文章编号:1000-5862(2013)04-0376-06

新浪微博名人堂用户关系网络分析

李永成,黄曙光,杨　斌,郭　浩
(解放军电子工程学院网络系,安徽 合肥　230037)

摘要:以新浪微博名人堂用户所形成的用户关系网络为研究对象,利用复杂网络的分析方法对该网络的度分布、小世界现象、度相关性等多个方面进行了分析.研究结果表明:该网络在度分布上存在背离幂律分布的现象,网络有效直径较短,各节点度之间不存在明显的相关性、网络不具有层次性等特点.

关键词:网络结构分析;名人堂网络;度相关性
中图分类号:TP 391　　**文献标志码**:A

0　引言

近年来,随着复杂网络分析方法的发展和应用,针对 Facebook,Twitter 等大型在线社交网络(Oline Social Networks,OSNs)的网络结构、网络特性等相关分析正逐渐成为众多学科领域的研究热点.中文微博自 2009 年以来迅猛发展,新浪与腾讯微博的注册用户均已超过 2 亿.但是,针对中文微博网络结构的研究还较少,尤其对基于关注行为的用户关系网络仍缺乏代表性研究.

人们研究 OSNs 的目的之一是获得真实人际关系网在 OSNs 中的投影,以便有效地研究大规模社会网络的形成机理及拓扑特性等问题.但原始的微博用户中存在大量的非活跃用户(注册后再未登录)、spam 用户(程序自动注册)等干扰因素,这类用户破坏了原始微博关系网的真实性,受限于当前的 spam 用户检测方法[1],很难从原始的微博用户网络中快速且准确地抽取出一个真实用户的关系网.因此,本文以"新浪名人堂"用户之间相互关注形成的网络为研究目标,该网络可以被看作是新浪微博真实用户网络的精简版.简便起见,下文将新浪名人堂的用户关注关系网络简称为"名人堂网络".

"新浪名人堂"是新浪微博的一个模块.名人堂用户均为已被新浪官方认证,且在各领域有一定的影响力的用户,这些用户在促进信息传播,引导意见走向等方面都具有重要的作用.因此,研究名人堂网络,对基于传播的口碑营销、信息监测等进一步研究具有重要的意义.

1　数据集及基本统计特征

新浪微博的开放 API 并不提供获取名人堂用户列表的接口,因此,首先使用结合 Cookie 的 Ajax 爬虫获取名人堂用户 ID,然后使用 API 接口获取用户的关注列表.在构建名人堂网络时,将指向名人堂用户集合外部的关注边删除.以有向图 $G(V,E)$ 代表名人堂网络,则节点 $v_i \in V$ 代表一个名人堂用户,如果用户 i 关注了用户 j,则 G 中存在一条由 v_i 指向 v_j 的有向边 $e_{ij} \in E$,同时用户 i 称为用户 j 的一个追随者(follower)或入度邻居(in-degree neighbour),用户 j 称为用户 i 的关注对象(followee)或出度邻居(out-degree neighbour).节点 i 的所有的出度邻居构成其出度邻居集,所有的入度邻居构成入度邻居集.本文的数据为 2012 年 2 月 17 日的名人堂用户及他们关注关系的快照.

如表 1 所示,对比名人堂网络与其它几个社交网络(Facebook 与 RenRen 网络是基于相互朋友关系的无向图)可以发现,尽管名人堂网络中的一些关注边被删除,但是名人堂网络仍表现出较高的平均度及相应的高网络密度,其原因一方面由于微博中关注连接建立的代价很低(不需要双方同意);另一方面说明名人堂用户的活跃性高,相互之间交流紧密.

收稿日期:2012-11-15
基金项目:国家自然科学基金(61202337)资助项目.
作者简介:李永成(1986-),男,山东淮坊人,工程师,博士,主要从事社会网络分析和数据挖掘的研究.

表 1　几种在线社交网络的基本信息

	名人堂	Facebook [2]	Renren [3]	Twitter [4]	Twitter [5]
网络类型	Directed	Undirected	Undirected	Directed	Directed
节点数	137 561	0.721B	42.115M	87 897	41.7M
边数	29.082M	68.7B	1 657.273M	829 247	1.47B
平均度	422	190	78.7	18.86	70.5
密度	1.537E-3	2.641E-7	1.869E-6	1.070E-4	8.450E-7

2　网络拓扑特征

2.1　度分布

对实际网络的研究中发现了节点度的幂律分布现象[6-7]. 图 1 描述了名人堂网络的入度(a ~ b)及出度(c ~ d)的分布情况,图 1 中可以发现网络节点入度分布的大部分符合幂律分布,通过线性拟合互补累计分布函数(Complementary Cumulative Distribution Function,CCDF)曲线,获得其入度概率分布的幂指数为 -2.56,这与实际网络幂律指数多在 2 ~ 3 之间的结论相符[8];但是,名人堂网络节点的出度分布具有明显的背离幂律分布现象,其概率密度函数(Probability Distribution Function,pdf)曲线仅随出度值的增加呈单调下降的趋势,类似的现象在 Twitter 和 Facebook[9] 的研究中同样出现. 同时,由于新浪微博将可关注的用户数量限制为 2 000,所以图 1(d)在达到 2 000 左右时发生急剧的下降.

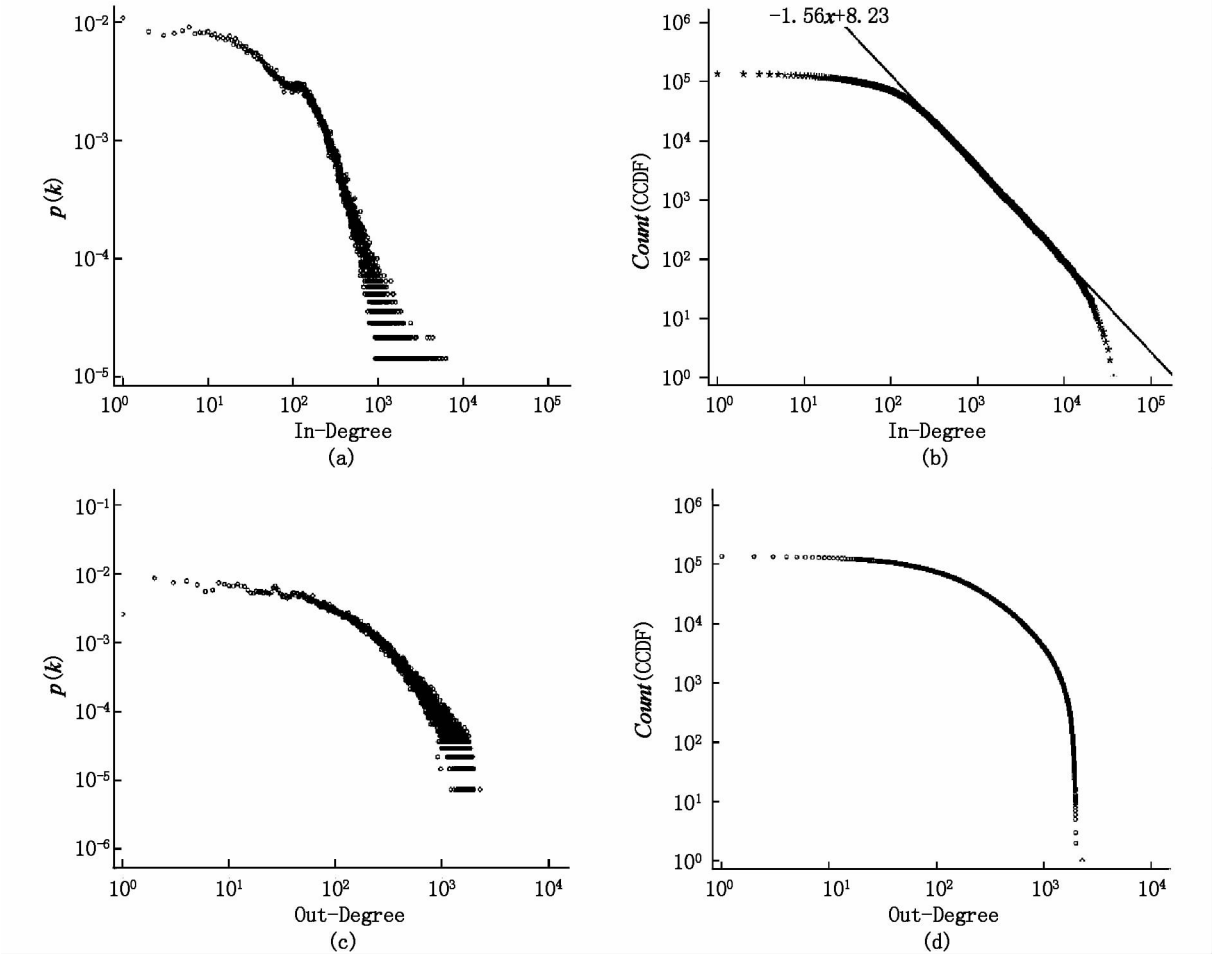

图 1　a,b 为名人堂节点入度值的 pdf 曲线及 CCDF 曲线; c,d 为名人堂节点出度值的 pdf 曲线及 CCDF 曲线

2.2　连通分量

连通分量用来反应图中的节点是否存在连接路径,由于有向图中节点间的路径存在方向,因此又可以分为强连通分量(Strongly Connected Component,SCC)和弱连通分量(Weakly Connected Component,WCC). 一个 SCC 指的是有向图 G 中的一个子图 G',且 G 中的任意节点 i 和 j 之间满足存在路径 $i \to j$ 和路

径 $j \to i$. 而 WCC 可以仅存在单向的连通. 分析发现整个名人堂网络出人意料的构成了一个完整的 WCC,这意味着所有节点都连接到了一起,且名人堂网络中存在一个占总节点数 98.58% 的 SCC,并有 1 897 个节点以单向边连接到这个 SCC 中.

2.3 互惠性

"互粉"代表了微博中的相互关注行为. 可以使用有向图中的互惠性(Reciprocity)来描述用户间这种互粉关系,网络的互惠率可以用"存在关注关系的用户对中具有双向关注关系的用户对所占的比例"来表示. 而本文通过计算名人堂网络的互惠率,得出 26.45% 的关注用户对之间存在了双向关注行为. 该数值接近于 Twitter 用户关系网 22.1% 的互惠率,远低于 Flickr 的 68% 和 Yahoo! 360 的 84%. 而具有相同节点数和相同密度的随机网络互惠率仅为 0.076 8%.

2.4 小世界特性

网络中连接 2 个节点的最短路径被称为节点间的距离,网络的平均路径长度定义为任意 2 个节点之间距离的平均值. 很多真实网络的研究证实 OSNs 呈现小世界特性,这反映为虽然网络规模很大,但网络的平均路径长度很短. 图 2(a)为名人堂网络节点间距离的分布图. 其中,出现次数最多的节点间距离为 3,网络的平均路径长度为 2.99. 这说明名人堂网络同样呈现出小世界特性. 网络的直径(Diameter)定义为网络中任意 2 个节点间距离的最大值,图 2(a)显示名人堂网络的直径为 8. 但是,为了反应绝大多数节点对之间的距离,一般使用有效直径[10]的概念来衡量 90% 的节点对所获得的网络直径. 名人堂网络的 Hop-Plot 图(图 2(b))反映了跳数与可达节点对的分布,其中虚线的位置是名人堂网络的有效直径为 3.784.

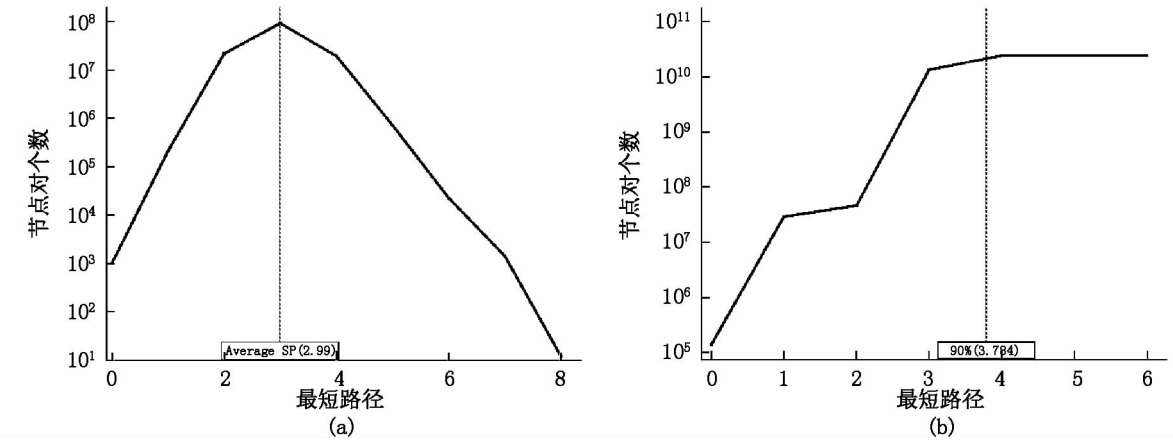

图 2 a 为名人堂网络节点距离的分布;b 为名人堂网络的 Hop-Plot 图

表 2 对比了名人堂网络与其它几个 OSNs 的平均路径长度和有效直径. 与网络规模无关,名人堂网络具有更短的平均最短路径和有效直径. 这说明名人堂网络用户之间连接更紧密,其原因是由于名人堂网络节点的平均度更高,网络表现出更高的密度.

表 2 几种在线社交网络中的最短路径

网络名称	节点数	平均路径长度	有效直径
名人堂	137 561	2.99	3.784
Facebook[9]	0.721B	4.7	<5
Renren[3]	42,115	5.38	–
Twitter[5]	41.7M	4.12	4.8
Cyworld[11]	12 048	3.2	–

3 基于节点度的相关性分析

在网络中心理论中,基于节点度数的度中心性可以直接作为衡量节点中心性的指标,而网络的中心节点,在网络的传播研究[12],网络结构健壮性研究[13]等方面都具有重要的意义. 因此本节主要分析不同度值的节点与多个特征的相关性.

3.1 度与聚类系数

聚类系数(Clustering Coefficient,CC)可以用来定量的衡量用户任意两朋友同样也是朋友的概率,它反映了被测量节点 i 的邻居网络(由节点 i 和与其直接相连的邻居组成的网络)的紧密程度. 图 3 是名人堂网络中节点度数与节点的平均聚类系数在双对数坐标系下的分布情况. 由图 3 可以看出,随着节点度数的增加,其聚类系数整体出现单调的下降趋势,该结论也与相关的研究相一致. 从图 3 中的拟合曲线可以看出,当节点度数较高时,聚类系数的下降趋势变得更加迅速,这说明,高度数的节点,用户

的追随者与关注对象的分布呈现出更大的任意性. 从整个网络来看,名人堂网络的平均聚类系数为 0.157,明显的高于 Twitter($CC=0.106$) 和 RenRen ($CC=0.063$) 的值,与 Facebook($CC=0.164$) 接近.

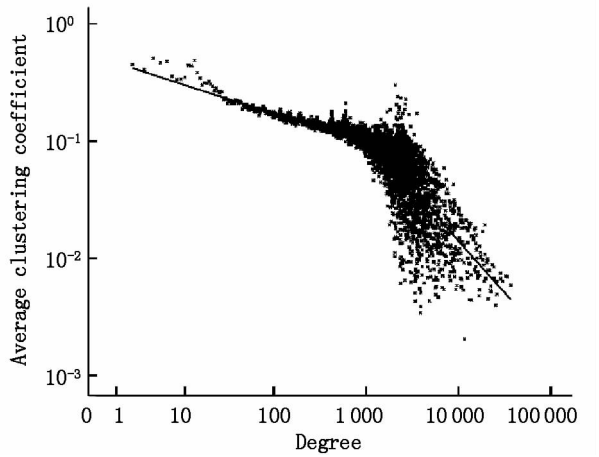

图 3　聚类系数和度数在双对数坐标下的分布

3.2　同配性分析

度相关指的是从网络中随机选择一条边,考察该边两端的节点的度数是否具有随机性,如果不是随机的,则网络具有度相关性,反之则无度相关性. 更进一步,如果网络是正相关的,称网络是同配(Assortative) 的,如果是负相关的,则网络是异配(Disassortative) 的. 同配网络的典型表现是“富人俱乐部”现象,即具有高度数的节点倾向于相互连接,彼此聚集形成所谓的“富人俱乐部”.

3.2.1　同配系数　所谓同配系数,是通过计算归一化的 Pearson 相关系数来刻画网络的同、易配情况[14],同配系数 r 的取值范围为 $r\in[-1,1]$. 若 $r>0$,则网路是同配的; 若 $r<0$,则网络是易配的. 有向网络的同配性问题可分为 4 种: (In,In) ,(In,Out) ,(Out,In) ,(Out,Out) [15],其计算公式为

$$r(\alpha,\beta)=\frac{E^{-1}\sum_i\left[j_i^{\alpha}-\overline{j^{\alpha}}\right]\left(k_i^{\beta}-\overline{k^{\beta}}\right)}{\sigma^{\alpha}\sigma^{\beta}},\quad(1)$$

其中 $(\alpha,\beta)\in[in,out]$ 为出度或入度类型,j_i^{α} 和 k_i^{β} 分别为有向边 i 的源节点和目标节点的 α 度和 β 度. $\overline{j^{\alpha}}=E^{-1}\sum_i j_i^{\alpha}$,$\sigma^{\alpha}=\sqrt{E^{-1}\sum_i\left(j_i^{\alpha}-j^{\alpha}\right)^2}$,$\overline{k^{\beta}}$ 及 σ^{β} 的计算类似于 $\overline{j^{\alpha}}$ 和 σ^{α},表 3 为名人堂网络 4 类同配系数的结果,由于(In,In) 型同配系数为负,因此名人堂网络没有出现“富人俱乐部”现象(本文将“富人”定义为具有较多的入度值,微博中具有较多的追随者) ,同时,因为(Out,In) 同配系数也为负相关,所以名人堂网络中并没有出现高出度节点倾向关注高入度节点的现象. 且四类同配系数的绝对这均小于 0.1,说明整个网络相关性较低,所以名人堂不呈现明显的层次结构,节点之间倾向于随机相连.

表 3　名人堂网络同配系数

同配类型	同配系数
(In,In)	-0.026 0
(In,Out)	0.083 6
(Out,In)	-0.096 0
(Out,Out)	0.056 7

3.2.2　不同邻居的入度同配性分析　为了进一步分析节点与其出(入) 邻居的连接情况. 图 4(a) 为节点的入度值 k 和其出度邻居的平均入度 $\langle k_{nn}^{Out}\rangle$ 值分布情况,即名人堂中用户的入度与其所关注对象的入度值. 图 4(a) 中的直线是 $f(k)=k$,

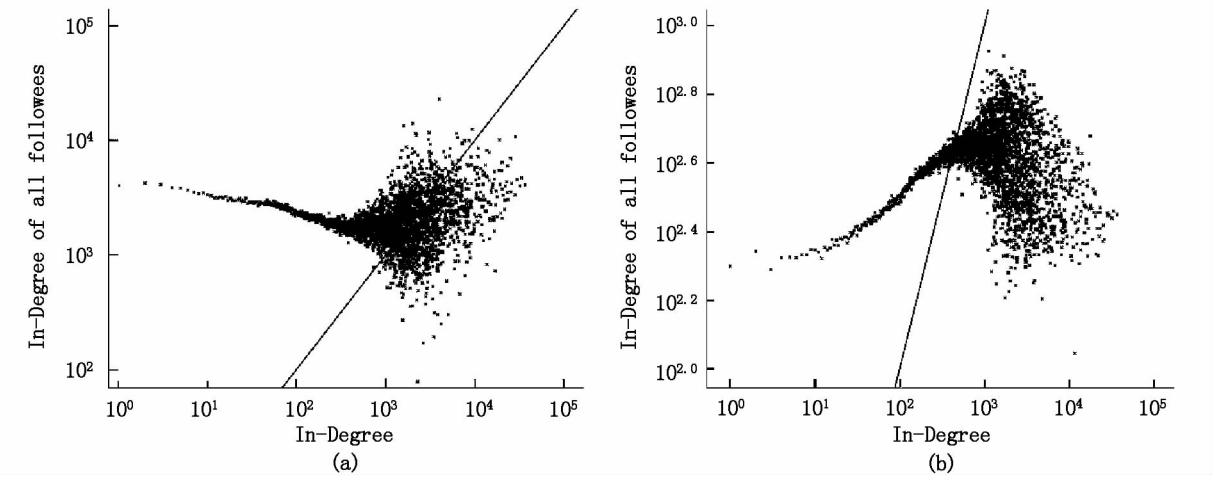

图 4　a 为出度邻居余平均(入) 度分布; b 为入度邻居余平均(入) 度分布

可以发现大多数节点倾向于关注比自己更有名(入度更高)的节点,当节点入度数较小时(约 < 1 000),节点随着自己入度值增大,其关注对象的入度值呈下降趋势,而当节点入度值较大时,其关注对象的入度值与节点的入度值无关. 图 4(b) 为节点的入度值和其入度邻居的平均入度值 $\langle k_{nn}^{ln} \rangle$ 分布情况,即节点入度与其的追随者的入度值. 图 4(b) 中的直线是 $f(k) = k$,可以发现大多数节点的追随者不如自身有名(更高入度). 当节点入度数较小时(约 < 1 000),节点随着自身入度值增大,其追随者的入度值倾向于上升,而当节点入度数较大时,其追随者的入度值呈现出与节点的入度值无关.

3.3 2 级影响范围

研究认为,节点的影响力不但取决于节点直接邻居的数量,还取决于由其邻居带来的间接邻居(Friends of Friends)的数量. 将从节点出发,使用广度优先的搜索策略在路径 $n(n \geqslant 2)$ 下的可达节点的数量称为节点的 n 级影响范围. 在网络的中心性理论中,Katz 中心性[16] 和 Semi-Local 中心性[17] 等本质上都是基于这种 n 级影响范围来衡量节点中心性的.

前述对名人堂网络直径的分析中知道,该网络有效直径小于 4 且网络中绝大多数节点连接成一个大的 SCC,这意味着如果路径 n 较大,几乎所有节点的 n 级影响范围均接近整个网络的节点数目. 因此,本节仅考虑节点的 2 级影响范围,即 $n = 2$,而微博转发传播中发现的 2 级传播现象[18],也进一步验证了这种 2 级影响范围的重要性.

因为微博信息传播是在有向网络中沿着边指向的反方向进行传播的,因此,名人堂网络中 2 级影响范围应当来自节点入度邻居带来的入度邻居(Follower of Followers,FoFs),而 FoFs 的数目计算可以包含 2 种:不考虑重复用户,称为不唯一的 FoFs(No-Unique FoFs) 和去除重复用户的唯一的 FoFs(Unique FoFs). 作为进一步的比较,一种简单的方式是假设节点的入度邻居具有与自身一样多得入度邻居,则节点的期望 2 级影响范围应为 $f(k) = k^2$,k 为节点的入度值.

图 5 为节点的入度值与具有该入度值的 3 种平均 2 级影响范围的分布情况,受限于网络的总节点数,实验仅使用节点入度值 $k \in [0,370]$ ($370^2 = 136\ 900$) 的范围. 3 种 2 级影响范围可以在 $k > 0$ 时拟合出 3 条曲线. 由图 5 可以发现,No-Unique FoFs 曲线的值始终大于 Except FoFs,这说明名人堂网络中,入度值为 $k \in [0,370]$ 的节点,其追随者的平均入度邻居要比自己的入度邻居多,这一点可以从图 4(b) 中得出同样的结论. 而社会学中一个类似的结论是你朋友的朋友比你多[19-20],这在近期对 Facebook 的研究中也得到类似结果.

另一方面 Unique FoFs 的拟合曲线为 2 次项系数小于 0 的 2 次曲线(该曲线的顶点为 $k \approx 478$),这意味随着节点入度数的增加,所带来不重复的间接入度邻居逐渐增加,但其增幅逐渐减少.

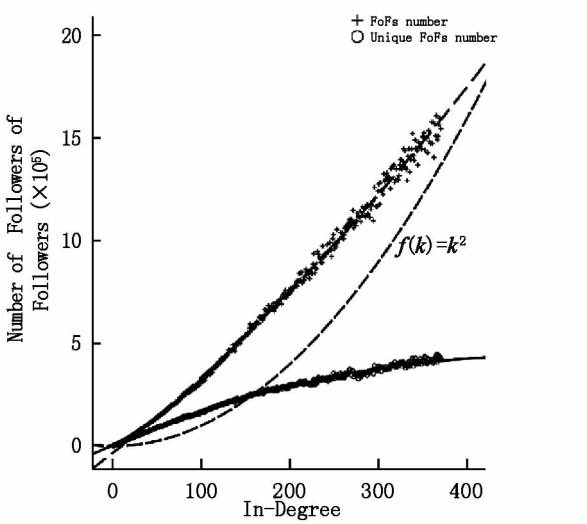

图 5 名人堂网络 2 级影响范围

4 总结及进一步工作

本文通过分析名人堂网络的结构,并发现了以下几个结论:

(i) 名人堂网络的节点出度分布上存在背离幂律分布的现象,入度分布则整体符合幂律分布.

(ii) 名人堂网络的节点平均度数较高,进而表现为网络密度更大、网络直径更短等特点.

(iii) 名人堂网络的同配系数接近 0,各节点的连接无明显的度相关性,同时,名人堂中拥有较低入度的用户所拥有的追随者的平均入度要高于自己,且他们倾向于关注比自己入度值更高的节点.

本文对名人堂网络的度分布仅进行了初步的描述,进一步地研究应包括对其度分布特别是出度分布的拟合,以及对名人堂网络社区结构的检测.

5 参考文献

[1] Benevenuto F, Magno G, Rodrigues T, et al. Detecting spammers on twitter [C]. Washington: IEEE Press, 2010.

[2] BackstromL, Boldi P, Rosa M, et al. Four degrees of separation [J/OL]. [2012-10-16]. http://arxiv.org/abs/

1111.4570.

[3] Jiang Jing,Wilson C,Wang Xiao,et al. Understanding latent interactions in online social networks [EB/OL]. [2012-10-17]. http: // conferences. sigcomm. org/imc/2010/papers/p369. pdf.

[4] Java A,Song Xiaodan,Finin T,et al. Why we twitter: understanding microblogging usage and communities [EB/OL]. [2012-11-12]. http: // www. citeulike. org/user/whym/article/2580227.

[5] Kwak H,Lee C,Park H,et al. What is Twitter,a social network or a news media? [EB/OL]. [2012-11-10]. http: // wenku. baidu. com/view/faeeef3f0912a2161479298e. html.

[6] Barabási A L,Albert R. Emergence of scaling in random networks [J]. Science,1999,286(5439): 509-512.

[7] Clauset A,Shalizi C R,Newman M E J. Power-law distributions in empirical data [J]. Siam Review,2009,51(4): 661-703.

[8] Newman M E J. Power laws,Pareto distributions and Zipf´s law [J]. Contemporary Physics,2005,46(5): 323-351.

[9] Ugander J,Karrer B,Backstrom L,et al. The anatomy of the facebook social graph [EB/OL]. [2012-10-17]. http: // www. citeulike. org/user/jamesgleeson/article/10056037.

[10] Tauro S L,Palmer C,Siganos G,et al. A simple conceptual model for the Internet topology [EB/OL]. [2012-10-17]. http: // www. cs. ucr. edu/ ~ michalis/PAPERS/jellyfish-GI. pdf.

[11] Chun H,Kwak H,Eom Y H,et al. Comparison of online social relations in volume vs interaction: a case study of cyworld [EB/OL]. [2012-10-17]. http: // dl. acm. org/citation. cfm? id = 1452528.

[12] Ghosh R,Lerman K. Predicting influential users in online social networks [EB/OL]. [2012-10-17]. http: // arxiv. org/abs/1005.4882.

[13] Nicosia V,Criado R,Romance M,et al. Controlling centrality in complex networks [EB/OL]. [2012-10-17]. http: // arxiv. org/abs/1109.4521.

[14] Newman M E J. Assortative Mixing in Networks. Physical Review [J]. Letters,2002,89(20): 208701.

[15] Foster J G,Foster D V,Grassberger P,et al. Edge direction and the structure of networks [J]. The National Academy of Sciences,2010,107(24): 10815.

[16] Aggarwal C C. Social network data analytics [M]. New York: Springer-Verlag Inc,2011: 177-209.

[17] Chen,D.,L. Lü,et al. Identifying influential nodes in complex networks [J]. Physica A: Statistical Mechanics and Its Applications,2012,391(4): 1777-1787.

[18] 张辉. 浅谈新媒介环境下微博二级传播优势 [J]. 大观周刊,2011,521(13): 109-110.

[19] 王博,彭玉涛,罗超. 基于模糊聚类广义回归补缀网络的网络入侵研究 [J]. 江西师范大学学报: 自然科学版,2012,36(3): 288-291.

[20] Feld S L. Why your friends have more friends than you do [M]. Chicago: The University of Chicago Press,1991,1464-1477.

A Research on Famous-User Network of Sina-Weibo

LI Yong-cheng,HUANG Shu-guang,YANG Bin,GUO Hao

(Department of Network,Electronic Engineering Institute,Hefei Anhui 230037,China)

Abstract: Aiming towards the user relation network oriented from the Famous-User of Sina-Weibo,many facets including the degree distribution,small-world phenomenon and degree correlation utilizing the complex network analysis algorithms has been studied. With experimental results showing for this network the existence of deviations from power-laws at the degree distribution level,the effective diameter being short,the nodes´ degrees being irrelevant,with the non-hierarchical intrinsic nature proved.

Key words: network structure analysis; famous-user network; degree correlation

(责任编辑: 冉小晓)