计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200400053

基于最大后验估计的谣言源定位器

鲍志强　陈卫东
华南师范大学计算机学院　广州 510631
(530420396@qq.com)

摘　要　随着互联网的普及,信息能够通过互联网以极快的速度被传播给大众。但同时,一些虚假信息比如谣言也借助网络的级联效应泛滥成灾,因此如何在传播网络中快速准确地确定谣言传播源成为一个亟待解决的问题。文章针对社交网络提出了一种谣言源定位的方法,与现有的基于最大后验(Maximum-a-posteriori,MAP)概率估计的方法不同,该方法首先考虑全局和局部感染点、非感染点的影响,使用效果更优的 MAP 先验概率估计(Prior Probability Estimation,PPE)计算方式。然后基于最小生成树贪心算法来稀疏化社交网络,让 MAP 中的似然估计(Likelihood Estimation,LE)计算更符合真实的传播结构。最后,采用新的 MAP 值来估计传播网络中节点为传播源的可能性,从而更准确地定位谣言源点。将所提方法与现有的几种方法分别在模型网络和真实网络中进行了对比,实验结果表明,所提方法优于现有的谣言源定位方法。

关键词:谣言源;源定位;稀疏化网络;最大后验概率估计;社交网络
中图法分类号　TP301

Rumor Source Detection in Social Networks via Maximum-a-Posteriori Estimation

BAO Zhi-qiang and CHEN Wei-dong
School of Computer Science,South China Normal University,Guangzhou 510631,China

Abstract　With the popularization of Internet,information can be transmitted to the public at an extremely rapid rate through Internet. But at the same time,some abnormal information,such as rumors,has been flooded with the cascade effect of Internet. How to quickly and accurately identify the source of a rumor spreading under a complex network becomes an urgent problem to be solved. This paper proposes a source localization algorithm in social networks. Different from some existing methods based on Maximum-a-posteriori(MAP) probability estimation,this method first considers the influence of global and local infected nodes and non-infected nodes,and proposes a better MAP prior probability estimation(PPE) calculation mode. Then,asocial network is sparsified through a greedy algorithm based on minimum spanning trees,which makes the likelihood estimation(LE) calculation in MAP more consistent with the real propagation structure. Finally,a new MAP value is used to estimate the possibility of a node as the source of propagation in the social network as to locate the source of the rumor more accurately. The proposed method is compared with some existing methods by an experiment on some model networks and real networks,and experimental results show that the proposed method is superior to these existing methods of locating the rumor source.

Keywords　Rumor source,Source location,Sparse network,Maximum-a-posteriori estimation,Social network

1　引言

当今,每个人都与技术网络、生物网络、信息网络、社会网络等各种复杂网络相关联。在利用这些网络传播和共享信息时,如果不对信息的正确性加以识别就予以转发或共享,会给虚假信息(谣言)传播可乘之机。识别复杂网络中的谣言来源,对于预防和控制网络中的虚假信息传播具有重要意义[1]。

近年来,谣言源定位问题吸引了越来越多学者的关注。谣言源定位问题,即给定一个传播网络,根据网络拓扑结构以及节点的感染情况,识别产生谣言的单个或者多个源头[2]。目前解决该问题的主要方法有基于最大似然估计(Maximum Likelihood Estimation,MLE)的方法和基于最大后验概率估计的方法。前者计算每个估计节点的 MLE 值,取 MLE 值最大的节点为传播源点,但不同方法中的 MLE 计算过程各不相同。这种传统的基于 MLE 的方法[2-5]仅仅考虑了传播网络中的感染节点,忽略了非感染点对节点 MLE 值的影响。后者结合先验概率估计和似然估计来计算被估计节点的 MAP 值[6-7]。文献[7]虽然提出使用不同的 PPE 计算方法,包括谣言中心性(Rumor Center,RC)、乔丹中心性、距离中心性[2,8]等,但是其忽略了全局节点对估计节点的 PPE 值计算的影响。其中,针对估计点的 LE 值的计算,文献[7]提出了暴力算法(Brute Force Search Approximation,BFSA)和贪心

到稿日期:2020-04-13　返修日期:2020-07-04　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61370003)
This work was supported by the National Natural Science Foundation of China(61370003).
通信作者:陈卫东 (CHENWD@SCNU.EDU.CN)

算法(Greedy Search Bound Approximation, GSBA)来遍历估计节点可能形成传播网络的所有序列,以每种序列可能性之和为 LE 值,但这两种算法有明显的缺陷。BFSA 算法尝试枚举所有允许的感染排列,计算复杂度高;GSBA 算法基于贪婪策略寻找形成感染排列的上限值,虽降低了计算复杂度,但由于谣言传播过程和网络结构都较复杂,其计算 LE 值的方式不符合真实传播的情况。

因此,本文在文献[7]的基础上,分别对 PPE 和 LE 计算过程做出改进,首先考虑全局感染和非感染点的影响,然后针对复杂感染网络提出基于传播结构的稀疏化网络方法,以解决传统方法从规则树网络扩展至复杂网络效果不佳的问题[9]。本文的贡献如下:

(1)基于文献[7]的工作,提出了比 RC 方法效果更优的 PPE 方法。该方法考虑了全局节点对估计节点的 PPE 值计算的影响,使用了节点覆盖率计算方法,并结合 EPA[10](Exoneration and Prominence based Age)中节点的免除和突出效应,以两者的乘积作为新的 PPE 值。

(2)根据传播结构特性,稀疏化复杂感染网络。本文基于最小生成树贪心算法[11]将传播边界点和估计节点作为终端节点,将复杂感染网络简化为稀疏传播子图,让 LE 值的计算过程更符合真实的传播结构,大大减少了算法复杂度。

2 相关工作

针对谣言在复杂网络中的传播,及时准确地定位谣言的传播源对信息传播控制具有重要意义。目前针对谣言源定位问题的研究仍然面临诸多挑战:1)从不同的谣言源开始传播可能造成多种传播模式,进而导致形成传播级联[12]和传播模型未知[13];2)由于未知传播时间,且可能只能观察到部分感染节点,无法掌握谣言传播的全局动态[14];3)当前多种算法将规则树实验数据集拓展至真实网络,由于真实网络的结构复杂性,将算法应用到复杂网络图的效果不佳;4)当将单源问题扩展至多源情形时,极易出现感染区域交叉的情况,降低了定源算法的性能[15]。

面对复杂网络中的谣言源定位问题,目前的解决方法主要有基于最大似然估计、中心性、样本路径、最大后验估计的算法等。Shah 等[2]提出了基于 MLE 的谣言源定位方法,该方法认为节点的 MLE 值与从该节点出发去感染其他节点的所有可能的顺序的计数成正比。他们首先让谣言分别在线性图、规则树以及复杂真实的网络中进行传播,然后基于网络拓扑结构计算每个感染节点的 MLE 值。该方法的缺点是忽略了非感染点的影响。Kai 等[8]提出的方法基于 Jordan 中心性在传播网络中寻找离心率最小的节点为源点,但忽略了信息传播的动力影响,以及传播过程的复杂性容易导致传播分支大小不一的情况[9]。Lokhov 等[16]采用 SIR 模型并提出了利用整个网络的所有节点来计算它们分别处于易感染、已感染、恢复这 3 种状态的边缘概率,由于真实网络的节点众多,这种方法耗时长且需要知道感染时间,不适用于真实情况。Fioriti 等[17]利用网络演化中增删节点导致的该网络最大特征值的变化范围来衡量该节点对整个传播过程的重要性,但其准确性不高。此外,基于最大后验估计,Chang 等在具有感染概率权重的网络图上利用 MAP 的 PPE 和 LE 这两者的乘积作为节点估计值[7],以 MAP 值最大的节点为谣言源,并通过实验证明了将 RC 作为 PEE 的计算方法效果最好。Ali 等提出了节点的免除和突出效应,并就感染节点的周边节点影响范围给出了理论基础[10]。但该方法有两个缺点:1)没有考虑节点之间传播概率不一的情况;2)对于复杂网络,仅仅以某感染点附近一定范围内的感染点、非感染点年龄来考虑该点的年龄,但由于传播模式多种多样,比较容易出现误差。

3 预备知识

3.1 传播模型

传播模型主要用于模拟信息在复杂网络中的传播过程。流行病模型是谣言源检测问题中适用较为广泛的模型[14],主要包括易感染-已感染病毒传播模型(简称 SI 模型)、易感染-已感染-恢复的病毒传播模型(简称 SIR 模型)、易感染-已感染-易受感染的病毒传播模型(简称 SIS 模型)等[18]。SI 传播模型中,网络节点只有两种状态,即当一个易感染状态节点 u 被感染后,它的邻居节点 v 从 u 处接收到信息的时间服从指数分布,同时节点 u 永远不会恢复,一直保持已感染状态。而 SIR 模型中,节点被感染后可以以一定概率或者在一定时间内恢复至易感染状态。SIS 模型中,一个被感染而后恢复的节点,会再次进入易感染状态。本文中谣言在复杂网络中的传播过程采用 SI 模型进行模拟。

3.2 问题定义

谣言传播的社交网络用带权图 $G=\langle V,E,ip\rangle$ 来建模,其中,V 是节点集,E 是边集,ip 是边的权函数,边的权重表示边的两端点之间的感染概率。本文考虑 G 是无向带权图的情形,假定只有一个谣言源点 v^*,并采用 SI 模型进行谣言传播。谣言在网络 G 中传播了一段时间后,可获取感染子图 $GI=(VI,EI,ip)$,即包含感染点集 VI 以及感染点间边的集合 EI 的子图。本文将根据特定时刻获得的网络快照(即感染子图 GI)来定位谣言传播源点。

3.3 EPA

Ali 等提出了基于节点免除和突出效应的 EPA 算法来解决谣言源定位问题[10]。通过考虑在感染点周围一定距离内感染点和非感染点的影响来计算每个感染点的年龄,其中节点年龄最大的就被认为是谣言源点。具体公式如下:

$$Age(u)=\sum_{l=0}^{r-1}\sum_{v\in T(u,l)}\left(\frac{dI(v)}{d(v)}\right)/\left(\frac{1}{1+\ln d(v)}\right) \tag{1}$$

其中,u 为待估计节点,r 是感染子图半径,$T(u,l)$ 为以 u 为树根构建的广度优先搜索树(Breadth First Search, BFS)的第 l 层的感染点集,$d(v)$ 是感染点 v 的邻域内所有节点的数目,$dI(v)$ 是感染点 v 的邻域内所有感染点的数目。

该算法通过递归过程来计算某点 u 周围 $r-1$ 半径内的所有点的邻域感染点所占比重,进而求和得到该点的 $Age(u)$ 值。

3.4 Qos 多播路由问题

对 Qos 多播路由问题的描述如下:给定无向图 $G=(V,E)$,任意边 $e\in E$ 有权重 $cost(e)$。设有点集 $A\subset V$,且集合 A 中某点 s 为组播源节点,$D\subseteq A-\{s\}$ 为组播目的节点集合,寻找一棵从源点 s 出发到所有目的节点 D 的组播树 T 并使其成本最小(成本即为形成该树的边成本总和)。该问题目标函数为:

$$\min C(T)=\sum_{e\in ET} cost(e) \tag{2}$$

其中，ET 为形成组播树 T 的边集。

Qos 多播路由问题已经被证明是 NP-Hard 问题，目前解决 Qos 多播路由问题的方法主要是组合优化以及智能优化这两类[19]。

3.5 算法简介

为了提高谣言源定位准确率，本文采用文献[7]所提出的 MAP 估计器。具体公式如下：

$$P(v^*=v|GI)=P(v^*=v)\times P(GI|v^*=v)\div P(GI) \tag{3}$$

$$P(v^*=v|GI)=P(v^*=v)\times P(GI|v^*=v)\div\left(\sum_{u\in GI}P(GI|v^*=u)\right) \tag{4}$$

$$\propto P(v^*=v)\times P(GI|v^*=v) \tag{5}$$

其中，$P(v^*=v)$是先验概率估计 PPE，若谣言源为 v，传播形成的传播子图为 GI，则 LE 值为 $P(v^*=v|GI)$。根据贝叶斯公式，如果分母 $P(GI)$是分子出现值之和，其可被认为是固定值而被移除，那么 $P(v^*=v)$和 $P(v^*=v|GI)$的乘积和 MAP 值成比例，即只需要计算每个估计点的 PPE 值和 LE 值的乘积即可，两者乘积最大的感染点就被检测为 MAP 值最大的源点。

尽管解决谣言源定位问题的众多算法都假定每个感染点有着相同的先验概率，但文献[20]的工作表明谣言源点很可能在传播子图中具有高中心性。本文在文献[7]的工作基础上，在 PPE 值计算过程中，以乘积方式融合节点覆盖率值，该值考虑了感染点周围所有感染点和非感染点对该点的影响，并以该值和 EPA 值乘积作为新的 PPE 值。本文将在 4.1 节对此进行详细描述。

在估计节点 LE 值的计算过程中，由于谣言传播过程和网络结构较复杂，文献[7]通过 GSBA 算法计算 LE 值的方法不符合真实传播过程，因此本文借助 Qos 多播路由问题的近似算法[11]，加入稀疏化复杂传播网络的过程，再基于 GSBA 算法计算 LE 值，能较为显著地提高 GSBA 算法的性能。其中，稀疏化过程中组播目的节点集合 D 为复杂传播网络边界点，而组播源节点 s 为待估计节点。边界点的概念定义如下。

定义 1 在传播子图所有感染点中，若某些感染点的一阶邻域内仅有一个点被感染，则这些感染点被认为是边界点。

以边界点和待估计节点共同组成近似算法的输入，利用该近似算法输出稀疏传播子图，即可基于 GSBA 算法计算估计节点 LE 值，本文将该算法称为贪心稀疏化近似算法，简称 SGSBA 算法。LE 值与之前 PPE 值乘积所得结果最大的节点就被认为是源点。

4 算法描述

4.1 先验概率估计的计算

PPE 值的计算公式如下：

$$P(v^*=u)=Age^+(u)\times C(u) \tag{6}$$

其中，$Age+(u)$为 EPA 算法在有权图上的计算扩展公式，$C(u)$为覆盖率公式。EPA 算法假定图中所有边传播概率一致，但本文中网络节点之间边的传播概率不一。因此本文在 EPA 算法的基础上提出了 Age^+ 公式，具体计算公式如下：

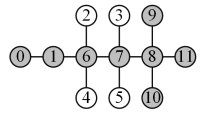
$$Age^+(u)=Age(u)\times\left(\sum_{u\in NI(v)}(1-ip(v,u))\div\left(\sum_{u\in N}(v)(1-ip(v,u))\right)\right) \tag{7}$$

其中，u 为该感染点，$NI(v)$为 v 点邻域中的感染点集，$N(v)$为 v 点邻域中的点集。$ip(v,u)$为 v 和 u 节点之间的边传播概率。式(7)计算该点邻域内感染点感染权重之和与该点邻域内所有节点感染权重之和的比值，将其与 $Age(u)$相乘，以将 EPA 算法扩展至有权图。

此外，当谣言传播出现传播分支大小不一的情况时，基于中心性的 $Age+(u)$值容易出现偏差。为了解决该问题，本文提出一种覆盖率($C(u)$)公式，其考虑了感染点周围所有感染点和非感染点对该估计节点 PEE 值的影响。具体公式如下：

$$C(u)=\left(\sum_{l=0} I/O\right)\div l \tag{8}$$

由式(8)可知，本文以任一 u 感染点为树根进行 BFS 树扩展直至树叶节点没有感染点。其中，l 为 BFS 树的层数，I 为该层感染点的数目，O 为该层所有节点数目。覆盖率的计算效果如图 1 所示。



注：灰色点表示被感染点，白点表示未被感染点

图 1 覆盖计算效果图

Fig. 1 Effect of coverage calculation

如图 1 所示，按照 $Age^+(u)$的计算方式，序号为 8 的节点 $Age^+(8)$值最大，约为 2.64。但如果按照 $C(u)$公式计算，序号为 7 的节点 $C(7)$值最大，约为 0.79。因此，直观上看，序号为 7 的点更可能是谣言传播源点。覆盖率在一定程度上可以弥补 EPA 算法在传播分支大小不一时的缺陷。本文采用这两种方法的乘积作为新的 PEE 值，简称该算法为 EC 算法。

4.2 似然估计的计算

基于 3.4 节，本文在每次 LE 值计算过程中，加入基于最小生成树的贪心算法[11]过程，以此稀疏化复杂传播网络，将该算法简称为 Sparse 算法，具体步骤如算法 1 所示。

算法 1 Sparse

```
Input:GI(VI,EI,ip),D,s /*GI为传播子图,D为边界点,s为估计源点*/
Output:Gs/*稀疏子图*/
Begin
1. M←(D∪s)/*M为终端节点集合*/
2. Gc←Complete(GI) /*生成由GI中节点构成的完全图,其中,Gc图中权重为节点间最短路径长度*/
3. Gc_sub←Subgraph(Gc,M) /*在Gc中获取M集合的子图*/
4. Gc_T←Min_tree(Gc_sub) /*获取Gc_sub图的最小生成树*/
5. Gs←Re(Gc_T) /*用GI节点之间最短路径替换Gc_T中的边*/
6. Retrun Gs
End
```

根据算法 1，首先获取由传播网络 GI 构建的完全图 Gc，其中，Gc 中边权重为图 GI 的两点之间最短路径长度；然后根据 Gc 生成包含终端节点的完全子图 Gc_sub，再获取该完全子图的最小生成树 Gc_T；最后将传播子图对应两点最短路径替换为该树的边，这样才能有效地精简传播网络 GI，得到稀疏子图 Gs。该稀疏子图在保留复杂传播网路 Gs 部分点的情况下，能够近似拟合真实传播结构，使得 LE 值计算过程更简洁有效。该算法简称 SGSBA 算法，具体步骤如算法 2 所示。

算法 2 SGSBA

Input: G(V,E,ip), GI(VI,EI,ip), D, s /* G 为网络，GI 为传播子图，D 为边界点，s 为估计源点 */

Output: LE(VI) /* 输出为每个节点的 LE 值 */

Begin

1. For node: V /* 用算法 1 精简传播子图 GI */
2. Gs←Sparse(GI) /* 用算法 1 精简传播子图 GI */
3. Queue←Ø/* 初始化空队列 */
4. LE(node)←0, W(node)=1 /* 初始化该估计节点 LE 值为 0，该点感染概率值为 1 */
5. Queue.add(node)/* 该估计节点加入队列 */
6. A←Ø /* 初始化集合，该集合存储已经被选择的节点 */
7. While A!=M: /* 初始化集合，该集合存储已经被选择的节点 */
8. u←Queue.pop() /* 队首元素出队，队首元素为感染点附近感染概率最大的点 */
9. LE(node)←LE(node)+W(u)/$\sum_{j\in Queue} W(j)$ /* 根据队首元素占整个队列概率总和值的比例，更新估计节点 LE 值 */
10. A.add(u), Queue.remove(u) /* 将计算过的 u 加入集 A，并在队列中移除 u */
11. For 每个 h∈N(u) /* N(u)为 u 的邻域 */
12. W(h)=1−(1−W(h))*(1−ip(u,h)) /* 更新 u 节点邻点 h 的概率值 */
13. Insert_descend_order(Queue,h) /* 将 u 节点每个邻居节点按照概率值大小降序插入到队列中 */
14. Return LE(VI)

End

由算法 2 可知，首先依次遍历每个感染点，利用算法 1 得到稀疏化传播子图并初始化 LE 值为 0；然后在该稀疏化子图上从估计节点开始每次贪心选择邻域感染概率最大的点，选择该点后，更新其邻域未感染点的感染概率值，直至遍历所有感染点得到稀疏化子图最大可能的感染序列；最后将形成该贪心序列得到的 LE 值进行累加，从而得到整个感染序列的 LE 值。综上所述，本文总体算法 EC-SGSBA(Effect of Coverage and Sparse for Greedy Search Bound Approximation)的步骤如算法 3 所示。

算法 3 EC-SGSBA

Input: G(V,E,W), GI(VI,EI,W), D, s /* G 为网络，GI 为感染传播子图 */

Output: estimate_source/* 估计源点 */

Begin

1. LE(VI) /* 基于 SGSBA 算法得到每个感染点的 LE 值 */
2. For node: PPE(VI) /* 基于 EC 算法计算每个感染点的 PPE 值 */
3. For node: VI
4. MAP(node)←PPE(node)*LE(node) /* 将 PPE 值和 LE 值乘积作为每个感染点的 MAP 值 */
5. Descend_order(VI) /* 将感染点按照 MAP 值大小进行降序排序 */
6. estimate_source←Max(VI) /* 取 MAP 值最大节点为估计源点 */
7. Retrun estimate_source

End

5 实验

5.1 数据集

考虑到本文利用稀疏化复杂传播网络来近似传播结构，因此本文所提算法对比其他算法更适用于复杂网络场景。本文使用 networks 库函数[21]生成标准无标度图和小世界网络，并在现实世界的真实社交网络数据集[22]上进行实验。因为本文算法针对的是无向有权图，所以本文对所有数据集都进行有向转无向以及给每条边随机赋予感染权重(0～1)的操作。数据集描述如表 1 所列。

表 1 数据集描述

Table 1 Description of dataset

数据集名	点数	边数
生成的无标度图	500	996
生成的小世界网络	3000	6000
Email-Eu-core	1005	25571
Face-book	4039	88234
Wiki-vote	7066	100736

实验针对数据集中每个算例网络，在网络中随机选择一点作为谣言源点，并使用 SI 模型进行模拟传播，当其传播到一定感染节点数目时停止。

5.2 算法复杂度对比

为了验证本文 EC-SGSBA 算法的有效性，将其与已有的多种算法进行对比，对比算法包括谣言中心(RC)、距离中心(Distance Center, DC)、乔丹中心(Jordan Center, JC)、反转感染中心(Reverse Infection, RI)、Dynamic Importance(DI)、GSBA-RC、EPA+、BFSA-RC 等。此外，知名的算法还包括 DMP 算法，但由于它时间复杂度较高，耗时长，实验没有将其与本文方法进行比较。首先进行各种算法的时间复杂度对比，结果如表 2 所列。

表 2 算法复杂度对比

Table 2 Complexity comparison of algorithms

算法	时间复杂度
RC	$O(n_I^3)$
DC	$O(n_I^3)$
JC	$O(n^3)$
RI	$O(n^3)$
DI	$O(n_I n^2)$
EPA+	$O(n^3)$
GSBA-RC	$O(2n_I^3)$
BFSA-RC	$O(n_I^3+k^{nl})$
EC-SGSBA	$O(3n_I^3)$

其中，n_I 表示感染节点数目，n 表示网络节点数目，m 表示边数目，t 表示传播时间，k 表示平均度。由表 2 可知，本文的 *EC-SGSBA* 算法在综合对比时间复杂度上没有明显优势，这是因为该算法基于 *GSBA-RC* 和 *BFSA-RC* 算法进行了改进，对于前者，该算法定源性能更好，对于后者，该算法时间复杂度更低，所以本文算法可以对性能和时间复杂度进行较好的均衡，本文将在下文阐述其改进措施的有效性。

5.3 源定位评价标准

为了使得评价更为合理，本文采用文献[7]中的 3 种源定位评价标准。

(1)准确率(Detection Rate)：准确率为多次运行算法时能正确定位谣言源的运行次数占算法总运行次数的比值。

(2)拓扑误差(Detection Error)：拓扑误差为真实源点与算法所定源点之间最短拓扑距离的平均值。

(3)归一化排序(Detection Ranking)：将感染节点估计值进行降序排列，标准化排名定义为$(Ranking(v^*)-1)/N$，其中，N 是感染点数目，$Ranking(v^*)$是 v^* 在降序序列中的排序序号。

5.4 实验结果

由于本文是在文献[7]的工作基础上分别对先验概率估计和似然估计的计算方式做出了改进，从而在 GSBA-RC 算法以及 BFSA-RC 算法的基础上提出了 EC-SGSBA 算法，因此首先就这 3 种算法进行对比实验，然后对整个 EC-SGSBA 算法与其他算法进行综合对比实验。

(1)EC-SGSBA 算法改进措施的有效性

为了验证 EC-SGSBA 算法改进措施的有效性，本文将其与原来的 GSBA-RC 算法以及 BFSA-RC 算法在相同的数据集上进行对比实验。实验在具有代表性的无标度网络、小世界网络中进行 100 次独立演算，因为 BFSA-RC 算法时间复杂度为指数级别，所以每次随机选择一点作为谣言源点并将其传播至 10 个感染点时就停止传播，得到具有代表性的拓扑误差指标和运行时间，并取实验平均值作为实验结果。具体结果如图 2 和表 3 所示。

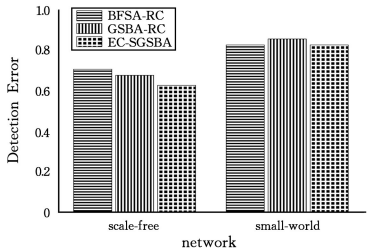


图 2 EC-SGSBA 算法的效果

Fig. 2 Effect of EC-SGSBA algorithm

表 3 算法运行时间的对比

Table 3 Comparison of running time of algorithms

算法	平均每次运行时间/s
BFSA-RC	63.75
GSBA-RC	0.01
EC-SGSBA	0.50

由图 2 和表 3 可知，本文提出的 EC-SGSBA 算法在感染节点较少的情况下，拓扑误差低于其他两种算法；在运行时间上，EC-SGSBA 算法的运行时间远远低于 BFSA-RC 算法，但略高于 GSBA-RC 算法，因此本文算法可以较好地均衡性能和时间复杂度。但由于感染节点较少，感染网络结构较为简单，EC-SGSBA 算法稀疏化复杂网络的优势尚未完全发挥，在之后的综合实验中，稀疏化复杂传播网络的过程会在一定程度上提高整个算法的性能。

(2)EC-SGSBA 算法有效性的综合验证实验

为了更全面直观地观察实验结果，本文将以 EC 算法所得的 PEE 值和 SGSBA 算法所得的 LE 值的乘积作为结果来综合对比其他方法，以此验证整个算法的优越性。其中，BFSA-RC 算法由于其指数级别的时间复杂度，并未被加入对比实验中。本文首先在小世界网络中进行以感染点数目为自变量的实验，然后扩展至其他算例网络。实验结果如图 3 所示。可以看出，在小世界网络的实验中，本文将感染点数目从 20 递增至 300，本文算法在 3 个指标上都超越了其他算法。其中，文献[7]的 GSBA-RC 算法在感染点变多的情况下，传播感染网络会变得较为复杂，LE 计算过程容易出现误差，导致整体效果变差。本文算法通过稀疏化传播子图的过程使得 LE 值的计算过程更贴近真实传播情况，因此即使感染点增多，本文算法仍然可以维持较好的性能。

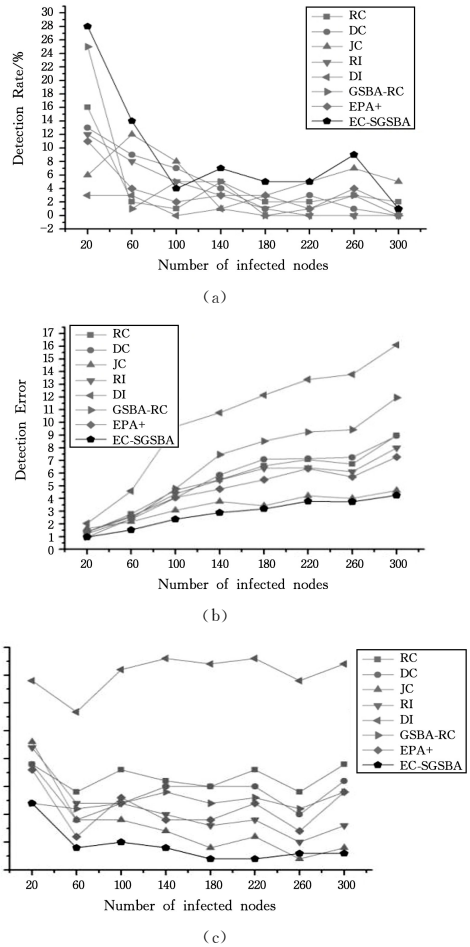


图 3 小世界数据集上的实验结果

Fig. 3 Experiment results on small world datasets

在其他 4 个数据集中，本文将感染点数目固定为 100，所得实验结果如图 4 所示。

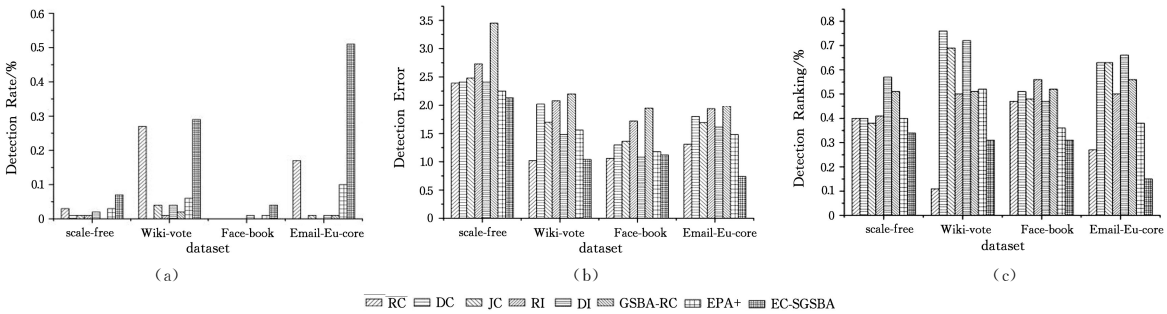


图 4 其他数据集上的实验

Fig. 4 Experiments on other datasets

在无标度网络以及真实社交网络的数据集中，感染点数目固定。由图4可知，本文算法在准确率、拓扑误差以及归一化排序指标上均优于其他法，同时在归一化排序指标上相比其他算法的稳定性更好，这是因为该算法综合考虑了全局节点对PEE值计算的影响，同时利用了稀疏化子图的方法，所以其只需要复杂网络的边界点即可完成运算，能够在一定程度上减小复杂网络给算法稳定性带来的影响。

结束语 针对复杂网络中的谣言定位源问题，本文首先基于MAP分别对PEE计算方式和LE计算方式做出改进。对于PEE计算方式，本文考虑了全局感染点和非感染点对估计节点的影响，并结合估计节点周围一定范围内节点的状态得出估计节点的PEE值。对于LE计算方式，本文设计了一种基于最小生成树算法的稀疏化复杂传播网络方法，让LE值的计算过程更符合真实传播过程。通过与现有方法在生成网络和真实网络中的对比实验，验证了本文所提方法的有效性和优越性。基于全局节点的PEE计算方式较为复杂，计算复杂度高，而LE值计算方式的稀疏传播子图虽然可以大幅度缩短LE值的计算时间，但其稀疏化过程耗时较长，因此，进一步的研究工作主要是精简算法流程、提高定位速度以及考虑更好的稀疏化传播网络方法。

参 考 文 献

[1] SHELKE S, ATTAR V. Source detection of rumor in social network-A review[J]. Online Social Networks and Media, 2019, 9: 30-42.

[2] SHAH D, ZAMAN T. Detecting sources of computer viruses in networks: theory and experiment[C]// Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. 2010: 203-214.

[3] DONG W, ZHANG W, TAN C W. Rooting out the rumor culprit from suspects[C]// 2013 IEEE International Symposium on Information Theory. IEEE, 2013: 2671-2675.

[4] FUCHS M, YU P D. Rumor source detection for rumor spreading on random increasing trees[J]. Electronic Communications in Probability, 2015, 20: 1-12.

[5] WANG Z, DONG W, ZHANG W, et al. Rumor source detection with multiple observations: Fundamental limits and algorithms [J]. ACM SIGMETRICS Performance Evaluation Review, 2014, 42(1): 1-13.

[6] CHANG B, ZHU F, CHEN E, et al. Information source detection via maximum a posteriori estimation[C]// 2015 IEEE International Conference on Data Mining. IEEE, 2015: 21-30.

[7] CHANG B, CHEN E, ZHU F, et al. Maximum a posteriori> estimation for information source detection[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018: 2242-2256.

[8] KAI Z, LEI Y. Information Source Detection in the SIR Model: A Sample Path Based Approach[J]. IEEE/ACM Transactions on Networking, 2012, 24(1): 408-421.

[9] JI F, TANG W, TAY W P. Properties and applications of Gromov matrices in network inference[C]// 2018 IEEE Statistical Signal Processing Workshop(SSP). IEEE, 2018: 478-482.

[10] ALI S S, ANWAR T, RASTOGI A, et al. EPA: Exoneration and Prominence based Age for Infection Source Identification[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 891-900.

[11] KOU L, MARKOWSKY G, BERMAN L. A fast algorithm for Steiner trees[J]. ACTA Informatica, 1981, 15(2): 141-145.

[12] SRIDHAR A, POOR H V. Sequential Estimation of Network Cascades[J]. arXiv: 1912.03800, 2019.

[13] ZHU L, WANG B. Stability analysis of a SAIR rumor spreading model with control strategies in online social networks[J]. Information Sciences, 2020, 536: 1-19.

[14] HETHCOTE H W. The mathematics of infectious diseases[J]. SIAM Review, 2000, 42(4): 599-653.

[15] SHELKE S, ATTAR V. Source detection of rumor in social network-a review[J]. Online Social Networks and Media, 2019, 9: 30-42.

[16] LOKHOV A Y, MÉZARD M, OHTA H, et al. Inferring the origin of an epidemic with a dynamic message-passing algorithm [J]. Physical Review E, 2014, 90(1): 012801.

[17] FIORITI V, CHINNICI M. Predicting the sources of an outbreak with a spectral technique[J]. arXiv: 1211.2333, 2012.

[18] BAILEYN T J. The mathematical theory of infectious diseases and its applications[M]. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE, 1975.

[19] OLIVEIRA C A S, PARDALOS P M. A survey of combinatorial optimization problems in multicast routing[J]. Computers & Operations Research, 2005, 32(8): 1953-1981.

[20] COMIN C H, LUCIANO D F C. Identifying the starting point of a spreading process in complex networks[J]. Physical Review E, 2011, 84(5): 056105.

[21] https://networkx.github.io/documentation/stable/news.html#networkx-2-4.

[22] https://snap.stanford.edu/data/.

BAO Zhi-qiang, born in 1995, postgraduate. His main research interests include graph theory, computation complexity, and algorithm.

CHEN Wei-dong, born in 1968, Ph. D, professor, Ph. D supervisor. His main research interests include graph theory, computation complexity, algorithm and machine learning.