Journal of Computational Science 10 (2015) 278–287

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Locating multiple sources in social networks under the SIR model: A divide-and-conquer approach☆

Wenyu Zang[a,*], Peng Zhang[b], Chuan Zhou[c], Li Guo[c]

[a] Institute of Computing Technology, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[b] Research Center on Quantum Computation and Intelligent Systems, University of Technology Sydney, NSW 2007, Australia
[c] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

ARTICLE INFO

Article history:
Received 11 October 2014
Received in revised form 11 May 2015
Accepted 13 May 2015
Available online 16 June 2015

Keywords:
Social network mining
Source locating
Community detection

ABSTRACT

Social networks greatly amplify the spread of information across different communities. However, we recently have observed that various malicious information, such as computer virus and rumors, were broadly spread via social networks. For better controlling the spread of malicious information, it is critical to develop effective methods to locate the diffusion source nodes in social networks. Many pioneer works have explored the source locating problem, but they mostly rely on the assumption that there is only a single source node. In this paper, we present an approximate multi-source locating algorithm by first introducing a new reverse propagation model to detect the recovered and unobserved infected nodes, and then developing a community detection method to cluster the extended infected nodes (including recovered nodes and infected nodes) into multiple infected communities. In doing so, we can identify the source nodes by using the maximum likelihood estimation on each infected community. Numerical simulations on both synthetic and real networks show the performance of the proposed method.

1. Introduction

Online social networks provide a vibrant platform for information dissemination. Known as a double-edged sword, social networks have both advantages and disadvantages in information propagation. On the one side, online advertising can be promoted by targeted recommendations in social networks. On the other side, rumors and virus can also be disseminated uncontrollably. In order to understand the dissemination of information and prevent the spread of malicious information, it is of utmost importance to identify the source nodes based on a given snapshot data of the observed infected nodes. This work may also be applied to criminal network disruption [1] and dynamical processes on complex networks [2].

To date, there have been some efforts in studying the source locating problem. First, some heuristic algorithms are proposed, such as the centrality measurement [3], Monte Carlo algorithm and spectral algorithm [4]. Then, a framework for source locating problem is proposed under the maximum likelihood estimator (MLE) [5]. All these works assume that a network only contains one single source node. Moreover, most existing works are based on this assumption [6–8,5]. However, in fact, information often be distributed from multiple sources. The spread of rumors in social networks, for example, always started from multiple sources. As shown in Fig. 1, previous single source locating solutions cannot handle the multi-source locating problem well. Recently, Prakash [9] and Luo [10] discussed the multi-source locating problem based on the SI propagation model in a restricted network. However, taking rumor propagation for example, besides the fact that people who received the rumor may forwarded it (infected) with certain probability, People who has been infected may also find the message is rumor and then delete it (recovered) with a certain probability. And these two works didn't consider the recovery phenomenon from infection.

Unlike the previous works, in this work we aim to develop a multi-source locating solution on general networks with sparsely observed infected nodes. The technical challenges of the proposed problem including the limited available information from the sparsely infected nodes, the uncertain number of source nodes behind a given snapshot of a network, and the stochastic recovery from infection during information propagation.

To solve these challenges, we first propose a reverse propagation model to detect recovered and unobserved infected nodes. Then,

☆ This work was supported by the NSFC (No. 61370025), 863 projects (No. 2011AA01A103 and 2012AA012502), 973 project (No. 2013CB329606), and the Strategic Leading Science and Technology Projects of Chinese Academy of Sciences(No. XDA06030200), Australia ARC Discovery Project (DP140102206).
* Corresponding author. Tel.: +86 1082546708.
E-mail address: zangwenyu@nelmail.iie.ac.cn (W. Zang).

http://dx.doi.org/10.1016/j.jocs.2015.05.002
1877-7503/

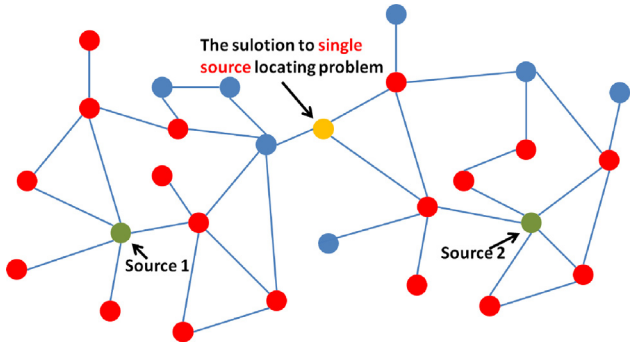

Fig. 1. The multi-source locating problem. There are two sources in this network, i.e., source 1 and source 2 (shown in green nodes). If we only consider it as a single source locating problem, the yellow node will be most likely misidentified as the source.

we cluster the infected nodes into predefined community partitions. At the last step, we estimate the most likely sources in each community partition. The technical contributions of the paper are as follows,

- The proposed score-based reverse propagation algorithm can approximately detect recovered and unobserved infected nodes in social networks, which can handle the difficulty of limited observations under the SIR model in source locating problem.
- We use the method of divide and conquer to transform multi-source locating problem into several single source locating problems, then to identify sources based on infected community partitions. It can largely reduce the computational complexity.

The rest of the paper is organized as follows. Section 2 gives a brief review of the related works. Section 3 formulates the multiple sources locating problem. Section 4 discusses the new multi-source locating algorithm. Numerical simulation results are presented in Section 5. Section 6 concludes the paper.

2. Related Work

The development of social networks makes information traveling around the world fast. On one hand, different kinds of information diffusion models [11] have been proposed to simulate the information propagation in social networks. These models can be roughly divided into two categories, Influence diffusion models (e.g. IC and LT [12] models) and Epidemic propagation models. In Epidemic propagation models, each node in the underlying network is in one of following states – Susceptible (S), Infected (I) or Recovered (R). According to the transitions between these states, epidemic propagation models can be further divided as the SI, SIR and SIS model. On the other hand, many works on influence maximization [12–15] have promoted the information dissemination significantly. Specifically, they formulate the problem as finding a small subset of nodes in social networks, such that by activating these nodes, the expected spread of the information can be maximized. By exploiting the submodular property of the spread function, greedy algorithms such as CELF [16], CELF++ [17], UBLF [18] and Upper Bounds [19] can achieve local optimal solutions under less computational cost (Monte-Carlo calls). Furthermore, many heuristic methods [20–22] greatly reduce the computational cost and still achieve a satisfactory performance.

The source locating problem has also been studied recently. First, there are works [7] using heuristic algorithms based on centrality measurement to locate the source. Then, Agaskar [23] using Monte Carlo algorithm based on geodesic distance to solve this problem. In this algorithm, additional observations during a period of time is needed. Other algorithms such as BP (Belief Propagation) [24] and mean-field-type approaches by DMP (Dynamic Message Passing) [8] also be proposed based on the assumption that propagation time t and a node j is in each of the three states S,I,R (at time t) are known. Unfortunately, these information is not easy to be observed. Moreover, spectral method [4] and Minimum Description Length (MDL) [9] also can be used in source locating problem.

However, all these algorithms lack a uniform frame for source detection. Shah [25,26] first defined the source locating problem as a maximum likelihood estimation problem and gave an exact solution on regular trees. Based on this work, a MAP estimator is proposed by Dong [3], and rumor source detection model with multiple observations was presented by Wang [27] to improve the source locating accuracy. Unfortunately, to obtain the information of prior distribution and multiple observations is even more difficult. In the framework of maximum likelihood estimation, Pinto [5] identified the source from measurements collected by sparsely placed observers, and Nino [6] proposed a statistical inference framework on an arbitrary network structure. However, all these works required tree properties and only worked under the SI model. Furthermore, some sample path based algorithms were proposed in recent years. Zhu [28] proposed an algorithm Jordan Center under the SIR model. Luo [29] proposed to find the source with limited observations based on the most likely infection path.

However, most existing algorithms work on the SI model. To date, some works can identify source node based on diffusion models rather than the SI model. For example, Saito [30] and Luo [31] presented source locating algorithms under the SIS model, and Chen [32], Zhu [28] identified the source under the SIR model. Nevertheless, all these works were focused on single source locating problem. They have not taken into consideration the multi-source locating problem. By investigating the number of source nodes, Prakash et al. [9] proposed to use the minimum description length (MDL) principle to identify seed nodes and virus propagation ripple that succinctly describes the given snapshot. However, it was a pity that this model works well only on grid networks. Other multiple source detection algorithms, like Luo [10] located sources with their infected regions together and Chen [32] detected multi-sources on tree-like networks, either on the SI model or the tree-like networks.

Contrary to these approaches, we proposed a solution to multi-source locating problem based on partially observed infected nodes in arbitrary networks under the SIR model. First, we proposed a score-based reverse propagation method to detect the recovered and unobserved infected nodes. Then, we present a divide and conquer method to handle the multi-source locating problem. Exactly, we employed community detection method to transform multi-source locating problem into several single source locating problems. We also proposed a heuristic method to estimate the number of sources. At last, we used modified centrality measurements to identify the sources in each detected community.

3. Problem formulation

In this section, we define a maximum likelihood estimator for the multi-source locating problem under the SIR model. The propagation model will be discussed in more details later in the Appendix A. Consider a network $\mathcal{G} = (V, E)$. Supposing the information sources, $\mathbf{v} \subseteq \mathcal{G}$, are the nodes that originate the information or initiate the diffusion. And a partial observed snapshot $\mathcal{O} \subseteq V$ is given. The task is to estimate the sources $\mathbf{v}$ based on the snapshot $\mathcal{O}$. Furthermore, the set V_w denotes the nodes in the state w, where w is one of the three states: I (Infected), R (Recovered) and S (Susceptible). It follows that $V_I \cup V_R \cup V_S = V$ and they are pairwise disjoint.