doi: 10.3969/j.issn.1005-3026.2017.02.014

# 一种动态社交网络上的传播源点定位方法

张锡哲, 孟庆虎, 张 斌
(东北大学 计算机科学与工程学院, 辽宁 沈阳 110819)

**摘 要**: 在线社交网络的拓扑会随时间而发生改变,使得确定潜在的传播源点非常困难. 为此,提出一种考虑网络动态变化的传播源点定位方法,通过对网络演化建模,推断传播拓扑,从而准确定位信息源点. 首先采用基于双曲几何学的链接分析方法,推断网络在传播过程中的拓扑变化,然后基于传播拓扑进行源点定位. 在实际网络及合成网络上进行了大规模的实验,结果证明了算法的可行性.

**关 键 词**: 社交网络;动态演化;信息传播;源点定位
**中图分类号**: TP 399 **文献标志码**: A **文章编号**: 1005-3026(2017)02-0219-05

# A Source Localization Method for Information Diffusion on Dynamic Social Networks

*ZHANG Xi-zhe*, *MENG Qing-hu*, *ZHANG Bin*
(School of Computer Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: ZHANG Xi-zhe, E-mail: zhangxizhe@mail.neu.edu.cn)

**Abstract**: The topology of social networks changes over time, which makes it very difficult to determine unknown spreading source. A localization method about diffusion source on dynamic networks is presented. The hidden source can be localized by means of modeling the network and deducing the spreading topology. First, the topological change of networks in the diffusion process is deduced based on the link analysis method given by hyperbolic geometry, and then the source based on the propagation topology is located. The large-scale experiments based on the actual networks and the synthetic networks show the feasibility of the proposed algorithm.

**Key words**: social network; dynamic evolution; information diffusion; source localization

随着以 Facebook,Twitter,微博和微信为代表的在线网络的飞速发展,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,如何分析、理解及利用社交网络上信息传播,已成为当前计算机科学、社会学、物理学等多个学科的热点研究方向[1-2].

监控社会网络上的信息传播过程,找出信息传播的源点,对于许多实际问题具有重要意义. 例如,寻找在线社交网络中的"谣言发布者"[3],设计在线广告的应对营销策略[4]等. 最早尝试解决这个问题的是 Shah 等[5],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;Shen 等[6]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[7]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. Pinto 等[8]提出了一种基于部分感知节点的源点定位方法,根据预先选择的部分感知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. 上述基于部分感知节点的方法避免了大规模网络数据的搜集,但是并没有考虑传播过程中网络拓扑动态改变的问题,因此并不

**收稿日期**: 2015-09-28
**基金项目**: 中央高校基本科研业务费专项资金资助项目(N140404011); 国家自然科学基金资助项目(71272216,91546110,61572117).
**作者简介**: 张锡哲(1978-),男,辽宁沈阳人, 东北大学副教授,博士; 张 斌(1964-),男,辽宁本溪人,东北大学教授,博士生导师.

能适用于快速变化的动态社交网络.

本文提出一种面向动态在线社交网络的信息传播源点定位方法,其基本思路是,在网络中预先部署少量观察点,负责监控并搜集网络中信息的传播状态. 在定位时刻网络拓扑结构的基础上,基于偏好性及相似性这两个网络演化的关键因素[9]推断传播时网络的拓扑结构;根据观察点搜集到的传播过程,基于传播时刻的网络拓扑估计潜在的信息源点. 上述过程的难点在于如何基于当前的拓扑结构推断以前传播时刻网络的拓扑结构,这本质上是链接预测的逆问题. 对此,基于网络演化模型,找出新近改变的边,从而还原网络的拓扑变化,进而准确地定位传播源点.

## 1 动态网络上的源点定位方法

将社交网络记为有向网络$G^t(N,E)$,表示网络在时刻 $t$ 的拓扑结构,其中 $N=S\cup O\cup V$ 是网络节点的集合,$S,O,V$ 分别为源节点、感知节点和其他节点的集合;$E=\{e_1,e_2,\cdots,e_L\}$ 是网络中 $L$ 条边的集合. 令未知源点 $s^*$ 在时刻 $t^*$ 向其所有的邻居节点发出消息 $m$,其对应的网络拓扑为 $G^{t^*}(N,E)$. 每个节点 $v_i\in V$ 有两种可能状态:激活状态,即已经接收到信息 $m$;未激活状态,即到当前时刻 $t$ 为止还未接收到信息 $m$. 若节点 $v_i$ 是第一次接收到消息 $m$,那么 $v_i$ 从未激活状态转变为激活状态. 消息在每条边上的传播延迟记为 $t_{v_iv_j}$,$t_{v_iv_j}$是随机变量,代表消息从节点 $v_i$ 到节点 $v_j$ 的延迟时间. 网络中所有边的传播延迟集合$\{t_{v_iv_j}\}$满足一个已知的任意联合分布,其均值为 $\mu$,方差为 $\sigma^2$. 对于任意节点 $v_i\in V$,$t_{v_i}$表示 $v_i$ 首次收到信息 $m$ 的时间. 信息的传播过程见图 1 和图 2.

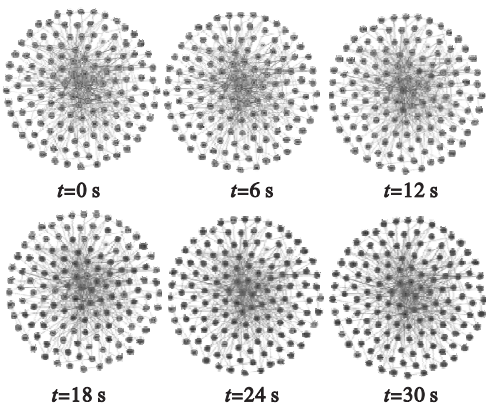


图 1 网络中一个信息的传播过程

Fig. 1 A process of information diffusion in the network

网络中存在一类特殊节点,称为感知节点. 感知节点记录了它收到消息的时间和方向. 即若 $o_k\in O$ 是感知节点,在 $o_k$ 首次收到信息 $m$ 时,记录传播信息$(o_k,v_i,t_{v_io_k})$,其中 $t_{v_io_k}$表示感知节点 $o_k$ 从其邻居节点 $v_i$ 接收到信息的时间. 令进行源点定位的时间为 $t^1$,因为一般都是在消息传播到一定范围之后才进行定位,所以定位时间 $t^1$ 大于传播时间 $t^*$. 在这个过程中,网络拓扑由 $G^{t^*}(N,E)$改变为 $G^{t^1}(N,E)$. 由于只能获取定位时的网络拓扑 $G^{t^1}(N,E)$,如何推断出传播时的网络拓扑 $G^{t^*}(N,E)$就成为关键问题.

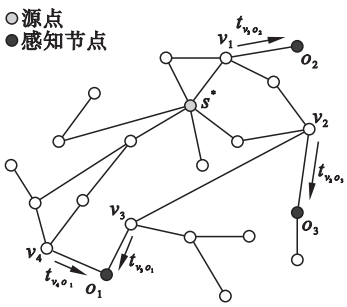


图 2 信息感知与传播过程示意图

Fig. 2 Schematic diagram of information perception and diffusion process

在线社交网络的一个显著特点是其结构不是稳定不变的,节点之间的连边会随着时间的推移不断改变. 例如,一个新用户加入社交网络后,会不断添加好友扩展朋友圈. 对于这个动态变化的过程,可以用网络的演化模型来描述. 目前经典的网络演化模型是 BA 无标度网络模型[9],其基本规则是偏好依附及增长,描述了网络在增长过程中节点的择优链接现象,其根源在于富者更富现象. 另外相似性也是网络一个重要特征,例如节点复制模型[10]等. 近期,Papadopoulos 等[11]结合了流行性及相似性这两个网络演化模型中的重要因素,提出了一个基于双曲几何学的复杂网络演化模型. 模型中将节点的流行性与相似性建模为双曲坐标系下的角坐标与半径坐标,能够生成与实际网络更符合的模型网络. 这个模型可以简单描述为:① 设置初始网络为空;② 在 $t(t\geqslant 1)$时刻,圆环的随机角位置 $\theta_t$ 上会出现新的节点 $v_t$;③ 为节点 $v_t$ 添加 $m$ 个链接,指向 $m$ 个网络现有节点,这些节点由任意节点 $v_s$ 的出现时间 $s(s<t)$与 $\theta_{st}$ 的乘积最小的 $m$ 个节点组成,其中 $m=\bar{k}/2$,$\bar{k}$ 为网络平均度,$\theta_{st}$是 $v_s$ 和 $v_t$ 之间的角距离.

为了推断源点传播消息时的网络拓扑,采用文献[11]的方法对网络动态演化进行建模,推断出传播过程中变化的拓扑连边,从而准确定位源

点. 对于网络 $G_1(N,E_1)$，令网络的邻接矩阵为 $\boldsymbol{A}=(a_{ij})$，如果节点 $v_i$ 与 $v_j$ 节点之间存在链接，则 $a_{ij}=a_{ji}=1$；否则 $a_{ij}=a_{ji}=0$，需要推断网络中在最终时间 $t$ 内所有节点 $v_i(i=1,2,\cdots,t)$ 的半径坐标 $r_i$ 和角坐标 $\theta_i$，即 $\{(r_i(t),\theta_i)\}=\{(r_1(t),\theta_1),(r_2(t),\theta_2),\cdots,(r_t(t),\theta_t)\}$. 为此，需要确定 $\bar{k}_i\sim \mathrm{e}^{r_t-r_i}$，其中 $\bar{k}_i$ 为节点 $v_i$ 的期望度；然后用真正的度 $k_i$ 代替 $\bar{k}_i$，进而推断节点的半径坐标. 每个节点的坐标值，可以通过最大化以下似然估计式[11]来完成：

$$L_{v_i}=\prod_{1\leqslant j<i}P(x_{ij}(v_i))^{a_{ij}}(1-P(x_{ij}(v_i)))^{1-a_{ij}}. \quad (1)$$

其中 $P(x_{ij}(v_i))=1/(1+\mathrm{e}^{(x_{ij}-R)/T})$ 是节点 $v_i$ 与 $v_j$ 连边概率，$R$ 是双曲圆半径，$T$ 是网络平均聚集系数.

在为网络中的每个节点拟合双曲坐标后，计算网络中所有连边的双曲距离 $x_{ij}$，并将链接按照分配的出现时间进行降序排序，在网络中去掉排序次序靠前的 $\Delta l$ 个链接，所得到的网络记为 $G_2(N,E_2)$，将其作为传播时的拓扑. 具体步骤如下：

1) 为了推断网络的半径坐标和角坐标，首先在当前网络 $G_1(N,E_1)$ 上运行双曲映射算法；

2) 给每个观察到的链接分配出现时间 $\varphi_{ij}=x_{ij}$，这里 $v_i,v_j$ 表示此链接的两个端点，$x_{ij}$ 是节点 $v_i$ 出现时与节点 $v_j$ 之间的双曲距离；

3) 对现存的链接按照分配的出现时间进行降序排序，删除当前网络序列中的前 $\Delta l$ 个链接，进而得到估计出的真正的传播拓扑 $G_2(N,E_2)$.

得到网络的传播拓扑后，采用文献[8]提出的最大似然估计器估计信息源点，算法的伪代码如下所示.

算法：动态源点定位(dynamic source localization algorithm，DLSA).

1) Input：观察量集合 $O=\{(o_i,v,t_{vo_i})\}$；网络 $G_1(N,E_1)$；

2) 对于网络 $G_1(N,E_1)$，所有节点按度值 $k_i$ 降序排序，$k_1>k_2>\cdots>k_M$，若节点的度相等则它们的顺序任意；每个拥有度 $k_i$ 的节点记为 $v_i$，$i=1,2,\cdots,M$；

3) 令 $v_i(i=1)$ 的初始坐标 $r_1=0$，角坐标 $\theta_1$ 为随机值，$\theta_1\in[1,2\pi]$；

4) for $i=2$ 到 $M$ do

5) 为节点 $v_i$ 分配初始坐标；

6) 按照公式 $r_j(i)=\beta r_j+(1-\beta)r_i$ 对所有现存节点 $v_i,v_j(j<i)$ 添加它们的半径坐标；

7) 给节点 $v_i$ 分配角坐标 $\theta_i$，使似然估计 $L_{v_i}$ (式(1))最大化.

8) endfor

9) 令 $v_i,v_j$ 链接的出现时间 $\varphi_{ij}=x_{ij}$，按 $\varphi_{ij}$ 降序排列观察到的链接；在 $G_1(N,E_1)$ 中删除序列中前 $\Delta l$ 个链接，得到估计的传播拓扑 $G_2(N,E_2)$；

10) 在 $G_2(N,E_2)$ 上，令收到消息的观察点子集为 $O_{\mathrm{a}}=\{o_k\}_{k=1}^{K_{\mathrm{a}}}\subseteq O$，$K_{\mathrm{a}}$ 为 $O_{\mathrm{a}}$ 的大小；

11) 选定 $O_{\mathrm{a}}$ 中任意观察点为 $o_1$，令其首次收到信息的时间为 $t_1$；

12) for $s\in S=N-O_{\mathrm{a}}$ do

13) 在 $G_2$ 上以 $s$ 为根生成广度优先搜索树 $T_{\mathrm{bfs},s}$；

14) 令 $\hat{s}=\underset{s\in S}{\operatorname{argmax}}\,\boldsymbol{\mu}_s^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\left(\boldsymbol{d}-\frac{1}{2}\boldsymbol{\mu}_s\right)$；

其中 $\boldsymbol{d}$ 是观察到的延迟，$\boldsymbol{\mu}_s$ 是确定性延迟，$\boldsymbol{\Lambda}$ 是协方差；

15) endfor

16) return 节点 $\hat{s}$ 即为信息源点

## 2　实验分析

为了验证算法的准确率及效率，定义了定位概率、定位误差距离和定位误差比例，通过这些指标对算法进行评价.

定位概率定义为 $L_{1\mathrm{p}}=\mathrm{Count}(s^*=\hat{s})/\mathrm{ExpTimes}$，其中 $\mathrm{Count}(s^*=\hat{s})$ 表示定位准确的实验次数，ExpTimes 表示总实验次数. 定位误差距离定义为 $L_{1\mathrm{el}}(h)=\sum_{i=0}^{i=L_{\max}} i\cdot(C_i/h)$，其中 $L_{\max}$ 表示估计量 $\hat{s}$ 距离真正源点 $s^*$ 的跳数在 $h$ 次实验中的最大值，$C_i$ 表示估计量 $\hat{s}$ 距离真正源点 $s^*$ 的跳数为 $i$ 的实验次数. 定位误差比例定义为 $L_{1\mathrm{ef}}(h)=\mathrm{CountH}(h)/\mathrm{ExpTimes}$，其中 $\mathrm{CountH}(h)$ 表示估计量 $\hat{s}$ 距离真正源点 $s^*$ 的跳数为 $h$ 的实验次数.

为验证算法的有效性，本文采用无标度网络[9]进行实验，生成的网络节点数 $M=1\,000$，边数 $L=3\,000$. 首先考察定位时间对于定位准确率的影响. 一般来说，当传播时间较长时，信息在网络中广泛传播，同时拓扑会发生较大的变化，可能会造成准确率降低. 因此，考察了不同定位时间差 $\Delta t=t^{\mathrm{o}}-t^{\mathrm{s}}$ 的源点定位准确率，其中 $t^{\mathrm{o}}$ 为定位时间，$t^{\mathrm{s}}$ 为源点传播时间. 令 $\Delta t=25,50,75,100,125,150$ s. 观察点的选择采用随机选取、高度节点和高 $k$-core 节点三种策略，观察点比例为 5%.

对每个时间差的每种部署策略进行 1 000 次随机传播实验,据此计算出定位概率. 选用文献[8]中使用的静态定位算法作为对比.

从图 3 中可以看出,在随机部署策略下,动态源点定位算法的定位概率与静态算法的定位概率大致相同,而对于高度节点策略和高 $k$-core 节点策略,动态源点定位算法的定位概率明显高于静态定位算法. 随着时间差的增大,动态源点定位算法的定位概率基本保持不变,而静态算法的定位概率大致呈下降趋势,除随机策略外.

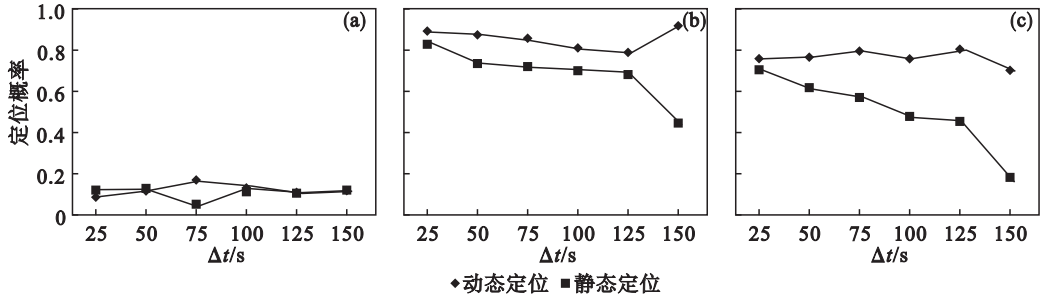


图 3 不同时间差下动态定位算法与静态定位算法比较

Fig. 3 Comparison of dynamic localization algorithm and static localization algorithm at different time differences

(a)—随机部署策略;(b)—高 $k$-core 节点策略;(c)—高度节点策略.

图 4 给出了动态定位算法不同观察点部署策略下的定位误差距离和定位误差比例. 令时间差为 $\Delta t = 25$ s. 从图 4 看出,对于定位误差比例,动态算法在高 $k$-core 节点策略中误差距离为 0 的比例最高;对于定位误差距离,在动态定位算法中,高 $k$-core 节点策略的误差距离主要集中在 0 附近,随机策略误差距离则集中在 2 附近. 对于不能准确定位的情况,真实源点与估计源点的距离也在 2 至 3 跳以内,这也说明了算法的可行性.

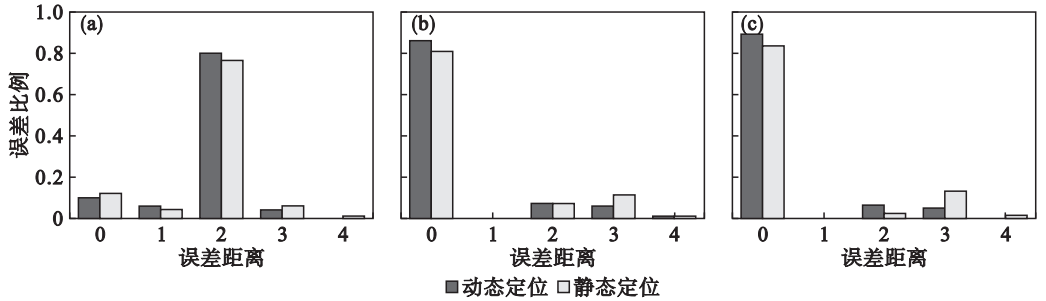


图 4 动态定位算法的定位误差比例

Fig. 4 Localization error frequency of dynamic positioning algorithm

(a)—随机部署策略;(b)—高度节点策略;(c)—高 $k$-core 节点策略.

## 3 结 语

本文提出一种面向动态社交网络的信息传播源点定位方法,算法基于偏好性及相似性这两个网络演化的关键因素,通过推断传播时网络的拓扑结构,从而准确地定位网络传播信息源点. 算法基于观察点搜集到的部分传播数据和传播时刻的网络拓扑,估计潜在的信息源点. 在模型网络上进行实验,结果表明,本文所提出的动态定位算法,相对于不考虑拓扑动态变化的静态定位算法,其定位准确率有较大提高.

本文的算法有一个前提假设,即动态社交网络只考虑边的变化,这对于以微博为代表的在线社交网络来说,在短时间内是符合实际情况的. 对于节点与边同时变化的情形,将在以后的工作中进行探讨.

**参考文献:**

[1] 李栋,徐志明,李生,等. 在线社会网络中信息扩散[J]. 计算机学报,2014,37(1):189 – 206.
(Li Dong, Xu Zhi-ming, Li Sheng, et al. A survey on information diffusion in online social networks[J]. *Chinese Journal of Computers*, 2014, 37(1): 189 – 206.)

[2] Guille A, Hacid H, Favre C, et al. Information diffusion in online social networks: a survey[J]. *ACM SIGMOD Record*, 2013, 42(2): 17 – 28.

[3] Difonzo N. Rumor research can douse digital wildfires[J]. *Nature*, 2013, 493(7431): 135 – 135.

[4] Centola D. The spread of behavior in an online social network experiment[J]. *Science*, 2010, 329(5996): 1194 – 1197.

[5] 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.

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

[7] Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena[J]. *Science*, 2013, 342 (6164): 1337 – 1342.

[8] Pinto P C, Thiran P, Vetterli M. Locating the source of diffusion in large-scale networks [J]. *Physical Review Letters*, 2012, 109(6): 1 – 5.

[9] Barabasi A L, Albert R. Emergence of scaling in random networks[J]. *Science*, 1999, 286(5439): 509 – 512.

[10] Kleinberg J M, Kumar R, Raghavan P, et al. The web as a graph: measurements, models, and methods [C]// International Conference on Computing and Combinatorics. [S. l.]: Springer-Verlag, 1999: 1 – 17.

[11] Papadopoulos F, Kitsak M, Serrano M A, et al. Popularity versus similarity in growing networks[J]. *Nature*, 2012, 489 (7417): 537 – 540.

(上接第 208 页)

对比仿真地震序列的时空域网络与真实地震时空域网络的累积度分布情况,发现改进后的 OFC 模型产生的仿真地震序列与真实地震数据在宏观网络拓扑上具有高度一致性,可以认为改进后的 OFC 模型生成的仿真地震数据与真实地震数据具有相似性.

## 参考文献:

[1] 谢周敏. 地震活动的网络拓扑结构和网络动力学行为[J]. 震灾防御技术, 2011, 6(1): 1 – 17.
(Xie Zhou-min. Network topology and network dynamical behavior of seismicity [J]. *Technology for Earthquake Disaster Prevention*, 2011, 6(1): 1 – 17.)

[2] Abe S, Suzuki N. Complex network of earthquakeity [J]. *Progress of Theoretical Physics Supplement*, 2006, 162: 138 – 146.

[3] Abe S, Suzuki N. Complex-network description of earthquakeity[J]. *Nonlinear Processes in Geophysics*, 2006, 13(2): 145 – 150.

[4] He X, Zhao H, Cai W, et al. Earthquake networks based on space-time influence domain [J]. *Physica A: Statistical Mechanics and Its Applications*, 2014, 407: 175 – 184.

[5] Wang X F, Chen G. Complex networks: small-world, scale-free and beyond[J]. *Circuits and Systems Magazine*, 2003, 3 (1): 6 – 20.

[6] Ferreira D S R, Papa A R R, Menezes R. On the agreement between small-world-like OFC model and real earthquakes [J]. *Physics Letters: A*, 2014, 379(7): 669 – 675.

[7] de Carvalho J, Prado C P C. Self-organized criticality in the Olami-Feder-Christensen model[J]. *Physical Review Letters*, 1999, 84(17): 4006 – 4009.

[8] Lise S, Paczuski M. Self-organized criticality and universality in a nonconservative earthquake model[J]. *Physical Review: E*, 2001, 63(3): 036111(1 – 5).

[9] Hergarten S, Neugebauer H J. Foreshocks and aftershocks in the Olami-Feder-Christensen model [J]. *Physical Review Letters*, 2003, 88(23): 238501(1 – 4).