文章编号：1003-0077(2018)08-0120-08

传播源估计中有效观察点部署策略研究

刘 栋[1,2]，赵 婧[1]，聂 豪[1]

(1. 河南师范大学 计算机与信息工程学院，河南 新乡 453007；
2. 教学资源与教育质量评估大数据河南省工程实验室，河南 新乡 453007)

摘 要：谣言或疾病的扩散均可模拟为传播源在网络中的传播，如何在网络中估计传播源位置是一项具有挑战性的任务。该任务往往根据部分观察点推断传播源的位置，故如何有效的选择观察点对准确定位传播源位置至关重要。该文分析了随机、度、聚类系数、特征向量、紧密度以及介数等观察点部署策略对传染源估计的影响。在实验中，采用 SI 传播模型和反向贪心算法估计传播源在三类合成网络和四个真实网络进行模拟仿真，实验结果表明采用特征向量的观察点部署策略更有利于提高传播源估计的精度。

关键词：复杂网络；传播源；观察点

中图分类号：TP391 **文献标识码**：A

A Study on Deployment Strategy of Efficient Observers for Locating Spreading Source

LIU Dong[1,2], ZHAO Jing[1], NIE Hao[1]

(1. School of Computer and InformationEngineering, Henan Normal University, Xinxiang, Henan 453007, China;
2. Big Data Engineering Laboratory for Teaching Resources and Assessment of Education Quality, Xinxiang, Henan 453007, China)

Abstract: Spread of a rumor or a disease can be modeled as propagation in a network. To estimate the spreading source in a network, the partial observers are necessary. To select effective observers, this work analyzes the influence of deployment strategies including random, degree-based, clustering-based, eigenvector-based, closeness-based and betweenness-based method. The experiments simulate three kinds of synthetic networks and four real networks using SI propagation model and reverse greedy algorithm. The results show that eigenvector deployment strategy is most contributive to the accuracy of estimating the spreading source.

Key words: complex network; spreading source; observers

0 引言

伴随着 Facebook、微信等在线社交网络服务的出现，社交网络已经成为当今社会人们信息交流的重要渠道和载体，且这些社交网络允许用户建立自己的“媒体”，对外发布、传播信息。可靠的信息能够给人们的生活带来便捷，但谣言在互联网上的传播所带来的负面影响也不可估量。传染病的传播也可看作网络上的传播现象，如 2009 年由流感病毒新型变体甲型 H1N1 流感所引发的全球性流行病疫情[1]，在很短时间内便蔓延全球。

谣言或疾病的传播都是依赖于具体网络而存在的。例如，谣言传播依赖于社交网络，疾病传播则依赖于人际交互网络。这些网络一般具有规模大，节点之间的联系复杂等特征。在此类网络上，如何快速准确地确定疾病的传染源、谣言的源头就显得十分重要。如果在短时间内确定出疾病的传染源，就能够更好地控制整个疾病的传播，缩小其传播范围，减少传染病对人们生命安全的威胁。如果能在散布谣言后的关键时期迅速定位谣言来源，就能减少谣言对政治、人类切身利益和社会稳定的负面影响。

收稿日期：2017-09-29 **定稿日期**：2017-10-10

基金项目：国家自然科学基金(U1404604)；河南省科技攻关项目(152102310313)

因此，在这个背景下，如何在一个网络上定位一个传染源或者谣言的源头就成为一件非常有意义且十分具有挑战性的事情。

对于网络上的源点定位问题，一种较为朴素的方法是通过获取网络中每个节点在网络传播中的状态，根据其相关记录，确定信息传播的源点。虽然这种方法能够准确定位信息源，但成本极高，可行性差，对于大规模网络来说几乎无法实施。因此，需要一种有效可行的方法来估计和预测信息源的准确位置。最新的研究趋势是通过在网络中部署少量观察点，根据其接收的节点信息，推断出信息源点在网络中的位置。在该任务下，如何有效地部署观察点，从而提高传播源定位精度，成为了亟待解决的关键问题。本工作针对上述问题展开研究，力图揭示网络中观察点部署策略与源点估算精度的关系。

本工作提出了几种观察点部署策略，包括随机、度、聚类系数、特征向量、紧密度和介数。并且采用 SI 传播模型和反向贪心溯源算法在合成网络和真实网络中进行模拟仿真。通过分析不同观察点部署策略对于传播源定位精度的影响，期望寻找到有效的观察点部署策略。

1　相关研究

目前，估计传播源在传染病控制、舆情控制等方面的应用越来越广泛，国内外研究人员对于估计传播源相关问题的研究也越来越深入。根据快照信息范围可以分为完整节点信息定位和部分节点信息定位。其中有的是关于定位传染病的源头、有的是关于定位谣言的源头等等，但根据快照信息范围大致可以归为以上两类。

Shah 等人[2]针对正则树上的 SI 模型提出一个基于传播中心性的极大似然估计算法来确定传播源。Fioriti 等人[3]提出了计算感染图中每个节点的动态年龄的溯源算法 DA。Comin 等人[4]在雪崩传播模型下提出了无偏中介算法。Prakash 等人[5]提出了基于最小描述长度的 NetSleuth 算法解决溯源问题。还有 Nino Antulov-Fantulin 等人[6]提出了基于蒙特卡洛模拟的定位算法。

以上工作都需要明确所有节点的传播状态，在大型网络中这将要消耗很大的成本。为了克服这些困难，Pinto[7]设计了一个基于少量观察点的精细极大似然估计算法来定位传播源。Shen 等人[8]开发了基于压缩感知的构建方法并提出了一种基于观察点反向传播的一种定位条件[9]。Luo 等人[10]提出了一种适用于一般网络的反向贪心算法。这些方法的基本思想都是基于选择的部分节点进行传播源的估计。定位传播源的精度在很大程度上取决于观察点的选择。

近几年来国内外研究人员在观察点部署策略上也取得了一定的进展。Brunella Spinelli 等人[11]提出了一种在线迭代的部署最佳观察点的方法来定位传播源。Brunella Spinelli 等人[12]还提出了基于传播延迟方差的最大值和最小值来部署观察点的方法。Celis 等人[13]提出基于双重解决集的部署观察点策略。张锡哲等人[14-17]基于 Pinto 的极大似然估计溯源算法提出的一系列观察点部署策略。上述研究均未基于反向贪心溯源算法提出观察点部署策略。

其中，张锡哲等人[17]分析了基于节点中心性的观察点部署策略对定位传播源的精度的影响，他们的研究结果表明基于节点中心性的观察点部署策略对于估计传播源的精度并无显著影响。与他们的研究结果不同的是，本工作同样采用基于节点中心性的观察点部署策略，但仿真结果表明采用特征向量的观察点部署策略更有利于提高传播源估计的精度。

分析与张锡哲等人研究的不同之处，主要在于他们的工作采用了基于少量观测节点的极大似然估计算法定位传播源，而本工作采用了基于传播路径的反向贪心算法。因此，估计传播源的溯源算法也是关系到基于节点中心性观察点部署策略适用性的重要因素。

本工作中，首先利用 SI 传播模型模拟传播。然后，根据传播信息利用反向贪心算法定位传播源，本文将在第二节中介绍所采用的传播模型和定位方法。我们的主要工作是在模拟传播完成后，提出了基于节点中心性的观察点部署策略，根据各种策略选择的部分观察点进行传播源定位，本文将在第三节介绍基于节点中心性的观察点部署策略。第四节介绍实验部分，最后在第五节做出结论。

2　所采用的传播模型与定位方法

2.1　SI 传播模型

网络可建模为一个无向图 $G=(V,E)$。其中 V 是顶点或节点的集合，E 是边的集合，代表两节点

之间有联系。在本工作中,假设在图 G 中最开始只有一个节点被“感染”,这个节点开始向它的邻居节点传播。本工作就是利用网络中部分感染节点信息来确定这个传播源。

在人群中的疾病传播模型已经在文献[18-20]中得到广泛的研究,这些模型也被用来模拟在线社交网络中信息的传播[21-24]。SI 模型是作为模拟疾病传播的一种自然现象出现的。在 SI 模型中,节点有三种状态:已感染状态、易感染状态和不易感染状态。已感染节点将会感染其他节点并且一直保持被感染的状态;易感染节点表示目前未被感染,但至少有一个邻居节点被感染;不易感染的节点也是未被感染的节点但是它的邻居都未被感染。类似于传染性疾病的传播,该模型也能描述谣言的传播,在模型中的单个人可分为两种类型:第一类人是不知情者,他们对谣言是浑然不知的;第二类人是传播者,他们是谣言的知情者,并且喜欢对谣言进行传播。

本工作中,采用的 SI 模型的时间是被划分了的离散时间段,在时间段 t 内节点 u 的状态用随机变量 $X(u,t)$ 表示。在时间 $t=0$ 时,假设只有一个节点 $v^* \in V$ 被感染,称这个节点为传播源。假设感染过程是一个在离散时间内概率测度为 P 的马尔可夫过程,且假设在下一个时间段开始。一个易感染节点被感染的概率为 p,其中 $p \in (0,1)$。假设每个易感染节点是否会被感染都是相互独立的,且不易感染节点始终都不会被感染。

在这个 SI 模型中,任何与被感染节点相邻的健康节点被感染的概率都相同,与被感染的邻居数量无关。传播一段时间后,在所有节点的状态信息中,健康节点的状态都能正确地显示出来,但是被感染节点的状态只有一部分能正确显示,称为显式状态,另一部分将与健康状态混淆,称为隐式状态。当易感染节点 u 被感染并且为显式状态时用 $X(u,t)=e$ 表示,并且被感染后为显示状态的节点集用 V_e 表示;当易感染节点 u 被感染并且为隐式状态时用 $X(u,t)=i$ 表示。

用 q_u 表示节点 u 被感染后为显式状态的概率,假设节点被感染后是否为显式状态与其他节点是相互独立的。对于所有节点 $u \in V$,如果 $q_u=1$,那么该模型本质上会和文献[25-26]中的一致;然而对于所有节点 $u \in V$ 如果 q_u 趋近于 0,则大多数节点将会表现为未感染状态。在这种情况下,估计传播源将会变的十分困难,并且在实际中估计传播源的精度将很低。直观地说,对于所有节点 u 如果 $q_u \geqslant p$,则将会在这个 SI 模型中观察到更多的被感染后为显式状态的节点,有利于提高估计传播源的精度。在本工作中提出一个较弱的假设,对于所有节点 $u \in V$,假设 q_u 满足

$$\max\left(0,2-\frac{1}{p}\right) \leqslant q_u \leqslant 1 \tag{1}$$

当 $p>1/2$ 时,对于任意节点 u,假设中的 q_u 需要满足足够大。这是因为在足够长的一段时间内,大部分节点将会被感染,若被感染的节点大都为隐式状态,那么将很难估计传播源;另一方面,如果 $p \leqslant 1/2$,那么假设将微不足道。此时,对于任意节点 u,该模型允许 $q_u=0$ 或者 1,当 $q_u=1$ 时,意味着将要观察所有被感染的节点,这种情况类似于文献[27]中的情况。故总结一个易感染节点的概率转换图如图 1 所示。

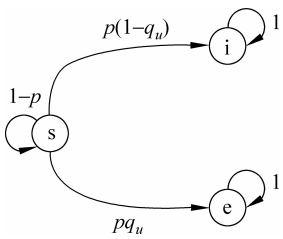


图 1 易感染节点的概率转换图

2.2 反向贪心溯源算法

本工作采用 Luo 等人[10]提出的方法估计传播源。首先寻找与被感染节点集合一致的感染树,然后计算可生成该感染树的最可能传播路径,最后估计该树的 Jordan 感染中心[28]作为网络的传播源。

在网络 G 中,假设一个易感染节点只能随机地被已感染邻居节点中的一个节点感染,则在传播源传播结束后,被感染的节点组成的所有路径将会是一个树。经过传播 t 时间后,任意一条感染路径用 $X^t=\{X(u,\tau):u \in V, 1 \leqslant \tau \leqslant t\}$ 表示,则由 X^t 形成的 $T(X^t)$ 为 G 的一棵子树。T_v 表示以节点 v 为根节点的树,则在 G 的所有子树中肯定存在一个与被感染节点集合组成的感染树一致的树。T_v 表示由已感染且为显式状态的节点 V_e 组成的感染树的集合。假设节点 v 为源点,该传播源的最可能感染路径,如式(2)所示。

$$\max_{v \in V} \max_{\substack{t \in \tau_v \\ X^t \in \chi_v}} P_v(X^t) = \max_{v \in V} \max_{T \in \mathrm{T}_v} \max_{\substack{t \in \tau_v \\ X^t : T(X^t)=T}} P_v(X^t) \tag{2}$$

在树中,Jordan 感染中心就是它估计的传播源。在网络 G 中,先找到一个从传播源 v 到节点集

V_e 中各个节点的最短路径树，再把在 G 中与该树中与节点集 V_e 中每个节点有联系的边相连，构成的子图称为 H_v，把 T_v 近似认为 H_v 的集合并用 $\hat{\mathrm{T}}_v$ 表示，则近似传播源，如式(3)所示。

$$\tilde{s} = \max_{v\in V} \max_{T\in \hat{\mathrm{T}}_v} \max_{\substack{t\in \tau_v \\ X^t:T(X^t)=T}} P_v(X^t) \tag{3}$$

下面的工作需在 $\hat{\mathrm{T}}_v$ 集合中找出最可能的感染树。对于任意传播源点 v 和任意一个感染树 $T\in \hat{\mathrm{T}}_v$，用 $D_T(u)=\bar{d}(u,T_u(v;T))$ 表示以 v 为根节点的树 T 的高度。

假设 G 为一般网络，且 $v\in V$ 是传播源。那么，对于任意一个以 v 为根节点且由节点集合 V_e 组成的感染树 T 有：

$$\begin{aligned}&\max_{\substack{t\in \tau_v \\ X^t:T(X^t)=T}} P_v(X^t)\\ &= p^{|T|-1}(1-p)^{\sum_{u\in T\setminus\{v\}}(D_T(pa(u))-D_T(u))-|T|+1}\\ &\prod_{u\in V_e} q_u \prod_{u\in T\setminus V_e}(1-q_u)\end{aligned} \tag{4}$$

其中 $pa(u)$ 表示节点 u 的父母节点，令

$$F_v^* = \min_{T\in \hat{\mathrm{T}}_v} \sum_{u\in T\setminus\{v\}} (D_T(pa(u))-D_T(u)) \tag{5}$$

如果 F_v^* 和 T_v^* 取最小值，则当

$|T_v^*|\log p+(F_v^*-|T_v^*|)\log(1-p)+$

$\sum_{u\in T_v^*\setminus V_e}\log(1-q_u)$ 取最大值时，就可以找到这个传播源 v。

为了取得 F_v^* 的最小值，Luo 等人提出令

$$F_v^* = \min_{T\in \hat{\mathrm{T}}_v} \sum_{u\in T\setminus\{v\}} (Deg_T(u)-2)D_T(u)+2D_T(v) \tag{6}$$

其中，$Deg_T(u)$ 表示在树 T 中节点 u 邻居节点的个数。从式(6)中看出为了得到 F_v^* 的最小值，如果 $D_T(u)$ 比较大，则节点 u 的度应该比较小。基于这个意图，该定位方法采用反向贪心技术不断调整感染树使得接近源点的节点有较小度，而远离源点的节点的度数比较大，具体算法可以参考文献[10]。

3　基于节点中心性的观察点部署策略

在网络 G 中，当传播源利用 SI 传播模型感染结束后，将得到为显式状态的节点集合 V_e。在文献[17]中，Luo 等人是在随机选择部分节点上部署观察点来收集信息。而本工作中，提出了基于节点中心性的观察点部署策略，含节点的度、聚类系数、特征向量、紧密度、介数。下面介绍这五种指标信息。

(1) 度

用于描述在静态网络中节点产生的直接影响力，其值为与该节点直接相连的节点数。则度定义为式(7)。

$$C_d(i) = d(i) \tag{7}$$

其中，$d(i)$ 表示与节点 i 直接相连的节点数。

(2) 紧密度

用于描述网络中的节点通过网络到达其他节点的难易度，其值为该节点到其他所有节点的距离之和的倒数。则紧密度定义为式(8)。

$$C_c(i) = \frac{1}{\sum_{j=1}^{n} d_{ij}} \tag{8}$$

其中，d_{ij} 表示节点 i 到节点 j 的最短距离。

(3) 介数

节点的介数反映了其在整个网络中的作用和影响力，具有比较强的实际意义。则介数定义为式(9)。

$$C_b(i) = \frac{\sum_{j=1,k=1}^{j=n,k=n} l_{jk}(i)}{\sum_{j=1,k=1}^{j=n,k=n} l_{jk}} \tag{9}$$

其中，l_{jk} 表示节点 j 到节点 k 的最短路径条数，$l_{jk}(i)$ 表示节点 j 与节点 k 之间经过点 i 的最短路径条数。

(4) 聚类系数

在复杂网络中，假设网络中的一个节点 i 有 k_i 条边将它和其他节点相连，这 k_i 个节点就称为节点 i 的邻居。由于在这 k_i 个节点之间最多可能有 $k_i(k_i-1)/2$ 条边，而这 k_i 个节点之间实际存在的边数 E_i 和总的可能的边数 $k_i(k_i-1)/2$ 之比就定义为节点 i 的聚类系数 $C_d(i)$，E_i 为这 k_i 个节点之间实际存在的边数总和即

$$C(i) = \frac{2E_i}{k_i(k_i-1)} \tag{10}$$

聚类系数用于描述节点邻居之间连接的紧密程度。整个网络的聚类系数 C_d 就是所有节点 i 的聚类系数 $C_d(i)$ 的平均值。显然 $0\leqslant C_d\leqslant 1$，当且仅当所有节点都为孤立节点，即没有任何连接边的情况下 $C_d=0$；当网络是全局耦合时，即网络中任意两个节点都直接相连此时 $C_d=1$。

(5) 特征向量

用于描述节点周围邻居节点的重要性对该节点产生的影响。若 λ 为网络邻接矩阵 A 的主特征值，$e=(e_1,e_2,\ldots,e_n)$ 为矩阵 A 对应于主特征值 λ 的特征向量，则特征向量定义，如式(11)所示。

$$C_e(i) = \lambda^{-1} \sum_{j=1}^{n} a_{ij} e_j \quad (11)$$

从上面可以看出度、聚类系数、特征向量、紧密度以及介数这些指标信息，在网络结构中的意义差别还是很大的。有的指标信息反映节点在整个网络中的地位，有的指标信息揭示周围邻居对中心节点的影响等。这些差别在信息扩散的过程中都有所体现，在传播过程都发挥了各自不同的作用，这也使得这几类节点成为网络中的重点研究对象。同时，正是这种在传播时所表现出来的差异，显得有必要按照这几种指标信息选择观察点。因它们能从各自不同的方面揭示网络中的信息传播过程，能够更全面地感知信息在网中的扩散，而这些对于准确定位网络中的信息源来说是十分必要的。所以可利用以上指标信息作为选择观察点的标准。下面的工作通过实验来分析这些部署策略对合成网络和实际网络的定位传播源精度的影响。

4 实验

4.1 实验设置

本工作选取两类网络进行实验。第一类是合成网络，分别为从BA模型构造的无标度网络、从WS模型构造的小世界网络和从ER模型构造的随机网络；第二类为真实网络，分别为属于社会网络的仓鼠网络[29]，即在仓鼠网站中用户的友谊与家庭之间的联系网络；属于技术网络的美国电力网络[30]；属于信息网络的合著关系网络[31]，选择了其中最大连通子图379个作者的关系网络；属于生物网络的线虫代谢网络[32]。表1列出了这些网络的规模。

本工作根据度、介数、紧密度、聚类系数以及特征向量五种网络节点指标信息就可以产生五种观察点部署策略。即在选择观察点之前，根据网络节点指标信息的定义，计算网络中每个节点的指标信息。之后将该指标进行从大到小排序，最后根据所需要的观察点数量按照指标从大到小选择节点作为观察点。除此五种观察点部署策略之外，还有一种就是随机选择节点作为观察点，将此种部署策略作为与其他部署策略的对比。之后提到的观察点部署策略也就是指的以上六种。

本工作中，设置选择节点集 V_e 中50%的节点作为观察点。如根据节点度大小来选择50%的节点作为观察点，则对网络中所有节点按照度的大小排序，最后选择排在前50%的节点作为观察点。如果是在随机选择观察点的情况下，就不需要排序，直接根据观察点的数量在网络中随机选择即可。

表1 实验网络的规模

网络类型	网络拓扑	节点数	连边数	直径
合成网络	小世界网络	200	6 098	2
	随机网络	200	2 038	3
	无标度网络	200	788	4
真实网络	仓鼠网络	2 000	16 097	10
	美国电力网络	4 941	6 594	46
	合著关系网络	379	914	6
	线虫代谢网络	453	2 298	7

在每种网络上，仿真模拟运行1 000次。在每次模拟运行中，先随机选择一个节点作为传播源，再使用SI模型模拟传播。对于任意节点 v，p 和 q_v 分别是在 $(0,1)$ 和 $[\max(0,2-1/p),1]$ 中均匀选择的。

4.2 实验结果和分析

通过研究在合成网络和真实网络上估计传播源精度的结果，探索了观察点部署策略与估计传播源精度的关系。从而发现观察点部署策略对于估计传播源精度产生的具体影响。

图2是小世界网络在各种观察点部署策略下溯源算法的错误距离分布图，错误距离是估计的传播源与真实的传播源之间的跳数。在观察点部署策略随机、度、聚类系数、介数、紧密度以及特征向量下估计传播源的精度分别为0.23、0.12、0.24、0.21、0.25以及0.27。实验结果表明采用特征向量观察点部署策略估计传播源的精度高于其他策略。

图3是随机网络在各种观察点部署策略下溯源算法的错误距离分布图。在观察点部署策略随机、度、聚类系数、介数、紧密度以及特征向量下估计传播源的精度分别为0.03、0.01、0.02、0.01、0.02以及0.08。实验结果表明采用特征向量观察点部署策略估计传播源的精度高于其他策略；另外，相比随机选择策略，采用特征向量观察点部署策略，准确定位传播源的精度可以提高到两倍多。

图4是无标度网络在各种观察点部署策略下溯源算法的错误距离分布图。在观察点部署策略随机、度、聚类系数、介数、紧密度以及特征向量下估计传播源的精度分别为0.10、0.21、0.05、0.15、0.21以及0.27。实验结果表明，采用特征向量观察点部

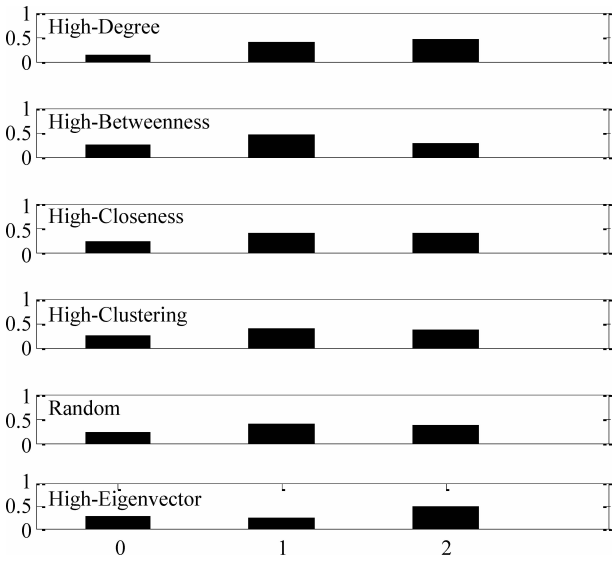


图 2　小世界网络在各种观察点部署策略下溯源算法的错误距离分布

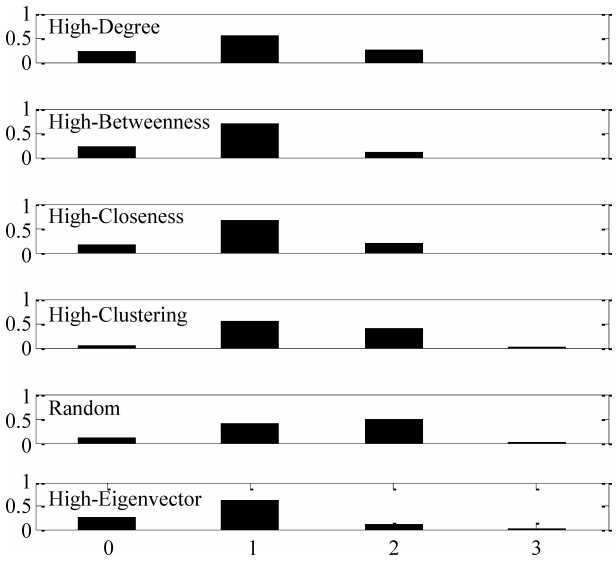


图 4　无标度网络在各种观察点部署策略下溯源算法的错误距离分布

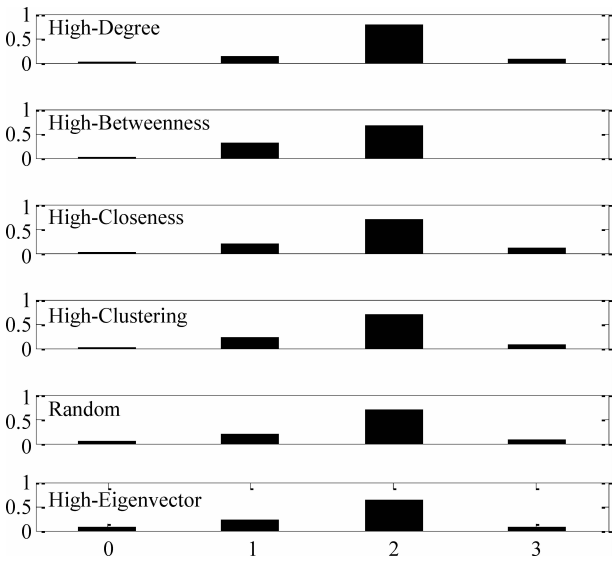


图 3　随机网络在各种观察点部署策略下溯源算法的错误距离分布

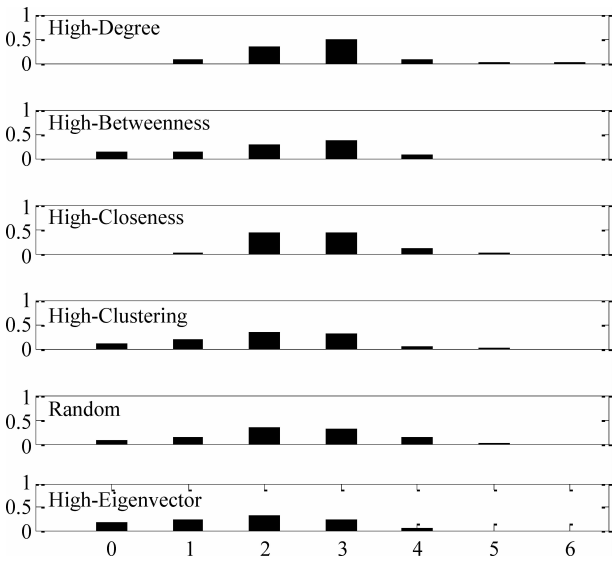


图 5　仓鼠网络在各种观察点部署策略下溯源算法的错误距离分布

署策略估计传播源的精度高于其他策略；另外，相比随机选择策略，采用特征向量观察点部署策略，准确定位传播源的精度可以提高到两倍多。

图 5 是仓鼠网络在各种观察点部署策略下溯源算法的错误距离分布图。在观察点部署策略随机、度、聚类系数、介数、紧密度以及特征向量下估计传播源的精度分别为 0.06、0.00、0.11、0.00、0.14 以及 0.16。实验结果表明，采用特征向量观察点部署策略估计传播源的精度高于其他策略；另外，相比随机选择策略，采用特征向量观察点部署策略，准确定位传播源的精度可以提高到两倍多。

图 6 是美国电力网络在各种观察点部署策略下溯源算法的错误距离分布图。需要说明的是，在文献[10]中指出在真实美国电力网络上采用基于传播中心性溯源算法的精度只有 3%，在实验中利用这六种观察点部署策略估计真实传播源的精度最高的也只有 1%。因为精度极低再加上错误距离最大的已经达到 20 跳，而美国电力网络的精度用的是错误距离在 3 跳以内的比例。在观察点部署策略随机、度、聚类系数、介数、紧密度以及特征向量下估计传播源的精度分别为 0.23、0.12、0.24、0.21、0.25 以及 0.27。实验结果表明，采用特征向量观察点部

署策略估计传播源的精度高于其他策略；另外，相比随机选择策略，采用特征向量观察点部署策略，准确定位传播源的精度可以提高到两倍多。

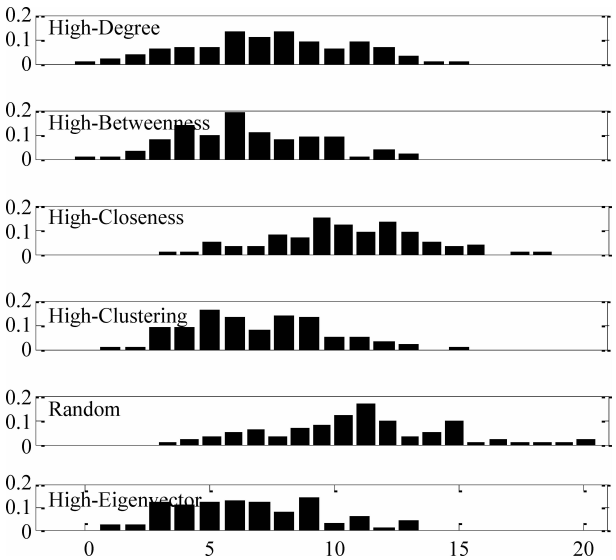


图6 美国电力网络在各种观察点部署策略下溯源算法的错误距离分布

图7是合著关系网络在各种观察点部署策略下溯源算法的错误距离分布图。在观察点部署策略随机、度、聚类系数、介数、紧密度以及特征向量下估计传播源的精度分别为0.02、0.02、0.01、0.00、0.02以及0.05。实验结果表明，采用特征向量观察点部署策略估计传播源的精度高于其他策略；另外，相比随机选择策略，采用特征向量观察点部署策略，准确定位传播源的精度可以提高到两倍多。

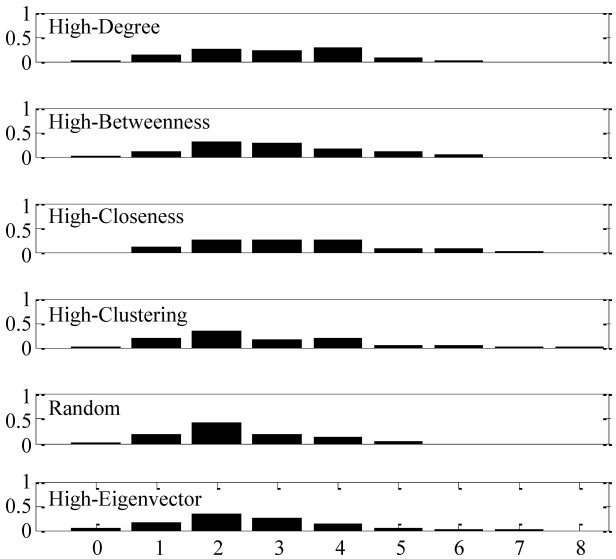


图7 合著关系网络在各种观察点部署策略下溯源算法的错误距离分布

图8是线虫代谢网络在各种观察点部署策略下溯源算法的错误距离分布图。在观察点部署策略随机、度、聚类系数、介数、紧密度以及特征向量下估计传播源的精度分别为0.06、0.01、0.05、0.00、0.00以及0.08。实验结果表明，采用特征向量观察点部署策略估计传播源的精度高于其他策略。

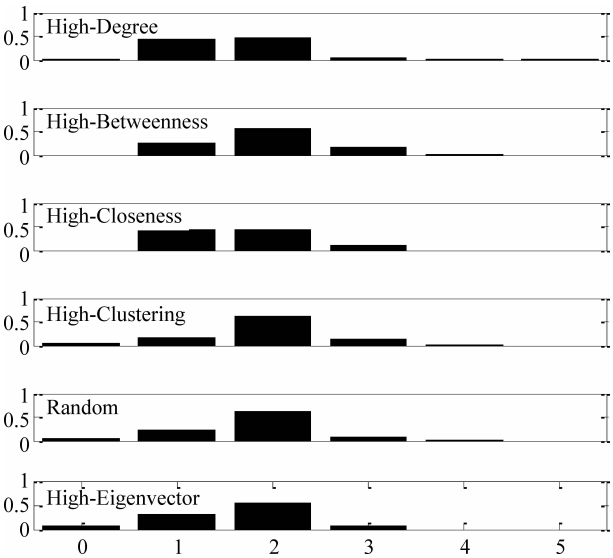


图8 线虫代谢网络在各种观察点部署策略下溯源算法的错误距离分布

本实验在上述七种网络拓扑结构上测试了这六种观察点部署策略对采用反向贪心溯源算法定位传播源精度的影响，将测试结果整理在表2中，并在表中将最优的性能值加粗表示。从表2可看出，采用特征向量观察点部署策略更有利于提高估计传播源的精度。另外，在无标度网络、随机网络、美国电力网络、仓鼠网络以及合著关系网络这些网络上。相比随机选择策略，采用特征向量部署策略，准确定位传播源的精度可以提高到两倍多。从这些错误距离分布图中可以看出每个网络都随着观察点部署策略的改变而改变，但是在各个网络中错误距离分布的趋势很相似。

表2 各种网络在六种观察点部署策略下定位传播源的精度

网络拓扑	随机	度	聚类系数	介数	紧密度	特征向量
小世界网络	0.23	0.12	0.24	0.21	0.25	**0.27**
随机网络	0.03	0.01	0.02	0.01	0.02	**0.08**
无标度网络	0.10	0.21	0.05	0.15	0.21	**0.27**
仓鼠网络	0.06	0.00	0.11	0.00	0.14	**0.16**
美国电力网络	0.06	0.07	0.11	0.07	0.13	**0.16**
合著关系网络	0.02	0.02	0.01	0.00	0.02	**0.05**
线虫代谢网络	0.06	0.01	0.05	0.00	0.00	**0.08**

5 总结

在网络中估计传播源是一项重要的研究课题。估计传播源的一种可行的方法，即是利用观察点收集的部分节点信息来定位传播源。故而，估计传播源与观察点的部署策略紧密相关。本研究评估了几种观察点部署策略，含度数、介数、紧密度、聚类系数、特征向量和随机。利用 SI 传播模型和反贪心溯源算法，在合成网络和真实网络上仿真模拟。实验结果表明，采用特征向量观察点部署策略更有利于提高估计传播源的精度，且与随机选择观察点部署策略相比，在某些网络上估计的传播源精度可以提高到两倍多。

基于节点中心性的观察点部署策略与采用的估计传播源的算法相关，而这些观察点部署策略对于其他溯源算法精度的影响需要进一步研究。是否能找到一种策略可以适用各种溯源算法，这些问题将会在以后的工作中做进一步的讨论。

参考文献

[1] 张玲霞，王永怡，陈文，等. 2009 年全球传染病疫情聚焦[J]. 传染病信息，2010，23(1)：4-7.

[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. ACM，2010：203-214.

[3] Fioriti V，Chinnici M. Predicting the sources of an outbreak with a spectral technique[J]. Computer Science，2012，8：6775-6782.

[4] Comin C H，Costa L F. Identifying the starting point of a spreading process in complex networks[J]. Physical Review E Statistical Nonlinear and Soft Matter Physics，2011，84(5 Pt 2)：056105.

[5] Prakash B A，Vreeken J，Faloutsos C. Efficiently spotting the starting points of an epidemic in a large graph[J]. Knowledge and Information Systems，2014，38(1)：35-59.

[6] Antulovfantulin N，LančićA，Šmuc T，et al. Identification of Patient Zero in Static and Temporal Networks：Robustness and Limitations[J]. Physical Review Letters，2015，114 (24)：248701.

[7] Pinto P C，Thiran P，Vetterli M. Locating the Source of Diffusion in Large-Scale Networks[J]. Physical Review Letters，2012，109(6)：068702.

[8] Shen Z，Cao S，Fan Y，et al. Locating the source of spreading in complex networks[J]. Computer Science，2015.

[9] Wang W X，Fan Y，et al. Reconstructing propagation networks with natural diversity and identifying hidden sources[J]. Nature Communications，2014，5(5)：4323.

[10] Luo W，Tay W P，Leng M. How to Identify an Infection Source With Limited Observations[J]. IEEE Journal of Selected Topics in Signal Processing，2014，8(4)：586-597.

[11] Spinelli B，Celis E L，Thiran P. Back to the Source：an Online Approach for Sensor Placement and Source Localization[C]//Proceedings of the 26th International World Wide Web Conference (WWW). Perth，Australia. 2017.

[12] Spinelli B，Celis L E，Thiran P. Observer Placement for Source Localization：The Effect of Budgets and Transmission Variance[C]//Proceedings of the 54th Annual Allerton Conference on Communication，Control，and Computing. Alabama ，USA. 2016.

[13] Celis L E，PavetićF，Spinelli B，et al. Budgeted sensor placement for source localization on trees[J]. Electronic Notes in Discrete Mathematics，2015，50：65-70.

[14] 张聿博，张锡哲，张斌. 基于观察点的信息源定位方法的精度分析[J]. 东北大学学报(自然科学版)，2015，36(3)：350-353.

[15] 张聿博，张锡哲，张斌. 面向社交网络信息源定位的观察点部署方法[J]. 软件学报，2014(12)：2837 - 2851.

[16] 晏迪. 面向网络扩散源点定位的观察点部署策略研究及定位算法优化[D]. 沈阳：东北大学硕士学位论文，2013.

[17] Zhang X，Zhang Y，Lv T，et al. Identification of efficient observers for locating spreading source in complex networks[J]. Physica A Statistical Mechanics and Its Applications，2016，442：100-109.

[18] Wood P H N. The Mathematical Theory of Infectious Diseases and its applications[M]. Griffin，1975.

[19] Allen L J. Some discrete-time SI，SIR and SIS epidemic models[J]. Mathematical Biosciences，1994，124(1)：83-105.

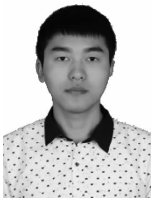
[20] Newman M E J. The structure and function of complex networks[C]//Proceedings of the SIAM Rev. 2003：167-256.

[21] Goldenberg J，Libai B，Muller E. Talk of the Network：A Complex Systems Look at the Underlying Process of Word-of-Mouth[J]. Marketing Letters，2001，12(3)：211-223.

(下转第 142 页)

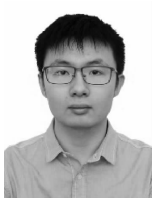
views[J]. WWW Workshop: NLP in the Information Explosion Era, 2008, 3 (3): 200-207.

[16] Yu J, Zha Z J, Wang M, et al. Aspect ranking: identifying important product aspects from online consumer reviews[C]//Proceedings of the Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011: 1496-1505.



陈放(1994—),本科生,主要研究领域为网络文本情感分析。
E-mail: f-chen17@mails.tsinghua.edu.cn

梁爽(1995—),本科生,主要研究领域为网络文本情感分析。
E-mail: s-liang13@tsinghua.org.cn



王颖(1996—),本科生,主要研究领域为网络文本情感分析。
E-mail: k-w17@mails.tsinghua.edu.cn

(上接第 127 页)

[22] Gruhl D, Guha R, Liben Nowell D, et al. Information diffusion through blogspace[C]//Proceedings of the International Conference on World Wide Web. ACM, 2004: 491-501.

[23] Cha M, Haddadi H, Benevenuto F, et al. Measuring User Influence in Twitter: The Million Follower Fallacy[C]//Proceedings of the International Conference on Weblogs and Social Media 2010, Washington, DC, USA, May. DBLP, 2010.

[24] Chou Y F, Huang H H, Cheng R G. Modeling Information Dissemination in Generalized Social Networks[J]. IEEE Communications Letters, 2013, 17 (7):1356-1359.

[25] Shah D, Zaman T. Rumors in a Network: Who's the Culprit? [J]. IEEE Transactions on Information Theory, 2011, 57(8):5163-5181.

[26] Luo W, Tay W P, Leng M. Identifying Infection Sources and Regions in Large Networks[J]. IEEE Transactions on Signal Processing, 2013, 61(11): 2850-2865.

[27] Pinto P C, Thiran P, Vetterli M. Locating the Source of Diffusion in Large-Scale Networks[J]. Physical Review Letters, 2012, 109(6):068702.

[28] Zhu K, Ying L. Information source detection in the SIR model: A sample path based approach[C]//Proceedings of the Information Theory and Applications Workshop. IEEE, 2013:1-9.

[29] Kunegis J, Hamsterster full network dataset[DB/OL]. http: //konect.unikoblenz.de/ networks/petster-hamster,KONECT. 2014.

[30] Watts D J, Strogatz S H. Collective dynamics of small-world networks[J]. Nature, 1998(393): 440-442.

[31] Duch J, Arenas A. Community identification using Extremal Optimization[J]. Physical Review E, 2005, 72.

[32] Newman M E J. Finding community structure in networks using the eigenvectors of matrices.[J]. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2006, 74(3 Pt 2):036104.

刘栋(1976—),博士,副教授,主要研究领域为复杂网络、机器学习等。
E-mail: liudong@htu.edu.cn

聂豪(1993—),硕士研究生,主要研究领域为复杂网络、网络中重要节点识别。
E-mail: niehao1993@gmail.com

赵婧(1989—),硕士研究生,主要研究领域为复杂网络、网络中定位传播源。
E-mail: zhaojinghn@gmail.com