Origin Estimation Model and Algorithm of Rumor Source in Social Networks

Keyi Pan*
University of Washington Department of Statistics, Seattle, United States
*Corresponding author e-mail: gerrykeyip@gmail.com

Abstract. In the network model, few put effort to study the methodology to find the source of diffusion while many topics on information diffusion are discussed in depth. To find the source of rumor is a different story from other problem like how to estimate the effect of rumor. The rumor source tracking is a substantially important problem since it is closely connected to the web security and social steadiness during convulsion. When the existing theorem based on diffusion problem cannot do much help on tracking rumor source, some novel methodologies using mathematical modeling and monte-carlo simulation can tackle the problem systematically. In this paper, we first introduce a target variable, a error function, to evaluate the estimation by comparing the difference of of infected spots which affected by the rumor. Then we proved the complexity of source backtracking problem. We imported the Approxmation Algorithm to optimize the monte-carlo trials for the convergence of minimized error function. For our experiments on email records of European research institutions, we implement our error function checking criteria to verify the effectiveness and suitability of our framework, as well as the evaluation of some special situation and advice for further improvement.

Keywords: The Rumor Source Tracking, Error Function, the Origin of the Rumor

1. Introduction
Currently, massive information are exchanged between different users on the Internet. While many online supervisions are made to guarantee the effectiveness and authenticity of news, many kinds of rumors are still spreading widely to mislead the public. Based on many ways of communication such as email, social media, and messaging Apps, rumor can spreading by users rapidly and unconsciously. Uncontrolled rumor will cause severe problems on the internet and society. Recently, the spreading of many unauthorized information about Covid-19 becomes a big challenge to the social security. Therefore, to find the origin of the rumor become an imperative problem.

Many affiliated researches focus on the social network model where emphasizes the detailed social model with high-degree realism to simulate the interactions among individuals in specific regions based on the considerations of their social properties such as population composition and group classification. Andrea et al [1]. provide a insightful research on information spreading via personal conversion focus on the time and locations of their meetings, with various social factors such as household income, age and social status. She provide an activities-density-based social network model, consist of a probability of interaction between individuals based on social relationship.
Following Blau and Schwartz [3], Damon Centola [2] develops a model based on the social structures to improve the social diffusion by moderating to high levels of consolidation. While most diffusion analysis concentrates on the real world condition supported by sociological structure, it is necessary to put a new sight on the information diffusion on the internet since it is more versatile and contagious. To deal with the relative unrestricted diffusion of rumor on the internet, a model specialized in finding the origin of a diffusion of network model based on the information exchange on the internet empower the network security by filtering the futile information.

We model this problem as diffusion backtrack problem. Given a diffusion network sample, which record the communication between each individuals, our goal is to find an optimal method to estimate the rumor source. The estimation should be the node with the maximum likelihood to form this network sample. In this paper, the problem is considered under the IC model.

2. Diffusion Backtrack Model

We use independent cascade(IC) model to analyze the origin of the false information diffusion. For an inactive node W, each of its neighbors \( V_i \) has a single chance \( P_{v_i,w} \) to activate the node W.

Consider a social network \( G(V,E,P) \), where \( V \) is the set of nodes and \( E \) is the set of edges. And \( P_{v,w} \) denotes the probability that w is activated by v in a single trial.

2.1 Origin Estimation

Define \( o \) as the origin of the rumor, and define rumor diffusion area \( A(o) \) as the set of nodes that all being activated(believed). Define \( \text{Sign}(v) = 1 \) as v has being activated. \( \text{Sign}(v) = 0 \) otherwise.

\[
A(o) = \{v | v \in V, \text{sign}(v) = 1\}
\]

2.2 Simple Mismatch Estimation

In the same network model \( G(V,E,P) \), suppose \( v_0 \) as the possible rumor source. Define the mismatch value of \( v_0 \) as

\[
\nu_0 = \mathbb{E}(n(A(v_0) + A(o) - A(v_0) \cap A(o)))
\]

Where \( \mathbb{E}(\cdot) \) is the expectation operator, \( n(\cdot) \) is the cardinality of the set.We estimate the origin as the \( \hat{o} \) such that

\[
\hat{o} = \arg \min_{o \in V} \nu_o
\]

2.3 Weighted Mismatch Estimation

We consider each mismatch nodes with different weights. One method for this is to divide the mismatching nodes into two groups. One group \( C_o \) as the set of nodes in \( A(o) \) but not in \( A(v_0) \). The other group \( C_x \) as the set of nodes in \( A(v_0) \) but not in \( A(o) \).

For a mismatching node \( m \in C_o \cup C_x \) such that \( m \in A(v_0) + A(o) - A(v_0) \cap A(o) \), \( A(o) \cdot P_{o,m} \) as the probability that m being activated with the origin o.

\[
W(m) = \max \left\{ \frac{P_{o,m}}{P_{v_0,m}}, \frac{P_{o,m}}{P_{v_0,m}} \right\}
\]

Then define the new mismatch value of \( v_0 \) as

\[
\nu^*(v_0) = \mathbb{E}( \sum_{m \in C_o \cup C_x} W(m) )
\]

We estimate the origin as the \( \hat{o}^* \) such that

\[
\hat{o}^* = \arg \min_{o \in V} \nu^*(v)
\]
3. Hardness Results

Theorem 1. Source backtracking problem is \#P-hard under the IC model.

Proof. We prove by showing one special case of source backtracking problem is \#P-hard, reducing from the influence maximization problem under IC model.

Given such a condition of source backtracking problem: A sample Graph \( G(V, E, P) \) and a source \( s_0 \), we suppose \( A(s_0) = V \). So our situation is all nodes have been activated by a single source. Our target is to find the source node \( s_0 \). Recall the influence maximization problem with a single seed formulated by Kempe, Kleinberg, and Tardos as follows: Under the independent cascade (IC) model, the influence maximization problem is to find a node in the graph such that the expected number of the nodes activated by this node is largest possible, which is proved as \#P-hard by Wei, Chi, and Yajun [4][5][6]. This problem is equivalent to our assumed situation with \( A(s_0) = V \). So to find \( s_0 \) to satisfy the situation is also \#P-hard. Therefore source backtracking problem is at least \#P-hard.

3.1 Sample Complexity

We will use the Approximation Algorithm \( \mathcal{A}_4 \) introduced by P. Dagum et al. [7] to restrict the estimate error of our monte carlo approximation by optimizing the number of trials required.

First we need to transform our mismatch error expression to satisfy \( \mathcal{A}_4 \). Recall that for \( v_0 \) as a possible rumor source,

\[
\mathcal{E}(v_0) = E(n(A(v_0) + A(o) - A(v_0) \cap A(o)))
\]

Denote \( Z_t \) as

\[
Z_{v_0,i} = \frac{n(A(v_0) + A(o) - A(v_0) \cap A(o))}{n(v)}
\]

Then \( Z_{v_0,1}, Z_{v_0,2}, Z_{v_0,3}, \ldots \) are independently and identically distributed according to \( Z \) in the interval \([0,1]\) with mean \( \mu_z \), our target is to find \( \mu_z \), a \( \delta \) approximation of \( \mu_z \). That is given arbitrary, \( \delta > 0 \)

\[
P[\mu_z(1-\varepsilon) \leq \hat{\mu}_z \leq \mu_z(1+\varepsilon)] > 1 - \delta
\]

First we need to introduce Stopping Rule Theorem [7],

3.1.1 Theorem 1. The Stopping Rule Theorem:

Let \( Z_1, Z_2, Z_3, \ldots \) be independently and identically distributed according to \( Z \) in the interval \([0,1]\) with mean \( \mu_z > 0 \). Given arbitrary, \( \delta > 0 \), define \( \lambda = (\varepsilon / 2) \), \( \gamma = 4 \lambda \lambda \ln(2/\delta) \varepsilon^{-2} \), and let \( \gamma_1 = 1 + (1 + \varepsilon^2) \gamma \). When \( S > \gamma_1 \) (Stopping Rule Algorithm stopped), then,

1. \( P[\mu_z(1-\varepsilon) \leq \hat{\mu}_z \leq \mu_z(1+\varepsilon)] > 1 - \delta \)
2. \( E[N_\gamma] \leq \gamma / \mu_z \)

Approximation Algorithm Framework

Precondition: \( \varepsilon \), \( \delta \) with \( 0 < \varepsilon \leq 1 \) and \( 0 < \delta \leq 1 \), \( \gamma_2 = 2(1 + \sqrt{\varepsilon})(1 + 2\sqrt{\varepsilon})(1 + \ln(\frac{3}{2})/\ln(\frac{2}{\delta})) \gamma_1 \) for small and \( \delta \). Set two sets of iid random variables \( Z_1, Z_2, Z_3, \ldots \) and \( Z_1^*, Z_2^*, Z_3^*, \ldots \)

**Step1.** Run the Stopping Rule Algorithm on \( Z_{v_0,1}, Z_{v_0,2}, Z_{v_0,3}, \ldots \) with \( \varepsilon_0 = \min\left\{ \frac{1}{2}, \sqrt{\varepsilon} \right\} \) and \( \delta_0 = \frac{1}{3} \delta \)

We will get the \( \varepsilon \), \( \delta \) estimate \( \hat{\mu}_z \) of \( \mu_z \)

**Step2.** Set \( N = \gamma_2 \times \varepsilon / \hat{\mu}_z \), and set \( S=0 \), For i from to \( N \), do: \( S = S + (Z_{2i-1}^* - Z_{2i}^*)^2 / 2 \), Set \( \hat{\mu}_z = \max\{S / N, \in \hat{\mu}_z \} \)
Step 3. Set $\hat{N} = \gamma_z \times \hat{\rho}_z / \hat{\mu}_z^2$ and set $S^* = 0$, For i from 1 to $N$, do: $S^* = S^* + Z_i$, Set $\hat{\mu}_z = S/N$.

3.1.2 Theorem 2. AA Theorem:
Let $Z_1, Z_2, Z_3, \ldots$ be independently and identically distributed according to $Z$ in the interval $[0,1]$ with mean $\mu_z > 0$. Let the variance of $Z$ be $\sigma_z^2$ and $\rho_z = \max\{\sigma_z^2, \mu_z\}$. Let $\hat{\mu}_z$ be approximation of $AA$ and $N_z$ be the number of trials derived from $AA$ Framework ($\hat{N} = \gamma_z \times \hat{\rho}_z / \hat{\mu}_z^2$) with the given $\varepsilon, \delta$. Then,

1. $P[\mu_z(1 - \varepsilon) \leq \hat{\mu}_z \leq \mu_z(1 + \varepsilon)] > 1 - \delta$
2. There exists an constant $c'$ s.t $P[N_z \geq c' \gamma \rho_z / \mu_z^2] \leq \delta$
3. There exists an constant $c'$ s.t $E[N_z] \leq c' \gamma \rho_z / \mu_z$

Therefore, using AA with $N = \gamma_z \times \hat{\rho}_z / \hat{\mu}_z^2$, we can get, $\delta$ approximation of $\mu_z = E\left(\frac{n(A(\nu_0) + A(\sigma) - A(\nu_0) \cap A(\sigma))}{n(V)}\right) = \frac{1}{n(V)} \in (\nu_0)$ by optimally using the minimal trials of experiments within a constant factor to produce the $\varepsilon, \delta$ approximation on each possible choice.

4. Experiments
In this section, we evaluate our algorithm via an experiment on a data set in real world.

4.1 Settings
The experiment is based on data set available at the SNAP website (http://snap.stanford.edu). It is a network data describing communications among members from European research institution via email. All nodes are anonymized institution members and the frequencies are the count of emails between nodes. All messages between each node and individuals outside the data set are not included. Considering situations that some node pairs might appear in data set repeatedly with different frequencies, we will randomly choose one row of such cases to keep uniqueness of the count between two nodes.

Table 1. The first 6 rows of the dataset on probabilities of nodes interaction

| node1 | node2 | prob    |
|-------|-------|---------|
| 1     | 126   | 0.0703077 |
| 1     | 155   | 0.0968189 |
| 1     | 26    | 0.2127439 |
| 1     | 113   | 0.2661979 |
| 1     | 136   | 0.2798723 |
| 1     | 154   | 0.3922953 |

All algorithms in this experiment are based on basic data analysis with Pandas and graph algorithms like BFS. We implement our basic data manipulation and algorithm in Python, the experiment was conducted on a laptop with a 2.9GHz Intel Core i7-7820HK processor and 16GB memory, running Microsoft Windows 10.

4.2 Data Manipulation
We take the cleaned data set with single frequency in each node pairs via random selection. The data have 1772 rows. We also consider the frequencies as a one-to-one correspondence to the probability that two nodes may communicate.
Table 2. The summary of probabilities to decide the default graph model

|       | Min. | First.Qu | Median | Third.Qu. | Max. |
|-------|------|----------|--------|-----------|------|
|       | 0    | 0.0273   | 0.1389 | 0.36633   | 1    |

To simulate the common situation in reality, we made a default graph with every edge is formed when its probability is greater than the median of all probabilities. And we define this default graph as every edge have same distance with unit one because we ignore the difference between the cost of emails to different addresses. So the default map, a directed graph with equal-length edge, will be used when we consider the distance between estimation and the origin of rumor.

In order to reduce the bias comes from the random sample of rumor diffusion, we only count the nodes that have probability to be contracted as the nodes that will trust the rumor. The specific method for this intuition is to record the frequencies of all appeared nodes with one hundred random samples of a single-source diffusion and only consider those who are 90% or higher possible to be infected.

4.3 Estimation

By checking every possible option for rumor source, we take every node to be the test source and record its trace of diffusion. Further, we simulate the common realization of their diffusion by Monte Carlo Methods. Then we compare the theoretical result of diffusion with all experimental diffusion origined from each node and we get the conclusion of the efficacy of our error function.

Count for the Accuracies of Model in All Possible Rumor Source

Figure 1. The possibility that our estimation is the rumor source is 58.6%, and there is 80.9% possible that our estimation is within the 2 unit distance of the true rumor source, which can also be a insightful guide for censoring.
Convergence of rumor source as 4

Figure 2. When the rumor source is node 4, all estimations based on different number of monte carlo trials are within 2 unit distance of the node 4, as the values of error function of node 4 is decreasing as the trials grow. Moreover, the difference between errors of node 4 and estimations are also decrease.

5. Conclusion
The estimation based on this error function have a high accuracy to be within the 2 unit distances of the true resource. It can help to target the rumor source more efficiently by reducing the scope of suspects. Since this model focus on the communication among the agencies or institutions of similar social group, improvements of this estimation might need the consideration of sociological considerations of historical credibility or information hierarchy, which can be factor to change the probability of diffusion.

Reference
[1] A. Apolloni et al., “A Study of Information Diffusion over a Realistic Social Network Model,” 2009 International Conference on Computational Science and Engineering, Vancouver, BC, 2009, pp. 675-682, doi: 10.1109/CSE.2009.240.
[2] D. Centola, “The Social Origins of Network and Diffusion,” American Journal of Sociology, vol. 120, no.5, 2015.
[3] P. M. Blau, “A Macrosociological Theory of Social Structure,” American Journal of Sociology, vol. 86, no.1, 1977.
[4] W. Chen, C. Wang, and Y. Wang, “Scalable influence maximization for prevalent viral marketing in large-scale social networks,” in SIGKDD. ACM, 2010, pp. 1029-1038
[5] Wang, C., Chen, W. & Wang, Y, “Scalable influence maximization for independent cascade model in large-scale social networks,” Data Min Knowl Disc 25, pp. 545–576, 2012. https://doi.org/10.1007/s10618-012-0262-1
[6] W. Chen, Y. Yuan, and L. Zhang, “Scalable influence maximization in social networks under the
linear threshold model,” in ICDM. IEEE, 2010, pp. 88-97

[7] M. Pradhan & P. Dagum. “Optimal Monte Carlo Estimation of Belief Network Inference,” Twelfth Conference on Uncertainty in Artificial Intelligence, UAI, 1996, arXiv:1302.3598v1