



论 文

基于子图抽取的在线社交网络多传播源点定位方法

张锡哲^{①*}, 张聿博^①, 吕天阳^{②③}, 付世海^①, 张斌^①

① 东北大学计算机科学与工程学院, 沈阳 110819

② 哈尔滨工程大学计算机科学与技术学院, 哈尔滨 150001

③ 审计署计算机技术中心, 北京 100830

* 通信作者. E-mail: zhangxizhe@mail.neu.edu.cn

收稿日期: 2015-12-01; 接受日期: 2015-12-31

国家自然科学基金 (批准号: 60903009, 71272216, 91546110, 61572117) 和中央高校基本科研业务费 (批准号: N140404011) 资助项目

摘要 有效定位在线社交网络上的信息传播源点, 对于谣言检测、传播控制等应用具有重要的意义. 在线社交网络规模巨大, 消息往往在多处同时发起传播, 这使得现有的传播源点定位方法很难适用. 本文针对多源点传播定位问题, 首先分析了具有明显社团结构的社交网络中的信息传播规律, 发现社团结构是分离多个源点影响范围的有效途径. 在此基础上设计了一种基于子图抽取的源点定位方法, 算法基于社团特征抽取网络的局部子图, 使得每个子图中的观察点首先受到同一个源点的影响, 然后分别在各子图中独立定位传播源点, 从而将网络中的多源定位问题转换为子图中的单源定位问题. 算法无需预先获取网络的全部拓扑结构, 而是基于局部子图进行定位, 在极大地降低了计算复杂度的同时, 隔离多个源点间的相互影响, 提高了定位的准确率. 在模型网络 and 实际网络上的实验结果表明, 算法具有较高的定位准确率及效率. 该方法为谣言定位、传染病控制等应用提供了一种有效的解决方案.

关键词 社交网络 信息传播 社团结构 多源定位 子图抽取

1 引言

近年来, 随着以 Facebook、Twitter、微博、微信为代表的在线社交网络和社交媒体的飞速发展, 使得信息在前所未有的广阔范围内, 以惊人的速度传递和分享. 社交网络上的用户依据社交关联 (关注、好友等) 形成了复杂的网络结构, 谣言等不实信息在这种高度互联的复杂网络上传播, 往往会带来不可预料的严重后果^[1]. 如何分析、理解及控制社交网络上信息传播过程, 已成为计算机科学、社会学、物理学等多个学科的热点研究方向^[2,3].

为了分析和控制信息传播造成的影响, 需要了解信息传播的动力学过程^[4~6]. 然而由于信息的突然爆发, 我们很难确定信息的来源; 由于网络的规模巨大, 我们无法监测所有节点的状态; 由于网络的

引用格式: 张锡哲, 张聿博, 吕天阳, 等. 基于子图抽取的在线社交网络多传播源点定位方法. 中国科学: 信息科学, 2016, 46: 496–510, doi: 10.1360/N112015-00190

结构复杂, 我们很难还原信息的传播过程. 已有很多工作致力于建立网络传播模型^[5,6]及分析传播动力学过程^[7,8], 但是, 对于如何定位在线社交网络上的潜在信息传播源点我们还知之甚少, 这对社交网络中的谣言控制和计算机病毒控制等问题具有重要研究价值^[9].

对于复杂网络上的传播源点定位问题, 一类方法是获取网络的传播子网, 即网络中所有收到消息的节点构成的子图, 通过基于拓扑中心性^[10]或传播路径^[11]的分析, 找到最有可能的传播源点. 然而, 获取网络中所有被感染节点的传播子网代价巨大, 对于大规模社交网络上的信息传播过程是不可行的. 另一类方法^[12~14]根据网络中部分节点的传播状态, 推断潜在的传播源点. 这种方法的优点在于降低了对数据的要求, 在分析单源点传播过程时取得了良好的效果. 但是, 由于观测数据的稀疏, 对于多个传播源在网络中进行传播的情况, 无法分辨节点收到的是哪个源点传播的消息, 从而很难准确找出网络中多个传播源.

现有的方法在传染病溯源等领域得到了很好的效果, 但是如何定位在线社交网络上的潜在信息传播源点, 还存在很多困难, 具体来说: 1. 为了最大化消息传播的效果, 影响最大化等算法^[8]会在网络中部署多个源点同时传播消息. 这些源点的影响范围互相重叠, 使得我们很难准确定位潜在的源点, 甚至无法估计源点的个数; 2. 以往的定位算法需要全网拓扑结构^[9]或全部节点的传播状态快照^[10], 计算复杂度在 $O(N^3)$ 或以上, 这对于网络规模极大的在线社交网络很难适用.

针对上述问题, 本文提出了一种基于子图抽取的社交网络多传播源点定位方法. 首先分析了具有明显社团结构的在线社交网络中多源点同时发起传播时的特征规律, 发现社团结构是分离多个源点传播影响范围的有效途径. 源点发出的消息一般首先会在社团内部密集传播, 然后扩散到其他社团. 由于社团结构对于传播的阻碍作用, 与源点在同一社团的节点应该优先收到本社团源点发出的信息. 因此, 可以将在线社交网络中多传播源点定位问题转化为在多个局部子图中的单源点定位问题, 以此来解决定位多源点的困难.

然而, 获取大规模社交网络的社团结构仍然是一个困难的问题, 即使是复杂度较低的快速算法^[12], 进行大规模的社团划分仍然是非常耗时的. 对此, 我们提出一种面向源点定位的子图抽取方法, 基于社团结构特征, 从较早收到信息的节点附近开始进行扩展, 用于获得潜在信息源点所处的局部拓扑结构. 算法无需了解网络的全部拓扑结构, 能够有效降低大规模网络上的计算复杂度, 同时分离多源点的影响, 提高网络的定位准确率. 我们在模型网络及实际网络上进行了大规模随机传播实验, 与前期方法相比, 本方法具有较高的性能及准确率.

本方法的创新之处在于, 提出基于网络局部子图结构进行源点定位的方法, 无需获取网络全部拓扑结构, 从而极大地降低了定位的计算复杂度; 同时, 社团结构分离了多源点传播之间的影响, 使算法可以准确估计并定位多个传播源点, 而不受到多源点传播的互相干扰.

2 相关工作

社交网络上的信息传播 (information spreading), 又称为信息扩散, 是社会学、传播学、心理学等领域研究的经典问题. 社会网络上存在着复杂的动态过程, 包括一系列的社会传播现象, 例如谣言、时尚、政治意见以及新技术的采纳等. 因此, 早期学者主要研究创新行为及传染病在真实社会网络中的传播, 并提出了很多具有表现力的模型, 如独立级联模型^[15]、线性阈值模型^[5,6]、传染病扩散模型^[16]等. 受限于真实社会网络数据获取的困难, 这些模型的建立和分析主要采用定性方法或仅仅利用少量的社会调查数据^[17], 其有效性和说服力仍存疑, 且遇到“霍桑效应” (Hawthorne effect)^[18]的干扰. 而在线社会网络拥有丰富的数据, 而且参与者意识不到自己被观察. 因此, 研究者一般认为基于

大量社交网络数据的实证分析是当前的必由之路^[2].

分析在线社交网络信息传播过程的一个重要问题是寻找网络中信息爆发的起始节点. 有效地找出信息传播的起始点可还原信息的传播过程, 控制信息传播造成的影响. 但无论是数学还是实际角度, 这都是一个困难的任务. 最早尝试解决这个问题的是 Shah 等^[11], 他们针对树形拓扑的计算机病毒的传播, 提出了一种面向 SIR 模型的定位方法, 通过最大似然估计推断源点的位置; 随后, 他们将该方法应用到社交网络的谣言定位, 提出谣言中心性度量^[19]. Prakash 等^[20] 针对易感染 (SI) 病毒传播模型也提出一种基于最小描述长度方法, 能够自动地确定传播源点的数目, 并识别出网络中的多个传播源点. Comin 等^[10] 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, 这种方法对于雪球传播模型非常有效, 因为源点多数趋向于具有最高的中心度. Dong 等^[21] 给出一种从可疑节点中推断出谣言源头的方法. Shen 等^[22] 将压缩感知理论 (compressed sensing) 与复杂网络理论结合, 提出了针对二元时间序列的网络重构方法, 并用于发现网络中的潜在源点. Brockmann 等^[23] 提出了传染病网络中有效距离的概念, 将复杂的传播拓扑还原为环状扩散, 而扩散圆心即为潜在的传播源点.

以上方法的一个前提是已知网络的传播快照, 即网络中所有被感染或收到消息的节点构成的子图. 对于小规模网络来说, 如以地理位置为节点的传染病网络, 上述方法可以容易地提取传播子图并定位源点. 但是, 对于社会网络等大规模网络, 监控并获取网络中所有节点的状态代价太高, 几乎是不可能实现的. 因此, 如何通过有限的节点数据, 估计网络中的传播源点, 就具有重要的意义.

Pinto 等^[12] 提出了一种基于部分观察点的源点定位方法, 根据预先选择的部分观察点收到消息的时间和方向, 采用最大似然估计方法估计信息源点. 这种方法能够有效地降低传播数据的需求, 因此具有很大的应用潜力. Luo 等^[24] 研究基于 SI 传播模型上部分节点的传播过程, 发现对于树形网络源节点与产生感染节点观察子集的最优传染路径相关. 上述基于部分观察点的方法避免了大规模网络数据搜集的困难, 但是其定位准确度依赖于传播模型及观察点的收集到的数据质量. 所以评估网络中节点在定位中的作用, 寻找重要的观察点是上述方法的关键点之一.

对于观察点对定位准确率的影响, 我们在前期工作^[25] 中分析了 6 种常见的部署策略对定位准确率的影响, 包括度、介数、聚集系数、特征向量、k-核及随机策略等, 发现这几种策略对定位准确率的影响基本相同. 进一步地, 我们发现观察点的覆盖率是影响定位准确率的关键因素, 并提出了一种基于集合覆盖的观察点优化部署策略^[26], 能够有效地提高定位准确度.

社团结构是在线社交网络的典型拓扑特征, 已有很多方法^[13] 能够找到在线社交网络的社团结构. 显然, 社团结构对于传播具有明显的影响. 为此, Weng 等^[14] 分析了消息传播范围与社团结构的关系, 发现消息类型与其在社团中的传播模式具有密切关联. Hu 等^[27] 提出了一种社团级别扩散方法, 用以提取并分析话题的传播结构. 但是, 对于源点定位问题, 目前利用社团结构进行分析的方法还较少见.

总的说来, 目前的传播源点定位算法依赖于很多假设, 如特定的传播模型、特殊的拓扑结构、特殊的传播过程, 且大多用于分析传染病或计算机病毒的传播. 而在线社交网络上的热点信息传播具有突发、多源、动态等特点, 如何针对社交网络的特点, 有效地感知监测热点信息的传播过程, 解决大规模网络上的计算困难, 寻找定位社会网络上的有效的源点定位算法仍然是一个尚未解决的问题.

3 面向多传播源的传播特征分析及定位算法

3.1 问题定义

考虑在线社交网络 $G(V, E)$ 上的传播过程, 其中 $V = \{v_1, v_2, \dots, v_N\}$ 为节点集, $E = \{e_1, e_2, \dots, e_L\}$

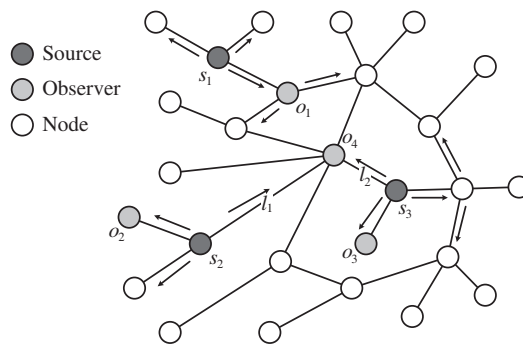


图 1 多源信息传播过程示意图

Figure 1 Diagram of multi-source information spread process

为边集. 在某一未知时刻 t^* , 网络中源点集合 $S^* = \{s_1, s_2, \dots, s_k\}$ 发出消息 m , 网络中的其他节点收到消息 m 并向其他邻居节点转发. 每条边 e_{ij} 上均有信息延迟 θ_{ij} , 代表消息 m 从节点 i 传播到节点 j 所花费的时间. 每个节点 $u \in G$ 有两种可能状态: 激活状态, 即已经接收到消息 m ; 未激活状态, 即到当前时刻 t 为止还未接收到消息 m . 若节点 u 是第一次接收到消息 m , 那么 u 从未激活状态转变为激活状态. 如果节点 u 首先收到源点 i 发出的消息, 则称 u 受到 i 的影响. 为了定位未知源点集合 S^* , 需要在网络中部署一些特殊节点, 称之为观察点. 观察点 o 在首次收到消息 m 时, 记录了收到消息的时间和方向 $(o, v, t_{v,o})$, 其中 $t_{v,o}$ 表示观察点 o 从其邻居节点 v 接收到消息 m 的时间. 所有观察点记录的数据称为观察向量 $O_k = \{(o_i, v, t_{v,o_i}), i = 1, 2, \dots, k\}$.

根据上述模型, 对于社交网络 $G(V, E)$ 上的传播源点定位问题, 可以表示为: 给定观察点集合 $\{o_1, o_2, \dots, o_k\}$, 基于观察向量 $O_k = \{(o_i, v, t_{v,o_i}), i = 1, 2, \dots, k\}$, 找到发出消息 m 的信息源点集合 S^* . 然而, 对于大规模网络, 源点与观察点之间存在众多可能的传播路径, 而我们仅知道观察点收到消息的时间及方向, 因此定位信息源点并不是一个简单的问题. Pinto 等^[12] 首先提出了一种面向单传播源的解决方法, 其主要思路是比较观察点实际收到消息的时间 t_{v,o_i} 与传播模型估计的接收时间之间的差值, 找到最有可能的节点作为信息传播源点. 然而, 这一方法的前提是观察点收到的消息都来自于同一个源点. 当网络中存在多个源点时, 不同的观察点可能首先接收到不同源点发出的消息, 因此利用所有的观察点进行信息源定位不再有效. 图 1 给出了一个典型示例, 其中 s_1, s_2, s_3 是三个信息源点. 观察点 o_1 首先收到源点 s_1 的消息, 观察点 o_2 首先收到源点 s_2 的消息, 观察点 o_3 和 o_4 首先收到源点 s_3 的消息. 由于每个观察点受到不同源点的影响, 当利用观察向量进行源点定位时, 不能准确的找到传播源点, 从而导致定位的失败.

3.2 传播特征分析

为了设计有效的定位算法, 我们首先分析典型社交网络上的传播过程, 希望找出网络上的多源传播规律. 选用了 Political blogs 网络^[28], 表示 2004 年美国总统选举期间的政治博客之间的互联情况, 其中节点代表博客, 边代表博客间的链接关系. 从图 2 中可以看出, 网络具有明显的社团结构, 分裂为两个主要社团, 分别代表奉行自由主义的民主党人和奉行保守主义的共和党人.

为了分析网络中多源传播的特征, 本文选用了社交网络中常用的 3 种传播模型, 分别是随机传播模型^[12]、线性阈值模型^[16]及独立级联模型^[15], 其区别在于消息在边上的传播. 在随机传播模型中, 消息在每条边上有传播延迟 θ_{uv} , θ_{uv} 是随机变量, 代表消息从节点 u 到节点 v 的传输时间. 在独立级

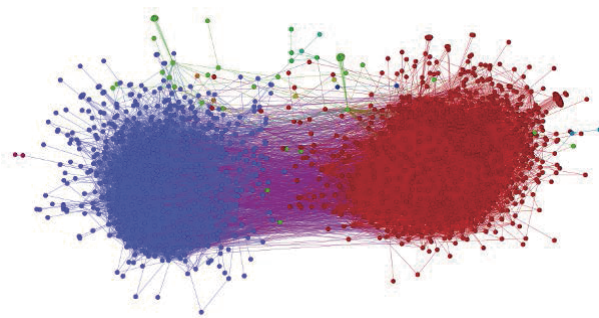


图 2 (网络版彩图) Political blogs 网络拓扑示意图
Figure 2 (Color online) Diagram of political blogs network topology

联模型中, 每条边上都有权值 P_{uv} , 代表了节点间相互激活成功的概率, 当一个节点 u 被激活时, 它会以概率 P_{uv} 对它未激活的邻居节点 v 尝试激活. 在线性阈值模型中, 每个节点都有激活阈值 $\theta_u \in [0,1]$, 当节点的已激活邻居节点对节点 u 的累积影响力大于 u 的激活阈值 θ_u 时, 将消息 m 发送给其邻居节点.

已有工作表明^[29] 度、介数、 k -核能够有效衡量网络中信息源点的传播效率, 以高度、高介数、高 k -核的节点为信息源, 能够得到较大的传播影响范围. 因此, 为了分析网络社团结构与传播范围之间的关系, 我们分别按照上述中心性度量指标, 在每个社团中选取 1 个节点作为传播源点, 根据随机传播、线性阈值及独立级联 3 种模型, 进行了 1000 次随机传播实验.

我们分析了每个源点在其所属社团内部及社团外部的影响范围, 具体结果如图 3 所示. 一个明显的现象是, 源点的影响范围起初均局限在所属社团内部, 当社团内部的大部分节点收到消息后, 才扩散到网络的其他部分. 在整个传播过程中, 源点在其所属社团内的影响范围总是大于其所属社团外的传播范围. 对于随机传播、线性阈值及独立级联 3 种模型, 无论源点选择高度、高介数或是高 k -核节点, 仅源点影响范围扩散到外部社团的时间略有差异, 而源点在社团内外的影响范围非常类似. 因此, 对于具有明显社团结构的网络, 在传播过程的早期, 源点的影响范围主要局限在其所属的社团内部, 社团结构是分离多个源点影响范围的有效途径.

为了分析不具有明显社团结构的网络中的传播范围特征, 我们使用具有社团结构的模型网络^[13]. 为了衡量源点在社团内外的传播范围, 定义以下 Range 度量, 表示传播源点在社团内的传播范围占其所有传播范围的比例, 即

$$\text{Range} = \sum_{i=1}^c \frac{1}{c} \times \frac{n_c}{N_c}, \quad (1)$$

其中 c 为传播源点的个数, N_c 表示传播源点 c 影响的节点个数, n_c 表示传播源点 c 在其所在的社团内影响的节点个数. 我们生成了一系列具有不同模块度^[13] 的网络, 在每个网络中进行了 1000 次随机传播实验, 结果如图 4 所示. 可以看出, 对于 3 种传播源点选取方法, Range 值均随着网络模块度的增大而增加, 这说明网络的社团结构越明显, 消息就会更多的在源点所在社团的内部进行传播. 因此, 社团特性对于多源传播范围的分离是一个普遍的性质.

3.3 基于子图抽取的多传播源定位算法

利用部署观察点对源点进行定位的方法, 对于单个信息源的定位具有良好的定位效果, 但是当网络中有多个源点同时开始进行信息传播时, 由于观察点所能够记录的信息有限, 很难区分其所收到的

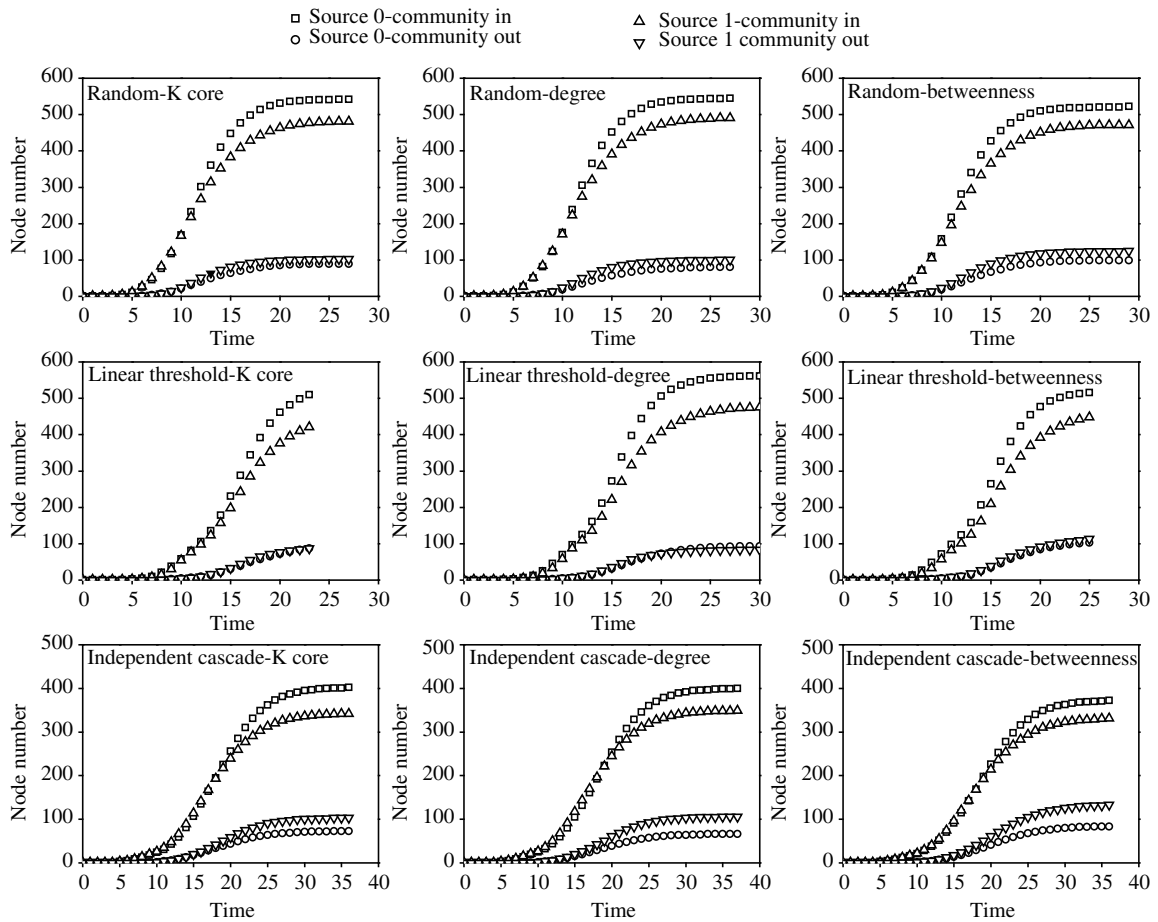


图 3 Political blogs 网络传播范围分析

Figure 3 Analysis of spreading range on political blogs network

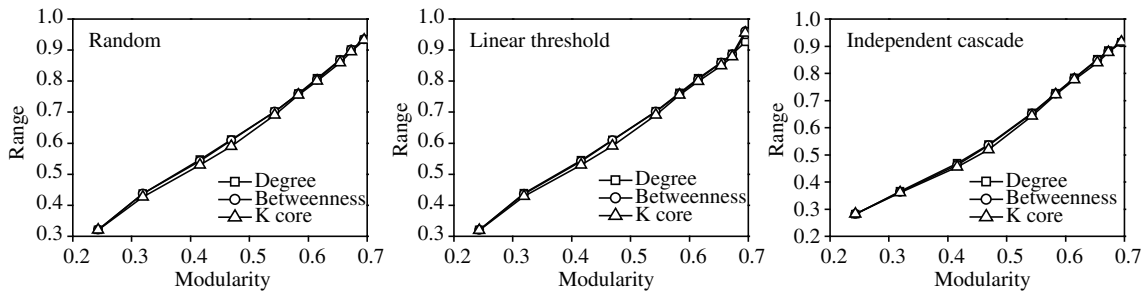


图 4 网络社团特征对源点在其所属社团内外传播范围的影响

Figure 4 Effect of the network community features on the spreading range of the source

信息是由哪个源点发出的, 无法对多个信息源进行有效地辨别和定位. 如果能够有一种方法, 可以将不同的信息源的影响范围区分开, 那么在每个信息源的传播范围内, 都只存在一个信息源, 这样就可以将多传播源定位问题, 转化为在多个子网内分别对单信息源进行定位, 这样就可以充分发挥部署观察点进行定位的方法的优势, 大大提高定位效果.

根据 3.2 小节中的分析, 社团结构可以有效地分离多个源点的传播范围. 因此, 我们提出一种基于子图抽取的多传播源定位算法 (locating multi diffusion source based on community structure, LSCS). 算法的基本思路是在网络中部署若干观察点, 用来监测网络中传播的热点信息; 以较早接收到信息的观察点为核心点, 抽取网络局部子图; 在每个局部子图中根据观察点收集到的数据, 独立进行源点定位. 该方法的核心出发点是根据抽取的局部子图对观察点集合进行划分, 使得同一子图中的观察点首先收到同一个源点的消息, 从而将全网中的多源定位问题转换为在多个子图中的单源定位问题, 以此来解决多传播源点定位的困难.

对于子图抽取问题, 首先需要选择抽取的初始节点. 一般来说, 收到消息较早的观察点往往位于源点的附近. 因此, 将观察点按照其收到消息的时间进行排序, 从较早收到消息的观察点出发, 抽取其所在的局部子图, 能够找到包括源点、观察点及其之间传播路径的子网结构, 从而能够有效地定位源点.

其次需要设计子图抽取的规则. 根据 3.2 小节的分析, 在具有明显社团结构的网络中, 消息会在社团内部传播, 然后扩散到其他区域. 因此, 通过对网络的社团划分来进行子图抽取是一个自然地思路. 然而, 网络的社团发现是一个非常耗时的过程, 复杂度往往在 $O(n^2)$ 以上. 而且, 对于信息源定位问题, 网络的完整社团结构并不是必要的, 我们仅需了解与传播范围相关的局部拓扑结构即可. 因此, 我们设计一个局部子图抽取方法, 从观察点出发, 按照社团结构进行子图扩展, 从而得到具有社团结构的局部子图. 为了评价子图的社团特征, 我们定义以下子图适应度函数:

$$Q(C) = |e_C| / (|e_C| + |e'_C|), \quad (2)$$

其中, $|e_C|$ 为子图 C 内部边的个数; $|e'_C|$ 为子图 C 指向外部节点的边的个数. 该函数通过子图内外连边的比例度量其“内聚”程度. 显然, 如果 C 的社团性质较明显, Q 值较大, 反之较小. 然后, 我们需要分析如何评价节点是否属于待抽取的子图? 对于一个节点来说, 当其连边大多在子图内部时, 则更有可能属于该子图; 而当节点大部分连边都指向子图外部节点时, 则不大可能属于该子图. 因此, 定义节点 n 的评价函数 f 为

$$f(n) = k_n^C / k_n^G, \quad (3)$$

其中 k_n^C 为节点 n 在子图 C 中的度; k_n^G 为节点 n 在整个网络 G 中的度.

综上, 子图抽取的过程为: 从较早收到消息的观察点 $\{o_1, o_2, \dots, o_k\}$ 出发, 沿着边扩展子图, 在扩展的过程中同时进行节点的筛选, 以保证所得子图的社团结构. 其中包括两个主要过程. 一个是寻找最有可能属于子图的节点, 另一个是判定加入该节点是否能增加子图适应度函数. 重复上述步骤直到子图达到指定规模 N 或子图适应度停止增长为止. 算法的示意图如图 5 所示.

在抽取的过程中, 从不同观察点出发抽取的子图有可能重叠. 这可能是由于观察点同处于一个源点的影响范围内造成的. 对于这种情况, 可以将其合并然后继续进行扩展. 判断重叠子图是否合并的原则在于合并后的子图的适应度 Q 是否增加. 由于扩展过程中存在子图合并, 所以子图抽取的初始节点数 k 应大于网络中的源点数, 从而保证能够抽取到所有的信息传播源点. 另外, 由于子图抽取的目标是准确的定位源点, 因此没有必要扩展到完整的社团. 当子图扩展到预先制定的期望大小 M 时即可停止. 子图大小 M 取决于观察点的部署密度, 如果观察点较多, M 可相对较小, 当观察点较稀疏时, M 可设置较大, 从而保证子图中包括尽可能多的观察点, 从而提高定位准确率.

当子图抽取完成后, 针对每个子图, 可以采用 Pinto 等^[12] 提出的最大似然估计器进行源点定位. 具体如下: 令 $\{t_k\}_{k=1}^K$ 表示观察点首次收到信息的实际时间集合, 用 $d = \{d_1, d_2, \dots, d_{K-1}\}^T$ 表示传播

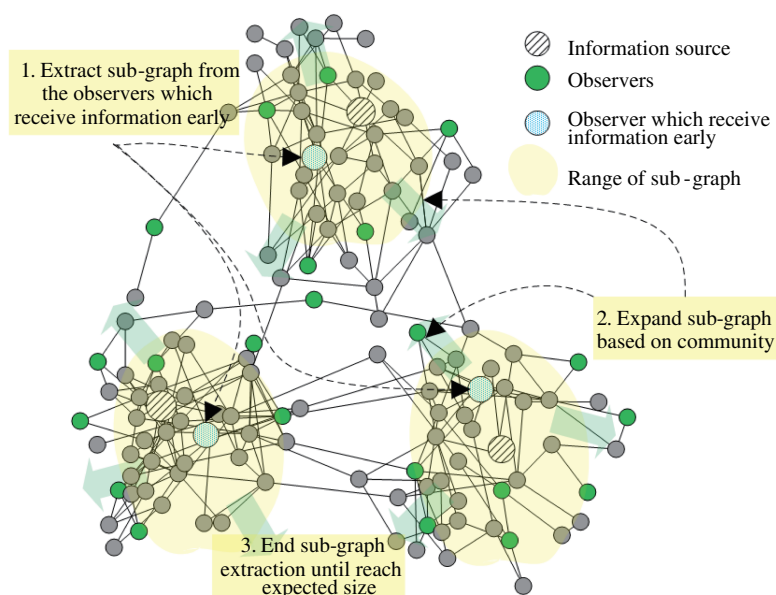


图 5 (网络版彩图) 面向多传播源点定位的子图抽取的过程示意图
Figure 5 (Color online) Diagram of extracting the sub-graph for multi-source

延迟向量, 其中, $d_{i=}=t_{i+1}-t_1$ 表示观测点 o_{i+1} 与观察点 o_1 首次收到信息的实际时间差, o_1 为第 1 个收到信息的观察点. 令 μ, σ^2 分别表示传播延迟的均值和方差. 令 $p(u, v)$ 表示 u 到 v 之间的最短路径, $|p(u, v)|$ 表示最短路径的长度, 则传播源点为使式 (4) 取得最大值的节点:

$$\hat{s} = \operatorname{argmax}_{s \in T_a} \mu_s^T \Lambda^{-1} \left(d - \frac{1}{2} \mu_s \right), \quad (4)$$

其中,

$$[\mu_s]_k = \mu \cdot (P(s, o_{k+1}) - P(s, o_1)), \quad (5)$$

$$[\Lambda]_{k,i} = \sigma^2 \cdot \begin{cases} |P(o_1, o_{k+1})|, & k = i, \\ |P(o_1, o_{k+1}) \cap P(o_1, o_{i+1})|, & k \neq i. \end{cases} \quad (6)$$

综上, 在线社交网络上的多源定位算法的伪代码如算法 1 所示.

注意到, 算法 1 的一个假设是网络每个局部子图中只有一个传播源点. 这是由于在一个局部子图中部署多个源点一般是低效且无用的, 很多真实传播场景都符合这种假设. 例如, 在在线广告中, 在同一位置附近部署多个源点, 会造成资源的浪费; 对于谣言发布者, 同时发布多个距离相近的源点, 既无助于扩大传播范围, 也会同时增大被监管机构发现的概率. 因此, 在同一子图中至多存在一个源点, 是一个合理的假设.

上述方法的一个核心思路是通过局部子图抽取估计源点范围、定位源点位置, 这可以有效地降低网络的源点定位复杂度. 与以往工作 [10~12] 不同, 方法不需要已知网络全部拓扑, 而是基于源点附近的局部子图结构进行定位. 令网络中源点个数为 c , 子图规模为 M , 则抽取子图的复杂度为 $O(c \cdot M)$; 抽取子图后, 在每个子图中单独定位源点的复杂度为 $O(M^3)$. 因此, 算法的总体复杂度为 $O(c \cdot M^3)$. 由于子图规模 M 一般远小于网络规模 N , 因此, 与论文 [12] 中的算法复杂度 $O(N^3)$ 相比, 基于局部子图结构的定位方法, 排除子图外的多数节点, 能够极大地提高效率.

Algorithm 1: Locating multi diffusion source based on community structure (LSCS)

```

1. Input: Observer vectors  $O = \{(o_i, v, t_{v,oi})\}$ ; number of initial nodes  $k$ ; expected sub-graph size  $M$ ;
2. Sort the observer vectors  $O = \{(o_i, v, t_{v,oi})\}$  based on receive time  $t_{v,oi}$ ;
3. Let the top- $k$  observers be  $O_k = \{o_1, \dots, o_k\}$ ; initialize sub-graph  $C_i = \{o_i\}$ ,  $i = 1, 2, \dots, k$ ;
4. repeat
    // Expand each sub-graph until they are large enough or did not increase
5.   for each  $C_i$  do
6.     if  $\text{size}(C_i) < M$ 
7.       //expand sub-graph  $C_i$ 
8.       Let  $N_{ci} \leftarrow \text{neighbor}(C_i)$ ;  $\text{increase}.C_i = \text{false}$ ;
9.       repeat
10.        let  $m = \text{argmax}_{m \in N_{ci}} h(m)$ ;
11.        if  $Q(C_i + \{m\}) > Q(C_i)$  then  $C_i \leftarrow C_i + \{m\}$ ;  $\text{increase} = \text{true}$ ;
12.         $N_{ci} \leftarrow N_{ci} - \{m\}$ ;
13.      Until  $\text{Size}(N_{ci}) = 0$ ;
14.     end
15.   end
16.   // merge sub-graphs if the result get high  $Q$  value
17.   if  $C_i \cap C_j \neq \emptyset$  and  $Q(C_i \cup C_j) > \max(Q(C_i), Q(C_j))$ 
18.     then  $C_i \leftarrow C_i \cup C_j$ ;  $C_j \leftarrow \emptyset$ ;
19.   until  $\text{size}(c_i) > M$  or  $\text{increase}.C_i = \text{false}$ ;
20.   // locate source node in each sub-graph
21.   for each  $C_i$  do
22.     // find node with max value based on equation 4
23.     Let  $\hat{s} = \text{argmax}_{s \in T_a} \mu_s^T \Lambda^{-1} (d - \frac{1}{2} \mu_s)$ ;
24.      $S^* \leftarrow S^* + \{\hat{s}\}$ ;
25.   end
26. Output: source node set  $\{S^*\}$ .

```

4 实验与分析

为了验证本文提出方法的有效性, 我们在具有明显社团结构的社交网络上进行了模拟传播实验. 采用文献 [13] 中的方法生成具有社团结构的网络. 具体而言, 首先生成网络的节点集 V , 并对节点集进行划分, 形成社团. 对于任意节点对 $i, j \in V$, 如果节点 i, j 在同一个社团内, 则以概率 p_{in} 在节点 i, j 间添加连边; 否则以概率 p_{out} 在节点 i, j 间添加连边. 重复该过程直到网络达到了指定的边数为止. 显然, 如果连边概率 $p_{\text{in}} > p_{\text{out}}$, 上述模型生成的网络社团内部连边个数会大于社团之间的连边个数, 从而形成了明显的社团结构. 根据以上方法, 我们生成了一系列的模型网络, 具体参数如表 1 所示.

为了验证算法的准确性, 我们在网络上进行了模拟传播及定位实验. 具体实验过程是, 对于网络 $G(V, E)$, 在网络中随机选择 k 个节点作为消息传播源点, 按照 3.1 小节所述传播模型进行模拟传播. 在网络中分别以高度优先 (high-degree)、高覆盖率优先^[25] (high-co)、随机选取 (random) 等 3 种观察点部署策略进行传播监测, 记录观察点的传播向量 $O_k = \{(o_i, v, t_{v,oi}), i = 1, 2, \dots, k\}$. 对于每种观察点部署策略和观察点部署比例, 分别进行 1000 次模拟传播实验并计算平均定位准确率. 令实际源点集合为 S_{real} , 预期源点集合为 S_{est} , 如果 $S_{\text{est}} = S_{\text{real}}$, 则认为定位命中. 如果 n 次实验中定位命中的次数为 m , 则定位准确率为 $P = m/n$.

为了准确地定位源点, 首先要准确地抽取网络中的传播社团. 因此, 我们首先考察算法是否能够准确地估计网络中的传播源点个数. 令网络规模 $N=20000$, 社团数 $C_{\text{num}}=20$, 每个社团的节点数均为

表 1 实验数据集
Table 1 Experimental data

Network name	N	L	Community	Modularity	p_{in}	p_{out}
Network1	20000	1733577	20	0.512	0.05	0.002
Network2	20000	1154479	20	0.269	0.02	0.002
Network3	20000	1636449	20	0.069	0.006	0.004
Network4	9947	25748	5	0.624	0.0012	0.00002
Network5	9958	20055	5	0.253	0.0008	0.0001
Political-blogs	1224	16718	2	0.431	—	—

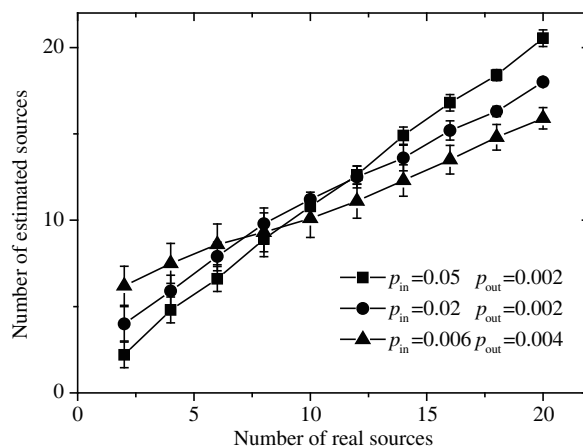


图 6 子网抽取个数与实际源点个数的之间的关系

Figure 6 Relationship between the sub-graph extraction number and real source number

$C_{size} = 1000$. 在该网络中随机生成 k 个源点, 随机部署到网络中, 进行传播并抽取子网. 观察点采用随机部署方式, 观察点占全部节点的比例为 0.4. 图 6 给出了实验结果. 对于社团结构明显的网络, 例如 $p_{in}=0.05$, $p_{out}=0.002$ 时, 算法基本能够准确地判定源点的个数. 当实际源点个数较少时, 估计数略大于实际源点数. 这主要是由于当源点位于社团边缘时, 信息会以较大概率传播至相邻社团, 从而造成源点数被高估. 对于社团结构不明显的网络, 如 $p_{in}=0.006$, $p_{out}=0.004$ 时, 当实际源点数较大时, 估计数会明显低于实际源点数, 这主要是由于社团结构不明显, 相邻社团中被合并抽取所造成. 因此, 子图抽取算法对于具有明显社团结构的网络, 可以有效并准确地估计传播源点的个数.

为了分析 LSCS 算法的定位准确度, 我们选取了论文 [12] 提出的定位算法 LS 算法作为对比. LS 算法的具体过程是, 根据部署在网络中的观察点所记录的收到消息的时间和方向, 采用最大似然估计方法, 计算网络中各候选源点 (网络中全部非观察点) 的似然估计值, 似然估计值最大的即为信息源. 在实验过程中, 选用 LS 算法得出的 k 个似然估计值最大的节点, 作为多源点定位结果. 模型网络及实际网络上的实验结果如图 7 和 8 所示. 可以看出, 对于具有较强社团特征的网络 Network4, 当网络中观察点比例上升时, LSCS 算法的准确率随之上升. 当观察点比例超过 30% 时, 网络的定位准确率均超过 80%, 这明显的优于论文 [12] 的定位算法. 对于社团特征较不明显的网络 Network5, 算法准确率有所下降, 但仍高于利用全部观察点进行定位的 LS 算法. 注意到, 随机部署的观察点的准确率略低于另外两种部署策略, 这主要是由于随机部署的观察点在搜集信息及抽取子图结构方面并不具备优

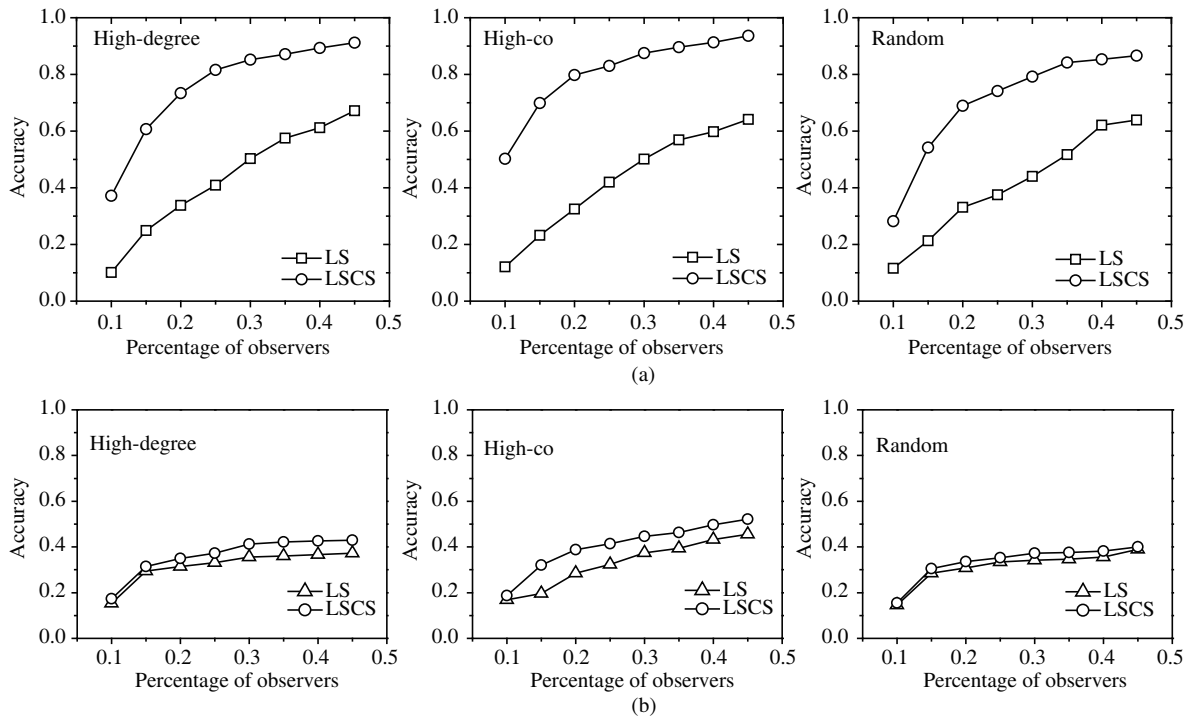


图 7 LSCS 算法在模型网络上的定位准确率

Figure 7 Localization accuracy of LSCS algorithm on model network. (a) Network size: $N=10000$, number of communities: $C_{\text{num}}=5$, group size: $C_{\text{size}}=2000$, modularity: 0.624; (b) network size: $N=10000$, number of communities: $C_{\text{num}}=5$, group size: $C_{\text{size}}=2000$, modularity: 0.253

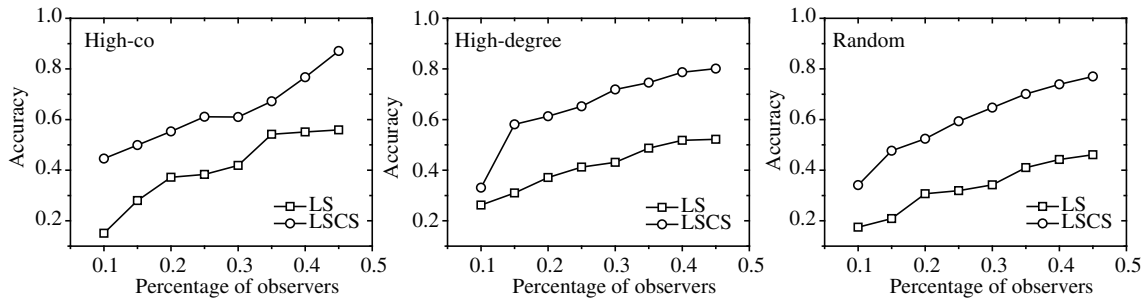


图 8 LSCS 算法在 Political-blogs 网络上的定位准确率

Figure 8 Localization accuracy of LSCS algorithm on political blogs network

势. 在 Political blogs 网络上也观察到了类似的结果. 因此, 可以得出结论, 在相同数量的观察点情况下, 本文提出的定位算法具有更高的定位准确率.

对于不能准确定位全部源点的情形, 我们分析了实际源点与估计源点之间的平均距离分布, 结果如图 9 所示. 可以看出, 即使算法在某些情况下无法准确定位真实的源点, 估计源点也与实际源点距离很近, 基本在距离 2 范围内, 有 20% 左右的情况是实际源点的直接邻居. 这说明算法可以有效地缩小候选源点的范围.

最后, 我们分析了 LSCS 算法的定位效率. 实验环境如下: 处理器为 Intel®Core™i7 3770k 3.2 Hz; 内存 32 GB RAM; 硬盘为 500 GB. 首先分析了社团大小对算法效率的影响. 首先随机生成若干网络,

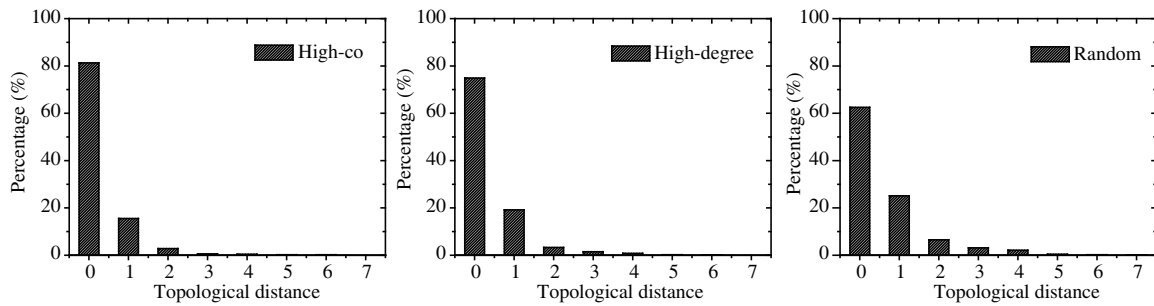


图 9 社团网络上的定位误差分布
Figure 9 Distribution of localization error on social network

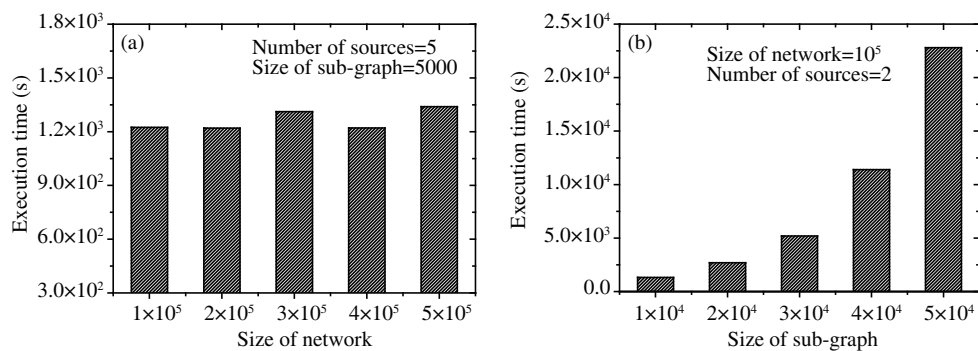


图 10 算法的效率分析

Figure 10 Efficiency analysis of algorithm. (a) Relationship of network size versus algorithm efficiency; (b) relationship of community size versus algorithm efficiency

具有不同大小的社团, 然后按照高度数优先部署策略选取观察点, 进行 1000 次随机传播实验并计算定位源点所用时间的均值, 实验结果图 10 所示. 可以看出, 算法所需时间随社团规模的增大而增大. 进一步地, 我们分析网络规模对于算法的影响. 固定社团大小为 $C_{\text{size}}=1000$, 源点个数 $c=2$, 增加网络中的社团个数. 实验结果如图 9 所示. 可以看出, 当社团规模固定时, 网络的规模基本不会影响算法的计算时间, 这是由于, 我们的算法本质上是基于局部拓扑的, 非传播区域不会对算法效率产生影响. 因此, LSCS 算法可以有效地用于大规模在线社交网络.

5 结论

快速准确地定位在线社交网络中的信息传播源点对于传播范围预测、传播过程分析、谣言控制的应用具有重要的应用价值. 然而, 在线社交网络的多源传播、规模巨大等特征使得以往用于传染病溯源的定位算法并不适用.

本文针对在线社交网络中信息传播的特点, 分析了社团结构在传播中的影响, 发现社团结构是分离多源点传播影响范围的有效途径. 在此基础上, 提出了一种基于子图抽取的多扩散源定位算法, 有效地将网络中的多扩散源定位问题转换为多社团内的单扩散源定位问题. 算法能够获得源点附近的局部社团结构, 而无需了解网络的全部拓扑结构. 该方法能够降低大规模网络上的计算复杂度, 同时分离多源点的影响, 提高网络的定位准确率. 我们在模型网络及实际网络上进行了大规模随机传播实验,

充分验证了本文所提出方法的准确性.

本文所提出的多扩散源定位算法, 主要是针对社团结构比较明显的网络, 对于社团结构不明显的网络, 当有多个消息源进行扩散时, 如何准确地定位扩散源点, 是下一阶段需要我们研究和解决的问题.

参考文献

- 1 Lee H W. Global Risks Report. World Economic Forum 301211. 2013
- 2 Li D, Xu Z M, Li S, et al. A survey on information diffusion in online social networks. *Chinese J Comput*, 2014, 37: 189–206 [李栋, 徐志明, 李生, 等. 在线社会网络中信息扩散. *计算机学报*, 2014, 37: 189–206]
- 3 Adrien G, Hakim H, Cecile F, et al. Information diffusion in online social networks: a survey. *ACM SIGMOD Record*, 2013, 42: 17–28
- 4 Ernesto E, Naomichi H, Michele B. The physics of communicability in complex networks. *Phys Reports*, 2012, 514: 89–119
- 5 Jurij L, Lada A A, Bernardo A H. The dynamics of viral marketing. In: *Proceedings of the 7th ACM Conference on Electronic Commerce*. New York: ACM, 2006. 228–237
- 6 Mark G. Threshold models of collective behavior. *American J Sociol*, 1978, 83: 1420–1443
- 7 Mislove A, Massimiliano M, Krishna P G, et al. Measurement and analysis of online social networks. In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, San Diego, 2007. 29–42
- 8 Kempe D, Jon K, Éva T. Maximizing the spread of influence through a social network. In: *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, 2003. 137–146
- 9 Easley D, Kleinberg J. *Networks, Crowds, and Markets*. Cambridge Univ Press, 2010, 6: 6.1
- 10 Comin C H, Costa L. Identifying the starting point of a spreading process in complex network. *Phys Rev E*, 2011, 84: 1903–1904
- 11 Shah D, Zaman T. Detecting sources of computer viruses in networks: theory and experiment. In: *Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, New York, 2010. 203–214
- 12 Pinto P C, Thiran P, Vetterli M. Locating the source of diffusion in large-scale networks. *Phys Rev Lett*, 2012, 109: 1–5
- 13 Fortunato S. Community detection in graphs. *Phys Rep*, 2010, 486: 75–174
- 14 Weng L, Menczer F, Ahn Y Y. Virality prediction and community structure in social networks. *Sci Rep*, 2013, 3: 618
- 15 Goldenberg J, Libai B, Muller E. Talk of the network: a complex systems look at the underlying process of word-of-mouth. *Marketing Lett*, 2001, 12: 211–223
- 16 May R M, Lloyd A L. Infection dynamics on scale-free networks. *Phys Rev E*, 2001, 64: 116–126
- 17 Xu K, Zhang S, Chen H, et al. Measurement and analysis of online social networks. *Chinese J Comput*, 2014, 37: 165–188 [徐格, 张赛, 陈昊, 等. 在线社会网络的测量与分析. *计算机学报*, 2014, 37: 165–188]
- 18 Rob M, James W, Steve L, et al. The Hawthorne effect: a randomised, controlled trial. *BMC Med Res Methodol*, 2007, 7: 1–8
- 19 Shah D, Zaman T. Rumors in a network: who's the culprit? *IEEE Trans Inf Theory*, 2011, 57: 5163–5181
- 20 Prakash B A, Vreeken J, Faloutsos C. Spotting culprits in epidemics: how many and which ones? In: *Proceedings of the IEEE 12th International Conference on Data Mining*, Belgium, 2012. 11–20
- 21 Dong W X, Zhang W Y, Tan C W. Rooting out the rumor culprit from suspects. In: *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Istanbul, 2013. 2671–2675
- 22 Shen Z, Wang W X, Fan Y, et al. Reconstructing propagation networks with natural diversity and identifying hidden sources. *Nature Commun*, 2014, 5: 4323
- 23 Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena. *Science*, 2013, 342: 1337–1342
- 24 Luo W Q, Tay Q P, Leng M. How to identify an infection source with limited observations. *IEEE J Sel Topics Signal Process*, 2013, 8: 586–597

- 25 Zhang X Z, Zhang Y B, Lv T Y, et al. Identification of efficient observers for locating spreading source in complex networks. *Phys A: Stat Mech Appl*, 2016, 442: 100–109
- 26 Zhang Y B, Zhang X Z, Zhang B. Observer deployment method for locating the information source in social network. *J Softw*, 2014, 25: 2837–2851 [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. *软件学报*, 2014, 25: 2837–2851]
- 27 Hu Z, Yao J, Cui B, et al. Community level diffusion extraction. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Melbourne, 2015. 1555–1569
- 28 Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks. *J Stat Mech Theory Exp*, 2008, 30: 155–168
- 29 Maksim K, Lazaros K G, Shlomo H, et al. Identification of influential spreaders in complex networks. *Nature Phys*, 2010, 6: 888–893

A multi-diffusion source-localization method for online social networks based on sub-graph extraction

Xizhe ZHANG^{1*}, Yubo ZHANG¹, Tianyang LÜ^{2,3}, Shihai FU¹ & Bin ZHANG¹

¹ *School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China;*

² *College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China;*

³ *IT Center, National Audit Office, Beijing 100830, China*

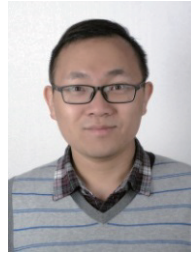
*E-mail: zhangxizhe@mail.neu.edu.cn

Abstract Locating the spreading sources of online social networks is of great importance in many applications, such as rumor detection and infectious-disease control. Existing work on source localization is not suitable here because of the huge size of online social networks and the multi-spreading nature of information diffusion. This paper focuses on the multi-source localization problem of social networks, and analyses the spreading principles of a network with a community structure. We find that the community is capable of separating the spreading influence area of multi-sources. Therefore, we present a multi-source localization algorithm based on the sub-graph extraction of complex networks. The algorithm extracts the local sub-graph based on the community structure, and ensures that the observers receive the same source in the same sub-graph. This converts the multi-source localization problem of the whole network into the single-source localization problem in a sub-graph. The algorithm does not require the whole topological structure of the network, and it can significantly decrease the computational complexity. Experimental results on real and model networks show the effectiveness and validity of our multi-source localization algorithm. The proposed localization method is of great significance for locating rumors and controlling infectious diseases.

Keywords social network, information diffusion, community structure, multi-source localization, sub-graph extraction



Xizhe ZHANG was born in 1978. He received a Ph.D. degree in computer application from Jilin University in 2006. Currently, he is an associate professor at Northeastern University. His research interest includes complex network analysis.



Yubo ZHANG was born in 1984. He is a Ph.D. candidate in computer application at Northeastern University. His research interests include complex networks, social networks and information spreading.



Tianyang LÜ was born in 1979. He received a Ph.D. degree in computer application from Jilin University in 2007. Currently, he is a senior engineer at the National Audit Office. His research interest includes complex network analysis.



Bin ZHANG was born in 1964. He received a Ph.D. degree in computer application from Northeastern University in 1997. Currently, he is a professor at Northeastern University. His research interests include web information processing, service computing and data mining.