complete knowledge of infected and recovered nodes (ii) that probabilities for infection and recovery of the underlying epidemic process are known in advance, as is the time from the start of the epidemic. We empirically demonstrate inference performance of the framework on different types of networks and for different contagion properties. We also investigate the impact on the performance of the framework in case when the assumptions are relaxed i.e. not complete knowledge on network status and when contagion parameters and time are uncertain. Finally, we demonstrate generality of the approach through solving source detection problem for different compartment models (SIR, SI and ISS).

Recently, the problem of estimating the initial source has gained a lot of attention due to its importance and practical aspects. Under different assumption on network structures or spreading process different source estimators have been developed [21],[6],[25],[18],[13]. However, we have made a significant contribution in problems of source detection for more general spreading processes on arbitrary network structures. In this work we cast this problem into a statistical inference framework based on the maximum likelihood estimation of the source of observed epidemic realization. This inference framework relies on a large scale simulation of contagion spreading processes from the set of potential source locations and subgraph similarity measures.

In section 1 we describe the SIR compartment model. Section 2 we describe our statistical inference framework and define different maximum likelihood estimators and subgraph similarity measures used to infer conditional probability of epidemic realizations from particular source locations. In section 3 we describe experiments that demonstrate network, contagion dynamics and noise effects and section 4 explains the related work.

Notation

Notation	Description
G	is a network with a set of nodes V and a set of edges E
Θ	general variable which identifies source nodes
θ	specific value for Θ variable, example: $\Theta = \theta_i$ the source node is the node i
p	probability of infection in one discrete time step
q	probability of recovery in one discrete time step
n	number of simulations for a specific SIR process
T	temporal threshold (random variable or constant)
$\vec{R}$	epidemic random vector $\vec{R} = (R(1), R(2), .., R(k))$
$R(i)$	Bernoulli indicator random variable for node i
$\vec{r}$	epidemic realization, example $\vec{r}_1 = (1, 0, 0, 1, ..., 1)$
$\vec{r}_*$	observed epidemic realization
$\vec{r}(i)$	i-th component of the realization vector $\vec{r}$, example $\vec{r} = (1, 0, 1, 1)$, $\vec{r}(2) = 0$
$\vec{R}_\theta$	random vector for realizations from node θ
$\vec{R}_{\theta,i}$	i-th sample realization vector from random vector $\vec{R}_\theta$
S	set of potential sources
$\varphi(\vec{r}_1, \vec{r}_2)$	similarity measure between two realizations $\vec{r}_1, \vec{r}_2$
$\varphi(\vec{r}_*, \vec{R}_\theta)$	random variable which measures the similarity between realization $\vec{r}_*$ and realizations from random vector $\vec{R}_\theta$
$\psi_\oplus(m_1, m_2)$	a bitwise XNOR function
$\psi_\vee(m_1, m_2)$	a bitwise OR function
$\psi_\wedge(m_1, m_2)$	a bitwise AND function
$\varphi_x(\vec{r}_1, \vec{r}_2)$	the similarity calculated with $\overline{XNOR}(\vec{r}_1, \vec{r}_2)$ function
$\varphi_J(\vec{r}_1, \vec{r}_2)$	the similarity calculated with $Jaccard(\vec{r}_1, \vec{r}_2)$ function
$\delta(x)$	the Dirac delta function

1. SIR compartment model

We define the contact-network as an undirected and non-weighted graph $G(V, E)$ (V-set of nodes or vertices, E-set of links). A link (u, v) exists only if two nodes u and v are in contact during the epidemic time. We also assume that the contact-network during the epidemic process is a static one. To simulate epidemic propagation through a contact-network, we use the standard stochastic SIR model. In this model each node at some time can be in one of the following states: susceptible (S), infected (I) and recovered (R). The spreading process is simulated using discrete time step model.

The SIR epidemic process is a stochastic process, which is simulated with n mutually independent simulation steps on the contact network G. At the beginning of each epidemic simulation all nodes from graph G are in the susceptible state except set of nodes which are initially infected. We assume in our treatment that epidemic parameters p and q are predefined, constant and known beforehand. The epidemic parameter p is the probability that an infected node u infects an adjacent susceptible node v in one discrete time step. The epidemic parameter q is the probability that an infected node recovers in one discrete time step. Set of initially infected nodes is denoted with the letter Θ. At the end of one full epidemic simulation, all nodes can be in one of two following states: susceptible or recovered. In our treatment however, we will limit epidemic spreading to a predefined number of discrete time steps, which basically means that we will deal with partially realized epidemic spreads and that this number of steps is also known parameter in the inference procedure for the source location estimation.

2. Statistical inference on epidemic propagation realizations

In this section we formulate the problem of the source localization in the network and develop related statistical inference framework.

Epidemic source location problem

Let us define the random vector $\vec{R} = (R(1), R(2), ..., R(N))$, that indicates which nodes got infected prior up to some predefined temporal threshold T (random variable or constant). The random variable $R(i)$ is a Bernoulli random variable, which assigns the value 1 if node i got infected before time T from the start of the epidemic process and the value 0 otherwise.

Let us assume that we have observed one spatio-temporal epidemic propagation realization $\vec{r}$ of SIR process defined by (p, q) and T, and we want to infer which nodes from the set S are the most likely source of realization $\vec{r}$ for the SIR process (p, q) and T. $S = \{\theta_1, \theta_2, ..., \theta_N\}$ is the finite set of possible source nodes that is defined by observed infected or infected and recovered set of nodes prior to moment T in the network.

In order to find a node or a small subset of infected nodes that have highest likelihood for being the source of the epidemic spread, we pose the following maximum likelihood problem.

$$\hat{\Theta} = \arg\max_{\Theta \in S} P(\Theta | \vec{R} = \vec{r}),$$

where $\Theta \in S$ is a set of all possible sources of epidemic.
By applying Bayes theorem, we get the following expression:

$$\hat{\Theta} = \arg\max_{\Theta \in S} \frac{P(\vec{R} = \vec{r} | \Theta = \theta) P(\Theta = \theta)}{\sum_{\theta_k} P(\vec{R} = \vec{r} | \Theta = \theta_k) P(\Theta = \theta_k)}.$$

If all Θ (apriori) are equally likely, this is equivalent to:

$$\hat{\Theta} = \arg\max_{\Theta \in S} P(\vec{R} = \vec{r} | \Theta = \theta).$$

Thus, the core of source location estimation problem is the determination of the likelihood of the observed epidemic realization being initiated at the source location Θ. We now proceed with description of the algorithms for determining the maximum likelihood for the observed epidemic realization.

2.1. The Maximum Likelihood source estimator

First, we give a pseudo-code (Algorithm 1) for the original problem of the maximum likelihood source estimation, where source can be any node from set S. In principle, this treatment can be extended to problem of multiple sources determination, but the necessary extensions are out of the scope of this work. Note that among algorithm parameters $(G, p, q, \vec{r}_*, T, S, n)$ the parameter n represents number of random simulations from a single candidate for the epidemic source node. In our framework, the number of random simulations n is very important from the perspective of the accuracy/stability of results and it is also a major determinant of the running time of the estimation procedure.

Algorithm 1 The Maximum Likelihood source estimator algorithm: $(G, p, q, \vec{r}_*, T, S, n)$

```
**Input:** Network structure $G$, SIR process parameters $(p, q)$, $S = \{\theta_1, \theta_2, ..., \theta_N\}$ a set of possible sources $\theta_i$, observed realization $\vec{r}_*$ ending at some temporal threshold $T$ , $n$ a number of simulations
**for** each $\theta_j \in S$ (apriori set of possible sources of epidemic) **do**
  Call likelihood estimation function $(G, p, q, \vec{r}_*, T, n)$
  Save $\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta_j)$
**end for**
**Output 1:** $\theta_k$ with maximum likelihood $\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta_k)$
**Output 2:** Ranked sources in $S = \{\theta_1, \theta_2, ..., \theta_N\}$ according to likelihoods $\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta_k)$
```

It is obvious that the Algorithm 2.1 is just a wrapper code that calls likelihood estimation function for each potential source of epidemics. We now proceed with the description of different algorithms for calculating the likelihood $P(\vec{R} = \vec{r} | \Theta = \theta)$.

2.2. Realization similarity matching

Let us define the function $\varphi(\vec{r_1}, \vec{r_2})$, which measures the similarity between two epidemic realizations or subgraphs of the underlying network: $\vec{r_1}$ and $\vec{r_2}$.

We first define new random variable $\varphi(\vec{r_*}, \vec{r_\theta})$, which measures the φ similarity between the fixed realization $\vec{r_*}$ and random realization that comes from SIR process with the source θ. We can calculate the unbiased estimator of the following cumulative distribution function as the empirical distribution function:

$$\hat{F}(x) = \hat{P}(\varphi(\vec{r_*}, \vec{R_\theta}) \leq x) = \frac{\sum_{i=1}^{n} \mathbf{1}_{[0,x\rangle} \left(\varphi(\vec{r_*}, \vec{R}_{\theta,i}) \right)}{n},$$

where $\mathbf{1}_{[0,x\rangle}$ is a characteristic function defined as:

$$\mathbf{1}_{[0,x\rangle}(y) = \begin{cases} 1 & : y \in [0, x\rangle, \\ 0 & : else. \end{cases}$$

Then, its probability density function is calculated like this:

$$PDF(x) = \frac{d}{dx}\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta \left(x - \varphi(\vec{r_*}, \vec{R}_{\theta,i}) \right),$$

where $\delta(x)$ is the Dirac delta function.
Central limit theorem states that pointwise, $\hat{F}(x)$ has asymptotically normal distribution. The rate at which this convergence happens is bounded by Berry–Esseen theorem. This implies that the rate of convergence is bounded by $O(1/\sqrt{n})$, where n is the number of random simulations.

Next, we define two measures ($XNOR$ and Jaccard) that are used to determine the similarity φ. The first one is a binary NOT XOR function or $XNOR(\vec{r_1}, \vec{r_2})$ counts the number of corresponding non-infected and infected nodes in realizations $\vec{r_1}$ and $\vec{r_2}$:

$$XNOR(\vec{r_1}, \vec{r_2}) = \sum_{k \in V} \psi_{\oplus}(\vec{r_1}(k), \vec{r_2}(k)),$$

,where $\psi_{\oplus}(m_1, m_2)$ function is defined as:

$$\psi_{\oplus}(m_1, m_2) = \begin{cases} 1 & : (m_1 = 1 \text{ and } m_2 = 1) \text{ or } (m_1 = 0 \text{ and } m_2 = 0)\ , \\ 0 & : \text{else.} \end{cases}$$

In other words, $\psi(m_1, m_2)$ is equal to one only if two nodes were infected or they did not get infected prior to temporal threshold T. We also define function: $\overline{XNOR}(\vec{r_1}, \vec{r_2})$, which is normalized $XNOR$ function over total number of nodes: $XNOR(\vec{r_1}, \vec{r_2}) * N^{-1}$.

The second similarity measure is a well known Jaccard measure, which in our case counts the number of corresponding infected nodes in $\vec{r_1}$ and in $\vec{r_2}$ normalized by the number of corresponding infected nodes in $\vec{r_1}$ or in $\vec{r_2}$.

$$Jaccard(\vec{r_1}, \vec{r_2}) = \frac{|\vec{r_1} \wedge \vec{r_2}|}{|\vec{r_1} \vee \vec{r_2}|} = \frac{\sum_{k \in V} \psi_{\wedge}(\vec{r_1}(k), \vec{r_2}(k))}{\sum_{k \in V} \psi_{\vee}(\vec{r_1}(k), \vec{r_2}(k))},$$

where $\psi_{\wedge}(m_1, m_2)$ and $\psi_{\vee}(m_1, m_2)$ functions are defined as:

$$\psi_{\wedge}(m_1, m_2) = \begin{cases} 1 & : (m_1 = 1 \text{ and } m_2 = 1), \\ 0 & : \text{else} \end{cases}$$

and where $\psi_{\vee}(m_1, m_2)$ function is defined as:

$$\psi_{\vee}(m_1, m_2) = \begin{cases} 1 & : (m_1 = 1 \text{ or } m_2 = 1)\ , \\ 0 & : \text{else.} \end{cases}$$

In the following text the $\varphi_x(\vec{r_1}, \vec{r_2})$ will denote the similarity calculated with $\overline{XNOR}(\vec{r_1}, \vec{r_2})$ function and $\varphi_J(\vec{r_1}, \vec{r_2})$ will denote the similarity calculated with $Jaccard(\vec{r_1}, \vec{r_2})$ function. In order to speed the similarity matching between realizations, we use the bitwise operations (XOR, NOT, AND) and bit count with Biran-Kernignan method.

2.3. Likelihood estimation functions

In this section we define three variants of likelihood estimation functions: AUCDF, AvgTopK, and Naive Bayes. First two functions, AUCDF and AvgTopK can use any of the similarity measures defined above, while Naive Bayes produces likelihood based on its own similarity measure.

As a first likelihood estimation function we define AUCDF (Area Under Cumulative Distribution Function) (see Algorithm 2), which can use any of the similarity measures defined above.

Algorithm 2 AUCDF estimation function $(G, p, q, \vec{r}_*, T, \theta, n)$

```
Input: $G$ - network structure , $(p, q)$ - SIR process parameters , $\vec{r}_*$ - observed realization prior to some temporal threshold $T$, $\theta$ - source for which likelihood is calculated, $n$ a number of simulations
for $i = 1$ to $n$ (number of simulations) do
    - Run SIR simulation $(p, q)$ with $\Theta = \theta$ and obtain epidemic realization $\vec{R}_{\theta,i}$, ending at the temporal threshold $T$;
    - Calculate and save $\varphi(\vec{r}_*, \vec{R}_{\theta,i})$ ;
end for
- Calculate empirical distribution function:
    $\hat{P}(\varphi(\vec{r}_*, \vec{R}_\theta) \leq x) = \frac{\sum_{i=1}^{n} \mathbf{1}_{[0,x\rangle}(\varphi(\vec{r}_*, \vec{R}_{\theta,i}))}{n}$
- Estimate likelihood using the area under the empirical cumulative distribution:
    $AUCDF_\theta = \int_0^1 \hat{P}(\varphi(\vec{r}_*, \vec{R}_\theta) \leq x) \mathrm{d}x$
Output: $\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta) = 1 - AUCDF_\theta$ likelihood for $\theta$;
```

Different sources θ produce different empirical cumulative distributions of similarities to $\vec{r}_*$. If we compare two empirical distribution functions CDF_1 and CDF_2 from two different sources θ_1 and θ_2 and if the $AUCDF_1 < AUCDF_2$ then sample of realizations from θ_1 source are more similar to fixed realization $\vec{r}_*$ than the sample realizations from θ_2 source. This is the primary reason, why we use value $1 - AUCDF$ to estimate source likelihood $\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta)$.

Algorithm AvgTopK represents a variant of the previous estimation function, which uses only k highest values from the tail of the probability density function of the random variable $\varphi(\vec{r}_*, \vec{r}_\theta)$:

$$PDF(x) = \frac{d}{dx}\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta\left(x - \varphi(\vec{r}_*, \vec{R}_{\theta,i})\right).$$

Algorithm 3 AvgTopK likelihood estimation function $(G, p, q, \vec{r}_*, T, \theta, n)$

```
Input: $G$ - network structure , $(p, q)$ - SIR process parameters , $\vec{r}_*$ - observed realization prior to some temporal threshold $T$, $\theta$ - source for which likelihood is calculated, $n$ a number of simulations
for $i = 1$ to $n$ (number of simulations) do
    - Run SIR simulation $(p, q)$ with $\Theta = \theta$ and obtain epidemic realization $\vec{R}_{\theta,i}$, ending at the temporal threshold $T$;
    - Calculate and save $\varphi(\vec{r}_*, \vec{R}_{\theta,i})$ ;
end for
- Sort the scores $\left\{\varphi(\vec{r}_*, \vec{R}_{\theta,i})\right\}$ in descending order;
- Average top $k$ highest scores:
    $$\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta) = \frac{1}{k} \sum_{i=1}^{k} \left\{\varphi(\vec{r}_*, \vec{R}_{\theta,i}))\right\}_{sorted}$$
Output: $\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta)$ likelihood for $\theta$;
```

In each simulation we calculate how similar $\vec{R}_{\theta,i}$ realization to observed $\vec{r}_*$ realization is by using φ function. Estimate $\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta)$ is the average score over top k highest similarities $\varphi(\vec{r}_*, \vec{R}_{\theta,i})$ in n simulations (tail of pdf).

Finally, we propose the third likelihood estimation function which is based on node probabilities for being infected from a particular source node. Main assumption of this approach is independence between nodes with respect to epidemic spreading.

The conditional probability that the node k in realization $\vec{r}_*$ is infected from source θ is:

$$\hat{P}(\vec{r}_*(k) = 1 | \Theta = \theta) = \frac{m_k + \epsilon}{n + \epsilon}, \forall k \in G,$$

where m_k is the number of times that node k got infected from the total of n simulations $SIR(p, q)$ from source node θ and ϵ is a smoothing factor. Smoothing factor ϵ is necessary to mitigate the problem of zero values, stemming from the finite number of simulations used to calculate $\hat{P}(\vec{r}_*(k) = 1 | \Theta = \theta)$.

Then we define the estimator for the likelihood of observing realization $\vec{r_*}$ from source node θ as:

$$\hat{P}(\vec{R} = \vec{r_*} | \Theta = \theta) = \prod_{\{k : \vec{r_*}(k) = 1\}} \hat{P}(\vec{r}_*(k) = 1 | \Theta = \theta) \prod_{\{j : \vec{r_*}(j) = 0\}} (1 - \hat{P}(\vec{r}_*(j) | \Theta = \theta)).$$

This equation uses the probability estimates that nodes $\{k : \vec{r_*}(k) = 1\}$ from realization $\vec{r_*}$ got infected and the probability estimates that nodes $\{j : \vec{r_*}(j) = 0\}$ from realization $\vec{r_*}$ did not get infected from source node θ.

In mathematical sense, probability of finding an infected node k at time t is dependent on other infected nodes prior to time t. Nevertheless, we use the same assumption of independence to estimate the rank of potential sources. There is obvious resemblance between this approach and the well known studied probabilistic classifier - Naive Bayes. Although Naive Bayes uses a strong assumption of independence, it has been shown that in practice its performance is comparable to more complex probabilistic classifiers [9].

In order to have more stable numerical likelihood estimations, we used the log likelihood variant for estimating $\hat{P}(\vec{R} = \vec{r}_*|\Theta = \theta))$ (see Algorithm 3).

Algorithm 4 Naive Bayes likelihood estimation function $(G, p, q, \vec{r}_*, T, \theta, n)$

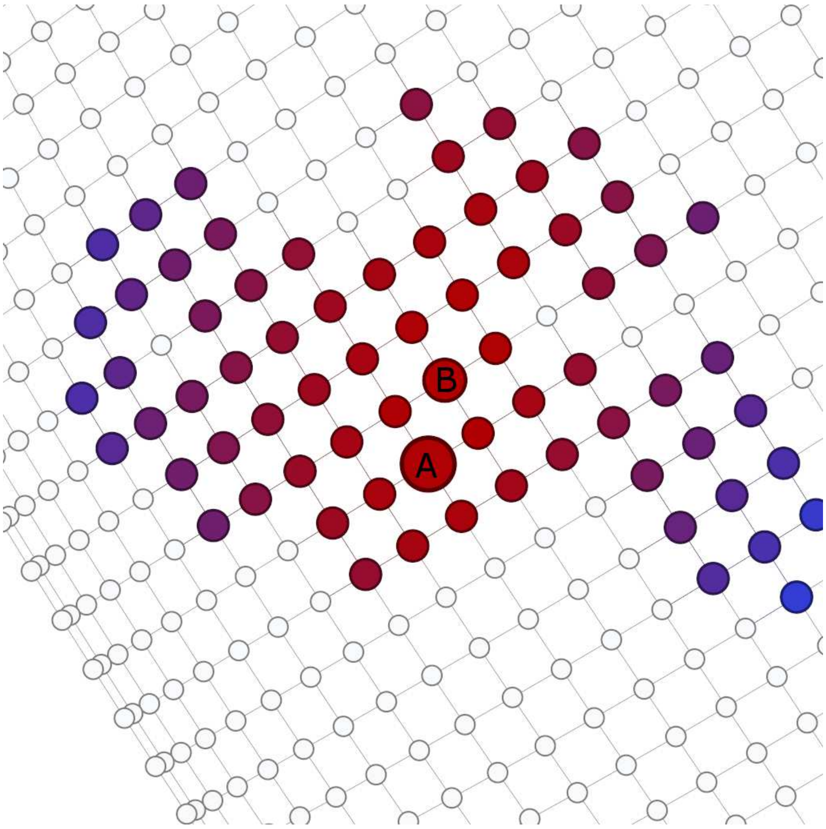
```
**Input:** $G$ - network structure , $(p, q)$ - SIR process parameters , $\vec{r}_*$ - observed realization prior to some temporal threshold $T$, $\theta$ - initial source for which likelihood is calculated
- $m_k = 0 : \forall k \in V$ from $G$;
**for** $i = 1$ to $n$ (number of simulations) **do**
    - Run SIR simulation $(p, q)$ with $\Theta = \theta$ and obtain realization $\vec{R}_{\theta,i}$ prior to the temporal threshold $T$;
    - Update: $m_k = m_k + 1$; $\forall k$ which were infected in $\vec{R}_{\theta,i}$;
**end for**
- Calculate:
    $\hat{P}(\vec{r}_*(k) = 1|\Theta = \theta) = \frac{m_k + \epsilon}{n + \epsilon}, \forall k \in G$
- Calculate log likelihood: $log(\hat{P}(\vec{R} = \vec{r}_*|\Theta = \theta)) =$
    $= \sum_{\{k:\vec{r}_*(k)=1\}} log(\hat{P}(\vec{r}_*(k) = 1|\Theta = \theta)) + \sum_{\{j:\vec{r}_*(j)=0\}} log(1 - \hat{P}(\vec{r}_*(j)|\Theta = \theta));$
**Output:** $log(\hat{P}(\vec{R} = \vec{r}_*|\Theta = \theta))$ likelihood for $\theta$;
```

3. Epidemic source location experiments

In this section, we describe the experiments along with the obtained results performed on different networks and in different epidemic settings. The experiments were designed to illustrate the overall predictability properties of the source detection problem with the introduced inference framework and compare the performances of individual algorithms.

We test the performance of source likelihood estimation algorithms on single source epidemic detection problems. In our experiments we observe one spatio-temporal epidemic propagation realization $\vec{r}_*$ and we want to infer the potential source of realization from the set S. In Figure 1, we illustrate one epidemic realization on a synthetic grid, where the color gradient from blue to red represents estimated source likelihood (blue - lower and red - higher). We have used a Naive SIR algorithm implementation [1] as efficient SIR process simulation on network structures.

Figure 1: One epidemic realization of SIR process ($p = 0.3, q = 0.7$) on a synthetic grid, where the color gradient from blue to red represents estimated source likelihood (blue - lower and red - higher). Node with the letter ”A” represents true source of epidemic realization and the node with the letter ”B” represents the Maximum Likelihood source estimate by the ”Naive Bayes” likelihood estimation function



Due to the strong stochastic nature of epidemic process, frequency of correct estimations of the source location in the network is not the best measure to test the predictability of algorithms. The topological distance of maximum likelihood node from true source can be a misleading low even for random estimations on networks with low average shortest path. Therefore, we measure the rank of true source in the output list of potential sources from set S in experiments on different network structures. The overall testing procedure is given in the following pseudo-code.

Let us assume that in some source location detection experiment we get realization $\vec{r}_*$ that has k infected nodes. We rank the nodes θ_i in a list of k potential nodes according to the likelihood $\hat{P}(\vec{R} = \vec{r}_*|\Theta = \theta_i)$. We express the rank of real source as a relative source rank, i.e. the rank of the true source node normalized to the list size (for example, if the rank of the true source node is at the position 10 in the list of 100 potential sources, then the relative source rank is 0.1). For the performed batch of experiments, we calculate cumulative source rank probability distribution, which tells us the probability that the relative rank of the source node is lower or equal to some specified value. By its nature cumulative source rank is very similar to the well

Algorithm 5 Source location experiments

```
for experiment = 1 to total number of experiments do
    - Sample random initial source $\theta_*$ from network $G$
    - Obtain realization $\vec{r}_*$ from SIR process $SIR(p, q, \Theta = \theta_*, T)$ that has at least 0.01 infected nodes in total network
    - Create set $S$ as the set of all nodes which were infected in realization $\vec{r}_*$
    - Call the Maximum Likelihood source estimator algorithm $(G, p, q, \vec{r}_*, T, S)$
    - Measure the rank of true source on ranked likelihood list of $S$
end for
```

known receiver operating characteristic (ROC), a measure frequently used in signal detection and machine learning for measuring the performance of classifier systems. Ideal estimator or classifier would have area under the cumulative source rank equal to one, exactly as in the case of ROC measure (AUC measure represents the area under the ROC curve). One can argue that other measures might have been appropriate as well, for example the distance of the maximum likelihood node to the true source node in a network. We opted for cumulative source rank, because it is a more versatile measure, due to its invariance to network size size and structure (e.g. for networks with different average shortest paths one would get grossly different results).

The influence of network structure on source localization performance has been tested on the following classes of networks: regular grid (figure 2) and lattice (figure 8 part A), Small-World networks (figure 8 part B), Erdös-Rényi networks (figure 8 part C), Albert-Barabasi network (figure 3 part A) and Western States Power Grid of the United States [24] (figure 3 part B).

In order to measure the performance of source localization we have done the following experiments:

- Comparison of different estimators: we compare performance of different algorithms for different epidemic conditions,
- Network structure experiments: this set of experiments illustrates the effects of network structure on the prediction performance over diverse network topologies,
- Process dynamics experiments: here we observe the effects of different process parameters like (p, q, T) on source localization performance and
- Uncertainty experiments: Performance degradation associated with uncertain epidemic parameters or incomplete knowledge about epidemic realization.

Comparison of different estimators

In Figures 4,5 we can see the results of the source location detection experiment for different likelihood estimation functions (AUCDF, AvgTopK and Naive Bayes) on different network structures. The cumulative probability function in these experiments measure the probability of ranking the true source at specific position. These results suggest that Naive Bayes and AvgTopK estimators have better performance than the AUCDF estimator. For instance, we can see that in Figure 5 the Naive Bayes estimator ranks the true source in approximately 80 % of experiments in top 10 % of the source list. We have also made a baseline solution which uses random likelihood estimation function to rank the potential sources (see Figures 4,5). Random likelihood

Figure 2: Visualization of regular grid of size $N = 30x30$

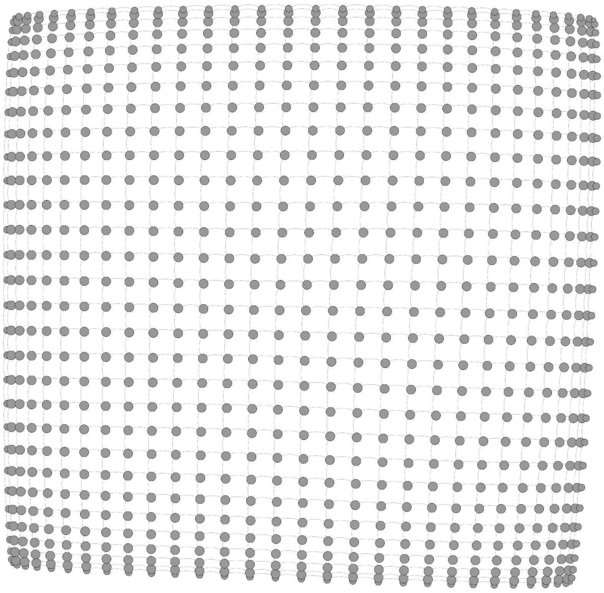
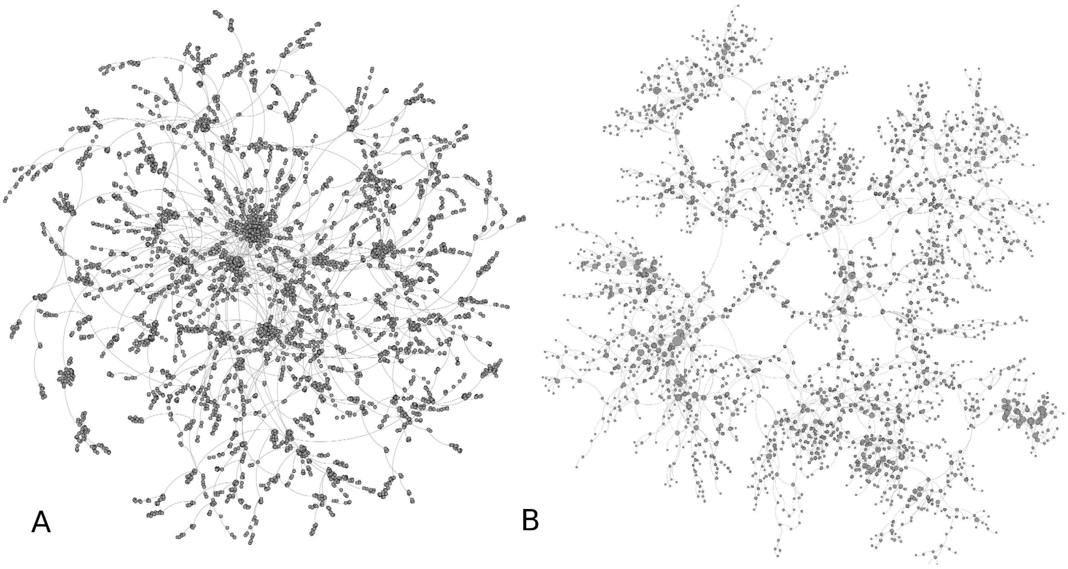


Figure 3: Visualization of Albert-Barabasi network (part A) of size $N = 5000$, with $m_0 = 5$ initial full connected core, and $m = 1$ added edges in preferential attachment. In part B: the visualization of power-grid network (Western States Power Grid of the United States [24]) of size $N = 4941$.



estimation function returns random uniform probability value $[0 - 1]$ for each node. Note, that the AvgTopK likelihood estimation function tends to give more accurate source localization performance than the Naive Bayes and AUCDF estimation functions. In our experiments we have used the top $k = 5\%$ of highest scores from pdf in AvgTopK likelihood estimation function.

Figure 4: Cumulative probability distribution of source relative rank based on 500 experiments with random initial source on synthetic grid $N = 30x30$ for $p = 0.3$, $q = 0.7$, $T = 10$ with different likelihood estimation functions.

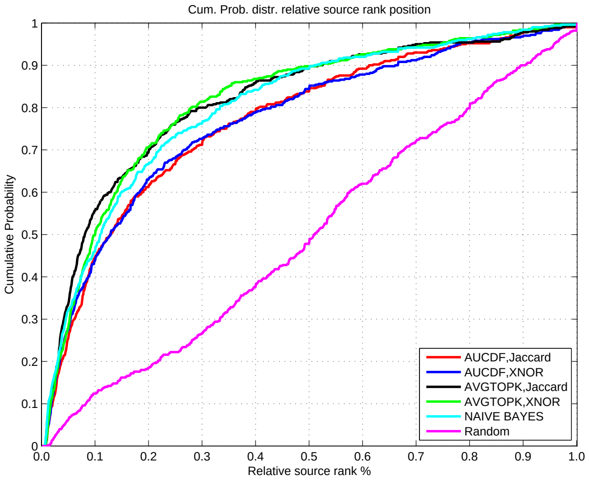


Figure 5: Cumulative probability distribution of source relative rank for 500 experiments with random initial source on power grid network of size $N = 4941$ for $p = 0.7$, $q = 0.6$, $T = 7$ and different likelihood estimation functions.

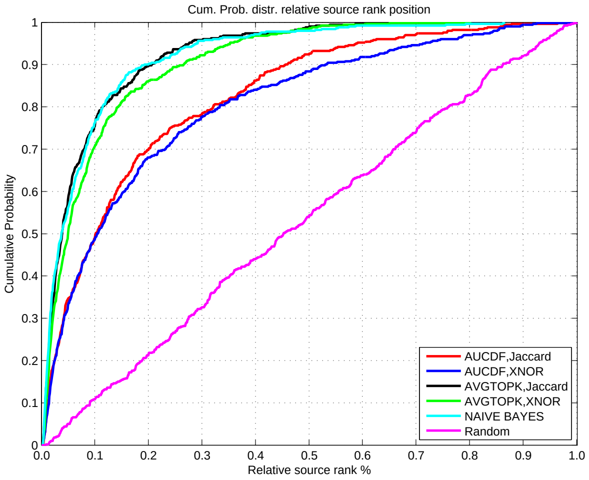
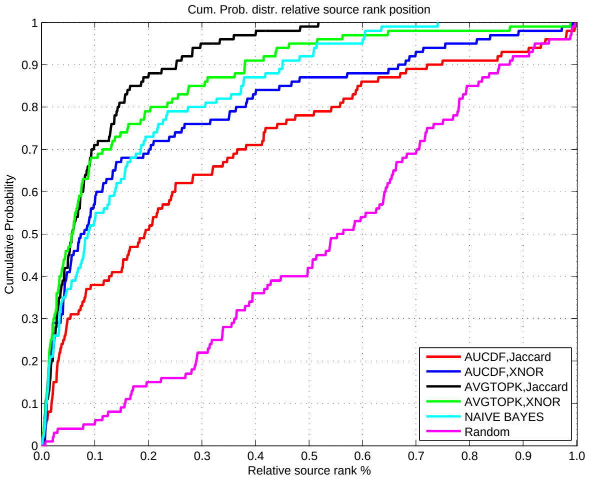


Figure 6: Cumulative probability distribution of source relative rank for 100 experiments with random initial source on Albert-Barabasi network ($N = 5000, M_0 = 5, m = 1$) for $p = 0.6$, $q = 0.2$, $T = 5$ and different likelihood estimation functions.



Network structure experiments

The effects of different network structures on source estimation performance is demonstrated with the following Small-World experiment. We are generating networks from regular lattice

($\beta = 0$) to random networks ($\beta = 1$) with Small-world networks in the middle and observing the performance of source estimators. We measure the area under the cumulative source rank function and observe that the performance of source estimator drops as the average shortest path of network decreases.

Figure 7: Source location aggregate performance value: area under the cumulative probability of relative source rank (AvgTopk estimator with $\varphi_X()$ similarity function) for 100 experiments on classes of networks (size: $N = 5000$) from regular lattice ($\beta = 0$) to random networks ($\beta = 1$) with Small-world networks in the middle. SIR process has parameters $p = 0.1$, $q = 0.8$ and $T = 7$. Average shortest path is normalized by average shortest path (≈ 120) in regular lattice. Average clustering coefficient is normalized by average clustering coefficient (≈ 0.7) in regular lattice.

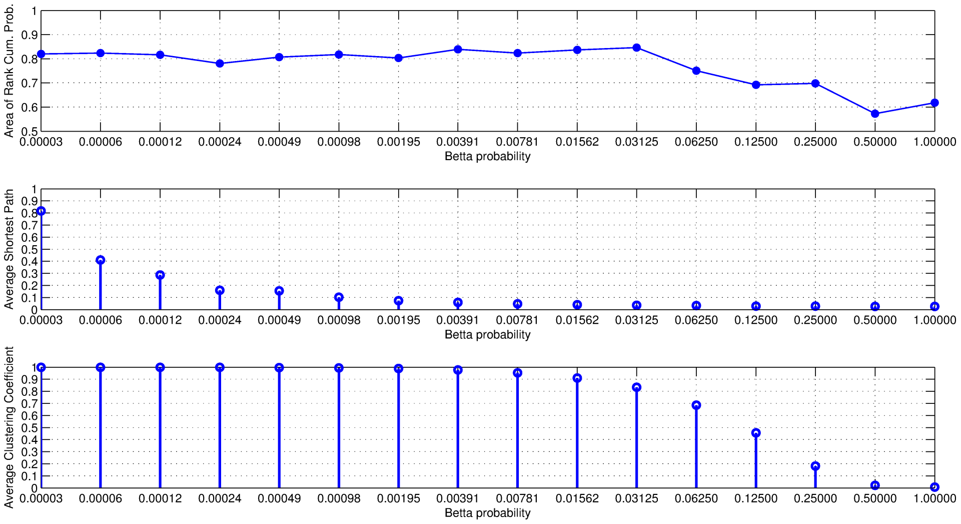
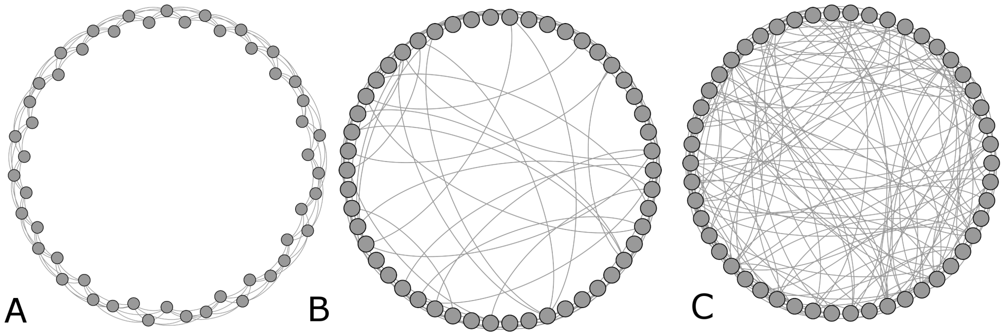


Figure 8: Classes of networks are generated according to the Watts-Strogatz small-world β model (size: $N = 5000$) from the regular lattice ($\beta = 0$ and 10 local edges) to random networks ($\beta = 1$) with small-world networks in the middle. Visualization is done on smaller networks (size: 50) from regular lattice to random networks (part C: $\beta = 1$) with Small-world networks in the middle (e.g. part B: $\beta = 0.1$).



Process dynamics experiments

Finally, we perform a set of experiments to put our source location inference framework into a perspective with recent models for diffusion-like processes published in the literature [21] [18]. We illustrate performance of our inference framework on diffusion like processes which can be understood as a limiting case of SIR process in which recovery parameter q is close to or equal zero. In Figure 9 and 10 we can observe that the performance of source estimation algorithms is highest in these conditions. This is expected behaviour which can be interpreted as a consequence of that initial conditions are preserved more in diffusion-like processes.

Figure 9: Cumulative probability of relative source rank for 100 experiments with random initial source on power-grid network ($N = 4941$) for different parameters q. Diffusion like processes are special case of SIR model where recovery parameter $q = 0$ (red line). Experiments were performed with AUCDF likelihood estimation function with φ_X similarity function

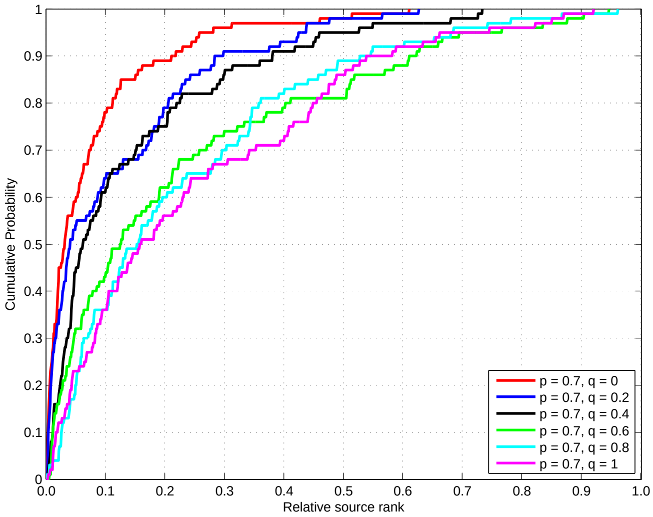
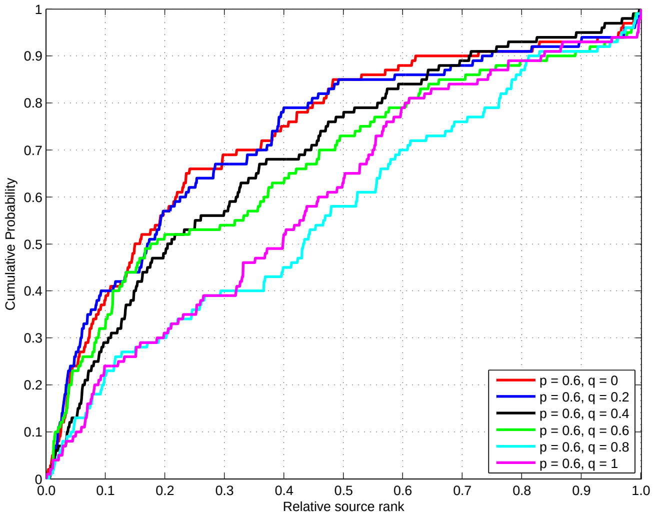


Figure 10: Cumulative probability of relative source rank for 100 experiments with random initial source on the Albert-Barabasi network ($N = 5000, M_0 = 5, m = 1$) for different parameters q. Diffusion like processes are special case of SIR model where recovery parameter $q = 0$ (red line). Experiments were performed with AUCDF likelihood estimation function with φ_J similarity function



Uncertainty experiments

Note that the previous experiments were performed on processes for which the parameters p, q and T were degenerative random variable i.e. constants. Now, we demonstrate the effects on performance when the exact values of p, q and T are sampled from probability distributions. We model the temporal threshold T as a random variable of the following form: $T = T_0 + \epsilon$, where ϵ represents the noise from some probability distribution. In Figure 11 we have made a series of experiments where the ϵ noise was modelled with the Geometric distribution with different parameters. As the variance of noise is increased, the performance of source localization is decreased. We have also made a series of experiments in which the parameters (p, q) were also modelled with the noise: $p = p_0 + \gamma$, $q = q_0 + \gamma$, where γ noise was distributed as a Normal distribution with parameters (μ, σ). In Figure 12, we observe that the performance of source location decreases as the noise of parameters p,q and T increases. This findings suggest that if the predictability is low for parameters p, q and T with no noise then predictability can only be lower when the noise is present. Furthermore, this implies certain limits of predictability for source localization on Small-World networks.

Figure 11: Cumulative probability of relative source rank for 300 experiments with random initial source on the on power-grid network ($N = 4941$) for $p = 0.7$, $q = 0.4$, $T = T_0 + \epsilon$, where $\epsilon \sim$ Geometric distribution with different parameters and $T_0 = 15$.

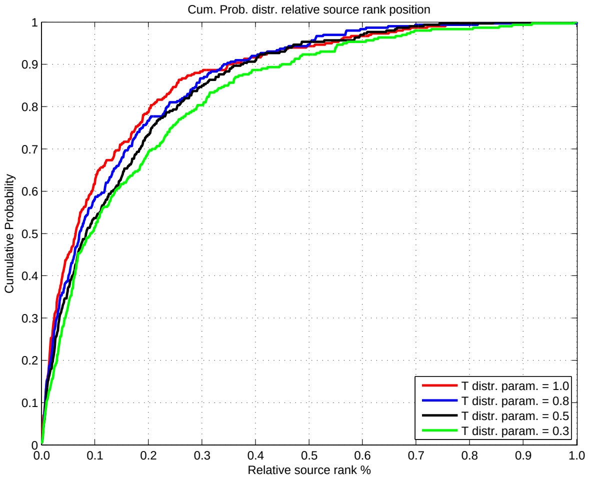
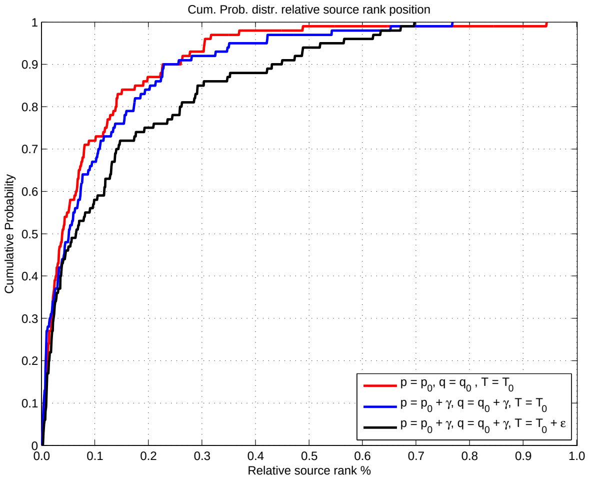
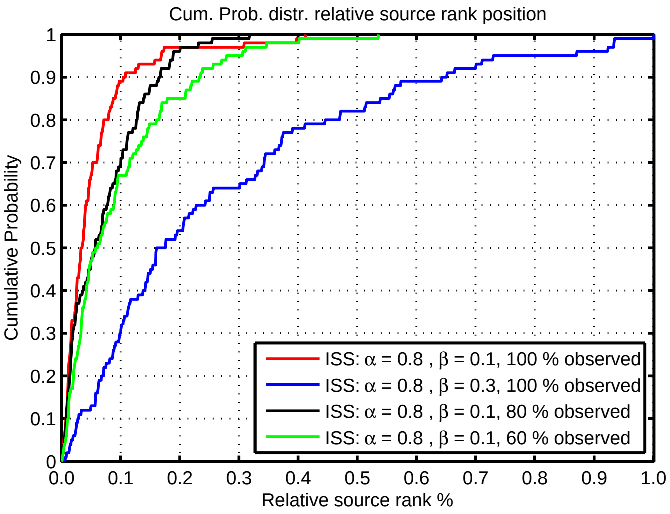


Figure 12: Cumulative probability of relative source rank for 100 experiments with random initial source on the on power-grid network ($N = 4941$) for $p = p_0 + \gamma$, $q = q_0 + \gamma$, $T = T_0 + \epsilon$, where $T_0 = 10$, $p_0 = 0.7$, $q_0 = 0.4$, $\epsilon \sim$ Geometric distribution with parameter 0.5 and $\gamma \sim$ Normal distribution with parameters ($\mu = 0, \sigma = 0.05$).



In order to demonstrate the applicability of statistical inference framework for general type of compartment contagion processes, we have made a localization experiments with the information/rumor spreading ISS (ignorant-spreading-stifler) model. The ISS model divide the individuals to three groups: ignorants who have not heard the information/rumor, spreaders who are propagating the information/rumor to ignorants and stiflers who know the information/rumor and are no longer propagating it. The probability of spreading the information/rumor from spreaders to ignorants is α in one discrete time step. If the spreader interacts with other spreader or stifler it turns to stifler state with probability of β. The infected nodes in the SIR model recovery according to its internal state contrary to the ISS model where spreaders becomes stiflers according to states of its neighbours. In figure 13 we can observe the localization performance of inference framework on ISS model on regular grid for different parameters (α, β). Even in case when a fraction of random nodes in a network can be observed the statistical inference framework can localize the initial source (see figure 13).

Figure 13: Cumulative probability of relative source rank for 100 experiments with random initial source on the on regular grid of size $N = 30x30$ for the ISS spreading process for different parameters $\alpha, \beta, T = 50$ and different fraction of observed nodes (100 % of realization or 80 % or 60 % of random nodes in a realization) in a network.



4. Related work

Although the research of epidemic processes on complex networks is very mature the problem of epidemic source detection was formulated very recently. Various researchers have proposed different solutions to the problem of epidemic source detection which are based on number of assumptions on contact network structures and spreading models.

Zaman et. al. formulated a problem, where the rumor spreads with the SI model over network structure for some unknown amount of time and observe information about which nodes got infected. They rise a question who is the most likely source of the rumor and when can they find him. As a solution to the problem of source detection, they developed a rumor centrality measure, which is the maximum likelihood estimator for a regular trees under the SI model. They also obtained various theoretical results about the detection probability on different classes of trees [21],[22]. But, when the rumor spreading happens at the general graphs they use the simple heuristics that the rumour spreads along the breadth first search rooted at the source. Dong et. al. also studied the problem of rooting the rumor source with the SI model and demonstrate similar results of asymptotic source detection probability on regular tree-type networks [6]. Comin et. al. studied and compared different measures like degree, betweenness, closeness and eigenvector centrality as estimators for source detection [5]. Pinto et. al. also formulated a similar problem of locating the source of diffusion in networks from sparsely places observers [18]. They also assume that the diffusion tree is a breadth first search, the model of spreading with no recovery and the exact direction and times of infection transfers. Spectral algorithms for detection of initial seed of nodes that best explain given snapshot under the SI model has also been derived [19].

Zhu et. al. adopted the SIR model and proposed a sample path counting approach for source detection [25]. They prove that the source node on infinite trees minimizes the maximum distance to the infected nodes. They assume that the infected and susceptible nodes are indistinguishable. Lokhov et. al. use a dynamic message-passing algorithm to estimate the probability that a given node produces an observed snapshot. They use a mean-field-like approximation to compute the marginal probabilities and an assumption of sparse contact network [13].

Contrary to these approaches, our source estimation approach reduces the assumptions on network structures and spreading process properties. Our statistical inference framework can also work on arbitrary network structures and with arbitrary compartment spreading processes (SI, SIR, SEIR, ISS, etc.)

5. Conclusion

In this paper we have constructed a statistical framework for detecting the source location of an epidemic or rumour spread from a single realization of a stochastic contagion model on an arbitrary network. Detecting the source of an epidemic or rumour spreading under a stochastic SIR discrete model, represents an extension of existing research methodologies, mainly focussed on diffusion-like processes. Furthermore, this statistical framework can be deployed for different kinds of stochastic compartment processes (ISS, SI, SIR, SEIR) on networks whose dynamical patterns can be described by probability distributions over similarities among realization vectors. We have also demonstrated that we can relax even the assumptions on complete knowledge about epidemic realization, contagion process parameters and time with uncertainty.

Acknowledgements

The authors would like to thank Professor Yaneer Bar-Yam (New England Complex System Institute, Cambridge, MA 02142, USA) for valuable help in the early stages of the research.

References

[1] N. Antulov-Fantulin, A. Lancic, H. Stefancic, M. Sikic, Fastsir algorithm: A fast algorithm for the simulation of the epidemic spread in large networks by using the susceptible-infected-recovered compartment model, http://arxiv.org/abs/1202.1639, Information Sciences (2013).

[2] A.L. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.

[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.U. Hwang, Complex networks : Structure and dynamics, Phys. Rep. 424 (2006) 175–308.

[4] C. Castellano, R. Pastor-Satorras, Thresholds for epidemic spreading in networks., Phys Rev Lett 105 (2010) 218701.

[5] C.H. Comin, L. da Fontoura Costa, Identifying the starting point of a spreading process in complex networks, Phys. Rev. E 84 (2011) 056105.

[6] W. Dong, W. Zhang, C.W. Tan, Rooting out the rumor culprit from suspects, http://arxiv.org/abs/1301.6312 (2013).

[7] S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes, Critical phenomena in complex networks, Reviews of Modern Physics 80 (2008) 1275–1335.

[8] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Structure of growing networks with preferential linking, Phys. Rev. Lett. 85 (2000) 4633–4636.

[9] D.J. Hand, K. Yu, Idiot's Bayes—Not So Stupid After All?, International Statistical Review 69 (2001) 385–398.

[10] H.W. Hethcote, The mathematics of infectious diseases, SIAM Review 42 (2000) 599–653.

[11] E. Kenah, J.M. Robins, Second look at the spread of epidemics on networks, Phys. Rev. E 76 (2007) 036113.

[12] W.O. Kermack, A.G. McKendrick, Contributions to the mathematical theory of epidemics, Proc. Roy. Soc. London Ser. A 115 (1927) 700–721.

[13] A.Y. Lokhov, M. Mzard, H. Ohta, L. Zdeborov, Inferring the origin of an epidemy with dynamic message-passing algorithm (2013).

[14] D. Mollison, Spatial contact models for ecological and epidemic spread, Journal of the Royal Statistical Society. Series B (Methodological) 39 (1977) 283–326.

[15] Y. Moreno, M. Nekovee, A.F. Pacheco, Dynamics of rumor spreading in complex networks, Phys. Rev. E 69 (2004) 066130.

[16] M.E.J. Newman, The Structure and Function of Complex Networks, SIAM Review 45 (2003) 167–256.

[17] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Phys. Rev. Lett. 86 (2001) 3200–3203.

[18] P.C. Pinto, P. Thiran, M. Vetterli, Locating the Source of Diffusion in Large-Scale Networks, Physical Review Letters 109 (2012) 068702+.

[19] B.A. Prakash, J. Vreeken, C. Faloutsos, in: ICDM'2012; Proceedings of the IEEE International Conference on Data Mining (2012).

[20] G. Serazzi, S. Zanero, Computer virus propagation models, in: M. Calzarossa, E. Gelenbe (Eds.), Performance Tools and Applications to Networked Systems, volume 2965 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2004, pp. 26–50.

[21] D. Shah, T. Zaman, Detecting sources of computer viruses in networks: theory and experiment, in: Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems, SIGMETRICS '10, ACM, New York, NY, USA, 2010, pp. 203–214.

[22] D. Shah, T. Zaman, Rumors in a Network: Who's the Culprit?, volume 57 of *IEEE Transactions on Information Theory*, pp. 5163–5181.

[23] A. Vespignani, Modelling dynamical processes in complex socio-technical systems, Nat Phys 8 (2012) 32–39.

[24] D. Watts, S. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (1998) 440–442.

[25] K. Zhu, L. Ying, Information source detection in the sir model: A sample path based approach, http://arxiv.org/abs/1206.5421 (2013).