Localizing the Information Source in a Network

Guanyu Nie
nieg@purdue.edu
Purdue University
West Lafayette, Indiana

Christoper Quinn
cjquinn@purdue.edu
Purdue University
West Lafayette, Indiana

ABSTRACT

Information and content can spread in social networks analogous to how diseases spread between organisms. Identifying the source of an outbreak is challenging when the infection times are unknown. We consider the problem of detecting the source of a rumor that spread randomly in a network according to a simple diffusion model, the susceptible-infected (SI) exponential time model. The infection times are unknown. Only the set of nodes that propagated the rumor before a certain time is known. Since evaluating the likelihood of spreads is computationally prohibitive, we propose a simple and efficient procedure to approximate the likelihood and select a candidate rumor source. We empirically demonstrate our method out-performs the Jordan center procedure in various random graphs and a real-world network.

CCS CONCEPTS

• **Mathematics of computing** → **Maximum likelihood estimation**; • **Applied computing** → **Sociology**.

KEYWORDS

Complex networks, information source, maximum likelihood (ML) estimator, sparse graph

ACM Reference Format:
Guanyu Nie and Christoper Quinn. 2019. Localizing the Information Source in a Network. In *Proceedings of Truth Discovery and Fact Checking: Theory and Practice (TrueFact 2019).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the advent of internet, and specifically online social networking platforms, information and ideas can spread rapidly. These decentralized diffusions can be beneficial, allowing citizens to circumvent traditional mediums such as radio and television to quickly broadcast information. However, they can also allow rumors to spread to a large audience before fact-checking can be performed and corrective information disseminated to mitigate any significant damage accidentally or intentionally false rumors can cause.

In this paper, we consider the problem of identifying the source of the information diffusion when timing information is not known. Instead, we know the set of nodes that propagated the rumor, e.g. the "infected" nodes, before a certain time. We also know the diffusion model, a stochastic susceptible-infected (SI) model with independent exponential spreading times. In principle, we could compute the likelihood of each infected node as the source. However, that is challenging even for simple network topologies such as chains and trees. For networks with non-tree topologies, for which there are many paths the diffusion could have taken, it is necessary to develop alternative procedures.

Shah and Zaman were the first to study this problem under the same diffusion model [1]. They focused on networks with tree topologies and proposed a novel centrality metric, known as rumor centrality, for ranking candidate rumor sources. They proved the best candidate according to rumor centrality had the highest likelihood for regular trees (uniform degrees). They proved their results to be asymptotically good for regular trees and geometric trees (where the tree grows polynomially). They extended their results to more general graphs using a bread-first search heuristic which performed well on many random graphs and real world networks.

In [2], [3] and [4], more complex propagation models were investigated, susceptible-infected-susceptible (SIS), susceptible-infected-recovered (SIR), susceptible-infected-recovered-infected (SIRI) respectively. It was also shown in [5] that under the SI, SIR and SIRI models, the Jordan center is the infection source in a tree-network associated with the most likely infection path with a single infection source. Hence, Jordan centers are considered "universal" information source estimators for trees. The Jordan center of a graph is the node whose longest shortest-path to any other node is minimal. Other methods in estimating the source of SIR model include dynamic message passing (DMP) [6] and belief propagation [7].

One generalization of this problem was to consider multiple sources of a single diffuion. The result of [3] is not restricted to assuming a single source. In [8], the number of infection sources was estimated. In [9–11], the problem of detecting multiple sources was studied. Another generalization was to allow for only a portion of the infected nodes to be observed, randomly in [3] and arbitrarily in [12]. The estimator proposed in [12] is optimal for geometric trees as described above when the observation rate (proportion of observed nodes) is greater than 0. See [13] for a detailed survey.

We propose a simple and efficient procedure to estimate the rumor source. To circumvent the computational difficulty of evaluating the likelihood of node infections on a general graph, we use a sparse approximation. For each candidate source, we approximate the likelihood of that source infecting or not each of the other nodes in the graph independently. We formally describe our problem setup in Section 2. Rumor source detection on tree graphs is studied in Section 3 and extended to general graphs in Section 4. Empirical results on random graphs and a subgraph of the Facebook network are presented in Section 5. We conclude the paper in Section 6.

TrueFact 2019, August 5, 2019, Anchorage, AK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 PROBLEM SETUP

In this section, we discuss the information spreading model and the maximum likelihood estimator. The network is represented as an undirected graph $G(V, E)$ where V represents the set of nodes (e.g. users) in the network and E represents the set of edges (e.g. user-defined "friend" relationships).

2.1 Information Spreading Model

We use the SI infection model where nodes in the graph are either "susceptible" (have not yet heard the information) or "infected" (have already heard the information). Once a node receives the information, the node can transmit the information along each of its edges to its neighbors. Let τ_{ij} denote the time it takes for the information to spread from from node i to another node j directly. We assume that τ_{ij} can be modeled as an exponential random variable with rate parameter λ for all $(i, j) \in E$, and all the τ_{ij}'s are independent and identically distributed. We assume only one node, denoted v_0, is the information source. An example network of this model is shown in Figure 1.

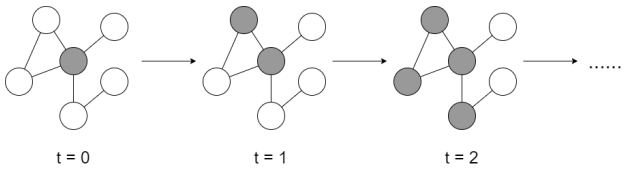


Figure 1: Example network illustrating information spreading model.

2.2 Maximum Likelihood (ML) Source Estimator

We consider the setting where one source v_0 initiates the information spreading on graph G at time 0 and after some time T we get a snapshot of the network. Let N denote the number of nodes infected by time T. We denote the snapshot (graph G with binary labels) as G_N. The key idea of maximum likelihood estimator is to find the node v that has the highest probability of resulting in G_N. We assume that there is no (informative) prior on which node might be the source. Thus, the maximum-a-priori estimate is also the maximum likelihood estimate, e.g. $\mathbb{P}(G_N|v) = \mathbb{P}(v|G_N)$ where $\mathbb{P}(G_N|v)$ is the probability of observing G_N given v being the source and $\mathbb{P}(v|G_N)$ defined similarly. In this setting, the maximum likelihood (ML) estimator of v_{ML} given G_N maximizes the correct detection probability, i.e.,

$$v_{\text{ML}} \in \arg\max_{v \in G_N} \mathbb{P}(G_N|v). \tag{1}$$

Evaluating $\mathbb{P}(G_N|v)$ is computationally hard. Assuming only one node can be infected at each time step. We need to calculate the probability of all possible permutations of infected nodes representing infection paths, which requires $O(N!)$ space and time in worst case. It is even more complicated when more vertices can be infected at one time step.

3 TREES

As discussed above, Evaluating $\mathbb{P}(G_N|v)$ is computationally expensive. We next describe a procedure to approximate $\mathbb{P}(G_N|v)$ for trees.

Suppose the underlying graph G is a tree. At time 0, node v_0 begins a rumor spread. At time T, we observe G_N. Let I denote the infected set in G_N. One approach to efficiently approximate $\mathbb{P}(G_N|v)$ is to calculate the probability of G_N assuming conditional independence given a candidate source v:

$$\begin{aligned}\mathbb{P}(G_N|v) &= \mathbb{P}(i \in I, j \in G \setminus I \,|v_0 = v) \\ &\approx \prod_{i \in I} \mathbb{P}(i \in I \,|v_0 = v) \prod_{j \in G \setminus I} \mathbb{P}(j \in G \setminus I \,|v_0 = v). \end{aligned} \tag{2}$$

Suppose for example, that G_N, observed at time T, corresponded to the network shown in Figure 2, with I corresponding to the shaded nodes. The actual likelihood of this G_N conditioned on node 1 being the source is

$$\begin{aligned}&\mathbb{P}(\{1,2,3,5\} \in I \text{ and } \{4,6,7\} \notin I \,|v_0 = 1) \\ &\quad = \mathbb{P}(\tau_{1,2} < T, \tau_{1,2} + \tau_{2,5} < T, \tau_{1,3} < T, \tau_{1,2} + \tau_{2,4} > T, \\ &\qquad \tau_{1,2} + \tau_{2,6} > T, \tau_{1,3} + \tau_{3,7} > T \,|v_0 = 1) \\ &\quad \approx \mathbb{P}(\tau_{1,2} < T \,|v_0 = 1)\, \mathbb{P}(\tau_{1,2} + \tau_{2,5} < T \,|v_0 = 1) \\ &\qquad \mathbb{P}(\tau_{1,3} < T \,|v_0 = 1)\, \mathbb{P}(\tau_{1,2} + \tau_{2,4} > T \,|v_0 = 1) \\ &\qquad \mathbb{P}(\tau_{1,2} + \tau_{2,6} > T \,|v_0 = 1)\, \mathbb{P}(\tau_{1,3} + \tau_{3,7} > T \,|v_0 = 1)\end{aligned}$$

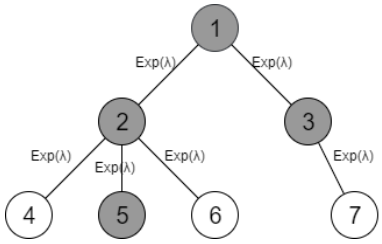


Figure 2: Example network

Now the problem left is to evaluate each term on the right-hand side in (2). Recall that the transition time between two neighboring nodes τ_{ij} is an exponential distribution with rate parameter λ. Also notice that the sum of n exponential random variables with with rate parameter becomes a gamma distribution random variable with shape parameter n and rate parameter λ. Thus the infection time of every node in G can be seen as a random variable with gamma distribution, i.e., the probability of one node receives rumor from another node n edges away within time t becomes

$$F(t; n, \lambda) = \frac{\gamma(n, \lambda t)}{\Gamma(n)}$$

Where $\gamma(n, \lambda t)$ is the lower incomplete gamma function defined as:

$$\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt$$

More precisely, we shall extend the original tree graph to a star graph with different transition time distribution. Again let's consider the simple example shown in Figure 2. Under the hypothesis

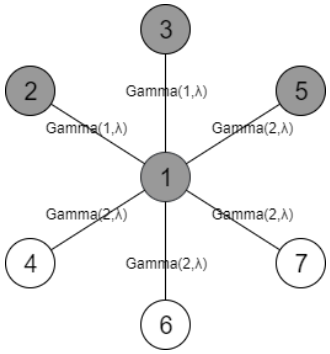


Figure 3: Converted star graph

that node 1 is the source, the original graph can be converted to the network shown in Figure 3.

Let Γ_{ij} be the random variable representing the time it takes for node j to receive rumor from node i in the star graph. For this approximation, the Γ_{ij}'s are independent random variables with gamma distribution with parameters d and λ, where d is the distance (number of edges) between two nodes i and j. Since the underlying graph is a tree, there is no ambiguity of parameter d (there is only one path from one node to another). We can calculate each node being infected or not accordingly and plug it into (2),

$$\mathbb{P}(\{1,2,3,5\} \in I \text{ and } \{4,6,7\} \notin I \,|v_0 = 1) \approx [F(T;1,\lambda)]^2\, F(T;2,\lambda)\,[1-F(T;2,\lambda)]^3$$

4 GENERAL GRAPHS

Now we want to generalize the procedure to general graphs. The challenging part is that there might be multiple, possibly overlapping paths from one node to another, which gives random variables with different gamma distributions. To take this into account, we will construct a simple heuristic.

Our heuristic is described in Algorithm 1, and based on the following simple idea. The distance between two nodes may take many different values, but it may be that a majority of them have a common value. We approximate the rumor transition between two nodes as always going through the shortest path between them, which correspond to the fastest or most probable spreading of the rumor in general. A simple example is shown in Figure 4. To approximate the likelihood of node i as the source, we will form a star graph analogous to the tree-graph case in Figure 2. For the edge (i, j) in the star graph, we use the distribution of infections propagating along the shortest path from i to j in the original graph G.

5 DATA ANALYSIS

5.1 Random Graphs

Many real world graphs have various statistical properties, such as small diameter, high clustering coefficients, modularity, and power-law degree distributions (or hubs) [14]. To evaluate the performance of our proposed method, we will examine its performance under various random graph models, where we will both know the ground truth and be able to identify for what graph properties our method works well.

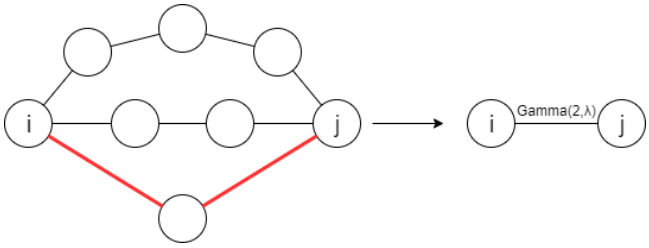


Figure 4: When there multiple paths from node i to j, we select shortest path when converting to star graph.

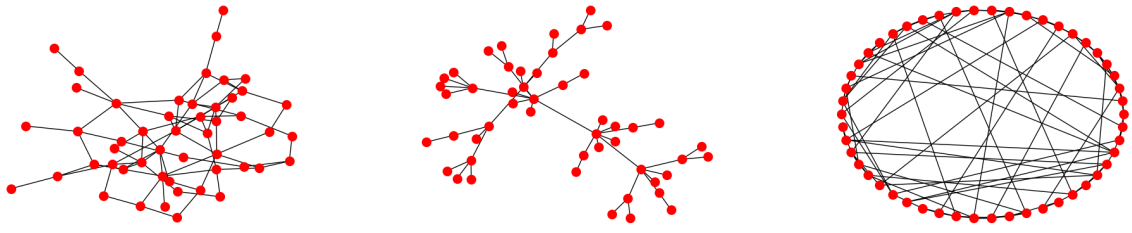
Algorithm 1: A source detection algorithm

Input: G (network graph), I (infected nodes), T (total propagation time)
Output: rumor source estimate

```
initialization;
p ← {};
source ← v ∈ I;
forall v ∈ I do
    p(v) ← 1;
    forall u ∈ G.nodes do
        n ← ShortestPath(v, u);
        if u ∈ I then
            p(v) ← p(v) * F(T; n, λ);
             /* F is the cdf of gamma distribution */
        else
            p(v) ← p(v) * (1 − F(T; n, λ));
        end
    end
end
return source ← arg max_{v∈I} ℙ(v)
```

5.1.1 Random graph models: Erdos-Renyi model. In the Erdos-Renyi model [15], a graph G is constructed by connecting nodes randomly. For each possible edge, an i.i.d. Bernoulli random variable with success probability p is drawn. If successful, the corresponding edge is included. Figure 5a shows an Erdos-Renyi Graph with 50 nodes and $p = 0.08$.

5.1.2 Random graph models: Barabasi-Albert model. The Barabasi-Albert (BA) model [16] is used to generate scale-free networks. The degree distribution follows a power-law and has a non-vanishing tail (e.g. some nodes have very high degree relative to the others). The network is constructed from m_0 initial nodes. Nodes are added one at a time. Each new node is connected to $m \leq m_0$ existing nodes with probability proportional to the degree of existing node, i.e., existing nodes with higher degree have a higher chance of getting connected by new nodes. Figure 5b shows a Barabasi-Albert graph with 50 nodes and $m = 1$.



(a) Erdos-Renyi (b) Barabasi-Albert (c) WattsStrogatz

Figure 5: Example random graphs

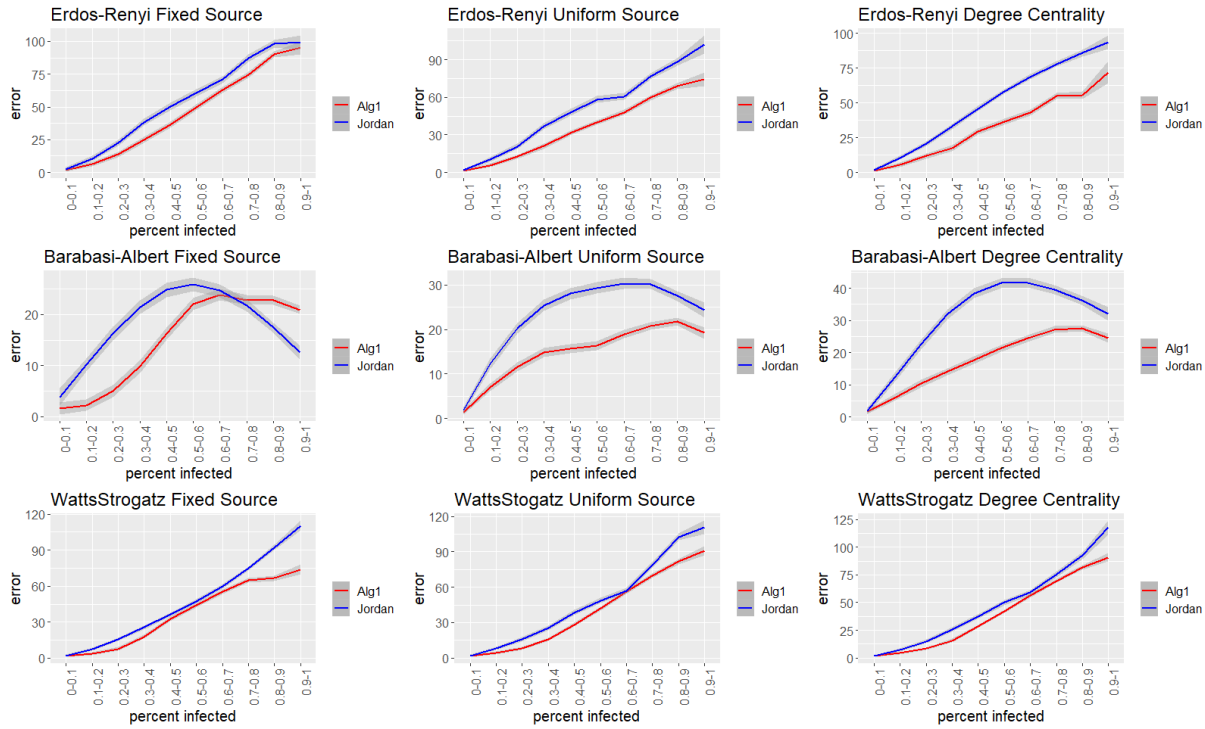


Figure 6: Three rows show results on Erdos-Renyi graph, Barabasi-Albert graph and WattsStrogatz graph, respectively. In first column, the source is fixed when doing simulation. For second column, we select true source uniformly at random. For the third column, the nodes with higher degree centrality have more chance to be selected as source.

5.1.3 Random graph models: Watts-Strogatz model. The Watts-Strogatz model [17] is used to generate networks with the "small-worlds" property. Roughly speaking, this kind of network has both short average path lengths and high clustering. Given the parameters N, K and β, The graph is generated as follows: construct a regular ring lattice, a graph with N nodes each connected to K neighbors, $K/2$ on each side; then for every node, take its rightmost $K/2$ edges, and rewire it with probabilty β. Figure 5c shows a Watts-Strogatz graph with $N = 50$, $K = 4$ and $\beta = 0.4$.

5.2 Rumor spreads on random graphs

5.2.1 Setup. We perform simulations on random graphs with 300 nodes. The parameters were selected so that the graphs all had similar average degree. We selected the ground-truth source using three methods: (1) randomly pick a source and use it throughout all diffusions; (2) uniformly choose sources from infected nodes before each diffusion; (3) chooses sources according to degree centrality (nodes with larger degree centrality have more chance to be picked) before each diffusion. We generated each plot using 8000 diffusion simulations.

5.2.2 Results. Figure 6 shows the results of Algorithm 1 against the Jordan center. Jordan center was calculated by built in algorithm of networkx [18]. We plotted error vs. percentage of infected nodes in the network, where error is measured by the order that the estimator is infected. For example, if in a trial one of the methods selected the 21st node that was infected as the rumor source candidate, that would be an error with value 20. The shaded are corresponds to 95% confidence area for the curve fit using the *geom_smooth* method in the R package ggplot2. Overall, Algorithm 1 performs better in almost all settings examined than the Jordan center. For scale-free networks (BA), our proposed method outperforms Jordan center by a wide margin. A possible explanation is that in scale-free networks, the high-degree "hub" nodes tend to have high Jordan-centrality. Thus, the Jordan center method would often pick hubs that were infected. Our method accounts not only for proximity to infected nodes but also distance from uninfected nodes.

5.3 Real-world graph

5.3.1 Data-set description: We also performed simulations on a data set collected from Facebook [19].[1] The network was collected from survey participants using a Facebook app. Two nodes are connected when the users they represent have same political affiliations. The network has 4039 nodes and 88234 links. The network is shown in Figure 7. For simplicity, we used community detection algorithms and used one community as our network. The community has 372 nodes and 2929 edges.

5.3.2 Setup and results: For each method, we ran 500 diffusions. The ground truth source was picked uniformly at random. The results are shown in Figure 8. Our proposed method performs similarly with Jordan center for small diffusions, but much better than Jordan center for larger diffusions. Similar to the performance of the scale-free graphs, we hypothesize the performance gain is because our proposed Algorithm 1 seeks to select a candidate source by balancing closeness to infected nodes with distance to uninfected nodes, while the Jordan central method only seeks the former.

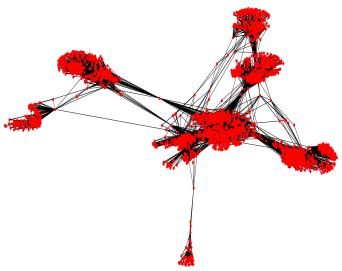


Figure 7: Facebook network

6 CONCLUSION

In this work, we proposed a novel, efficient information source estimator using star-graph approximations. We verified on both random graph models and on a subgraph of the Facebook network

[1] Available at http://snap.stanford.edu/data/ego-Facebook.html

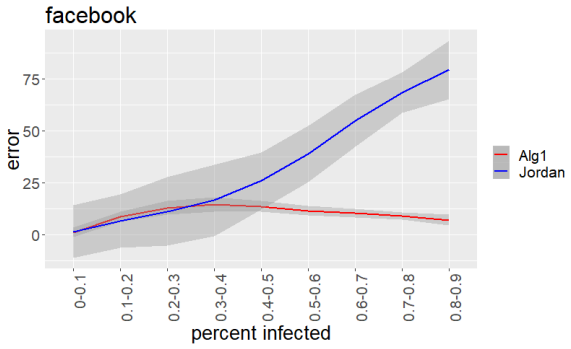


Figure 8: Simulation result on real-world data

that our approach in some situations is slightly better than, and in other cases significantly outperforms, the Jordan center.

REFERENCES

[1] D. Shah and T. Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory*, 57(8):5163–5181, Aug 2011.

[2] W. Luo and W. P. Tay. Finding an infection source under the sis model. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2930–2934, May 2013.

[3] Kai Zhu and Lei Ying. Information source detection in the sir model: A sample-path-based approach. *IEEE/ACM Trans. Netw.*, 24(1):408–421, February 2016.

[4] W. Hu, W. P. Tay, A. Harilal, and G. Xiao. Network infection source identification under the siri model. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1712–1716, April 2015.

[5] W. Luo, W. P. Tay, and M. Leng. On the universality of jordan centers for estimating infection sources in tree networks. *IEEE Transactions on Information Theory*, 63(7):4634–4657, July 2017.

[6] Andrey Y. Lokhov, Marc Mézard, Hiroki Ohta, and Lenka Zdeborová. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Phys. Rev. E*, 90:012801, Jul 2014.

[7] Fabrizio Altarelli, Alfredo Braunstein, Luca Dall'Asta, Alejandro Lage-Castellanos, and Riccardo Zecchina. Bayesian inference of epidemics on networks via belief propagation. *Phys. Rev. Lett.*, 112:118701, Mar 2014.

[8] F. Ji, W. P. Tay, and L. R. Varshney. Estimating the number of infection sources in a tree. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 380–384, Dec 2016.

[9] F. Ji and W. P. Tay. Identifying rumor sources with different start times. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–5, June 2016.

[10] W. Luo and W. P. Tay. Identifying infection sources in large tree networks. In *2012 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, pages 281–289, June 2012.

[11] W. Luo, W. P. Tay, and M. Leng. Identifying infection sources and regions in large networks. *IEEE Transactions on Signal Processing*, 61(11):2850–2865, June 2013.

[12] N. Karamchandani and M. Franceschetti. Rumor source detection under probabilistic sampling. In *2013 IEEE International Symposium on Information Theory*, pages 2184–2188, July 2013.

[13] J. Jiang, S. Wen, S. Yu, Y. Xiang, and W. Zhou. Identifying propagation sources in networks: State-of-the-art and comparative studies. *IEEE Communications Surveys Tutorials*, 19(1):465–481, Firstquarter 2017.

[14] Mark Newman. *Networks: an Introduction*. Oxford University Press, 2010.

[15] P. Erdös and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

[16] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, Jan 2002.

[17] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[18] Aric A. Hagberg, Daniel A Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *In Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena, CA USA*, pages 11–15. SciPy, 2008.

[19] Jure Leskovec and Julian J. Mcauley. Learning to discover social circles in ego networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 539–547. Curran Associates, Inc., 2012.