Topic-aware Source Locating in Social Networks

Wenyu Zang[1,2], Peng Zhang[3,1], Chuan Zhou[1], and Li Guo[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100086, China
[3]Quantum Computation and Intelligent Systems, University of Technology, Sydney (UTS), Australia
zangwenyu@nelmail.iie.ac.cn, {zhangpeng,zhouchuan,guoli}@iie.ac.cn

ABSTRACT

In this paper we address the problem of source locating in social networks from a topic modeling perspective. From the observation that the topic factor can help infer the propagation paths, we propose a topic-aware source locating method based on topic analysis of propagation items and participants. We evaluate our algorithm on both generated and real-world datasets. The experimental results show significant improvement over existing popular methods.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

source locating, topics, interests

1. INTRODUCTION

Locating the source of observed information in social networks has attracted much attention in recent years, which is crucial for controlling and preventing rumor risks. But quite surprisingly, existing work on diffusion source locating merely relies on network structure analysis and doesn't fully use the topic information which, based on our observation, is the key component for source locating.

Actually the information propagation in social networks presents many topic characters, for example, people prefer to spread information they are interested in and information with different topic trends usually propagate along different paths. Hence, introducing the topic factor into source locating problem improve the detection accuracy, which makes source locating in topic-aware social networks a practically useful tool.

In this paper, we study the source locating problem in social networks by unleashing the power of topic information. First of all, we present a topic-aware susceptible-infected (T-SI) propagation model. In this model, information diffuses faster between nodes with similar factor (or interests). Then, we present a proximate maximum likelihood estimator to identify the source based on the T-SI model. Furthermore a sample path based heuristic algorithm is also presented for fast source locating.

1.1 Related work

Early solutions to source locating problem are based on heuristic topological centrality measures. Later, solutions with maximum likelihood (ML) estimator [2] and maximum a posteriori (MAP) framework [1] were proposed. Other methods based on spectral analysis and Monte Carlo algorithm were used improve the accuracy. Most of these methods are mainly applicable for tree-like network structures under the SI or SIR model.

2. TOPIC-AWARE SI MODEL (T-SI)

Different from the common SI model, the infected probability on each edge would change to according to different topics in the topic-aware SI model (T-SI). For each edge $(u, v) \in E$ and each item i, $p^i_{u,v}$ represents the infected probability between user u and v on item i. More specifically, for each item i spreading on social network, we view it as a topic distribution $\gamma^z_i = p(z|i)$ for each topic $z \in [1, K]$ with $\sum_{z=1}^{K} \gamma^z_i = 1$. And for each user in the social networks, we describe it as mixture distribution of interests p^z_u for each interest $z \in [1, K]$ with $\sum_{z=1}^{K} p^z_u = 1$. Then we construct the infected probability $p^i_{u,v}$ for item i based on threefold: 1) the interest similarity of neighbor users; 2) the user's enthusiasm for this item; 3) the time-sensitive of this item, i.e., new things usually spread faster in social networks.

In the T-SI model, each node in the network is in one of two states - Susceptible (S) or Infected (I). Once a node is infected, it will stay infected forever. With the diffused item i and its existence time t_i, an infected node u tries to infect each of its susceptible neighbors v independently with probability $p^i_{u,v}$, as shown in Eq. (1),

$$p^i_{u,v} = \sum_z p(z|u, i, t_i) p^z_v \tag{1}$$

where $p(z|u, i, t_i)$ is the following logistic selection function

$$p(z|u, i, t_i) = \frac{exp(p^z_u + \gamma^z_i f(t_i))}{1 + exp(p^z_u + \gamma^z_i f(t_i))} \tag{2}$$

with $f(t_i) \propto 1/t_i$. Let τ_{uv} be the time for node v to receive the item i from node u. In the T-SI model, $\{\tau_{uv}\}_{(u,v)\in E}$ are independent and all with exponential distribution with parameter λ. Without loss of generality, we assume $\lambda = 1$ in this paper.

3. APPROXIMATE ML ESTIMATOR

We consider a network as a graph $G(\mathcal{V}, \mathcal{E})$, where the vertex set $\mathcal{V}$ has N nodes, and $\mathcal{E}$ is the set of edges. We assume that one user in the network $G(\mathcal{V}, \mathcal{E})$, $v^\star$, spreads a message i to the whole network under the T-SI model at time 0. And at some time t, an infected users set $\mathcal{I}$ is observed. Our goal is to set an estimator $\hat{v}$ of the real source $v^\star$ based on the infected set $\mathcal{I}$ and the network $G(\mathcal{V}, \mathcal{E})$. Without loss of generality, we assume a uniform prior probability for the

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742745.

source among all nodes in $G(\mathcal{V}, \mathcal{E})$. Then the ML estimator can be formulated as follows:

$$\hat{v} \in \arg\max_{v \in \mathcal{I}} P(\mathcal{I}|v) \tag{3}$$

where $P(\mathcal{I}|v)$ is the probability that infected set $\mathcal{I}$ will be observed under T-SI model with the source v.

If the infection probability $p^i_{u,v}$ is constant, we can use the rumor center $R(v, T_{bfs}(v))$ [2] to calculate $P(\mathcal{I}|v)$, where $T_{bfs}(v)$ is a breadth first search (BFS) tree rooted at v. The estimator in this case is formulated as:

$$\hat{v} \in \arg\max_{v \in G_{\mathcal{V}}} R(v, T_{bfs}(v)) \tag{4}$$

where $R(v, T_{bfs}(v)) = N! \prod_{u \in T_{bfs}(v)} \frac{1}{T^v_u}$ and T^v_u is the number of nodes in the subtree rooted at u with the source v. When the infected probabilities vary according to the different diffusing items, we can rewrite the $P(\mathcal{I}|v)$ as follows:

$$P(\mathcal{I}|v) = \sum_{\sigma \in \Upsilon} P(\sigma|v) \tag{5}$$

where Υ is the set of all possible spreading paths of observed infected users $\mathcal{I}$ with the source v.

As the rumor center $R(v, T_{bfs}(v))$ can be interpreted as the number of possible paths of the spreading information with the source node $v \in T_{bfs}(v)$. Then,

$$P(\mathcal{I}|v) \approx E(P(\sigma|v))R(v, T_{bfs}(v)) \tag{6}$$

In general, $E(P(\sigma|v))$ is difficult to evaluate. However, the probability expectation depend mostly on the path with the highest probability. Thus the estimator can be approximated as follows:

$$\hat{v} \in \arg\max_{v \in G_{\mathcal{V}}} H(v, \mathcal{I})R(v, T_{bfs}(v)) \tag{7}$$

where $H(v, \mathcal{I})$ is the probability of the path with most probability. Here we employ the dynamic message passing algorithm [2] to approximate the optimal problem (7). We call this algorithm as *TopicCenter* in the experimental part.

4. SAMPLE PATH BASED SOLUTION

In addition, we here propose a sample path based solution to simulate the most likely source that triggers the existing infected subgraph. Specifically, each infected user $u \in \mathcal{I}$ spreads its ID though the network under the T-SI model. Meanwhile, each user v in the network record the time t^u_v (when he is first receive the ID of u for each $u \in \mathcal{I}$), and then broadcasts this ID to its neighbors. Moreover, the user who first receives IDs of all infected users is regarded as the information source. Furthermore, if there are multiple users received all IDs at the same time, the cumulative infected time($\sum_{u \in \mathcal{I}} t^u_v$) is used to select the information source and we choose the smallest one as the source. We summarize the *SamplePath* algorithm in Algorithm 1.

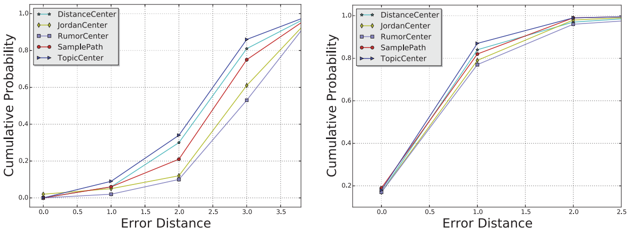
5. EVALUATION

We evaluate our algorithms (TopicCenter and SamplePath) on both generated and real-world datasets. The synthetic dataset is generated by *MMSB* model (a generated model that can generate network structure and node attributes at the same time), which contains 3000 nodes and 17477 edges. While the real-world dataset is Author collaboration[1], which is crawled from the Internet on several specific topics containing link and text information (1377 nodes and 4480 edges). Furthermore, the nodes attributes are learned from the text information by the common *LDA*.

Here we evaluate the precision of source locating algorithms. We show the cumulative probability distribution of the average distance between actual source and estimated source in Fig 1. All reported results are averaged over 100 independent runs on different sources. We can find that our TopicCenter achieves superior results on both generated and real-world networks.

Algorithm 1: Sample Path based Solution to Source Locating

```
Initialization: all users set $\mathcal{V}$ which contains infected users
                set $\mathcal{I}$, set $STOP = 0$, infection time $t = 0$
for $v \in \mathcal{V}$ do
    for $u \in \mathcal{I}$ do
        $t^u_v = N$, $flag^u_v = 0$
for $u \in \mathcal{I}$ do
    $u$ spreading its ID to its neighbors following the T-SI model
while $STOP == 0$ do
    for $v \in \mathcal{V}$ do
        if $v$ received the ID of $u$ for the first time then
            $t^u_v = t$, $flag^u_v = 1$
            $v$ spreading the ID of $u$ to its neighbors
        if $flag^u_v = 1$ for all $u \in \mathcal{I}$ then
            STOP = 1
return $\hat{v} = arg\ min_{v \in \mathcal{A}} \sum_{u \in \mathcal{I}} t^u_v$ where $\mathcal{A}$ is the set of users with $flag^i_a = 1$ when algorithm terminates.
```



MMSB **Author Collaboration**

Figure 1: Cumulative probability distribution of average distance between real source and estimated source.

Acknowledgement. This work was supported by the NSFC (No. 61370025), and the Strategic Leading Science and Technology Projects of CAS (No.XDA06030200), 973 project (No. 2013CB329605) and Australia ARC Discovery Project (DP140102206).

6. REFERENCES

[1] W. Dong, W. Zhang, and C. W. Tan. Rooting out the rumor culprit from suspects. In Proceedings of *IEEE International Symposium on Information Theory (ISIT)*, pages 2671–2675, 2013.

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

[3] Peng Zhang, Chuan Zhou, Peng Wang, Byron J. Gao, Xingquan Zhu, Li Guo: E-Tree: An Efficient Indexing Structure for Ensemble Models on Data Streams. IEEE Trans. Knowl. Data Eng. 27(2): 461-474 (2015)

[1] http://arnetminer.org/heterinf