Locating the Source of Diffusion in Complex Networks via Gaussian-Based Localization and Deduction

Xiang Li 1, Xiaojie Wang 1,*, Chengli Zhao 1, Xue Zhang 1 and Dongyun Yi 1,2

1 College of Liberal Arts and Sciences, National University of Defense Technology, Changsha 410073, China; lixiang13a@foxmail.com (X.L.); chenglizhao@nudt.edu.cn (C.Z.); xuezhang@nudt.edu.cn (X.Z.); dongyunyi@nudt.edu.cn (D.Y.)
2 State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha 410073, China

* Correspondence: wangxiaojie0817@gmail.com

Received: 24 July 2019; Accepted: 6 September 2019; Published: 8 September 2019

Abstract: Locating the source that undergoes a diffusion-like process is a fundamental and challenging problem in complex network, which can help inhibit the outbreak of epidemics among humans, suppress the spread of rumors on the Internet, prevent cascading failures of power grids, etc. However, our ability to accurately locate the diffusion source is strictly limited by incomplete information of nodes and inevitable randomness of diffusion process. In this paper, we propose an efficient optimization approach via maximum likelihood estimation to locate the diffusion source in complex networks with limited observations. By modeling the informed times of the observers, we derive an optimal source localization solution for arbitrary trees and then extend it to general graphs via proper approximations. The numerical analyses on synthetic networks and real networks all indicate that our method is superior to several benchmark methods in terms of the average localization accuracy, high-precision localization and approximate area localization. In addition, low computational cost enables our method to be widely applied for the source localization problem in large-scale networks. We believe that our work can provide valuable insights on the interplay between information diffusion and source localization in complex networks.

Keywords: source localization; optimization algorithm; data mining; complex networks

1. Introduction

Diffusion dynamics taking places on complex networks has been a long-term hot topic with importance value to help us better understand many ubiquitous natural phenomena and social behaviors. In general, many different diffusion processes that occur in daily life are harmful and may result in huge losses to society. Prototypical examples include outbreaks of epidemics among humans [1], the spreading of rumor over social networks [2], cascading failures of power grids around country wide [3], etc. If decision-makers, such as managers and politicians, can identify the diffusion sources as early as possible, they are more likely to make the right decision in time and avoid economic losses and social panic due to the associated tragedies. In order to better resist the potential terrible consequences induced by those diffusion processes, there is a great need for us to develop efficient strategies to locate the source of diffusion and devise control methods as early as possible.

Despite great effort that has been made in this field, the problem of identifying the zero patient is still challenging work. The main difficulty lies in two aspects, i.e., the incomplete information and the stochastic nature. In one aspect, the exact number of sources and the zero time at which diffusion first occurred are usually unknown to us. In addition, even if we can get full access to such information, the
inevitable randomness of the diffusion process still weakens our ability to accurately locate the source. For example, in the epidemic spreading process, such as the susceptible-infected process, different initial conditions may lead to the same observations, making it extremely hard for us to identify the real source.

Models and methods for source localization have been studied in a number of literature works. Shah and Zaman [4] were two pioneers who first provided a systematic study of the problem about how to detect the origins of a computer virus in a network. They modeled the process of virus spreading within a network via a susceptible-infected (SI) model and presented the Rumor centrality as the maximum likelihood estimator for a class of graphs. Inspired by their work, Zhu and Ying [5] developed a sample-path-based approach named Jordan centrality to detect the information source in a network under susceptible-infected-recovered (SIR) dynamics. Later, two Bayesian inference solutions, namely dynamic message-passing [6] and belief propagation [7], were successively designed to measure the probability distribution for the observations and identify the zero patient of the networks. Brockmann and Helbing [8] studied global disease dynamics and proposed a new index named effective distance to locate the origin of complex spatiotemporal patterns. Zhu and Ying [9] developed the Short-Fat Tree algorithm to locate the diffusion source within an independent cascade model. Hu et al. [10] developed a framework for optimal source localization in arbitrarily weighted networks with arbitrary distributions of sources based on controllability theory and compressive sensing. Several other methods for source localization [11–16] and observation selection [17,18] have also been proposed based on different considerations.

Those excellent works are mainly based on the knowledge of the diffusion status for a proportion of nodes at specific snapshots. However, the time information or the timestamp when diffusion first arrives at some nodes has not been fully investigated by many scholars. In [19], Pinto et al. presented the Gaussian heuristic as a maximum likelihood estimator of source localization problem on arbitrary trees, but its performance could not be guaranteed in general graphs. Zhu et al. [20] formulated the source localization problem as a ranking problem on graphs and proposed two ranking algorithms, namely cost-based ranking and tree-based ranking, to locate the diffusion source in networks. Recently, Shen et al. [21] developed a time-reversal backward spreading algorithm to efficiently locate the source of a diffusion-like process and proposed a general locatability condition. For multiple sources detection, Fu et al. [22] investigated a maximum-minimum strategy based on backward diffusion, which has been further extended by Hu et al. [23] via integer programming.

In this paper, we present the Gaussian-based localization and deduction (GLAD) as a simple and efficient framework for locating the diffusion source in networks based on partial timestamps. We mainly considered a simple diffusion process associated with time delay and provided a probabilistic method to locate the source of diffusion via parameter estimation and maximum likelihood estimation. To be more concise, we derived an optimal solution of the source localization problem on arbitrary trees and extended it to general graphs via approximations and simplifications. Experimental results were conducted on synthetic networks (including arbitrary trees and general graphs) and real networks, and the results all verified the good performance of our algorithm.

2. Methods

In this section, we provide a brief introduction about diffusion models and source localization problems at first. Then, on arbitrary trees, we derive the GLAD framework as a maximum likelihood estimator that can simultaneously estimate the probability that a node is the diffusion source and provide the corresponding diffusion parameters. Finally, we discuss the difficulties for source localization problem on general graphs, and present an approximate method with low computational cost for those cases.
2.1. Problem Definition

The underlying network in which the diffusion occurs was modeled as an undirected simple graph $G = (V, E)$, where $V$ denotes the set of nodes and $E$ denotes the set of edges. Specifically, we mainly focused on static networks whose topology never changes during the diffusion process. The diffusion source $s^* \in V$ represents the only node at which the information originates and it triggers diffusion at some unknown initial time $t_0$. Compared with many previous literature works based on stochastic epidemic processes such as SI, SIS and SIR, a simple diffusion model associated with a time delay along edges was employed in our study.

The diffusion process is modeled as follows. At time $t$, each node can be in one of two states: informed and uninformed. An informed node represents the individual who is aware of the information and will propagate it to its neighbors, whereas one uninformed node represents the individual who has not been informed yet. Let $\Gamma(v)$ denote the neighbors of node $v$ and suppose that $v$ is in the uninformed state. After $v$ receives the information from one of its neighbors for the first time $t_v$, it will change to the informed state and propagate the information. Then, each of its neighbors $u \in \Gamma(v)$ receives the information from $v$ at $t_v + \theta_{uv}$, where $\theta_{uv}$ denotes the diffusion delay associated with edge $uv$. Specifically, the diffusion delays $\{\theta_{uv}\}$ along each edges were modeled as an independent and identically distributed random variables that follows a Gaussian distribution $N(\mu, \sigma^2)$. As is often the case in many real situations, such as the spread of a computer virus on the Internet via a cable, the mean diffusion delay $\mu$ is easy to evaluate, whereas the standard deviation $\sigma$ is hard to obtain. We further assumed that only $\mu$ is known, and $\sigma$ needed to be estimated.

Let $O = \{o_k\}_{k=1}^K$ denote a group of observers whose informed times $T_O = [t_{o_1}, t_{o_2}, ..., t_{o_K}]^T$ can be accessed by us. Then, the source localization problem could be described as follows: given the network topology and several observers $O$, our goal was to find the diffusion source $s^*$ in the network. To simplify the derivation and avoid a non-invertible matrix in our method, we assumed that $s^* \notin O$. There was no loss of generality since one can generalize our method by adding an extra step to calculate the probability of being the source for those $o_k \in O$ with a known initial time $t_{o_k}$. In reality, the rumor source will not report its informed time, since this information will likely expose itself as the first to report the rumor.

2.2. Source Localization on Arbitrary Trees

Consider the case of an arbitrary tree. In graph theory, a tree is defined as a connected, undirected graph that contains no closed loops. Obviously, describing the diffusion process occurring in trees presents a natural advantage because there is only one path between the source and each observer. Recall that $O = \{o_1, o_2, ..., o_K\}$ is a set of $K$ observers with diffusion timestamps $T_O = [t_{o_1}, t_{o_2}, ..., t_{o_K}]^T$. Since the diffusion delays $\{\theta_{uv}\}$ along edges are random variables, the arrival times $\{t_{o_k}\}$ for observations can also be viewed as random variables.

Suppose that $s$ is the source node and $t_0$ is the initial time that $s$ starts diffusion, and let $P(s, o_k) \subset E$ be the path from the node $s$ to the observer $o_k$ with a length $d_{sk} = |P(s, o_k)|$. As all diffusion delays $\{\theta_{uv}\}$ along edges are i.i.d Gaussian variables $\theta_{uv} \sim N(\mu, \sigma^2)$, the arrival time $t_{o_k}$ of each observer $o_k \in O$ can be inferred to follow a Gaussian distribution $t_{o_k} \sim N(t_0 + d_{sk} \cdot \mu, d_{sk} \cdot \sigma^2)$. Therefore, the joint distribution of the arrival time $T_O = [t_{o_1}, t_{o_2}, ..., t_{o_K}]^T$ satisfies a K-dimensional Gaussian form $T_O \sim N(\mu_s, \Lambda_s)$. Taking the above, the likelihood probability of observing the diffusion timestamps $T_O$ at arrival time $T_O$ with respect to the source node $s$, the initial time $t_0$ and the delay standard deviation along edges $\sigma$ is expressed as follows:

$$p(T_O | s, t_0, \sigma) = \frac{1}{(2\pi)^{K/2}|\Lambda_s|^{1/2}} \exp\left(-\frac{1}{2}(T_O - \mu_s)^T \Lambda_s^{-1}(T_O - \mu_s)\right).$$  \hspace{1cm} (1)
The mean vector $\mu_s$ and the covariance matrix $\Lambda_s$ of the joint Gaussian distribution are as follows:

\[
\begin{align*}
[\mu_s]_k &= t_0 + d_{sk} \cdot \mu \\
[\Lambda_s]_{i,j} &= \sigma^2 \cdot [\Lambda_{ps}]_{i,j},
\end{align*}
\]  

where $[\Lambda_{ps}]_{i,j} = |\mathcal{P}(s,o_i) \cap \mathcal{P}(s,o_j)|$ denotes the path intersection matrix, and the element of matrix represents the number of joint edges on two paths from node $s$ to observer $o_i$ and $o_j$.

Since we consider that no prior information is available on the location of the source node, the optimal estimator is the maximum likelihood estimator (MLE):

\[
(\hat{s}, \hat{t}_0, \hat{\sigma}) = \argmax_{s \in V \setminus O} P(T_O|s, t_0, \sigma).
\]  

Let $w_s = [w_{s1}, w_{s2}, ..., w_{sk}]^T$ represent a K-dimensional column vector of the differences between the observation time and the mean arrival time for the observers:

\[
w_{sk} = t_{0k} - \mu \cdot d_{sk}.
\]  

For any estimator $\hat{s}$, by maximizing Equation (1) with respect to $t_0$, the MLE of $t_0$ can be represented as

\[
\hat{t}_0 = \argmin_{t_0} (T_O - \mu_s)^T \Lambda_s^{-1} (T_O - \mu_s)
\]  

\[
= \argmin_{t_0} (w_s - t_0I)^T \Lambda_s^{-1} (w_s - t_0I)
\]  

\[
= I^T \Lambda_{ps}^{-1} w_s / I^T \Lambda_{ps}^{-1} I,
\]

where $I = [1, 1, ..., 1]^T$ denotes a K-dimensional column vector full of 1s.

To better describe the optimization procedure, we created an auxiliary variable

\[
z_s = (w_s - \hat{t}_0 I)^T \Lambda_{ps}^{-1} (w_s - \hat{t}_0 I).
\]

Substituting Equation (6) into Equation (1) and following a similar procedure by maximizing $P(T_O|s, \hat{t}_0, \sigma)$ with respect to $\sigma$, the MLE of $\sigma$ had the following form

\[
\hat{\sigma} = \argmax_{\sigma} P(T_O|s, \hat{t}_0, \sigma)
\]  

\[
= \argmax_{\sigma} \sigma^{-k} \exp\left(-\frac{1}{2} z_s^2 \sigma^{-2}\right)
\]  

\[
= (z_s / K)^{1/2}.
\]  

Finally, the optimal estimator for the source node $s$ is

\[
\hat{s} = \argmax_{s \in V \setminus O} P(T_O|s, \hat{t}_0, \hat{\sigma})
\]  

\[
= \argmax_{s \in V \setminus O} \hat{\sigma}^{-k} |\Lambda_{ps}|^{-1/2}
\]  

\[
= \argmin_{s \in V \setminus O} \{K \log z_s + \log \Lambda_{ps}\}.
\]

The equations above constituted the core of our source localization algorithm in arbitrary trees. Since our optimizations were based on the assumption of Gaussian distribution on edge delays, and the diffusion parameters were deduced from observers, we named our method Gaussian-based localization and deduction (GLAD).

Figure 1 demonstrates a diffusion process on a toy tree with 11 nodes and 10 edges. In this model, the diffusion delay along edges were sampled from a Gaussian distribution $\mathcal{N}(1, 0.25)$. We assumed that one node $s$ was the diffusion source and three nodes $O = \{o_1, o_2, o_3\}$ were observers. Only the
network topology, the mean diffusion delay, and the informed times of observers were accessible to us, where the informed time of observers was $T_0 = [2.319, 0.662, 2.488]$.

![Figure 1](image_url)

Figure 1. Illustration of the diffusion process on a representative tree. (a) topology of the tree and the underlying edge diffusion delay sampled from a Gaussian distribution $\theta_{uv} \sim N(1, 0.25)$; (b) diffusion process from one source node $s$ with initial time $t_0 = 0$, where $\{o_1, o_2, o_3\}$ denote three observers.

Now, we introduce the calculation process of a GLAD algorithm, taking node $B$ for example. From Figure 1, it is easy to know the path from node $B$ to three observers $A$, $J$, $F$:

$$
P(B, A) = \{BA\}, \quad P(B, J) = \{BD, DI, IJ\}, \quad P(B, F) = \{BD, DG, GF\}.
$$

Thus, mean arrival time is $d_s = [1, 3, 3]^T$ and then the corresponding time difference vector is $w_s = T_0 - d_s = [1.319, -2.378, -0.512]^T$.

According to the definition of path intersection matrix $[\Lambda_{ps}]_{ij} = |P(s, o_i) \cap P(s, o_j)|$, we have

$$
\Lambda_{ps} = \begin{pmatrix}
1 & 0 & 0 \\
0 & 3 & 1 \\
0 & 1 & 3
\end{pmatrix}, \quad \Lambda_{ps}^{-1} = \begin{pmatrix}
1 & 0 & 0 \\
0 & 0.375 & -0.125 \\
0 & -0.125 & 0.375
\end{pmatrix}.
$$

Combining Equation (6) and Equation (8), the MLE of $t_0$ and $\sigma$ is represented as

$$
\begin{align*}
\hat{t}_0 &= I^T \Lambda_{ps}^{-1} w_s / I^T \Lambda_{ps}^{-1} I = 0.404, \\
\hat{z}_s &= (w_s - \hat{t}_0 I)^T \Lambda_{ps}^{-1} (w_s - \hat{t}_0 I) = 3.343, \\
\hat{\sigma} &= (\hat{z}_s / K)^{1/2} = 1.056.
\end{align*}
$$

Finally, the objective function of node $s$ in Equation (9) is calculated:

$$
OBJ_s = K \log \hat{z}_s + \log |\Lambda_{ps}| = 5.700.
$$

Following the optimization process in GLAD, the diffusion parameters $(t_0, \sigma)$ estimated by Equations (6) and (8) and the corresponding objective function value in Equation (9) for each candidate source $s \in V \setminus O$ are illustrated in Table 1. From the table, we can see that the real source $I$ has been successfully identified, and the estimation errors for $t_0$ and $\sigma$ are fairly acceptable.
Table 1. MLE of the diffusion parameters \((t_0, \sigma)\) and corresponding objective function for each candidate source computed by GLAD on the toy tree.

| MLE | B | C | D | E | G | H | I | K |
|-----|---|---|---|---|---|---|---|---|
| \(\hat{t}_0\) | 0.404 | −0.596 | −0.177 | −1.177 | 0.489 | −0.511 | −0.424 | −1.424 |
| \(\hat{\sigma}\) | 1.056 | 1.056 | 0.583 | 0.583 | 1.108 | 1.108 | 0.099 | 0.099 |
| OBJ | 5.700 | 6.617 | 2.133 | 3.050 | 5.989 | 6.906 | −8.499 | −7.583 |

2.3. Source Localization on General Graphs

When information is diffused on general graphs, the source localization problem will be more difficult. In general graphs, a spanning tree grows naturally with the diffusion process based on the sequence at which information first reaches each node. Unfortunately, although the real diffusion process in the network is a deterministic process, we cannot ascertain which spanning tree is the actual diffusion tree because the exact diffusion delays along edges are unknown. For that reason, to find the MLE for the diffusion parameters \((t_0, \sigma)\) and the source \(s\), we need to optimize the likelihood function in Equation (1) over all possible spanning trees rooted at each candidate source. In the following, we introduce an approximation method to efficiently locate the source in general graphs.

In graph theory, an important feature is that for any BFS tree with the same root \(s\), the distance \(d_{sk}\) between \(s\) and each observer \(o_k\) never changes. Recall the diffusion parameters \(\mu_s\) and \(\Lambda_s\). As can be seen in Equation (2), the mean vector \(\mu_s\) does not change in different BFS trees. On the other hand, from Equation (3), we could find that, although the covariance matrix \(\Lambda_s\) may vary in different BFS trees, the diagonal elements in each \(\Lambda_s\) remain unchanged. In addition, matrix \(\Lambda_s\) is diagonally dominant and sparse in large networks so that the diagonal matrix of \(\Lambda_s\) can replace it with little effect on parameter estimation and likelihood maximization. Inspired by these phenomena, we proposed to perform the optimization with a modified function:

\[
\hat{P}(T_0|s, t_0, \sigma) = \frac{1}{(2\pi)^{K/2}|D_s|^{1/2}} \exp\left(-\frac{1}{2}(T_0 - \mu_s)^T D_s^{-1}(T_0 - \mu_s)\right),
\]

where \(D_s = \text{diag}(\Lambda_s)\) denotes the diagonal matrix of the original covariance matrix \(\Lambda_s\). As \(D_s\) does not change in different BFS trees, one can randomly build a BFS tree and obtain it.

In the new optimization problem, the corresponding MLE for \(t_0\) can be given by

\[
\hat{t}_0 = \hat{T}^T D_p^{-1} w_s / \hat{T}^T D_p^{-1} I = \sum k d_{sk}^{-1} w_{sk} / \sum k d_{sk}^{-1},
\]

where \(w_{sk} = t_{0k} - \mu_s \cdot d_{sk}\) is the same as in Equation (5).

Then, the auxiliary variable should be

\[
\hat{z}_s = (w_s - \hat{t}_0 I)^T D_p^{-1}(w_s - \hat{t}_0 I) = \sum k (w_{sk} - \hat{t}_0)^2 d_{sk}^{-1}.
\]

The MLE of \(\sigma\) remains the same as in a previous subsection:

\[
\hat{\sigma} = (\hat{z}_s / K)^{1/2}.
\]
Finally, the optimal estimator for the source node $s$ becomes

$$\hat{s} = \arg\min_{s \in V \setminus O} \left\{ K \log \tilde{s} + \log |D_{ps}| \right\} = \arg\min_{s \in V \setminus O} \left\{ K \log \tilde{s} + \sum_k \log d_{sk} \right\}. \quad (18)$$

Generally, we located the diffusion source by optimizing the naïve likelihood function (Equation (1)) and the modified function (Equation (14)) on general graphs. In order to distinguish between these two cases, we named them GLAD-naïve and GLAD-modified, respectively. However, for GLAD-naïve, we had to randomly build a BFS tree for each node firstly and optimize the naïve likelihood function (Equation (1)).

2.4. Computational Complexity Analysis

The computational cost of GLAD-naïve consists of three parts: the building of BFS tree rooted at one candidate source $v \in V \setminus O$, the calculation of path intersection matrix $\Lambda_{ps}$, and the estimation of diffusion parameters $t_0, \sigma$ and objective function. Suppose that the numbers of nodes and edges in the network $G(V, E)$ are $N$ and $M$ and we have $K$ observers. Generally, when the topology of the network is known, building a BFS tree rooted at one node $v$ will cost $O(M)$. In addition, in the meantime, the paths from $v$ to each nodes on the tree can be gained naturally along with the tree building scheme. As each element in the path intersection matrix $[\Lambda_{ps}]]_{i,j}$ denotes the length of the common path from $v$ to observers $i$ and $j$, $O(N)$ should be the cost to check how many edges on the tree lay on the common path $P(s, o_i) \cap P(s, o_j)$. Thus, the computational cost of this step is $O(NK^2)$. The main costs in the third step of parameter estimation and objective function calculation were the computation of $\Lambda_{ps}^{-1}$ and $|\Lambda_{ps}|$, which both cost $O(K^3)$ by typical algorithms in linear algebra. In total, the whole computational complexity of GLAD-naïve is $O(NM + N^2K^2 + NK^3)$.

The calculation time could be further reduced in GLAD-modified because the path intersection matrix $\Lambda_{ps}$ did not need to be explicitly constructed, since we only care about its diagonal elements $d_{sk}$. This process could be performed in a batch mode with only $O(KM)$ in which we started the BFS process for each observer $o_k$ and then recorded the distance $d_{sk}$ from it to all candidate sources $s \in V \setminus O$. According to Equations (15)–(18), the parameter estimation and objective function calculation can be finished in a linear time $O(K)$ for each candidate source. Consequently, the whole computational complexity of GLAD-modified is linear, i.e., $O(KM + KN) = O(KM)$, which enables its wide application in large-scale networks.

3. Experiments and Analysis

To quantify the validity of the proposed algorithm, we present numerical results on the success rate of source localization on arbitrary trees and general graphs. In real implementations, the diffusion delay $\theta_{uv}$ along each edge is sampled via an independent and identically truncated Gaussian distribution to ensure that $\theta_{uv} > 0$. Since no prior knowledge is available on the diffusion source $s$ and observers $O$, they are chosen randomly among the network. The results are obtained by averaging over 100 independent realizations.

3.1. Metrics and Benchmark Methods

The algorithm performance is evaluated using three metrics, namely average ranking, $\gamma$%-accuracy, and average error distance. The average ranking is the average location of the actual source in the list of nodes sorted in increasing order by the objective function value. We focus on the average value after many simulations. The $\gamma$%-accuracy represents the proportion of simulations for which the real source is ranked within the top $\gamma$% among all nodes. In particular, ties are broken randomly for nodes with the same ranking. As for the average error distance, it is defined as the average distance between the estimated source and the real source. Among those metrics, the average ranking
reflects the average accuracy of source localization, the $\gamma$%-accuracy focuses more on high-precision localization, and the average error distance mainly considers the approximate area of the source.

The performance of GLAD-naive and GLAD-modified are compared with two well-known benchmark methods, namely Gaussian heuristic and time-reversal backward spreading:

1. Gaussian heuristic (GAU). In [19], Pinto et al. first showed the possibility of estimating the location of the source from measurements collected by sparsely placed observers. They modeled the diffusion delays along edges with Gaussian distribution $N(\mu, \sigma)$, and built an MLE as the optimal solution for the source localization problem in arbitrary trees. Compared with our method, Pinto assumed that both $\mu$ and $\sigma$ were known parameters, whereas we allowed $\sigma$ to be unknown and determined it via estimation process.

2. Time-reversal backward spreading (TRBS). The time-reversal backward spreading algorithm proposed by Shen et al. [21] was an efficient method to infer the diffusion source based on a weighted network structure and partial timestamps. In their method, the variance of the differences between the true arrival times and the expected arrival from a node to all observers was calculated, as a measurement to evaluate the extent to which it is the diffusion source.

Note that several topological-based methods, such as Rumor centrality and Jordan centrality, are not employed as benchmark methods since they cannot well exploit the timestamp information. (We have implemented these methods in experiments, and the numerical results suggest that they are far less accurate than the timestamp-based methods.)

3.2. Results on Arbitrary Trees

We first perform simulations on two types of arbitrary trees, i.e., ER trees and BA trees. These two tree networks are generated by random network (ER) and scale-free network (BA), respectively, which are all connected networks.

To generate a BA tree, we firstly make the initial network contain only one isolated node. At time $t$, add one new node to the network and meanwhile add an edge linking the new node with one existing node according to the preference attachment rule in the BA model. Relatively speaking, it is very difficult to directly use an ER model to generate a connected network with $N$ nodes. Although $n - 1$ edges can be randomly generated among the $N$ nodes, there is no guarantee that the generated network is a connected network. Thus, a compromise is adopted here: Firstly, we generate a ER network with $1.5N$ nodes and its average degree is 1. If the giant component in the network has $N$ nodes (through many simulations, we do find the giant component), take the maximum spanning tree as the approximation of ER tree; if not, repeat the first step, generating another ER network, until we find the ER tree.

Figure 2 shows the average ranking under different methods with different fraction of observers. In the experiments, the number of nodes in the tree is fixed to be 100, and the signal-to-noise ratio of the diffusion delay along edges $\mu/\sigma$ are chosen from \{4, 3, 2\}. A lower signal-to-noise ratio implies larger uncertainty or noise on edges. The average ranking of the source for GLAD-naive and GAU is slightly lower than that of the other two methods, which is consistent with the previous discussion, indicating that they are all optimal solutions in arbitrary trees. Since GLAD-modified does not fully exploit the path intersections between the source and each observer, it performs strictly worse than GLAD-naive. The same reason leads to the poor results of TRBS in one aspect. On the other hand, as TRBS does not model the diffusion delay along edges, its performance is not robust to large noise. As can be seen from left to right in the figure, with the increase of the noise, TRBS performs worse and worse compared to other methods.
Figure 2. Average ranking of the real source in different source localization algorithms on synthetic trees. In each tree, the number of nodes is 100 and the diffusion delay along edges follows \( N(\mu, \sigma^2) \). Only the mean delay \( \mu \) is used to identify the source.

Figure 3 demonstrates the \( \gamma \)-accuracy of different source localization algorithms on arbitrary trees. Compared with the previous simulations, the results of the four methods under this metric are more distinguishable that GLAD-naive > GLAD-modified > GAU > TRBS. Although the average ranking of GLAD-naive and GAU are nearly equivalent, the \( \gamma \)-accuracy of the former algorithm is obviously better than that of the latter. This finding reveals one disadvantage of GAU that it cannot precisely locate the diffusion source in trees. In contrast, the proposed GLAD-naive and GLAD-modified estimate the diffusion parameters \( t_0 \) and \( \sigma \) for each node, which will improve the ability to distinguish between the real source and its neighbors. As the signal-to-noise ratio decreases, the elements of the covariance matrix \( \Lambda_s \) in Equation (3) become increasingly significant. As GLAD-modified ignores the non-diagonal elements of \( \Lambda_s \), the performance gaps between GLAD-naive and GLAD-modified will also increase.

Figure 3. \( \gamma \)-accuracy of different source localization algorithms on synthetic trees. Because the number of nodes is 100, this figure corresponds to a case in which the estimated source is also the real source.

The average distance between the estimated source and the real source is shown in Figure 4. Under this metric, GLAD-naive and GAU show the best performance among all methods, which is similar to the findings shown in Figure 2. More precisely, the performance of GLAD-naive is slightly better than GAU, especially in the trees with larger noise. A strange phenomenon is observed in which
the probability of identifying the real source for GLAD-modified is far larger than TRBS, whereas the average error distance of the two methods is similar. In BA trees, GLAD-modified performs even worse than TRBS. Compared with TRBS, the performance of GLAD-modified is more sensitive to the relative locations between the source and observers. In the discussion section, we will elaborate on the relationship between these two methods.

![Figure 4](imageurl) **Figure 4.** Average distance between the estimated source and the real source in different source localization algorithms on synthetic trees. If more than one node is ranked first, we randomly choose one as the estimated source.

The mean square errors (MSE) of the estimated parameters $t_0$ and $\sigma$ are illustrated in Figure 5. In the experiments, the initial time is $t_0 = 0$ and the parameters are $\mu = 4$ and $\sigma = 1$. We assume that only $\mu$ is known, and try to estimate the other two parameters by GLAD-naive and GLAD-modified. Since GLAD-naive is the optimal solution for the source localization problem in arbitrary trees, the MSEs of $t_0$ and $\sigma$ obtained by GLAD-naive are obviously lower than that of GLAD-modified.

![Figure 5](imageurl) **Figure 5.** MSE of the diffusion parameters in synthetic trees. In the diffusion process, the initial time $t_0 = 0$ and the diffusion delay along edges i.i.d follow $\mathcal{N}(4, 1)$. Note that the MSE results of $t_0$ and $\sigma$ are not calculated based on estimated source, but rather on the real source.
3.3. Results on General Graphs

Simulations on general graphs are performed on synthetic networks and real networks. For synthetic networks, we consider three common types, i.e., Erdös–Rényi random network $ER(N,k)$, Barabási–Albert scale-free network $BA(N,k)$ and Watts–Strogatz small-world network $WS(N,k,p)$, where $N$ denotes the number of nodes, $k$ denotes the average degree, and $p$ denotes the rewiring probability in WS networks. A brief introduction of some basic topological features is shown in Table 2.

| Type        | Name            | $N$ | $M$ | $\langle k \rangle$ | $k_{\text{max}}$ | $H$  | $\langle d \rangle$ | $\langle C \rangle$ |
|-------------|-----------------|-----|-----|---------------------|------------------|------|---------------------|---------------------|
| Random      | ER(100,4)       | 100 | 200 | 4.000               | 9.620            | 1.211| 3.418               | 0.038               |
| Scale-free  | BA(100,4)       | 100 | 200 | 4.000               | 21.756           | 1.723| 3.104               | 0.102               |
| Small-world | WS(100,4,0.1)   | 100 | 200 | 4.000               | 6.452            | 1.047| 4.263               | 0.286               |

Figure 6 shows the performance on three synthetic networks with an average degree $\langle k \rangle = 4$ and signal-to-noise ratio $\mu/\sigma = 3$. Unlike the case on arbitrary trees in previous section, the algorithm performances vary greatly on general graphs. The results demonstrate a uniform ranking of the four methods under all three metrics that GLAD-modified $>$ TRBS $>$ GLAD-naive $>$ GAU. Although GLAD-naive presents satisfactory results in arbitrary trees, its performance is just normal in general trees. However, one advantage of GLAD-naive is maintained in that it performs consistently better than GAU, thus illustrating the necessity of estimating diffusion parameters for different nodes. In arbitrary trees, GLAD-naive always performs better than GLAD-modified, whereas the situation is completely reversed in general graphs, which explains once again the great need to design a special algorithm on general graphs. Theoretically, if one BFS tree rooted at one node is indeed the diffusion tree, GLAD-naive will definitely outperform GLAD-modified, which is consistent with the results in arbitrary trees. Nevertheless, that condition is difficult to satisfy in dense networks because there are so many BFS trees in such networks. In addition, the uncertainty of the diffusion delay along edges further weakens our ability to determine which tree is the true diffusion tree. Taken together, although the likelihood function for GLAD-modified is just an approximation of GLAD-naive, it can present a much better performance on general graphs.

Another important finding is that it is often easier to locate the diffusion source in a homogeneous network like ER and WS than in a heterogeneous one like BA. The greatest obstacle of source localization problem in a heterogeneous network is the information redundancy of observers. Consider a simple heterogeneous case of a star-like network with one center node connected with many peripheral nodes. In such a network, regardless of which node is the real source, it cannot be uniquely identified since we do not know the exact initial time of diffusion. Actually, the information provided by peripheral observers is equivalent to the information provided by the center node, thus increasing the difficulty of locating the source in a heterogeneous network. However, such situations do not occur in homogeneous networks because there are only a few peripheral nodes in such networks.

Next, we consider the results of source localization in real networks. Nine networks from different fields are used to compare the performances. States [24] is a network of 48 contiguous states and the District of Columbia of USA, where an edge exists if two states share a border. Dolphins [25] contains the frequent associations between 62 dolphins in a community. Polbooks [26] represents the frequent co-purchasing books about US politics sold by the online bookseller Amazon.com. Football [27] is a network of American football games between Division IA colleges during regular season Fall 2000. Enron [28] email communication network covers the communications among the employees in Enron corporation between 1999 and 2003. Jazz [29] is a social collaboration network where nodes are Jazz
musicians and an edge implies that two musicians have played together in a band. USAir [24] is a network of flights between US airports in 1997. Netscience [30] represents a co-authorship network of scientists working on network theory and experiment. Celegans [31] is a biological network where nodes are substrates and the edges are metabolic reactions. All networks can be downloaded from the Internet. (Dolphins, Polbooks, Football and Netscience are available on Newman’s website [32]. States, Enron, Jazz, USAir and Celegans are available on Network Repository [24].) In the following experiments, we only consider the largest connected component and remove self-loops and multiple edges. A brief introduction of some basic topological features is shown in Table 3.

Figure 6. Source localization performance of different methods on synthetic networks with average degree \( \langle k \rangle = 4 \) and signal-to-noise ratio \( \mu/\sigma = 3 \).

Table 3. Basic topological features of nine real networks. \( N \) and \( M \) are the numbers of nodes and edges. \( \langle k \rangle \) and \( k_{\text{max}} \) are the average degree and the maximum degree. \( H \) is the degree heterogeneity. \( \langle d \rangle \) denotes the average path length between nodes pair. \( \langle C \rangle \) is the average local clustering coefficient of the network.

| Type          | Name       | \( N \) | \( M \) | \( \langle k \rangle \) | \( k_{\text{max}} \) | \( H \) | \( \langle d \rangle \) | \( \langle C \rangle \) |
|---------------|------------|---------|---------|--------------------------|----------------------|--------|----------------------|-------------------|
| Geography     | States     | 49      | 107     | 4.367                    | 8                    | 1.130  | 4.163                | 0.507             |
| Animal        | Dolphins   | 62      | 159     | 5.129                    | 12                   | 1.327  | 3.357                | 0.303             |
| Co-purchasing | Polbooks   | 105     | 441     | 8.400                    | 25                   | 1.421  | 3.079                | 0.488             |
| Sport         | Football   | 115     | 613     | 10.661                   | 12                   | 1.007  | 2.508                | 0.403             |
| Email         | Enron      | 143     | 623     | 8.713                    | 42                   | 1.483  | 2.967                | 0.453             |
| Social        | Jazz       | 198     | 2742    | 27.697                   | 100                  | 1.395  | 2.235                | 0.633             |
| Transport     | USAir      | 332     | 2126    | 12.807                   | 139                  | 3.464  | 2.738                | 0.749             |
| Co-authorship | Netscience | 379     | 914     | 4.823                    | 34                   | 1.663  | 6.042                | 0.798             |
| Biological    | Celegans   | 453     | 2025    | 8.940                    | 237                  | 4.485  | 2.664                | 0.655             |

Figure 7 demonstrates the average performance of source localization in real networks with signal-to-noise ratio \( \mu/\sigma = 4 \). Results show that GLAD-modified outperforms any other methods in all
real networks, which is consistent with its performance in synthetic networks. One remarkable feature is that, with the same number of observers, the source in Football is easier to determine compared with that in other networks. The homogeneous structure of Football may be the reason. Considering a highly heterogeneous network, the nodes in network can be divided into center nodes and edge nodes. If the center node is the diffusion source, it is difficult to identify the source among this center node and its neighbor nodes; if the edge node is the diffusion source, it is also hard for us to locate the source among this edge node, the first-order center nodes and the second-order center nodes. Linking with Table 3, it is easy to locate the diffusion source when the degree heterogeneity remains low. Actually, the degree heterogeneity of Football is even lower than that of an ER random network with the same number of nodes and edges. As previously discussed, the homogenous nature of Football improves the accuracy of locating the diffusion source under the same number of observers. In comparison, it is not easy to locate the source in heterogeneous networks such as Celegans and USAir.

Figure 7. Average ranking of the real source by source localization algorithms on nine real networks with signal-to-noise ratio $\mu/\sigma = 4$.

The $\gamma$%-accuracy and average error distance of source localization algorithms on real networks are shown in Figures 8 and 9. Under both metrics, GLAD-modified presents stable and satisfactory performance, while GAU performs the worst. For GLAD-naive and TRBS, their performance varies considerably in different networks. Generally, GLAD-naive performs slightly better than TRBS in small-scale networks with low average degree, whereas TRBS gradually performs better when the number of nodes increases. This situation is similar to that of GLAD-naive and GLAD-modified in arbitrary trees and general graphs. In addition, the performance of TRBS and GLAD-modified is hard to separate in several networks like Football, USAir and Celegans. Thus, we guess that there may be some potential connection between the two methods. In the discussion section, we will introduce the relationship between TRBS and GLAD-modified in detail.
Figure 8. $\gamma$%-accuracy of the source localization algorithms on nine real networks with signal-to-noise ratio $\mu/\sigma = 4$.

Figure 9. Average distance between the estimated source and the real source by source localization algorithms on nine real networks with signal-to-noise ratio $\mu/\sigma = 4$. If more than one node is ranked first, we randomly choose one as the estimated source.
4. Discussion

In this section, we will discuss some interesting features of GLAD-modified, including the effects of different observer placement strategies, and the internal relationship between GLAD-modified and TRBS.

4.1. Different Strategies for Observer Placements

Although we have proposed GLAD-modified as an effective source localization algorithm that can achieve nearly 90% accuracy in most real networks with 50% observers, it is often not easy to monitor such a huge number of individuals in the real world due to limited resources and privacy issues. Consequently, how to choose the most informative nodes as observers should also be one of the important counterparts in the research field of source localization. In the following, we will discuss the performance differences for GLAD-modified under several centrality-based observer placement strategies, including small degree, large degree, large betweenness [33] and large closeness [34].

The $\gamma$-accuracy of GLAD-modified on synthetic networks with the basic random strategy and the other four observer placement strategies is displayed in Figure 10. For the sake of simplicity, we focus more on high precision localization and do not distinguish among the cases when the $\gamma$-accuracy is lower than 0.9. Results show that the effects of different observer placement strategies are very similar in homogeneous networks ER and WS, whereas they vary dramatically in heterogeneous network BA. Compared with center nodes with larger degree, betweenness and closeness, the information provided by peripheral nodes with a lower degree is less useful for identifying the real source. This result is consistent with our intuitive assessment that experts are usually more important than ordinary individuals in decision-making.

![Figure 10. $\gamma$-accuracy of GLAD-modified on three types of synthetic networks under different observer placement strategies with signal-to-noise ratio $\mu/\sigma = 4$. The number of nodes is 100 in all networks, and the rewiring probability in WS is 0.1.](image)

4.2. Relationship with the TRBS Algorithm

In previous experiments, the results demonstrate that the performances of GLAD-modified and TRBS are closely correlated. Generally, the performance of GLAD-modified is slightly better than that of TRBS under all metrics in most synthetic networks and real networks. In particular, we notice that the performance of the two methods for Football and Celegans is indistinguishable, which inspires us to expose whether certain potential relationships occur between GLAD-modified and TRBS.
In [21], Shen et al. developed the TRBS algorithm to locate the source of a diffusion-like process. TRBS starts the reversed diffusion process from each observer \( o_k \) to all nodes in the networks along the reversed direction of links. At each node \( s \), the reversed arrival time \( t_{sk} - \hat{t}(s,o_k) \) from each observer \( o_k \) should be recorded, and a vector \( T_s = [t_{s_1} - \hat{t}(s,o_1), t_{s_2} - \hat{t}(s,o_2), ... , t_{s_K} - \hat{t}(s,o_K)]^T \) is then obtained. Afterward, the node \( s^* \) with the minimum variance of \( Var(T_s) \) is detected as the source. Along with the notations in our paper, the reversed arrival time vector \( T_s \) in TRBS is identical to \( w_s \) in Equation (5), thus the optimization problem of TRBS can be given by

\[
\hat{s} = \arg\min_{s \in V \setminus O} \{K \log \hat{z}_s + \sum_k \log d_{sk}\}
\]

\[
\approx \arg\min_{s \in V \setminus O} \{K \log d_s^{-1} + K \log \sum_k (w_{sk} - \bar{w})^2 + \sum_k \log d_k\}
\]

\[
\approx \arg\min_{s \in V \setminus O} \{\frac{1}{K} \sum_k (w_{sk} - \bar{w})^2\}
\]

\[
\approx \arg\min_{s \in V \setminus O} Var(w_s).
\]

By adding this prior knowledge, the optimization problem in GLAD-modified is similar to that of TRBS, which explains from the side of why GLAD-modified often performs better than TRBS. Compared with TRBS, the performance of GLAD-modified is more sensitive to the relative locations between the source and observers, which also provides insights on the strange phenomenon in Figure 4 that the average error distance of GLAD-modified is larger than TRBS in BA trees. In BA trees, the distance between two nodes changes considerably, which severely limits the effectiveness of GLAD-modified.

However, it is worth noting that this prior knowledge is not rigorous. Even if all observers \( \{o_k\} \) can be uniformly distributed around one node \( s \), they are almost impossible to be uniformly distributed around another node at the same time. For example, the observers can never be uniformly distributed around another node at the same time. For example, the observers can never be uniformly distributed around one node

5. Conclusions

A fundamental question in modern systems that undergo a diffusion-like process such as information propagation and epidemic spreading is where the origin is located. In this paper, our main purpose was to devise a method that can locate the diffusion source in complex networks with limited observations. To do so, we modeled the diffusion delay along each edge as a random variable that follows a Gaussian distribution \( \mathcal{N}(\mu, \sigma) \), and derived the corresponding likelihood function for the informed timestamps of observers. Thus, the source localization problem could be transformed to
a parameter estimation and likelihood maximization problem. Since our optimizations were based on Gaussian assumption, and the diffusion parameters were deduced from observers, we named our method Gaussian-based localization and deduction (GLAD). We obtained the GLAD-naive as the optimal solution on arbitrary trees, and further derived the GLAD-modified as an approximate solution on general graphs.

We compared the algorithm performances with two benchmark methods in terms of three types of abilities: average localization accuracy, high-precision localization, and approximate area localization. Extensive experiments on synthetic trees showed that GLAD-naive was superior to other methods, which was consistent with our conclusions because GLAD-naive was indeed an optimal solution in this case. The results on general graphs demonstrated that GLAD-modified performed the best among all methods. Furthermore, we studied the effects of different observer placement strategies for GLAD-modified, and the underlying relationship between GLAD-modified and TRBS.

The main contribution of this work is that we employ the parameter estimation process into the optimization problem of source localization, which enables us to build a complex model with more parameters and achieve better results. Compared with one well-known source localization algorithm named GAU, our framework applies the same Gaussian assumption on diffusion delays and utilizes the same known information, while it outperforms GAU significantly on both arbitrary trees and general graphs. In addition, the computational complexity of GLAD-modified is just linear with the scale of networks. The extremely low cost will enable the wide application of our method in large-scale networks for source localization within a reasonable time. Although we only present the optimization process for undirected networks, the corresponding formula for directed networks can be easily achieved via slight modifications. Upon we finish our work, we notice a recent paper proposed by Tang et al. [35] that also combines parameter estimation with source localization. Their approach is more complicated than ours since they treat more diffusion parameters as unknown and perform the optimization over a parameterized family of Gromov matrices. However, high computational cost restricts its further usage on large-scale networks.

Although our proposed algorithm provides a new perspective for the problem of source localization in complex networks, considerable work remains to be done. Our proposed method relied strongly on diffusion delay that all diffusion delays \( \{\theta_{uv}\} \) along edges are i.i.d Gaussian variables \( \theta_{uv} \sim N(\mu, \sigma^2) \), which indicates to us to consider other delay distributions. The most intuitive task is how to identify multiple diffusion sources in networks. Compared with the single source localization problem, the case of multiple source localization is far more complicated because it is a combinatorial optimization problem. Several studies [36–38] have addressed the problem via a two-step strategy that first obtains a set of source candidate clusters and then applies single source localization algorithms to identify the source in each cluster. However, a general framework for simultaneously determining the number of sources and their locations in a large complex network is still lacking. In addition, to the best of our knowledge, few theoretical and practical studies have focused on source localization in multi-layer networks [39] and temporal networks [40].

Author Contributions: Conceptualization, C.Z., X.L. and D.Y.; methodology, X.L. and X.W.; software, X.L. and X.W.; validation, X.W. and X.L.; formal analysis, X.W.; investigation, X.Z.; resources, X.W.; data curation, X.W.; writing—original draft preparation, X.L.; writing—review and editing, X.Z.; visualization, X.L. and X.W.; supervision, X.W. and X.L.; project administration, C.Z.; funding acquisition, C.Z. This paper was prepared the contributions of all authors. All authors have read and approved the final manuscript.

Funding: This work is supported by the National Key R&D Program of China (Grant No. 2017YCF1200301).

Acknowledgments: We thank Yangyang Liu for valuable discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pastor-Satorras, R.; Castellano, C.; Van Mieghem, P.; Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* 2015, 87, 925–979. [CrossRef]
2. Chierichetti, F.; Lattanzi, S.; Panconesi, A. Rumor spreading in social networks. *Theor. Comput. Sci.* 2011, 412, 2602–2610. [CrossRef]  
3. Albert, R.; Albert, I.; Nakarado, G.L. Structural vulnerability of the North American power grid. *Phys. Rev. E* 2004, 69, 25103. [CrossRef] [PubMed]  
4. Shah, D.; Zaman, T. Rumors in a Network: Who’s the Culprit? *IEEE Trans. Inf. Theory* 2011, 57, 5163–5181. [CrossRef]  
5. Zhu, K.; Ying, L. Information source detection in the SIR model: A sample-path-based approach. *IEEE ACM Trans. Netw.* 2016, 24, 408–421. [CrossRef]  
6. Lokhov, A.Y.; Mézard, M.; Ohta, H.; Zdeborová, L. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Phys. Rev. E* 2014, 90, 12801. [CrossRef]  
7. Altarelli, F.; Braunstein, A.; Dall’Asta, L.; Lage-Castellanos, A.; Zecchina, R. Bayesian inference of epidemics on networks via Belief Propagation. *Phys. Rev. Lett.* 2014, 112, 118701. [CrossRef] [PubMed]  
8. Brockmann, D.; Helbing, D. The Hidden Geometry of Complex, Network-Driven Contagion Phenomena. *Science* 2013, 342, 1337–1342. [CrossRef]  
9. Zhu, K.; Ying, L. Source Localization in Networks: Trees and Beyond. *arXiv* 2015, arXiv:1510.01814.  
10. Hu, Z.L.; Han, X.; Lai, Y.C.; Wang, W.X. Optimal localization of diffusion sources in complex networks. *R. Soc. Open Sci.* 2017, 4, 170091. [CrossRef]  
11. Luo, W.; Tay, W.P.; Leng, M. Identifying Infection Sources and Regions in Large Networks. *IEEE Trans. Signal Process.* 2013, 61, 2850–2865. [CrossRef]  
12. Wang, Z.; Dong, W.; Zhang, W.; Tan, C.W. Rumor source detection with multiple observations: Fundamental limits and algorithms. In Proceedings of the 2014 ACM International Conference on Measurement and Modeling of Computer Systems, Austin, TX, USA, 16–20 June 2014; Volume 42, pp. 1–13.  
13. Luo, W.; Tay, W.P.; Leng, M. How to Identify an Infection Source With Limited Observations. *IEEE J. Sel. Top. Signal Process.* 2014, 8, 586–597. [CrossRef]  
14. Antulov-Fantulin, N.; Lancic, A.; Smuc, T.; Stefancic, H.; Sikic, M. Identification of Patient Zero in Static and Temporal Networks: Robustness and Limitations. *Phys. Rev. Lett.* 2015, 114, 248701. [CrossRef] [PubMed]  
15. Louni, A.; Subbalakshmi, K.P. A two-stage algorithm to estimate the source of information diffusion in social media networks. In Proceedings of the 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, 27 April–2 May 2014; pp. 329–333.  
16. Li, X.; Wang, X.; Zhao, C.T. Locating the Epidemic Source in Complex Networks with Sparse Observers. *Appl. Sci.* 2019, 9, 3644. [CrossRef] [PubMed]  
17. Zejnilovic, S.; Gomez, J.P.; Sinopoli, B. Network observability and localization of the source of diffusion based on a subset of nodes. In Proceedings of the 2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 2–4 October 2013; pp. 847–852.  
18. Zejnilovic, S.; Xavier, J.M.F.; Gomez, J.P.; Sinopoli, B. Selecting observers for source localization via error exponents. In Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, China, 14–19 June 2015; pp. 2914–2918.  
19. Pinto, P.C.; Thiran, P.; Vetterli, M. Locating the Source of Diffusion in Large-Scale Networks. *Phys. Rev. Lett.* 2012, 109, 68702. [CrossRef] [PubMed]  
20. Zhu, K.; Chen, Z.; Ying, L. Locating the contagion source in networks with partial timestamps. *Data Min. Knowl. Discov.* 2016, 30, 1217–1248. [CrossRef]  
21. Shen, Z.; Cao, S.; Wang, W.-X.; Di, Z.; Stanley, H.E. Locating the source of diffusion in complex networks by time-reversal backward spreading. *Phys. Rev. E* 2016, 93, 32301. [CrossRef]  
22. Fu, L.; Shen, Z.; Wang, W.-X.; Fan, Y.; Di, Z. Multi-source localization on complex networks with limited observers. *EPL* 2016, 113, 18006. [CrossRef]  
23. Hu, Z.-L.; Shen, Z.; Tang, C.-B.; Xie, B.-B.; Lu, J.-F. Localization of diffusion sources in complex networks with sparse observations. *Phys. Lett. A* 2018, 382, 931–937. [CrossRef]  
24. Rossi, R.A.; Ahmed, N.K. The Network Data Repository with Interactive Graph Analytics and Visualization. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.  
25. Lusseau, D. The emergent properties of a dolphin social network. *Proc. Biol. Sci.* 2003, 270, S186–S188. [CrossRef]
26. Social Network Analysis Software & Services for Organizations, Communities, and Their Consultants. Available online: http://www.orgnet.com (accessed on 8 September 2008).
27. Girvan, M.; Newman, M.E.J. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 2002, 99, 7821–7826. [CrossRef] [PubMed]
28. Leskovec, J.; Lang, K.J.; Dasgupta, A.; Mahoney, M.W. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Internet Math. 2009, 6, 29–123. [CrossRef] [PubMed]
29. Gleiser, P.M.; Danon, L. Community structure in jazz. Adv. Complex Syst. 2003, 6, 565–573. [CrossRef]
30. Newman, M.E.J. Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 2006, 74, 36104. [CrossRef] [PubMed]
31. White, J.G.; Southgate, E.; Thomson, J.N.; Brenner, S. The structure of the nervous system of the nematode Caenorhabditis elegans. Philos. Trans. R. Soc. B 1986, 314, 1–340. [CrossRef] [PubMed]
32. Network Data. Available online: http://www-personal.umich.edu/~mejn/netdata/ (accessed on 19 April 2013).
33. Freeman, L.C. A set of measures of centrality based on betweenness. Sociometry 1977, 40, 35–41. [CrossRef]
34. Bavelas, A. Communication patterns in task-oriented groups. J. Acoust. Soc. Am. 1950, 22, 725–730. [CrossRef]
35. Tang, W.; Ji, F.; Tay, W.P. Estimating Infection Sources in Networks Using Partial Timestamps. IEEE Trans. Inf. Forensics Secur. 2018, 13, 3035–3049. [CrossRef]
36. Zang, W.; Zhang, P.; Zhou, C.; Guo, L. Locating multiple sources in social networks under the SIR model: A divide-and-conquer approach. J. Comput. Sci. 2015, 10, 278–287. [CrossRef]
37. Zang, W.; Zhang, P.; Zhou, C.; Guo, L. Discovering Multiple Diffusion Source Nodes in Social Networks. Procedia Comput. Sci. 2014, 29, 443–452. [CrossRef]
38. Tang, W.; Ji, F.; Tay, W.P. Multiple sources identification in networks with partial timestamps. In Proceedings of the 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, Canada, 14–16 November 2017; pp. 638–642.
39. Kivelä, M.; Arenas, A.; Barthelemy, M.; Gleeson, J.P.; Moreno, Y.; Porter, M.A. Multilayer networks. J. Complex Netw. 2014, 2, 203–271. [CrossRef]
40. Holme, P. Modern temporal network theory: A colloquium. Eur. Phys. J. B 2015, 88, 1–30. [CrossRef]