基于部分路径的社交网络信息源定位方法

张聿博，张锡哲，徐 超

(东北大学计算机科学与工程学院 沈阳 110819)

【摘要】随着微博、微信等在线社交网络的快速发展，社交网络上的不实信息呈现爆发式的传播，往往会引起严重的后果，如何寻找谣言等不实信息在社交网络中的传播源头具有重要的应用意义。该文提出一种面向在线社交网络的信息源点定位方法，与现有的基于观察点的定位方法不同，该方法考虑了传播过程中信息普遍带有的部分传播路径，并以此重构传播过程，修正传播子图，从而更准确地定位信息源点。在模型网络及实际网络上进行实验，说明了该算法的有效性。

关 键 词 信息传播； 部分路径； 社交网络； 源点定位

中图分类号 TP391 **文献标志码** A doi:10.3969/j.issn.1001-0548.2017.01.012

Source Localization Algorithm Based on Partial Paths for Social Networks

ZHANG Yu-bo, ZHANG Xi-zhe, and XU Chao

(College of Computer Science and Engineering, Northeastern University Shenyang 110819)

Abstract With the rapidly growth of online social networks such as microblog and WeChat, the false information breaks out on the social network and often brings serious consequences. How to locate the rumor source is of great importance for many applications. This paper proposes a source localization algorithm on online social network. We consider the characteristic that the information often contains some partial spreading, and design a more accurate algorithm to locate the information source. The results show that the improved algorithm can provide a more accurate spreading trees and improve the localization accuracy. Experiments on model and real network show the effectiveness of the improved algorithm.

Key words information diffusion; partial paths; social network; source localization

随着社交网络规模激增，信息的受众面不断扩大，社交网络已经成为一种非常重要的信息传播平台。用户在分享信息的同时，也要面临谣言等有害信息带来的不良影响。因此，如何控制谣言信息的传播已经成为当前的研究热点之一[1-2]。

现有的控制谣言信息传播方法，一类是采用链接预测[3]，通过现有的谣言传播信息，预测谣言的进一步传播趋势，并将可能产生的传播链路切断，来达到抑制谣言传播的目的。另一类有效的方法是找到谣言的源头[4-6]，这对于有效地控制谣言传播具有重要帮助。

对于传播源点定位问题，一种典型的方法是基于网络传播快照进行源点定位。如文献[7]针对易感染病毒传播模型(SI模型)提出一种基于最小描述长度的定位方法，能够自动地确定传播源点的数目，并识别网络中的多个传播源点。文献[8]给出了基于度、介数、紧密度和特征向量等拓扑度量的传播源点定位算法，由于源节点多数趋向于具有最高的中心度，因此这种方法对于雪球传播模型非常有效。文献[9]采用样本路径方法，寻找最有可能形成网络快照中样本路径的根节点作为信息源点。

虽然基于网络快照的定位方法具有不错的定位精度，但其需要获取网络全部节点的传播状态，对于在线社交网络这类大规模网络，很难实现。不同于此类方法，文献[10]提出了一种可以用于大型复杂网络的信息源定位方法，在网络中选取少量节点作为观察点，记录这些节点的传播状态，利用最大似然估计找出传播源点。该方法可以有效地减小数据需求，能够用于大规模网络上的源点定位。

但是，这种基于观察点的信息源定位方法，只考虑了网络中节点只能记录与其直接相关的传播信息(包括接收消息的时间、传入方向等)的情况，通过

收稿日期：2015－03－30；修回日期：2016－03－14
基金项目：中央高校基本科研业务费(N140404011)；国家自然科学基金(60093009)
作者简介：张聿博(1984－)，男，博士，主要从事社交网络方面的研究.

这些信息很难了解消息的真实传播过程。在线社交网络中，由于用户间存在信息互动，信息在传播的过程中可能会附加上一些与该信息传播过程有关的附加内容。

本文提出一种基于部分传播路径的信息源定位算法，与现有的基于观察点的定位方法不同，该方法考虑观察点所收到信息中附加的部分传播路径，并以此重构传播过程，筛选可能的信息源点，从而准确定位信息传播源点。并在大规模网络上进行实验，结果表明，利用部分传播路径可以有效地提高定位准确率。

1 信息传播模型

在线社交网络的信息传播具有方向性，即只有“关注者”用户节点才能从“被关注者”用户节点收到信息，因此本文采用有向图$G(V,E,W)$对社交网络进行建模。其中，V为节点集合，E为边集，W为权集，表示每条边上信息传播的延迟时间。

对于节点$u\in V$，$\gamma(u)$表示其邻居节点集合，t_u表示u首次收到某一指定信息的时间；对于边$e_{ij}\in E$，e_{ij}为节点i和节点j之间的连边；$w_{ij}\in W$为随机变量，表示边e_{ij}的传播延迟，本文考虑传播延迟服从均值为μ和方差为σ^2的高斯分布的情况[10]。

在某一未知时刻t^*，未知源点$s^*\in V$发送消息M给其邻居节点$\gamma(s^*)$。节点u有两种可能状态：

1) 知情状态，该节点已经接收到信息。

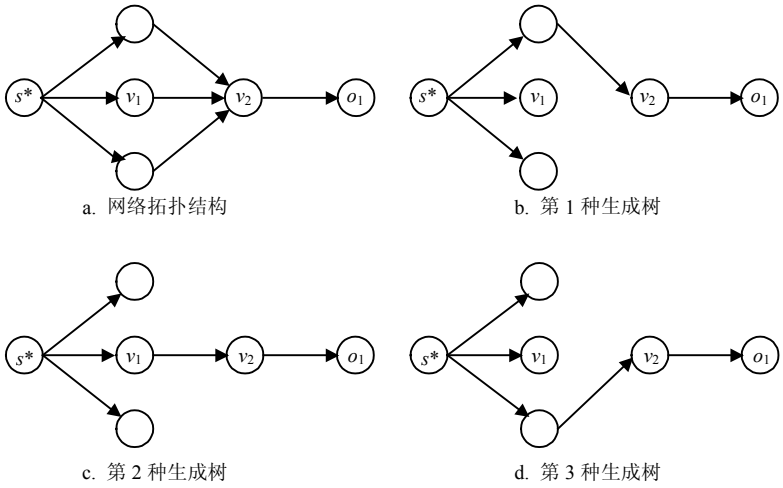
2) 不知情状态，该节点没有接收到信息。当节点u收到指定消息M时，若此时u为知情状态，则不做变化；否则由不知情状态变为知情状态，并将消息M发送给$\gamma(u)$。

选取社交网络上部分节点记录其传播状态，称这些节点为观察点，记为$O=\{o_k\}_{k=1}^{K}$，观察点会记录3类信息：1) 发送者，表示消息是从哪个节点传入的。2) 时间，信息首次到达该节点的时间。3) 消息内容，包括消息之前传播所附加的部分路径信息。

2 基于部分传播路径信息源定位算法

2.1 算法思路

对于基于部分观察点集的传播源点定位方法，文献[10]提出了一种最大似然估计方法，利用观察点搜集的时间及发送者信息定位传播源点。该方法的一个基本假设是信息沿着最短路径进行传播，所以对每个潜在源点(网络中除观察点外的节点)n，以n为中心构造广度优先搜索树，以此重构信息的传播过程，然后计算其似然估计值，似然估计值最大的潜在源点即为信息源。



a. 网络拓扑结构　b. 第 1 种生成树　c. 第 2 种生成树　d. 第 3 种生成树

图1 构建广度优先生成树举例

上述方法的一个缺点在于，很多情况下，节点间的最短路径不只一条，信息会沿着哪条路径进行传播，与很多复杂因素如节点间的交互强度、可信任性、权威性、影响力等密切相关。因此，上述方法在真实在线社交网络中，很难构建符合真实传播过程的搜索树，也得不到准确的定位结果。经过观察，社交网络中传播的信息，往往会附加其经过的部分传播路径。有效的利用这些路径，更准确的重构信息传播过程，可以提高定位的准确率。

本文算法的基本思路是，在最短路径传播假设的基础上，结合观察点搜集到的部分传播路径，建立更准确的广度优先搜索树，从而提高源点定位的准确度。具体做法是，当从潜在源点构造广度优先树时，如果遍历到记录了部分传播路径的观察点时，

首先判断观察点记录的接收信息节点方向与生成树中待连接的父节点是否为同一节点，然后判断从树根节点到当前待加入节点的路径是否与部分传播路径完全一致。如果不一致，则需从路径不一致的起始处选择其他路径构建生成树路径，直到符合观察点记录的路径信息为止。如果某一潜在源点无法构建出符合要求的生成树，则将该节点排除，不再进行似然估计值计算。

如图1所示，s^*是信息源点，o_1是观察点。观察点o_1记录从v_2接收到消息。在此情况下，以s^*为根可以构建3种可能的生成树，如图1b～图1d所示。若o_1收到的信息记录了部分路径$v_1 \to v_2$。那么，在判断生成树是否满足o_1记录的信息时，除比较接收信息的方向是否为v_2，还需判断已经加入生成树的节点是否符合部分传播路径。按照上述规则，图1b和图1d两种情况被排除，只有图1c符合。在构建生成树时参照部分路径的策略能够筛除一些不符合实际传播情况的生成树，保留符合实际传播情况的生成树。

2.2　算法描述

基于上节中提出的算法思路，本文在文献[10]的信息源定位方法基础上，提出了基于部分传播路径的信息源定位方法，该算法充分利用了社交网络中传播的信息会记录传播路径这一特征，对现有算法的生成树方法进行了改进。

社交网络中，信息附加的部分路径可以包括一个或者若干个节点，可以是一段或者多段，并且不会所有用户在转发过程中都会附加传播路径信息。本文考虑观察点只记录一段部分传播路径的情况，假设只有部分观察点会记录部分传播路径，用f_p表示记录部分路径的观察点占所有观察点的比例。f_p的取值与信息的内容和长度都有关，由于在线社交系统会限定信息的长度，对于本身内容较长的信息，用户会删除部分传播路径以符合系统的要求。令observers为观察点集合，fragment为部分路径集合，tree为构建的生成树。对于某一潜在源点n，其构造生成树的具体过程如下：

1) 遍历observers集合，将所有观察点记录的部分路径存入fragment。

2) 令n为生成树tree的根节点；获取n的所有外邻节点，并将其存入queue中。

3) 从queue中取出节点m，令m的父节点为q，做如下判断：

① 若fragment中不包括节点m，则将边e_{qm}加入生成树tree；计算m的外邻节点并将其存入queue；

② 若fragment中包括节点m，且边e_{qm}也在fragment中，则将边e_{qm}加入生成树tree；计算m的外邻节点并将其存入queue；

③ 若fragment中包括节点m，但边e_{qm}不在fragment中，则不做任何操作。

4) 重复执行过程3)，直到queue为空。

5) 若tree包含所有收到消息的观察点，则返回以n为根的生成树tree，否则返回null。

构建了以n为根的广度优先树之后，可以利用文献[10]的最大似然估计进行源点定位，找出具有最大似然估计值的节点作为传播源点。最大似然估计的计算为：

$$s^* = \arg\max_{s\in T_a} \boldsymbol{\mu}_s^{\mathrm{T}} \boldsymbol{\Lambda}^{-1}\left(d - \frac{1}{2}\boldsymbol{\mu}_s\right)$$

式中，

$$[\boldsymbol{\mu}_s]_k = \mu(|P(s,o_{k+1})| - |P(s,o_1)|)$$

$$[\boldsymbol{\Lambda}]_{k,i} = \sigma^2 \begin{cases} |P(o_1,o_{k+1})| & k = i \\ |P(o_1,o_{k+1}) \bigcap P(o_1,o_{i+1})| & k \neq i \end{cases}$$

式中，$k,i = 1,2,\cdots,K$；$|P(u,v)|$表示连接节点u和节点v的最短路径长度。

3　实验及结果分析

3.1　实验数据

表1　网络数据集

名称	N	L	<k>	d	Apl
ERNETWORK1	500	3 748	7.496	10	3.089
ERNETWORK2	500	5 041	10.082	10	2.81
ERNETWORK3	500	6 258	12.516	9	2.633
ERNETWORK4	500	7 403	14.806	8	2.489
BANETWORK1	500	3 750	7.5	7	3.339
BANETWORK2	500	5 000	10.082	6	2.974
BANETWORK3	500	6 300	12.5	5	2.761
BANETWORK4	500	7 400	14.806	5	2.628
SinaWeibo	3 861	3 528	0.914	3	1.109
Twitter	3 656	188 712	51.617	12	3.764

为了验证算法的有效性，本文采用模型网络与实际网络进行模拟实验。其中，模型网络采用了ER模型网络[11]和BA模型网络[12]。具体实验数据如表1所示。表中N表示网络节点数，L表示边数，<k>表示网络平均度，d表示网络直径，Apl表示网络平均路径长度。ERNETWORK1～ERNETWORK4是ER模型网络，BANETWORK1～BANETWORK4是BA模型网络，SinaWeibo和Twitter是通过新浪微博和Twitter公开的API分别抓取一条微博信息在用户中传播所经过的所有节点进而得到的实际网络数据。

3.2 实验过程及结果分析

本文按照节点在网络中的入度排列，选取入度最大的部分节点为观察点，观察点的比例为0.05～0.40(每次递加0.05)。能够记录部分路径的观察点比例 f_p 的取值为0.1～0.3(每次递加0.1)，F取值为2。在网络中随机选取一个非观察点为信息源进行信息传播，然后分别应用原算法(文献[10]提出的方法)和本文的方法进行比较，对不同的观察点比例和 f_p，在每个网络上进行2 000次定位实验，取均值得到命中率，模型网络实验结果如图2所示。图2a～图2d是ERNETWORK1-4 的 实 验 结 果 ， 图 2e ～ 图 2h 是BANETWORK1-4的实验结果。

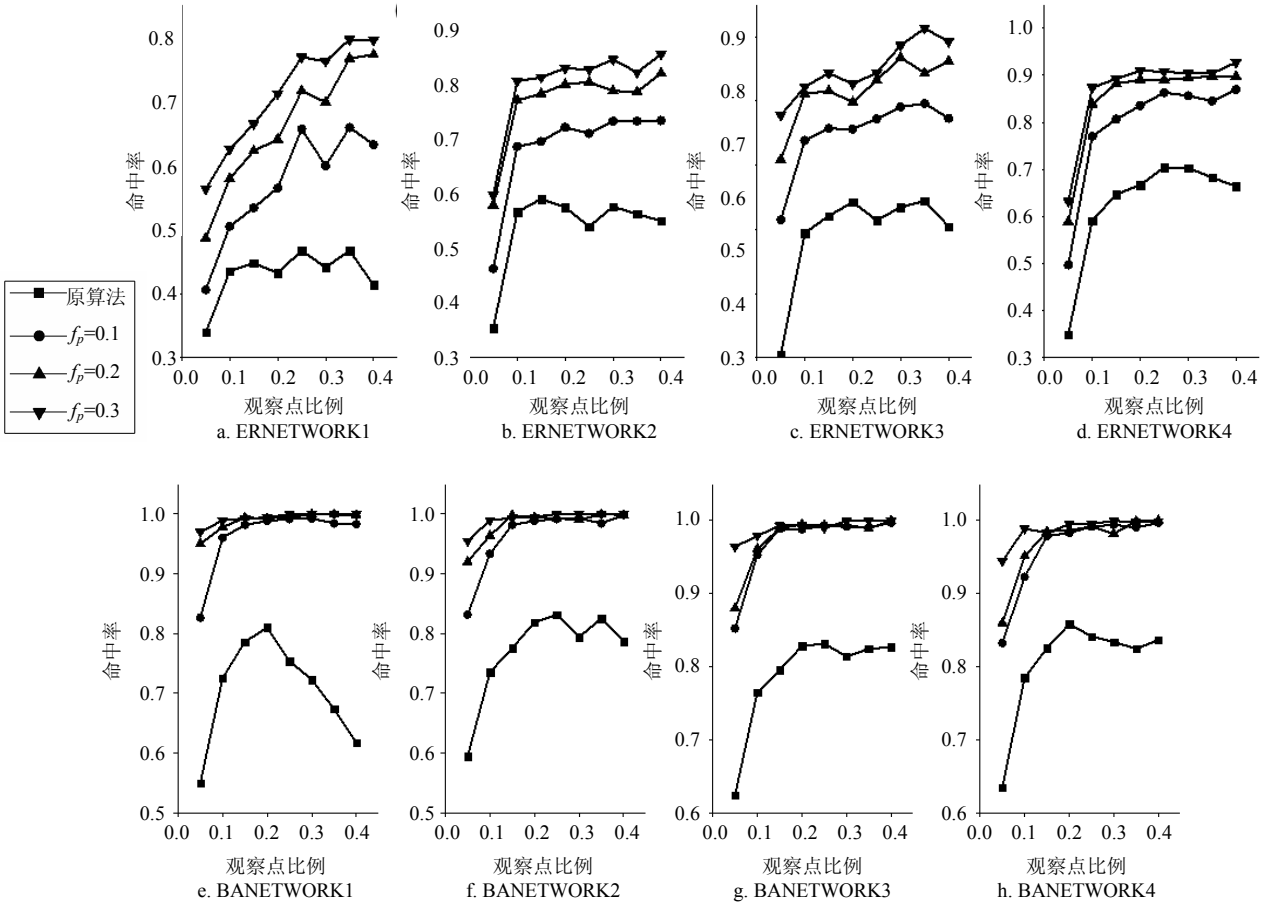


图2 模型网络命中率比较

可以看到各个观察点比例下改进算法的定位命中率均高于原算法。分析图中具体数据，当网络的观察点比例从0.05提高到0.10时，各个模型网络的定位命中率都有较大的提升，但是随后观察点比例的增加并没有对定位命中率提高造成较大影响，当观察点达到0.15后ER和BA网络的命中率上升缓慢并趋于平滑，接近最大值。可以得出，对于同一观察点比例，f_p 越大定位命中率越高。这是因为 f_p 越大，部分路径越多，在以潜在源点为根构建传播生成树时，能够得到更加真实反映信息传播过程的广度优先生成树，进而提高了算法的定位准确性。

从曲线中观察点变化引起定位命中率变化的趋势看，随着观察点比例的增加，定位命中率随之提高，因为观察点比例的增加使更多的节点记录真实传播过程，同时也增加了部分传播路径的数量，进而提高定位准确性。

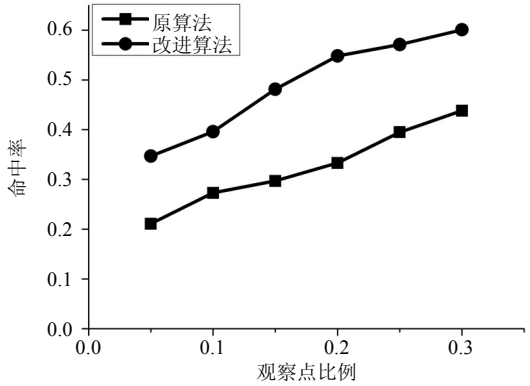


图3 SinaWeibo网络命中率比较

在SinaWeibo和Twitter两个实际网络上也进行

了对比实验，取F值为2，f_p为0.2，结果如图3和图4所示。

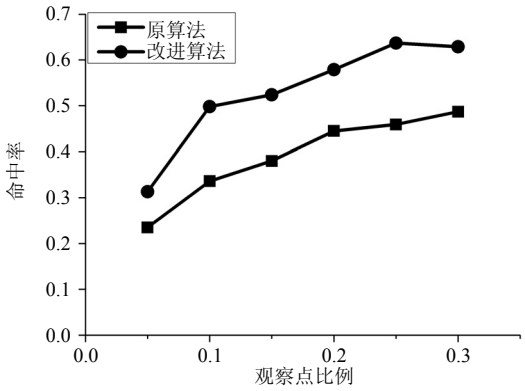


图4　Twitter网络命中率比较

从两个网络的实验结果来看，在不同观察点比例下改进算法都比原算法定位命中率高。其中，SinaWeibo网络的定位命中率改进算法平均高于原算法17%，Twitter网络的定位命中率改进算法平均高于原算法14%。可见改进的源点定位算法对于实际网络是有效的，可以提升一定幅度的定位命中率。

进一步，通过实验，对部分路径长度和部分路径比例对定位命中率的影响进行分析。选取ERNETWORK1作为实验数据集。对于ER网络，改变观察点比例，得到在指定观察点比例下，源点定位命中率随f_p变化的实验结果，如图5所示。其中图5a～图5f分别为观察点比例从0.05～0.30的实验结果，横坐标为部分路径比例f_p，纵坐标为源点定位的命中率。由图5可以得出，在不同观察点比例及不同f_p下，F=2均比F=1时定位命中率高，由此可知部分路径越长，源点定位命中率越高。原因是观察点能够记录的部分路径长度越长，观察点获取真实的传播路径信息则越多，可以构造更符合真实传播路径的广度优先生成树，定位命中率更高。

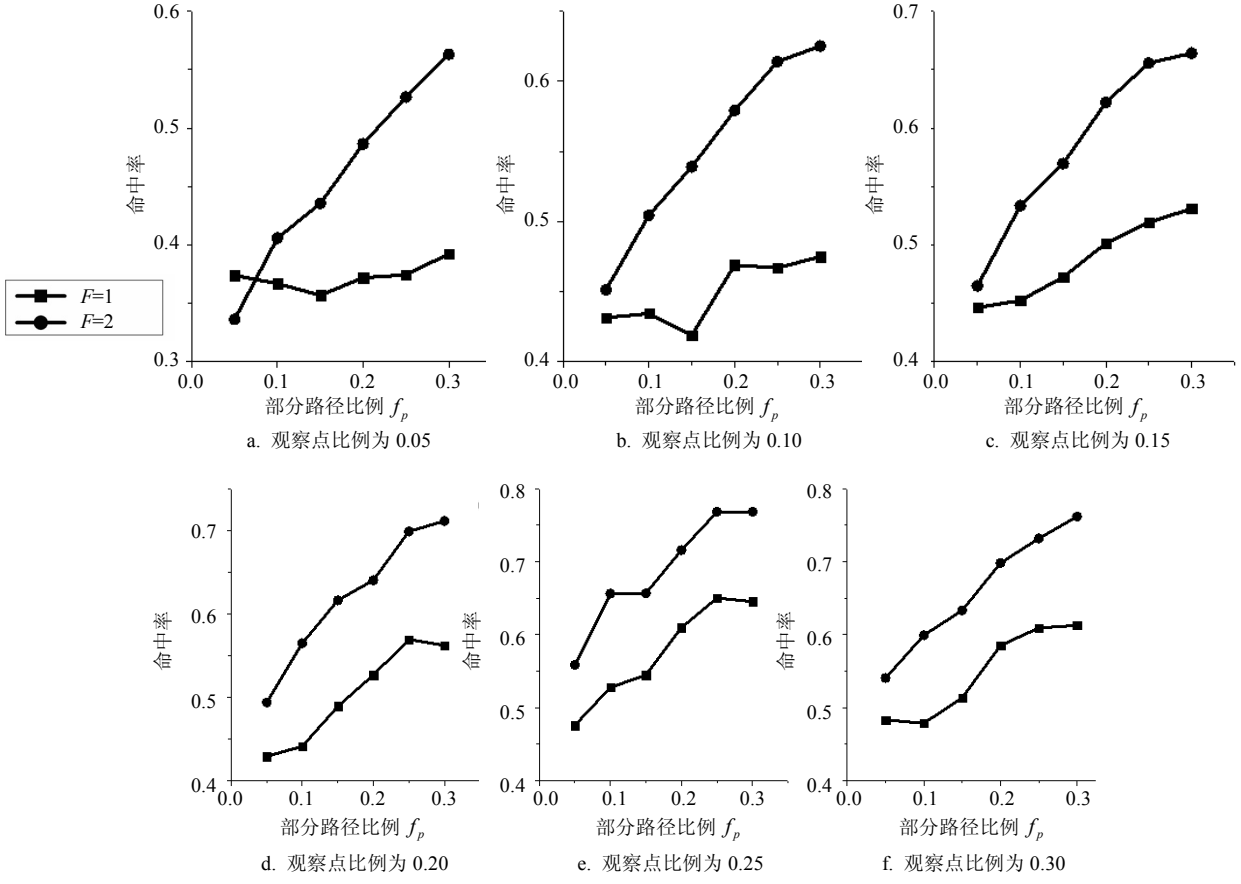


图5　不同部分路径长度的定位准确率

4　结 束 语

本文分析了在线社交网络中的信息传播过程，发现消息中携带的部分传播路径可以有效地提高信息源点定位的准确度。因此，提出一种基于部分传播路径的源点定位算法，根据观察点记录的信息，提取部分传播路径，并基于部分传播路径的约束，建立更接近信息传播过程的广度优先搜索树，从而定位信息源。

本文提出的方法利用了在线社交网络上的信息

传播特点，有效地提高了信息源点的定位准确率。在模型网络和实际网络上的实验结果表明，改进算法的定位准确率明显高于原有算法，充分验证了该方法的有效性。并且，从部分路径长度和观察点记录部分路径比例两个方面，分析了对该方法定位准确率产生影响的因素。本文提出的方法，对于有效定位在线社交网络中谣言等信息源点具有重要的应用意义。

参 考 文 献

[1] BUDAK C, AGRAWAL D, EI ABBADI A. Limiting the spread of misinformation in social networks[C]// Proceedings of the 20th International Conference on World Wide Web. [S.l.]: ACM, 2011: 665-674

[2] BUDAK D, EL ABBADI A. Information diffusion in social networks: Observing and influencing societal interests[J]. Proceedings of the VLDB Endowmen, 2011(4): 1512-1513.

[3] BUDAK C, AGRAWAL D, EL ABBADI A. Structural trend analysis for online social networks[J]. Proceedings of the VLDB Endowment, 2011, 4(10): 646-656.

[4] SHAH D, ZAMAN T. Detecting sources of computer viruses in networks: theory and experiment[J]. ACM Sigmetrics Performance Evaluation Review, 2010, 38(1): 203-214.

[5] BROCKMANN D, HELBING D. The hidden geometry of complex, network-driven contagion phenomena[J]. Science, 2013, 342(6164): 1337-1342.

[6] LOKHOV A Y, MEZARD M, OHTA H. Inferring the origin of an epidemic with dynamic message-passing algorithm[J]. Phys Rev E, 2014, 90(1): 012801.

[7] PRAKASH B A, VREEKEN J, FALOUTSOS C. Spotting culprits in epidemics: How many and which ones?[C]// ICDM. [S.l.]: IEEE, 2012, 12: 11-20.

[8] COMIN C H, COSTA L F. Identification of starting points in sampling of complex networks[J]. Phys Rev E, 2011, 84(5): 056105.

[9] ZHU K, YING L. Information source detection in the SIR model: a sample path based approach[J]. IEEE/ACM Transactions on Networking, 2013, 24(1): 408-421.

[10] PINTO P C, THIRAN P, VETTERLI M. Locating the source of diffusion in large-scale networks[J]. Physical Review Letters, 2012, 109(6): 068702.

[11] ERDOS P, RENYI A. On the evolution of random graphs[J]. Publ Math Inst Hungar Acad Sci, 1960, 5: 17-61.

[12] BARABASI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.

编 辑 漆 蓉