微博双向“关注”网络节点中心性及传播影响力的分析*

苑卫国[1)2)] 刘云[1)†] 程军军[1)] 熊菲[1)]

1) (北京交通大学, 通信与信息系统北京市重点实验室, 北京 100044)

2) (中国科学院计算机网络信息中心, 北京 100190)

(2012 年 6 月 7 日收到; 2012 年 9 月 6 日收到修改稿)

根据新浪微博的实际数据, 建立了两个基于双向“关注”的用户关系网络, 通过分析网络拓扑统计特征, 发现二者均具有小世界、无标度特征. 通过对节点度、紧密度、介数和 *k*-core 四个网络中心性指标进行实证分析, 发现节点度服从分段幂率分布; 介数相比其他中心性指标差异性最为显著; 两个网络均具有明显的层次性, 但不是所有度值大的节点核数也大; 全局范围内各中心性指标之间存在着较强的相关性, 但在度值较大的节点群这种相关性明显减弱. 此外, 借助基于传染病动力学的 SIR 信息传播模型来分析四种指标在刻画节点传播能力方面的差异性, 仿真结果表明, 选择具有不同中心性指标的初始传播节点, 对信息传播速度和范围均具有不同影响; 紧密度和 *k*-core 较其他指标可以更加准确地描述节点在信息传播中所处的网络核心位置, 这有助于识别信息传播拓扑网络中的关键节点.

关键词: 微博, 中心性, 复杂网络, 信息传播, *k*-core

PACS: 89.75.−k, 29.85.−c, 89.75.Da **DOI:** 10.7498/aps.62.038901

1 引 言

微博 (MicroBlog) 作为一种在线社会网络应用形式, 在最近几年中得到了迅猛的发展, 如 Twitter 和新浪微博就是其中的典型代表. 与社交网站 (Social Networking Sites) 和博客 (Blog) 相比, 微博兼具“社会网络”和“媒体平台”的特点. 而且微博短文本快速发布的方式极大地方便了用户分享和讨论感兴趣的信息, 这种方式有助于形成一种新型社会关系网络. 此外, 微博的受众数量巨大, 信息交互频繁、传播迅速, 使其成为反映社会舆情的主要网络载体来源之一. 综上, 微博在舆情事件发酵、信息传播过程中起着极其重要的作用.

复杂网络理论研究的兴起 [1−3] 为在线社会网络拓扑结构和信息传播特性研究提供了有效方法; 在线社会网络的迅速发展, 又为复杂网络研究提供了丰富的数据来源. 实际社会网络与在线社会网络具有显著差异, 同时又具有密切的内在联系, 早期研究者更关注于实际社会网络 [4]. 近期, 在线社会网络的研究吸引了包括统计物理学、社会学和计算机信息技术等领域研究者的广泛关注. 文献 [5—10] 研究了几种典型在线网络的拓扑结构, 发现多数在线社会网络具有无标度、小世界、层次性和社团结构等共性. 微博作为一种典型的在线社会网络也具有复杂网络特性, 研究其网络结构和传播特性同样具有重要意义. 文献 [11] 对早期 Twitter 的数据集进行了研究, 发现用户度分布服从严格幂率分布, 并具有小世界特性, 以及较高的互惠性. 文献 [12] 针对近乎全网的 Twitter 用户关系数据集进行分析, 发现其虽然符合无标度和小世界网络结构特征, 但节点的出入度并非严格遵循幂率分布; 用

* 国家自然科学基金 (批准号: 61172072, 61271308)、北京市自然科学基金 (批准号: 11DA1454) 和中央高校基本科研业务费专项资金 (批准号: 2011YJS215) 资助的课题.

† 通讯作者. E-mail: liuyun@bjtu.edu.cn

© 2013 中国物理学会 Chinese Physical Society

户之间连接是非对称的, 而双向 “关注” 用户节点之间却表现出同质性.

复杂网络的中心性研究, 起源于社会网络 [13], 并应用于生物网络 [14]、航空网络 [15]、公交网络 [16]、通信网络 [17,18]、城镇网络 [19] 等. 为了更加准确地描述网络中心性, 研究者提出了不同的中心性指标进行测量, 文献 [20] 给出了不同中心性指标的特点及其应用场合. 文献 [21] 分析了网络传播中最有影响力的节点, 发现核数 (*k*-core) 描述节点的中心性和影响力比节点度和介数更为准确. 文献 [22] 认为在信息传播过程中, 选择核数不同的初始节点虽然对信息传播最终范围大小没有影响, 但是核数高的节点对信息传播阻断能力更强. 文献 [23] 提出一种基于邻域信息的半局部中心性识别方法用于识别无向网络中最具影响力的节点, 在速度和效果上取得了很好的平衡. 文献 [24] 在综合考虑节点自身在网络中所处的位置和相邻节点的重要程度, 提出一种利用重要度评价矩阵来确定网络关键节点的方法. 文献 [25] 结合具体的模型, 分析了网络中心性与网络脆弱性的关系, 由此论证了对网络中的中心节点进行破坏的意义. 文献 [26] 构造了一个基于在线社交网络的信息传播模型, 发现初始传播节点的度越大, 信息越容易在网络中迅速传播. 文献 [27] 进一步提出一种在线社交网络的观点传播模型, 发现模型中信息传播速度与六度分隔理论 [28] 十分符合, 且稳定时观点分布与源节点的度值密切相关. 不难看出, 通过对不同类型的在线社会网络进行中心性指标分析, 发现网络中关键节点, 对于揭示在线社会网络的拓扑结构, 认识信息传播规律、研究信息传播控制方法具有重要理论和应用价值. 然而, 该方面研究并未引起充分关注.

本文根据新浪微博的一些实际用户数据, 构造了两个基于双向 “关注” 的用户关系网络; 通过分析网络拓扑统计特征发现上述两个网络都具有小世界和无标度的特征; 然后分别对两个网络的四种中心性指标 (节点度、紧密度、介数和 *k*-core) 及其相关性进行分析; 在此基础上, 借助基于传染病动力学的 SIR 信息传播模型, 分别分析两个网络中具有不同中心性指标的初始传播节点对信息传播速度和范围的影响. 结果表明, 紧密度和 *k*-core 较其他指标可以更加准确地描述节点在信息传播中所处的网络核心位置. 进一步的分析可知上述两个指标有助于识别信息传播拓扑网络中的关键节点.

本文结构如下: 第 2 节介绍所获取的双向关注网络和统计特性指标, 第 3 节分析网络节点的四个中心指标及其相关性, 第 4 节建立传播模型并进行仿真, 第 5 节给出结论.

2 数据集

一份准确的数据集是研究网络拓扑结构的必要条件. 本文基于新浪微博开放 API 的方式完成用户基本信息和用户关系数据的获取, 数据抓取过程采用的是滚雪球的方法和宽度优先的策略. 为了弱化不同类型的采样起点对分析结果可能造成的影响, 本文分别选取两类用户 (普通用户和名人用户) 为采样起点, 最终获取到的两个基本数据集如表 1 所示.

表 1 自新浪微博采集的两个基本数据集

数据集	用户数	关注数	时间区间	爬行起点
A	118517	2728213	2012-01-19 到 2011-01-27	作者
B	169428	2699411	2012-01-24 到 2012-02-03	姚晨

本文将用户表示为复杂网络中的节点, 用户间的关注关系表示为节点间的有向边, 例如节点 *i* 指向节点 *j* 的边表示用户 *i* 是用户 *j* 的粉丝 (Follower), 反过来用户 *j* 是用户 *i* 的关注好友 (Friends), 双向 “关注”(BiConnect) 关系意味着节点 *j* 也同时指向节点 *i*. 文献 [12] 统计 Twitter 中双向 “关注” 关系比例为 22.1%, 本文采集到的新浪微博中双向 “关注” 关系比例为 10.98%, 因此, 微博用户之间双向关注数远低于单向关注数. 本文主要研究双向 “关注” 的用户关系, 忽略大量单向 “关注” 用户关系的影响, 如名人和媒体用户, 以求更加准确地反映微博中普通用户之间的社会网络结构特征.

由于新浪接口对获取用户的关系数据有个数限制, 初始采集到的用户关系数据无法形成完整和封闭的网络, 因此, 需要对原始数据集做进一步处理, 分别构造了两个基于用户双向 “关注” 关系的封

闭网络. 具体构造过程如下: 找出用户边关系和用户信息 ID 交集, 反复删除不在交集的边关系, 直到剩余边关系的节点都在该集合内; 去除孤立点的边关系, 找到网络最大连通子图. 构造的两个双向 “关注” 网络统计特性数据如表 2 所示.

从表 2 所描述的平均聚类系数、直径和平均路径长度, 可以看出这两个网络都符合小世界网络特征.

表 2 微博双向 “关注” 网络的基本数据

网络	节点数	边关系	平均度	直径	平均聚类系数	平均路径长度
A	5906	26582	4.67	12	0.071	4.57
B	8578	32390	3.78	11	0.083	4.25

3 中心性指标分析

3.1 节点度中心性分析

节点度中心性 (degree centrality) 指节点的度数, 适用于对局部网络节点的中心地位和影响力进行刻画. 设网络有 n 个节点, k 为节点度, 可以定义节点 x 的节点度中心性为

$$C_D(x) = k(x), \tag{1}$$

对两个网络的节点度分布进行分析, 结果如图 1 所示.

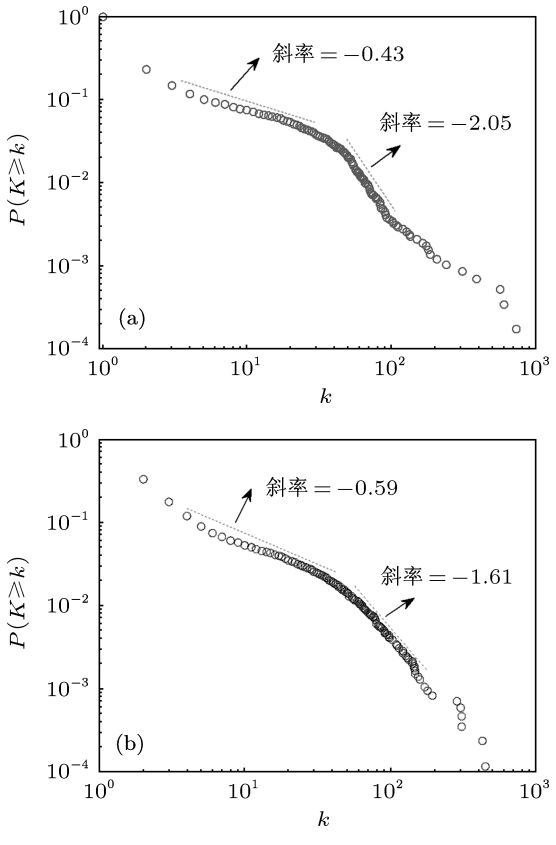

图 1 节点度累计分布图 (a) 网络 A; (b) 网络 B

从图 1 可知, 节点度累积分布都遵循分段幂率分布, 幂率指数分别为

$$P(k) \sim k^{-(\gamma-1)}, \begin{cases} \gamma - 1 = 0.43, & \text{if} \quad k < 40, \\ \gamma - 1 = 2.05, & \text{if} \quad k > 40, \end{cases}$$

$$P(k) \sim k^{-(\gamma-1)}, \begin{cases} \gamma - 1 = 0.59, & \text{if} \quad k < 40, \\ \gamma - 1 = 1.61, & \text{if} \quad k > 40. \end{cases}$$

根据已有研究对社会网络和自然生物网络等复杂网络的实证分析可知, 节点度都服从幂率分布 [4], 即 $P(k) \sim k^{-\gamma}, \gamma \in [1,3]$. 两个网络的节点度分布情况与此符合, 表明其具有复杂网络的无标度特性.

3.2 紧密度中心性分析

紧密度中心性 (closeness centrality), 是刻画节点通过网络到达其他节点难易程度的指标, 相比节点度指标更能反映网络的全局结构. 节点的紧密度越高, 则离其他节点越近, 传播信息时难度越低, 所需借助的节点越少, 反之亦然.

可以定义节点的紧密度中心性为

$$C_C(x) = \frac{n-1}{\sum_{y=1}^{n} d_{xy}}, \tag{2}$$

其中 d_{xy} 表示节点 y 到节点 x 的最短路径距离, n 表示网络节点总数, $n-1$ 表示最大可能的邻点数.

对两个网络的节点紧密度进行分析, 结果如图 2 所示.

从图 2 可以看出, 两个网络节点的紧密度值均在 0.1 到 0.5 之间, 节点连接很紧密, 说明信息可以快速在节点之间进行传播.

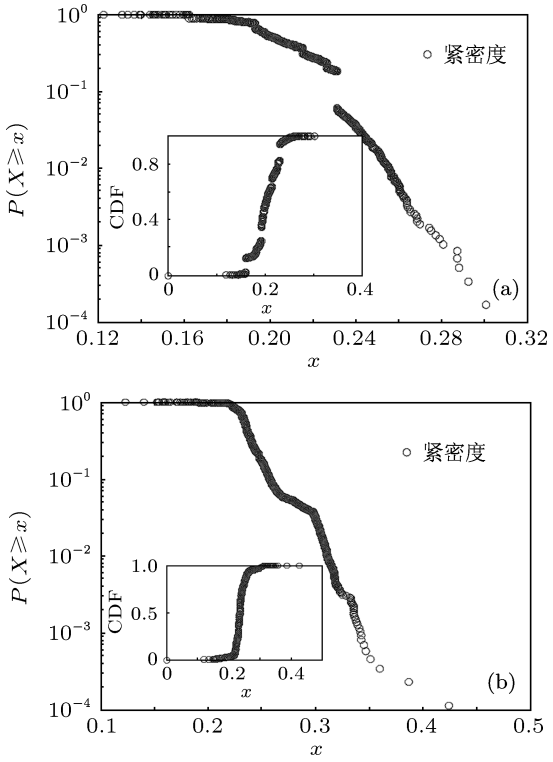

图 2 节点紧密度累计分布图 (a) 网络 A; (b) 网络 B

两个网络的节点度与紧密度之间的相关性分析结果如图 3 所示.

从图 3 可以看出, 网络多数节点的紧密度随节点度值增大而增大, 即度小的节点紧密度值较小, 反之亦然, 说明通常度大的节点处于网络中心位置, 度小的节点处于网络边缘. 但是也可以发现少数度非常大的节点, 紧密度却很小, 说明该类节点虽然拥有大量连接, 但仍处于整体网络的边缘地区.

3.3 介数中心性分析

介数中心性 (betweenness centrality), 是描述网络动态的全局中心性指标.

可以定义节点的介数中心性为

$$C_{\mathrm{B}}(x)=\frac{2\sum_{j<k}g_{jk}(x)}{(n-1)(n-2)g_{jk}}, \tag{3}$$

其中 g_{jk} 表示节点 j 与节点 k 之间的最短路径条数, $g_{jk}(x)$ 表示节点 j 与节点 k 之间经过节点 x 的最短路径条数 (节点 x 的介数), $(n-1)(n-2)/2$ 表示最大可能的节点介数 (任意其他两节点最短路径都经过节点 x). 根据节点介数中心性定义, 处于网络中心位置的节点是信息在网络上传输时负载最重的节点, 也就是经过此点的最短路径条数最多的节点.

对两个网络的介数中心性进行分析, 结果如图 4 所示.

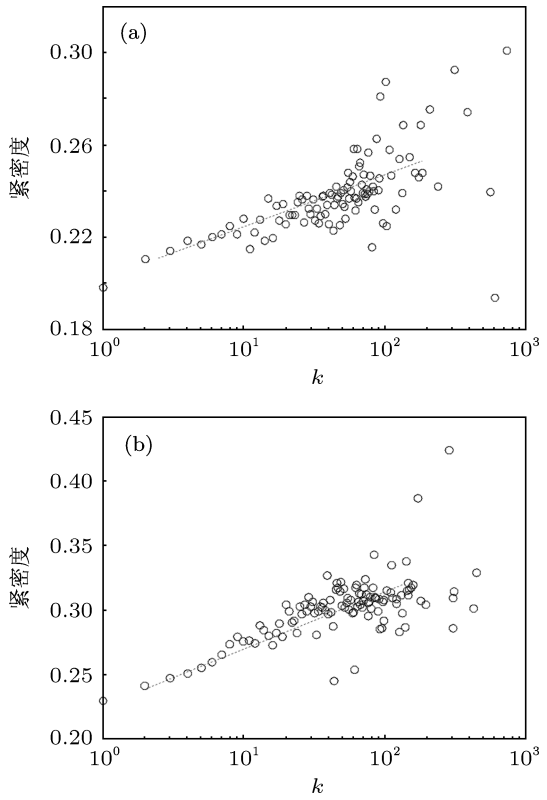

图 3 节点度与紧密度相关性 (a) 网络 A; (b) 网络 B

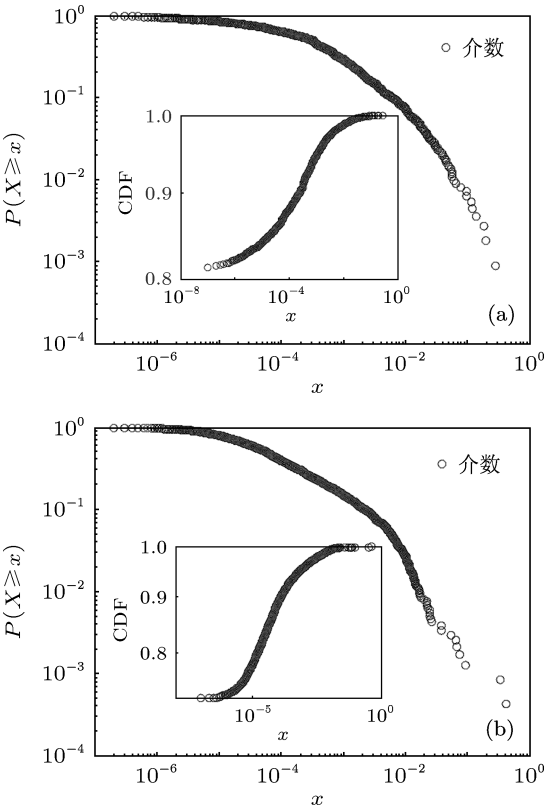

图 4 节点介数累计分布图 (a) 网络 A; (b) 网络 B

从图 4 可以看到, 网络 A 的节点介数累计分布近似服从广延指数分布, 网络 B 的节点介数累计分布更近似服从幂率分布. 网络 A 中有 81%节点介数为 0, B 网络中有 73%节点介数值为 0. 网络 A 中仅有 1 个节点介数值介于 0.2—0.4, 网络 B 中仅有 2 个节点介数值介于 0.2—0.4. 因此, 节点的介数相比紧密度和节点度差异更显著.

对两个网络的节点度与介数之间的关系进行分析, 结果如图 5 所示.

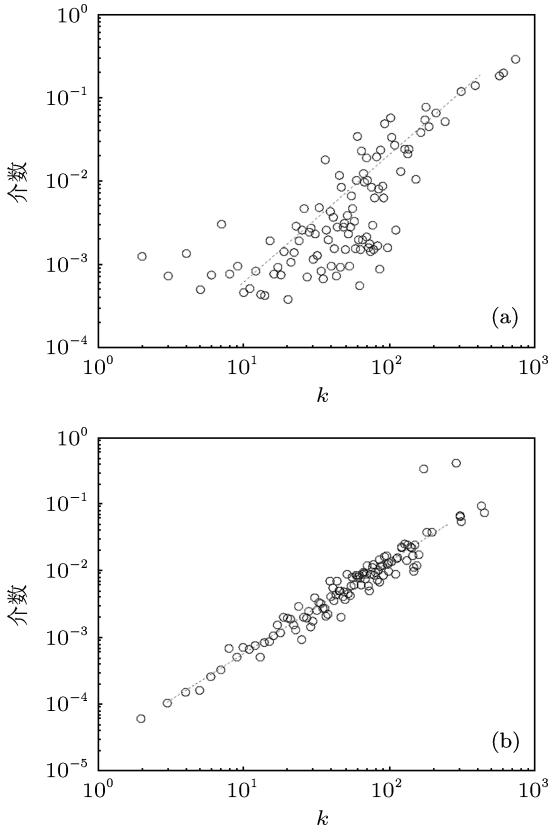

图 5 节点度与介数相关性 (a) 网络 A; (b) 网络 B

从图 5 可以看出, 通常随着节点度值的增加介数值逐渐增加, 度大的节点介数值也较大, 处于网络中心; 度小的节点介数值也较小, 处于网络边缘, 网络 B 这种趋势更为明显. 从网络 A 发现, 也存在少数节点度虽小, 但介数很大, 经查看这些节点一般是连接两个或者多个社区的 HUB 节点.

3.4 *k*-核中心性分析

网络的 *k*- 核 (*k*-Core) 是指反复去掉度小于或等于 *k* 的节点及其连接的边之后, 所剩余的子网. 节点的核数表示节点在核中的深度, 描述了网络拓扑的层次性, 即节点存在于 *k*-core 中, 但在 $(k+1)$-core 中被移去, 也就是说核数为 *k* 的节点存在于所有度都大于 *k* 的子网中. 节点核数中最大的值为网络的核数 (k-core$_{\max}$).

对两个网络的节点核数进行分析, 结果如图 6 所示.

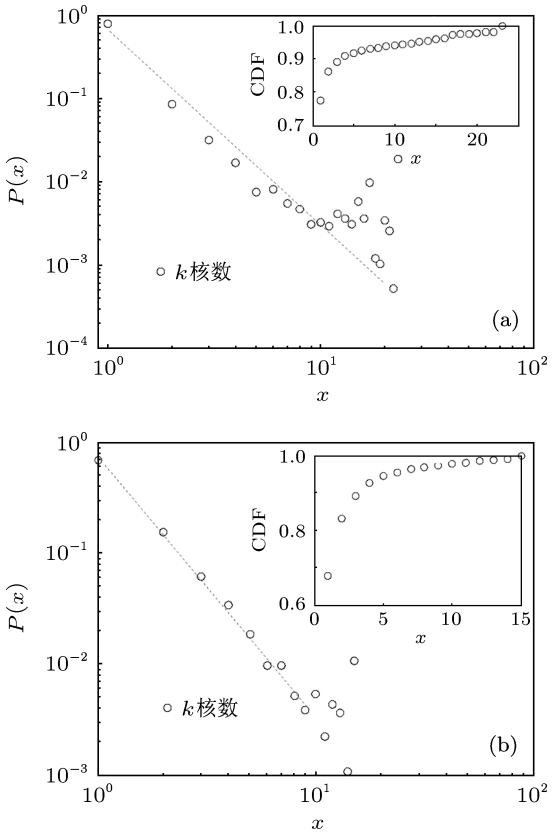

图 6 节点核数分布图 (a) 网络 A; (b) 网络 B

从图 6 可以看到, 两个网络的节点核数分布都近似服从幂率分布, 核数大的节点所占比例逐渐减少, 但核数最大的几个节点所占比例却有所上升. 说明虽然大多数节点的核数很小, 处于网络边缘位置; 但也有相当多的节点达到最大核数值, 集中处于网络的中心层次.

对两个网络的节点度与核数的关系进行分析, 结果如图 7 所示.

从图 7 可以看出, 当节点度值较小时, 两者之间具有正相关性; 当节点的度值较大时, 核数开始发散, 出现一些度值很大但是核数很小的节点.

使用文献 [29] 中 *k*-core 分解方法, 分别生成网络节点度和核数可视化图, 如图 8 所示.

从图 8 可以看出, 两个网络具有明显的层次性, *k*-core 值大的节点聚集在网络的核心, *k*-core 值小

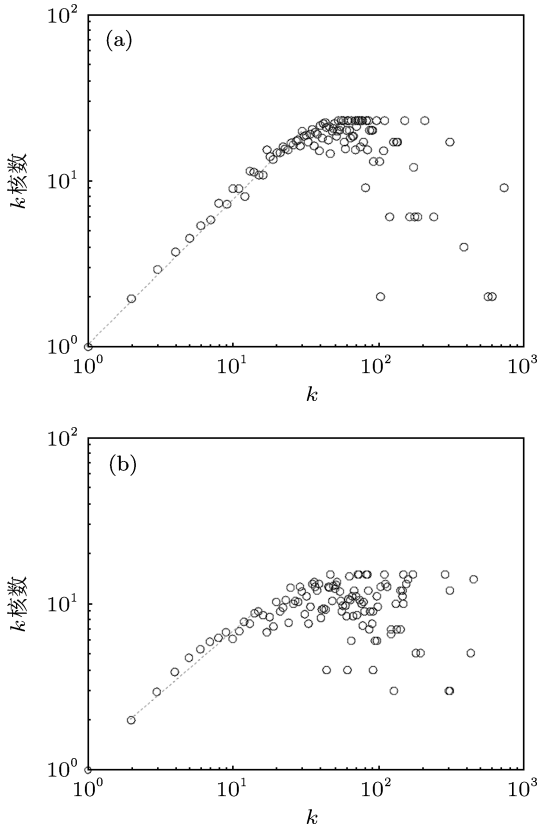

图 7 节点度与核数的相关性 (a) 网络 A; (b) 网络 B

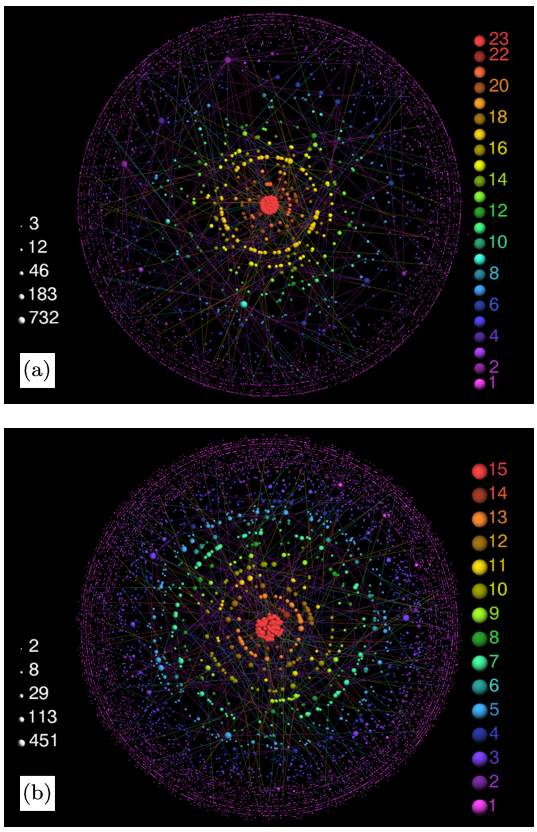

图 8 节点度与核数可视化图 (a) 网络 A; (b) 网络 B

的节点聚集在网络边缘; 一些度大的节点分散在网络的不同区域, 表明并不是所有度值大的节点核数也大.

3.5 中心性指标相关性分析

为了分析各中心性指标的相关程度, 本文采用 Spearman 相关系数 (Correlation coefficient) 计算. 该系数是等级相关系数, 反映两个变量间相关的密切程度与方向, 适用范围更广.

Spearman 相关系数定义为

$$r_R = 1 - \frac{6\sum D^2}{N(N^2-1)}, \tag{4}$$

其中, D 为两个变量每对数据的等级之差, N 表示样本容量.

本文对两个网络的节点度、紧密度、介数和 k-Core 4 种中心性指标的相关性进行计算, 结果如表 3 所示.

表 3 节点中心性特征指标相关性网络 (A)

相关系数	All	Top 1%	Top 10%
度 VS 介数	0.8976	0.6082	0.5646
度 VS 紧密度	0.4072	0.2208	0.4879
度 VS k-core	0.9853	−0.4705	0.7614
介数 VS 紧密度	0.4046	0.5281	0.6497
介数 VS k-core	0.8713	−0.8074	0.1243
紧密度 VS k-core	0.4129	−0.1416	0.2903

网络 (B)

相关系数	All	Top 1%	Top 10%
度 VS 介数	0.9055	0.7457	0.8039
度 VS 紧密度	0.6063	0.1062	0.7768
度 VS k-core	0.9886	−0.0744	0.6281
介数 VS 紧密度	0.6376	−0.0831	0.6772
介数 VS k-core	0.8857	−0.3689	0.2690
紧密度 VS k-core	0.6136	0.8120	0.6231

从表 3 可以看出, 在两个网络的全部节点中, 节点度与介数、核数三者之间相关性非常高, 而紧密度与其他中心性指标之间的相关性相对较低. 在度值排名前 1%的节点中, 除网络 A 的紧密度与介数、网络 B 的紧密度与核数之间相关性有所增强外, 其他节点中心指标间相关性均减弱, 特别是核

数与其他中心性指标之间的相关性甚至为负; 两个网络的核数和紧密度之间的相关性差异明显. 因此, 虽然在全部节点中各种中心性指标之间存在较强的相关性, 但在衡量度值大的节点时, 不同中心性指标相关性差异明显, 同时各中心性指标之间的相关性因所在具体网络不同而不同.

3.6 网络社区结构和中心性分析

为了更直观地研究网络社区结构和节点中心性的关系, 本文使用 Gephi 工具 [30], 对两个网络的社区结构按照颜色分类进行可视化, 图中节点显示的大小与度值呈正比, 并对度值大的节点使用 ID 标注, 结果如图 9 所示.

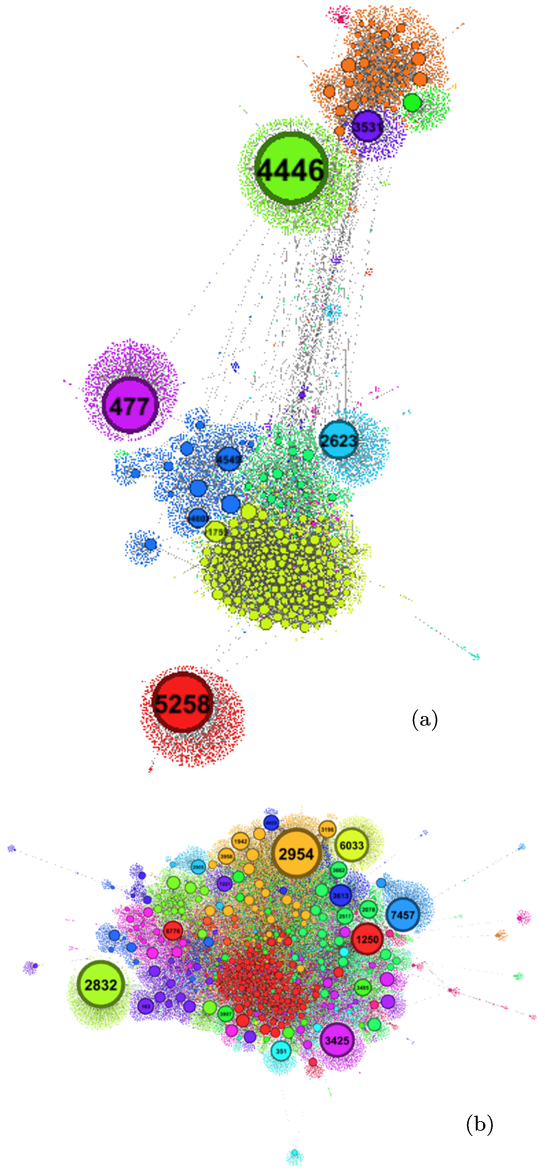

图 9 网络社区结构图 (a) 网络 A; (b) 网络 B

从图 9 可以看出, 两个网络都具有明显的社团结构, 有些社团是由一个度值大的 HUB 节点和众多度为 1 的节点构成, 有些社团则是由多个度值较大的 HUB 节点和众多度小的节点构成. 两个网络中都存在度值很大但并不处于整体网络中心的节点, 这些节点和其周围的节点同处于网络的边缘, 且与其它社区相连接的边关系很少, 如网络 A 中节点 v 5258, 网络 B 中节点 v 2832; 两个网络中也存在度值虽然较小但处于整体网络中心的节点, 该节点与其周围的节点共同处于网络的中心, 与其他多个社区相连接的边关系很多, 如网络 A 中节点 v 3531, 网络 B 中节点 v 2954.

4 信息传播分析

4.1 理论模型

在微博网络中, 用户发表微博后会以一定概率被好友看到, 好友若对内容感兴趣, 会以一定概率转发, 若对内容不感兴趣, 则不会传播, 因此, 信息沿用户好友关系进行传播. 为了进一步研究微博网络中节点的传播影响力和其中心性指标之间的关系, 本文采用传染病理论中的 SIR 模型, 将用户节点分为传播节点、免疫节点、未感染节点, 并定义以下传播规则: 1) 如果一个传播节点与一个未感染节点接触, 则未感染节点会以概率 α 成为传播节点; 2) 如果一个传播节点与一个免疫节点接触, 则传播节点会以概率 β 成为免疫节点; 3) 传播节点会以速度 v 变为免疫节点, 无需与其他节点接触. $S(t)$, $R(t)$ 和 $I(t)$ 为分别表示 t 时刻传播节点、免疫节点、未感染节点的密度.

4.2 仿真结果

初始网络中只有一个传播节点, 其余全部为未感染节点, 设置模型参数为 $\alpha = 1, \beta = 0.1, v = 0.05$, 迭代次数为 $T = 300$ 次. 网络中传播节点、免疫节点和未感染节点的密度随时间演化的结果, 如图 10 所示.

从图 10 可以看出, 由于网络高度联通, 信息传播速度非常快, 免疫节点密度 $R(t)$ 初期快速上升后, 逐渐平稳, 趋向于 1; 传播节点密度 $S(t)$ 初始阶段快速上升, 到达最大值后, 平稳下降, 趋向于 0; 未感染节点密度 $I(t)$ 迅速衰减直到下降趋向于 0.

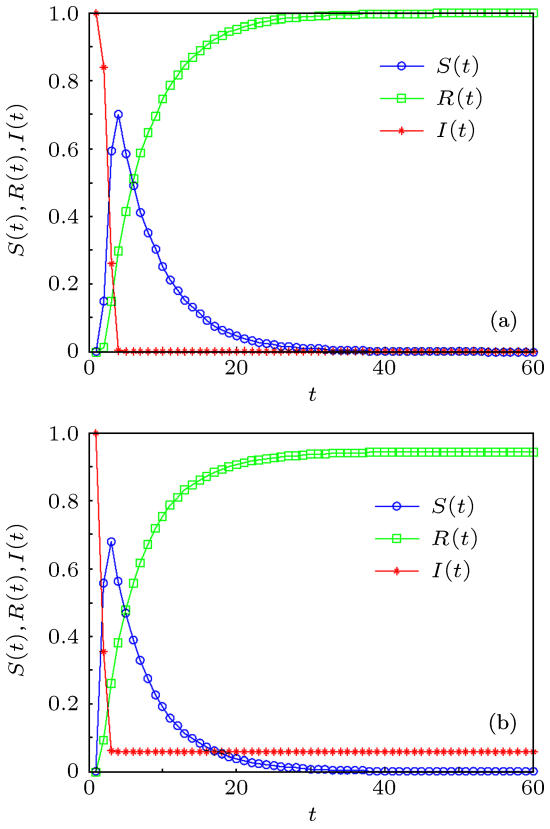

图 10 不同节点密度随时间变化关系 (a) 网络 A; (b) 网络 B

为分析不同初始节点对信息传播速度和规模的影响, 分别选择两个网络中度值排在前 5 名的节点作为初始传播节点进行仿真, 节点的中心性特征指标计算结果如表 4 所示.

表 4 度值前 5 名节点中心性特征指标网络 *A*

编号	k	紧密度	介数	k-core
ID1 = 4446	732	1	1	9
ID2 = 5258	606	3782	2	2
ID3 = 477	564	211	3	2
ID4 = 2623	387	11	4	4
ID5 = 3531	311	2	5	17

网络 B

编号	k	紧密度	介数	k-core
ID1 = 2954	451	26	4	14
ID2 = 2832	426	239	3	5
ID3 = 3425	310	67	7	12
ID4 = 6033	307	110	6	3
ID5 = 7457	306	386	5	3

从表 4 也可以看出, 两个网络的度值排在前五名节点的度值和介数, 以及紧密度和 k-core 数之间, 分别具有明显的相关性.

两个网络的免疫节点密度 $R(t)$ 和传播节点密度 $S(t)$ 随时间变化的结果, 如图 11 和 12 所示.

从图 11 和 12 可以看出, 初始传播节点为不同度值的情况下, 免疫节点密度和传播节点密度随时间的变化情况不同, 度值大的节点不一定是传播速度快的节点. 如网络 A 中, 节点 ID5 虽然度值最小, 但传播速度最快; 节点 ID2 虽然度值很大, 但传播速度最慢. 从图 9 和表 4 可以看出, 节点 ID5 的紧密度值排名靠前且 k-core 数很大, 处于整体网络的中心位置; 节点 ID2 的紧密度值排名靠后且 k-core 数很小, 处于整体网络靠近边缘的位置. 通过对网络 B 的仿真, 观察到类似结果. 实际仿真结果表明, 越靠近整个网络中心位置的节点作为初始节点进行信息传播时速度越快, 这与文献 [21] 在其他社会网络中进行仿真的结果一致. 因此, 从信息传播范围和速度角度看, 实际网络中最有影响力的节点不一定是度值或介数大的节点, 而更可能是处于整体网络核心位置, 且紧密度和 k-core 核数较大的节点.

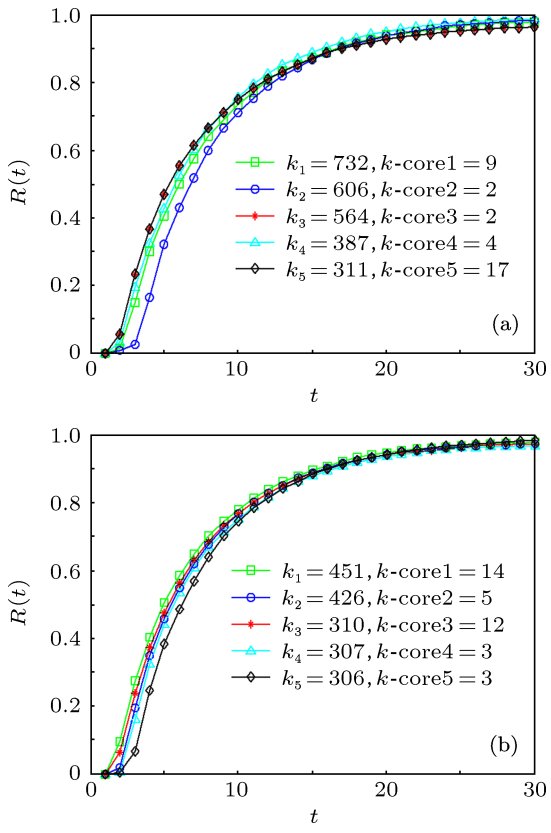

图 11 初始传播节点为不同度值的情况下免疫节点密度 $R(t)$ 随时间变化关系 (a) 网络 A; (b) 网络 B

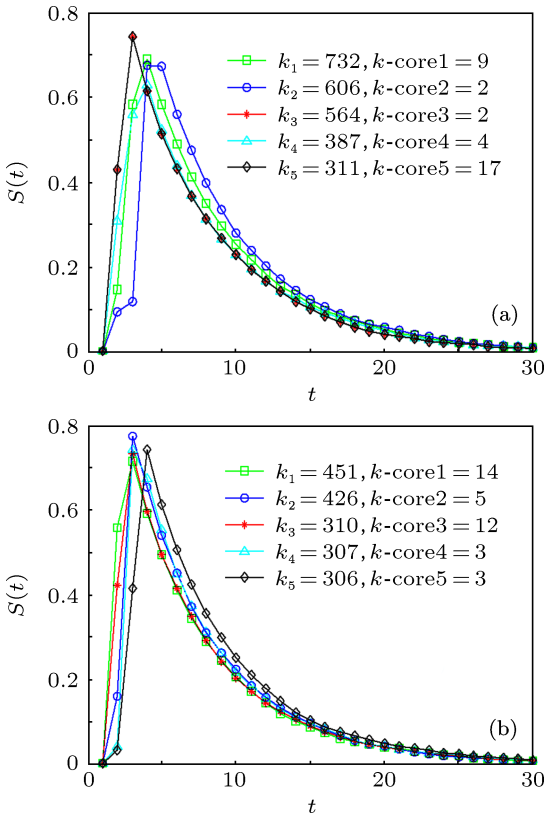

图 12 初始传播节点为不同度值的情况下传播节点密度 $S(t)$ 随时间变化关系 (a) 网络 A; (b) 网络 B

5 结 论

本文根据获取到的新浪微博中用户的实际关系数据, 构造了两个双向 "关注" 网络, 并对描述网络中心性的四个指标进行了实证分析, 发现节点度遵循分段幂率分布; 介数相比其他中心性指标差异更显著; 两个网络都具有明显层次性, 节点核数分布近似服从幂率分布, 但不是所有度值大的节点核数也大. 通过对比各中心性指标之间相关性, 发现在全部网络节点中节点度与介数、核数三者之间相关性非常高, 而紧密度与其他中心性指标之间的相关性相对较低; 虽然在全部节点中各种中心性指标之间存在较强的相关性, 但在衡量度值大的节点时, 不同中心性指标相关性差异明显, 同时各中心性指标之间的相关性因所在具体网络不同而不同. 本文通过结合节点度、核数及网络社区结构的可视化显示, 将网络中度值最大的几个节点作为初始传播节点并分析对网络信息传播速度和范围的影响, 发现相对节点度值和介数, 紧密度和 k-core 数的测量结果可更加准确地发现网络中信息传播的关键节点. 本文工作有助于揭示微博用户双向 "关注" 网络的结构与传播功能关系, 进而对微博中信息传播进行有效控制. 如何对移除关键节点后的网络传播功能进行分析是今后的研究方向.

感谢审稿人对本文提出的宝贵意见和建议.

[1] Watts D J, Strogatz S H 1998 *Nature* **393** 440

[2] Barabási A L, Albert R 1999 *Science* **286** 509

[3] Newman M E J 2003 *SIAM Rev.* **45** 167

[4] Newman M E J, Park J 2003 *Phys. Rev. E* **68** 036122

[5] Kumar R, Novak J, Tomkins A 2006 *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Philadelphia, USA, August 20–23, 2006 p611

[6] Ahn Y Y, Han S, Kwak H, Moon S, Jeong H 2007 *Proceedings of the 16th International Conference on World Wide Web* Banff, Canada, May 8–12, 2007 p835

[7] Mislove A, Marcon M, Gummadi K P, Druschel P, Bhattacharjee B 2007 *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement* San Diego, USA, October 24–26, 2007 p29

[8] Fu F, Liu L, Wang L 2008 *Physica* A **387** 675

[9] Hu H B, Wang X F 2009 *Phys. Lett.* A **373** 1105

[10] Si X M, Liu Y 2011 *Acta Phys. Sin.* **60** 78903 (in Chinese) [司夏萌, 刘云 2011 物理学报 **60** 78903]

[11] Java A, Song X, Finin T, Tseng B 2007 *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Snalysis* San Jose, USA , August 12, 2007 p56

[12] Kwak H, Lee C, Park H, Moon S 2010 *Proceedings of the 19th International Conference on World Wide web* Raleigh, USA, April 26–30, 2010 p591

[13] Wasserman S, Faust K 1994 *Social Network Analysis*: *Methods and Applications* (New York: Cambridge Univ. Press) p169

[14] Koschôtzki D, Schreiber F 2004 *Proceedings of the German Conference on Bioinformatics* Bielefeld, Germany, October 4–6, 2004 p199

[15] Guimerà R, Mossa S, Turtschi A, Amaral L A N 2005 *Proc. Natl. Acad. Sci. USA* **102** 7794

[16] Zheng X, Chen J P, Shao J L, Bie L D 2012 *Acta Phys. Sin.* **61** 190510 (in Chinese) [郑啸, 陈建平, 邵佳丽, 别立东 2012 物理学报 **61** 190510]

[17] Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E 2007 *Proc. Natl. Acad. Sci. USA* **104** 11150

[18] Cai K Q, Zhang J, Du W B, Cao X B 2012 *Chin. Phys.* B **21** 28903

[19] Paolo C, Vito L, Sergio P 2006 *Phys. Rev.* E **73** 036125

[20] Wang L, Zhang Q Q 2006 *Complex Systems and Complexity Science* **3** 13 (in Chinese) [王林, 张倩倩 2006 复杂系统与复杂性科学 **3** 13]

[21] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E, Makse H A 2010 *Nat. Phys.* **6** 888

[22] Borge-Holthoefer J, Moreno Y 2012 *Phys. Rev.* E **85** 026116

[23] Chen D B, Lü L Y, Shang M S, Zhang Y C, Zhou T 2012 *Physica* A **391** 1777

[24] Zhou X, Zhang F M, Li K W, Hui X B, Wu H S 2012 *Acta Phys. Sin.*

61 50201 (in Chinese) [周漩, 张凤鸣, 李克武, 惠晓滨, 吴虎胜 2012 物理学报 **61** 50201]

[25] Holme P, Kim B J, Yoon C N, Han S K 2002 *Phys. Rev.* E **65** 056109

[26] Zhang Y C, Liu Y, Zhang H F, Cheng H, Xiong F 2011 *Acta Phys. Sin.* **60** 50501 (in Chinese) [张彦超, 刘云, 张海峰, 程辉, 熊菲 2011 物理学报 **60** 50501]

[27] Xiong X, Hu Y 2012 *Acta Phys. Sin.* **61** 150509 (in Chinese) [熊熙, 胡勇 2012 物理学报 **61** 150509]

[28] http://en.wikipedia.org/wiki/Six_degrees_of_separation

[29] Alvarez-Hamelin J I, Dallásta L, Barrat A, Vespignani A 2006 *Advances in Neural Information Processing Systems* 18 (Cambridge: MIT Press) p41

[30] http://gephi.org/

Empirical analysis of microblog centrality and spread influence based on Bi-directional connection*

Yuan Wei-Guo[1)2)] Liu Yun[1)†] Cheng Jun-Jun[1)] Xiong Fei[1)]

1) (*Key Laboratory of Communication & Information Systems, Beijing Jiaotong University, Beijing* 100044*, China*)

2) (*Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100190*, China*)

(Received 7 June 2012; revised manuscript received 6 September 2012)

Abstract

The identifying of the most influential nodes in the complex network is of great significance for information dissemination and control. We collect actual data from Sina Weibo and establish two user relationship networks based on bi-directional "concern". By analyzing the statistical characteristics of the network topology, we find that each of them has a small world and scale free characteristics. Moreover, we describe four network centrality indicators, including node degree, closeness, betweenness and *k*-Core. Through empirical analysis of four-centrality metric distribution, we find that the node degrees follow a segmented power-law distribution; betweenness difference is most significant; both networks possess significant hierarchy, but not all of the nodes with higher degree have the greater *k*-Core values; strong correlation exists between the centrality indicators of all nodes, but this correlation is weakened in the node with higher degree value. The two networks are used to simulate the information spreading process with the SIR information dissemination model based on infectious disease dynamics. The simulation results show that there are different effects on the scope and speed of information dissemination under different initial selected individuals. We find that the closeness and *k*-Core can be more accurate representations of the core of the network location than other indicators, which helps us to identify influential nodes in the information dissemination network.

Keywords: microblog, centrality, complex network, information transmission, *k*-Core

PACS: 89.75.−k, 29.85.−c, 89.75.Da **DOI:** 10.7498/aps.62.038901

* Project supported by the National Natural Science Foundation of China (Grant Nos. 61172072, 61271308), the Beijing Natural Science Foundation (Grant No. 11DA1454) and the Fundamental Research Funds for the Central Universities (Grant No. 2011YJS215).

† Corresponding author. E-mail: liuyun@bjtu.edu.cn