A Two-stage Algorithm to Estimate the Source of Information Diffusion in Social Media Networks

Alireza Louni
IEEE Student Member
Department of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ, USA.
Email: alouni@stevens.edu

K. P. Subbalakshmi
IEEE Senior Member
Department of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ, USA.
Email: ksubbala@stevens.edu

Abstract—We study the important problem of source localization in the context of information spreads in large social networks. Specifically, we design a Maximum-Likelihood source localization algorithm that is especially suited to large social networks. Our proposed algorithm requires about 3% fewer sensor nodes than other single stage algorithms for the same level of accuracy in detection. For practical social networks, which are typically large, this translates to a significantly fewer number of sensor nodes.

I. INTRODUCTION

Social networks play a crucial role in the diffusion of information and adoption of new technologies. About 54% of Internet users in the United States are on Facebook [1]. 48% of their real world contacts happens on Facebook and many of these users intermittently forward various pieces of information about new technologies/services through their friendship/subscription networks. Due to this large amount of online interaction, social networks are well suited to viral marketing and rapid information diffusion [2]. Therefore, it is of great interest to understand and model how information disseminates through social networks.

Some researchers have proposed models inspired by the spread of contagion in a population to model the spread of information in networks. Two such models include the susceptible-infected-susceptible (SIS) and the susceptible-infected-recovered (SIR) models [3]–[5]. In these models, information spreads among infected nodes (prior information adopter) and susceptible nodes. For example, in Twitter, the nodes that have received a tweet are considered infected whereas those that have not yet received the tweet are considered susceptible. Most of these works study the conditions under which a piece of information spreads across the entire network when not all nodes are infected and when some nodes are capable of recovering from infection. It is shown that information spreads to the entire network if the probability that an infected node will affect a susceptible node exceeds a threshold value (known as the epidemic threshold). In one of the models, a node adopts a piece of information if the fraction of neighbors who have already adopted it is greater than a threshold value. However, in many situations, people tend to observe the outcomes of prior adopters for a period and then decide on whether they will adopt the information or not. This subtlety is incorporated in other models in other works [3], [6].

Social networks can also be used to spread malicious gossip, untruthful information or computer malware. For example, a recent fake tweet about an explosion in the White House, caused the Dow Jones industrial average to drop 152 points within seconds [7]. Although significant research has been done on how a rumor becomes viral on Twitter, or a malware proliferates on the entire Facebook network, the reverse problem of identifying the source of diffusion has not been addressed very well. In this paper we will present a technique to determine the source of diffusion of information, given the structure of the network. We first review the existing literature on source localization in the next subsection before we present out motivation and model.

A. Related Works

The first generation techniques to find the source of information assumes that the source of information is likely to be a node with a high degree of centrality, where centrality is appropriately defined. For example, in [8], closeness, betweenness, and eigenvector centralities are used to locate the origin of the information. However, high central node as the source of diffusion may not be always true. In general, and without prior knowledge about the source of diffusion, we can only assume that the source is uniformly distributed over the network.

Another line of approach uses information from a snapshot of the infected nodes to identify a single source of diffusion. In this approach a maximum likelihood detector is designed to determine the single source of diffusion using information gathered about the nodes in the network (whether infected or susceptible) [9]–[13]. Similar approaches to detect multiple sources of information have also been proposed [14], [15]. While some works assume that all nodes in the network monitor and report their status, [16] only uses a subset of nodes (called sensors) in the network to determine the source location. A source is located by analyzing the arrival times of information at different sensor locations. It is shown in [16] that by monitoring 20% of the network, an average localization error of less than 4 hops can be achieved.
B. Our Motivation

Although the method described in [16] achieves good source localization accuracy, it requires a large percentage of nodes in the network to function as sensors. This will work for small networks, but when the network is very large, this will involve the willing participation of a large number of nodes as sensors. For example, Twitter has 41.7 million users. Twenty percent of this will amount to about 8 million nodes acting as sensors. This is, in general, not practical for large networks like Twitter. It is even more inefficient to observe the entire network as in [9]–[15]. It is therefore important for us to look for alternative solutions that will not require such a high percentage of nodes to act as sensors for the same level of accuracy in source localization. This paper proposes such an algorithm and is based on the fact that most real social networks exhibit a highly clustered topography. We propose a two-stage source localization algorithm where, in the first stage, a candidate cluster that is most likely to contain the source is identified. At the second stage, the source is located within the candidate cluster. It is later shown that when the desired performance in terms of the correct detection probability is given, our algorithm decreases the percentage of sensors significantly.

C. Organizations

This paper is organized as follows. In Section II-A, the information diffusion model is described. The ML estimator is discussed in Section II-B. The two-stage algorithm is designed to localize the single source, in Section III. Finally, numerical results and conclusion are presented in Section IV and V respectively.

II. Problem Formulation

A. Information Diffusion Model

In this subsection, we describe our social network model and assumptions. The information diffusion network is modeled by an undirected graph $G = (V, E)$, where $V$ is the set of nodes and $E$ is the set of edges in the graph. For example, in a social network, $V$ is the set of individuals in the network, and an edge exists between any two individuals if they are engaged in any kind of social relationship such as friendship or followership. The graph $G$ is a weighted graph, with $w(e)$, the weight associated with edge $e \in E$ representing the value of social relationship between the nodes connected by the edge $e$. The value of $w(e_{i,j})$ indicates the strength of the $i$ and $j$ ($i, j \in V$). It is assumed that the diffusion graph, $G$, is known.

Furthermore, we assume the susceptible-infected (SI) model to study information diffusion in the network. Any susceptible node can become infected independently of other nodes. The source of diffusion ($s^*$) is modeled as a uniformly distributed random variable over the set $V$. The unknown source of information, $s^*$, starts to spreads the information through the network at unknown time $t^*$. The time delay with which an infected node, say $m$, can infect a susceptible neighbor, say $n$, is Gaussian distributed $N(\mu_{m,n}, \sigma^2_{m,n})$ [17]. To preserve generality of analysis, we assume that the underlying network is non-homogenous. Hence, the average time delay $\mu_{m,n}$ takes on different values for different edges in the network. In general, $\mu_{m,n}$ is a function of the strength of social ties between $m$ and $n$ or $w(e_{i,j})$. For example, when the relationship between $m$ and $n$ is strong, it would take less time for $n$ to retweet a tweet from $m$ to its own followers. We also assume that information disseminates from the source $s^*$ to each node $v \in V$ along the shortest path connecting them [3]. Let $L = (l_1, l_2, ..., l_k)$ be a set of sensors used to estimate the source location where $l_i \in V$. The network graph is composed of close-knit clusters (or communities) of strong within-cluster ties and sparse and weak between-clusters ties (Fig. 1). Social influence between the nodes will obviously affect the spread of information within a network. In other words, nodes with strong ties interact frequently and influence each other more, so, average time delay $\mu_{m,n}$ along strong ties is smaller than weak ties.

B. Diffusion Source Estimator Design

Recent works propose an estimator to identify the most likely information source based on the knowledge of infection status of all nodes in the network. However, it is practically impossible to track the status of all nodes on the social network. Firstly, the computational complexity of the estimator increases greatly with the number of nodes, making it impractical for a typical social media network with millions of nodes. Secondly, we cannot extract the status of just any node on the social network due to privacy concerns. So, we are limited to track the status of a specific set of a limited number of nodes in the network. This small percentage of willingly participating nodes, hereafter called sensors, are chosen to estimate the source location. Sensors observe the arrival times of any piece

![Fig. 1. An example of a modular network graph $G$ with five clusters.](image-url)
of information from all their test candidates. A Maximum Likelihood estimator is then used to determine the most likely source of information. Since the time that the source starts to spread information, \( t^* \), is typically unknown, time difference between sensor pairs, \( d_i = (t_i + t^*) - (t_1 + t^*) = t_i - t_1 \), can be used for estimation, where \( t_i \) is the time at which information is received at the \( i^{\text{th}} \) sensor. Therefore, the new estimator becomes

\[
\hat{s} = \max_{s \in \mathcal{V}} f(\mathbf{D}|s) \tag{1}
\]

where \( \mathbf{D} = (d_1, d_2, ..., d_{k-1}) \) is the arrival time difference between sensor pairs. \( f(\mathbf{D}|s) \) is the pdf of \( \mathbf{D} \), given \( s \) is the source of information. The random vector \( \mathbf{D} \) is a multivariate Gaussian distribution, since the individual distributions of the time delay are independent Gaussian themselves.

So, the optimal ML estimator can be written as

\[
\hat{s} = \max_{s \in \mathcal{V}} \frac{1}{\det(\Lambda_s)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{D} - \mu_s)^T(\Lambda_s)^{-1}(\mathbf{D} - \mu_s)\right) \tag{2}
\]

where \( \mu_s = [\mu_1, \mu_2, ..., \mu_1, ..., \mu_{k-1}] \) and \( \mu_i \) represents the mean value of difference in arrival times between the first and the \( i^{\text{th}} \) sensors. \( \Lambda_s(i, j) \) is the cross-correlation of arrival time difference between the \( i^{\text{th}} \) and the \( j^{\text{th}} \) sensors.

The number of sensors, \( k \), trades off computational complexity for accuracy of the source location estimation. It is due to this fact that the amount of information for locating the source increases, as the number of sensors grows.

### III. Source Localization Algorithm

In this section, we propose an algorithm to identify the source of diffusion. Our proposed algorithm (See Table I) consists of two stages. First, the most likely candidate cluster is identified using the ML estimator given in Eqn. 2. In the second stage, the ML estimator is applied again using only the nodes inside the cluster flagged by the first stage as the most likely cluster.

First we need to discover the clusters existing in this network. We adopt the community (or cluster) finding method described in [18], which is faster than other existing algorithms, with a complexity of \( O(N^2) \), where \( N \) is the number of nodes in the network. We then construct a new graph, \( \mathcal{G}_{gw} \), of nodes connecting clusters via between-cluster ties (called the gateway nodes). For each cluster, we need to select sensors, preferably, the nodes that receive the largest amount of information from the source. Since, it is assumed that information flows along the shortest path into the network, the most appropriate measure of centrality in this case would be betweenness centrality\(^2\). Fig. 2 illustrates an example of the two-stage algorithm.

\^2Betweenness centrality refers to the number of times that a node lies on the shortest paths between any pair of nodes in the network.

![Fig. 2. The left side: the graph \( \mathcal{G}_{gw} \) of the gateway nodes, in which the green nodes represent the sensors \((m = 2)\) at the first stage. At the end of this stage, e.g., the node circled is selected as gateway node that guides us to the candidate cluster \((s_{gw}^*\rangle\). The right side: the cluster that most likely contains the diffusion source. The green nodes represent the sensors \((n = 2)\) which measure the arrival times of messages and then estimate the source location \((s^*)\).](image)

### IV. Simulation Results

In this section, we present the numerical results for the performance analysis of the proposed algorithm. We apply the proposed algorithm to a clustered network with variable topologies. We simulate information spread on several different network topologies using the SI model. Since there is no prior knowledge of the source of diffusion, we generate a uniformly distributed source in \([1, N]\) where \( N \) is the number of nodes in the network. It is assumed that the inter-arrival times are independent and identically Gaussian distributed.
3 The detection rate is defined to be the fraction of experiments in which the estimator coincides with the actual source.
In this paper, the problem of locating the source of information in large-scale social networks is considered. We use the SI model to study information diffusion in the network. A two-stage algorithm is proposed in which at the first, the mostly likely candidate cluster is identified. In the second stage, the source is located within the candidate cluster. We evaluate the performance of the proposed algorithm on several different network topologies. Performance is measured in terms of the percentage of nodes in the network that need to act like sensors. We observe significant performance by using our two stage localization algorithm. We also see that as the heterogeneity of the network increases, the localization accuracy increases greatly.

REFERENCES

[1] “Social network ad spending to approach $1.7 billion this year,” eMarketer.


