在线社交网络结构与区域经济关联性研究

任晓龙[1,2]，朱燕燕[1]，王思云[3]，廖 好[2]，韩筱璞[1]，吕琳媛[1]

(1. 杭州师范大学阿里巴巴复杂科学研究中心 杭州 311121；2. 弗里堡大学物理系 瑞士 弗里堡 CH-1700；3. 成都理工大学商学院 成都 610059)

【摘要】通过腾讯QQ数据集详细分析了中国在线社交网络中用户的统计特征，交友、聊天等行为规律和地理分布规律。同时，对比《中国城市统计年鉴》等权威统计资料，发现在线社交网络的用户和用户行为数据，例如各地区的用户数、聊天数与该地区发展指标呈正相关，这表示在线社交网络数据能够在一定程度上反映各地经济、交通、通信等城市建设状况，相关结果对于区域经济研究具有启发意义。

关 键 词 行为特性； 经济预测； 地理分布； 网络结构； 社交网络

中图分类号 TP399 **文献标志码** A doi:10.3969/j.issn.1001-0548.2015.05.001

Online Social Network Analysis and the Relation with Regional Economic Development

REN Xiao-long[1,2], ZHU Yan-yan[1], WANG Si-yun[3], LIAO Hao[2], HAN Xiao-pu[1], and LÜ Lin-yuan[1]

(1. Alibaba Research Center for Complexity Sciences, Hangzhou Normal University Hangzhou 311121;

2. Physics Department, University of Fribourg Fribourg Switzerland CH-1700;

3. Business School, Chengdu University of Technology Chengdu 610059)

Abstract In this paper, we investigate a sample of Tencent QQ online social network and analyze in detail the users' basic characteristics, users' behavior patterns of making friends and chatting, and the users' geographical distribution. In addition, comparing with an authoritative data acquired from *China City Statistical Yearbook*, we find positive correlations between the regional development indicators and the characteristics of online social networks, such as the number of online users, the total chatting days of a city. This indicates that users' behaviors in online social networks could reflect the development of local economy, transport, telecommunications, etc. Our findings are instructive for regional economic research.

Key words behavior pattern; economic prediction; geography distribution; network structure; social networks

社交是人类最重要的行为之一，在人际交往与日常生活中时刻存在着。人们通过社交的方式传递信息、交流思想，以达到某种目的。这种人与人之间的社交关系就可以用社交网络进行刻画。社交网络是由社交活动参与者(如个人或组织)以及表征这些参与者之间相互关系的连边组成[1]。对人类行为的研究离不开对社交网络的分析。早年间，由于离线社交数据的收集成本极高，人类行为的研究常常囿于数据缺乏，因而只能在小规模范围进行。随着互联网的快速发展，可访问互联网的智能设备的增多，以及无线网络等相关基础设施的大力建设，在线社交网络已经从方方面面融入了人们的生活[2-3]。在线社交网络能够方便、及时且全面的记录用户的各种信息与行为，这为人们深入研究提供丰富的数据资源，并使得针对大规模人类行为的研究成为可能[4-7]。

近几年，在线社交网络[8]在人们社交活动中扮演的角色也发生了质的变化：由一开始作为离线社交网络的补充，逐渐演变为与离线社交并重。越来越多离线社交的活动内容被在线社交替代，用户体验的提升使人们通过网络交流与互动越来越得心应手[9]，在线社交逐渐可以给人们带来与离线社交完全相同的情感体验[10]。一方面，离线社交的时间成本、经济成本等一直居高不下；另一方面，在线社交的门槛越来越低，互动形式多样化，趣味性与实效性大大增加。可以预见，未来，在线社交将在很

收稿日期：2014 – 12 – 25；修回日期：2015 – 04 – 22

基金项目：国家自然科学基金(11205042,11205040)；2014年浙江省大学生科技创新活动计划(新苗人才计划)项目(2014R421062)

作者简介：任晓龙(1988 –)，男，硕士生，主要从事网络科学与工程、信息挖掘等方面的研究.

大程度上从内容上、形式上超越离线社交，在社交活动的很多方面占据主导地位，只是当在线社交无法完全满足用户需求的时候，人们才进行少量的、必要的离线社交行为[11]。

在线社交网络的兴起也引起学术界的关注，催生了一系列重要的研究成果。2003年，文献[12]分析了斯坦福大学早期的在线社交网络，印证了在线社交网络上同样会表现出小世界现象[13]和显著的局部聚类现象；文献[14]发现了朋友关系和地理位置之间有着非常强的关联，文献[15]更细致地分析了随着用户年龄的增长以及地理位置的变化，亲密朋友关系变化的规律；文献[2]发现稍具规模的社交网络都含有一个包含大多数用户的强联通片；文献[16]发现在线社交网络往往会倾向于形成紧密的社团；文献[17]详细分析了用户和用户之间的关系是如何一步步形成社团的，发现一个人加入某个社团不仅仅取决于他有多少个朋友在这个社团中，还取决于这些朋友的连接方式如何；文献[18]统计分析了中国最大的在线社交网络腾讯QQ中的群结构；文献[19]发现在含时网络中，社交网络中只有大约30%的链接在以月为单位的情况下持续互动，并且虽然两个朋友之间的互动情况随时间变化非常大，但是整个网络的基本性质不易改变；文献[20]研究了社交网络中的谣言传播问题；文献[21-22]分别讨论了无向和有向社交网络中的链路预测问题；文献[23-26]对社交网络上的用户影响力问题进行了深入的探讨。这些研究成果吸引了包括社会学、网络科学、统计学及图论等领域的学者的广泛关注。特别地，网络科学在社交网络分析中的应用进一步促进了该领域的快速发展。目前，已有研究成果只是冰山一角，其内涵和外延有很大的拓展空间，还有更多有趣的问题值得探讨。例如本文后面将讨论的在线社交网络和区域经济发展水平的关系。

中国在线社交网络发展迅速，其中腾讯QQ(简称“QQ”)是中国最受欢迎的在线社交软件。截止2014年10月1日，QQ历史最高同时在线人数为216 503 678人，是中国最具有代表意义的在线社交网络。本文分析了QQ在线社交网络数据的一个子集，包括了111万用户以及他们的年龄、性别、最常登录地点、聊天天数等信息，详细分析QQ在线社交网络的基本结构与特性。同时，发现社交网络中用户的行为模式与各地区经济、交通、通信等发展状况密切相关。通过对比分析《中国城市统计年鉴》等权威公开的数据，发现社交网络的用户数、聊天天数与地区多种发展指标呈正相关，这些结果对于区域经济的研究具有启发意义。

1 QQ在线社交网络的数据描述

本次分析中的QQ数据集的采集方法是：1) 随机抽取一万个用户作为种子用户，要求这些用户比较活跃(即抽取前30天内至少登录一次的用户)，并且注册时间超过一年；2) 取所有与种子用户有朋友关系的节点加入数据集，并将他们与种子用户之间存在朋友关系的信息也加入数据集；3) 补充当前数据集之中所有节点之间存在、但尚没有加入数据集的连边；4) 抽取数据集中所有用户的年龄、性别、30天内最常登录地点、30天内登录天数、注册时间等用户的个人信息。

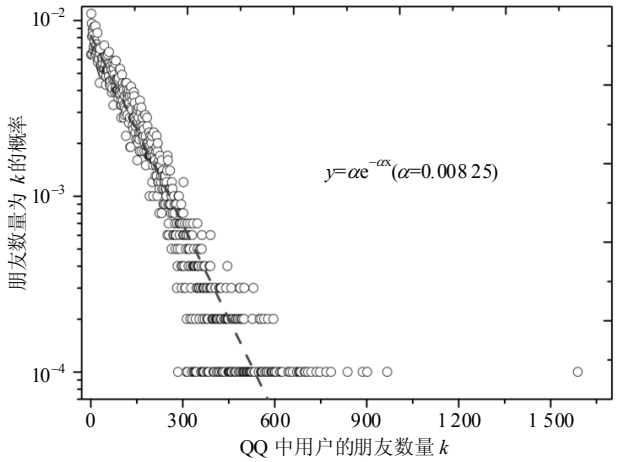


图1 用户朋友数量的概率分布

第一步抽取后，网络中有10 000个注册时间超过一年且比较活跃的用户，没有边；经过第二步抽取，网络中有1 113 435个用户，1 117 512条边；经过第三步的补充，网络中有1 113 435个用户和8 022 535条边。实际上，第二步抽取的1 117 512条边都是1万个种子用户和他直接相连的朋友的边，即这1万个种子用户的朋友数之和。由此可估算出QQ社交网络中注册超过一年且最近30天内至少登录一次的用户的平均朋友数约为111个。据统计，2006年11月流行的社交网络Orkut的用户平均朋友数量为106.1个[5]，2009年Facebook上用户的平均朋友数量约为120个[19]。对比说明，在平均朋友数量方面中国的在线社交网络与国际其他社交网络基本一致，平均朋友数量不因人口密度、地理位置等外界因素而产生较大差异。

在网络科学相关理论中，用户朋友数的多少是该用户影响力大小的体现[23]，拥有越多的朋友意味着一个用户的直接影响力越大，转发信息的能力越强。图1是QQ网络中1万个种子用户的朋友数量的概

率分布图。通过拟合可以看出，QQ社交网络中用户朋友数量的概率分布服从指数分布。

2　QQ在线社交网络的可视化

图2　QQ网络可视化(将网络外层节点剥离之后，取shell > 40的网络核心部分，用Gephi[27]进行画图)

网络可视化能够使人们对网络结构有一个直观的印象。K-shell网络分解法[28]可以将网络最外层的节点像“剥洋葱”一样一层层剥离，使网络只剩下最核心的部分，这种方法在分析大规模网络时经常用到，详细过程请参考综述[23]。QQ数据集中节点数过多，难以直接显示，通过分析我们发现，用K-shell分解法对QQ社交网络进行分解，最多可以将网络分成101层，其中最内层是一个由454个节点组成，相互之间连边最小为452的高密度连通片。剥去外面的40层外围节点后，网络的核心部分剩余8 308个节点，616 297条连边，其示意图如图2所示。从图中可以看出非常明显的群落结构，很多规模稍小的群落分布在一个大型群落周围，大型群落包含网络中大部分的节点和连边，不同的小型群落有不同的活跃程度。真实的社交关系也呈现出很明显的群落结构[16]，与分析结果一致。

3　QQ用户性别和年龄结构特征

着重分析了QQ社交网络中用户的性别和年龄特征。在整个数据集中，有年龄信息的有效用户总数为973 263人，占全体数据集中总用户数的87.4%. 有年龄信息的用户中，已知性别信息的用户占99.4%, 其中男女比例为54.1:45.9, 女性用户比例要高于由中国互联网络信息中心(CNNIC)于2014年1月发布的《第33次中国互联网络发展状况统计报告》(以下简称“报告”)中提到的2013年中国互联网用户男女比例56:44。如图3所示，在所有用户中，20～29岁年龄段的用户一共占所有用户的58.2%, 是QQ社交网络用户的主力军。

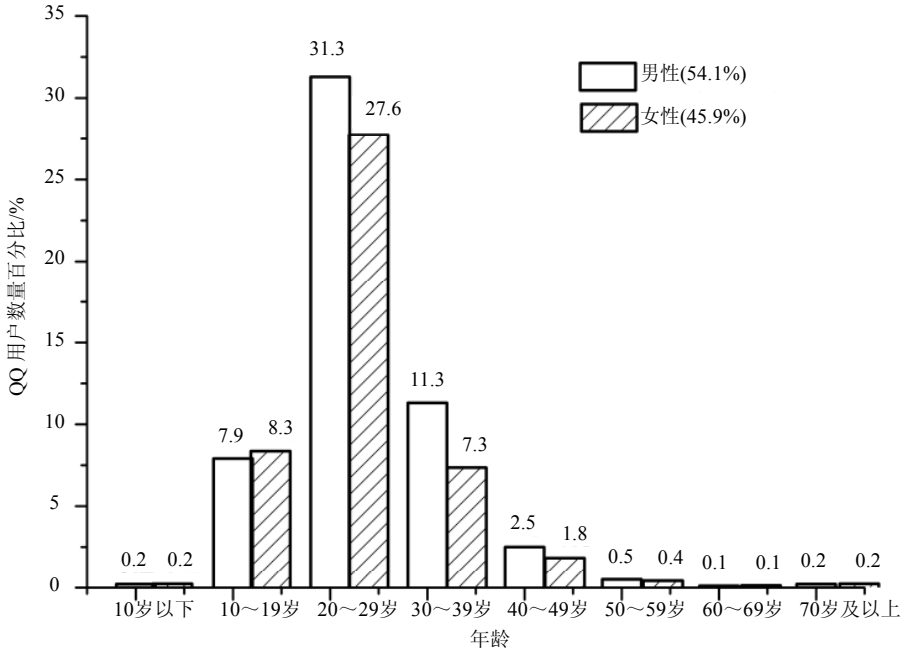
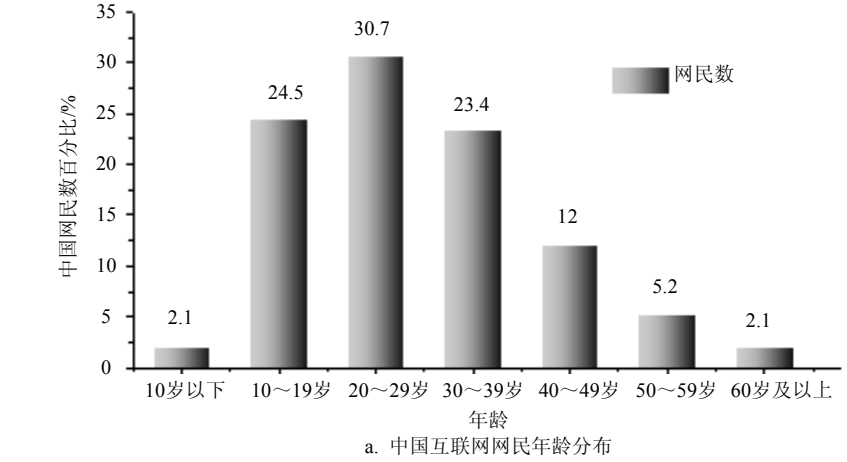


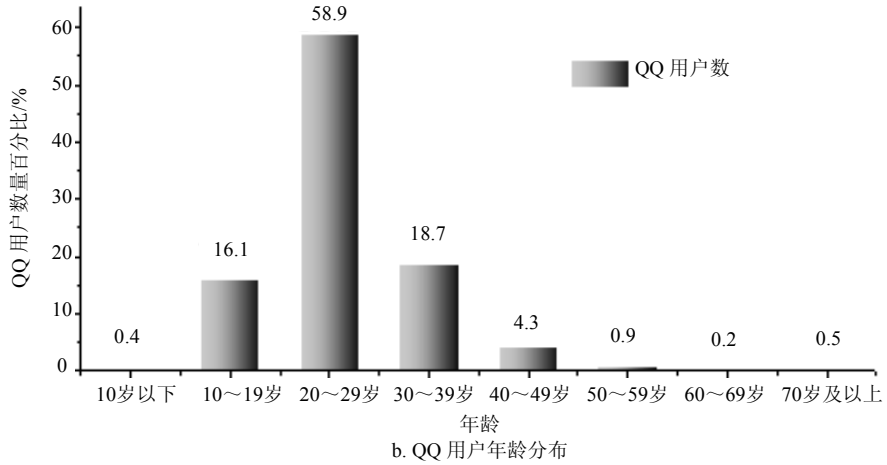
图3　QQ用户的性别年龄结构图

在QQ社交网络中，不同年龄阶段的用户数量差异非常大，如图4所示，20～29岁年龄段的用户数量占总量的58.9%，10～39岁年龄段的用户数量占总量的93.7% (图4的QQ用户中包含男性、女性和未知性别用户，图3只包含前两者)。与此形成鲜明对比的是，《报告》中相应的用户数量比例分别为30.7%和

78.6%. 这说明，青年用户更易掌握需要较为复杂操作的社交网络软件(如腾讯QQ，新浪微博等)，相对而言，仅成为互联网用户则技术门槛低很多，只需要掌握点击与浏览等基本技能。



a. 中国互联网网民年龄分布



b. QQ 用户年龄分布

图4 中国互联网网民分布与QQ用户分布的对比，前者明显较后者均匀

4 QQ在线社交网络用户交友、聊天信息的地理分布

由于数据缺乏，在以往的社交网络分析中很少涉及用户的地理位置信息。本次采集的QQ社交网络数据集中，含有用户最近30天内的登录地点，为深入分析中国社交网络用户与位置相关的行为规律提供了便利。为了对QQ社交网络的地理分布有一个直观印象，本文在中国地图上统计343个较大的城市中用户的分布图(由于数据信息缺失，北京、上海、天津、重庆四个直辖市各视为一个城市)。显然，图5中经济发达地区和人口聚集地区社交网络用户明显多于西北地区。直观上看，人数分布的分割线与中国经济/人口分界线十分相似。

社交网络中朋友之间的交友和互动情况是一个研究热点[29]。图6是中国31个省级行政区(本文中31个省级行政区的序号依次是：1～5：北京、天津、河北、山西、内蒙古；6～10：辽宁、吉林、黑龙江、上海、江苏；11～15：浙江、安徽、福建、江西、山东；16～20：河南、湖北、湖南、广东、广西；21～25：海南、四川、贵州、云南、西藏；26～31：陕西、甘肃、青海、宁夏、新疆、重庆。不包含港、澳、台地区)之间的交友和聊天信息，其中图6a是不同省份之间交友数量的关系。QQ社交网络中不同省份交友数量之间差异非常大，如坐标 $(x,y)=(11,11)$ 处即“浙江-浙江”之间交友数达到393 954对，而坐标 $(x,y)=(11,2)$ 处即“浙江-天津”之间交友数仅为2 732对，不足前者的1%。为了达到更好的显示效果，本文给图6a中所有数据取log10。从图中可以看出省内交友数量远远大于不同省份间交友数量，距离对于朋友关系的建立与保持影响巨大。有若干个地理、经济联系非常紧密的省份之间的交友数量明显多于其他省份，比如序号为1、2的北京、天津，序号为9、10、11的上海、江苏、浙江。序号为25、

28的西藏、青海则由于人口、地理原因与其他省份联系较弱，交友数量也明显比较少。

对QQ数据集中每个用户在30天内与每个朋友的聊天天数数据进行分析，本文将不同省之间的社交网络用户的总聊天天数累加，得到了图6b。同时，还计算出省份之间的平均聊天天数，如图6c所示。在图6b中，同样由于数据量级相差太大，给所有数据都取了lg。相比于图6a的朋友数量分布，朋友的聊天天数分布更能反映一个人的强关系[29]。对比两图可以发现，朋友之间的强关系的建立和保持更加受限于地理因素，大多数人的强关系都是分布在一个局域的范围内，这与文献[15]的结果相吻合。

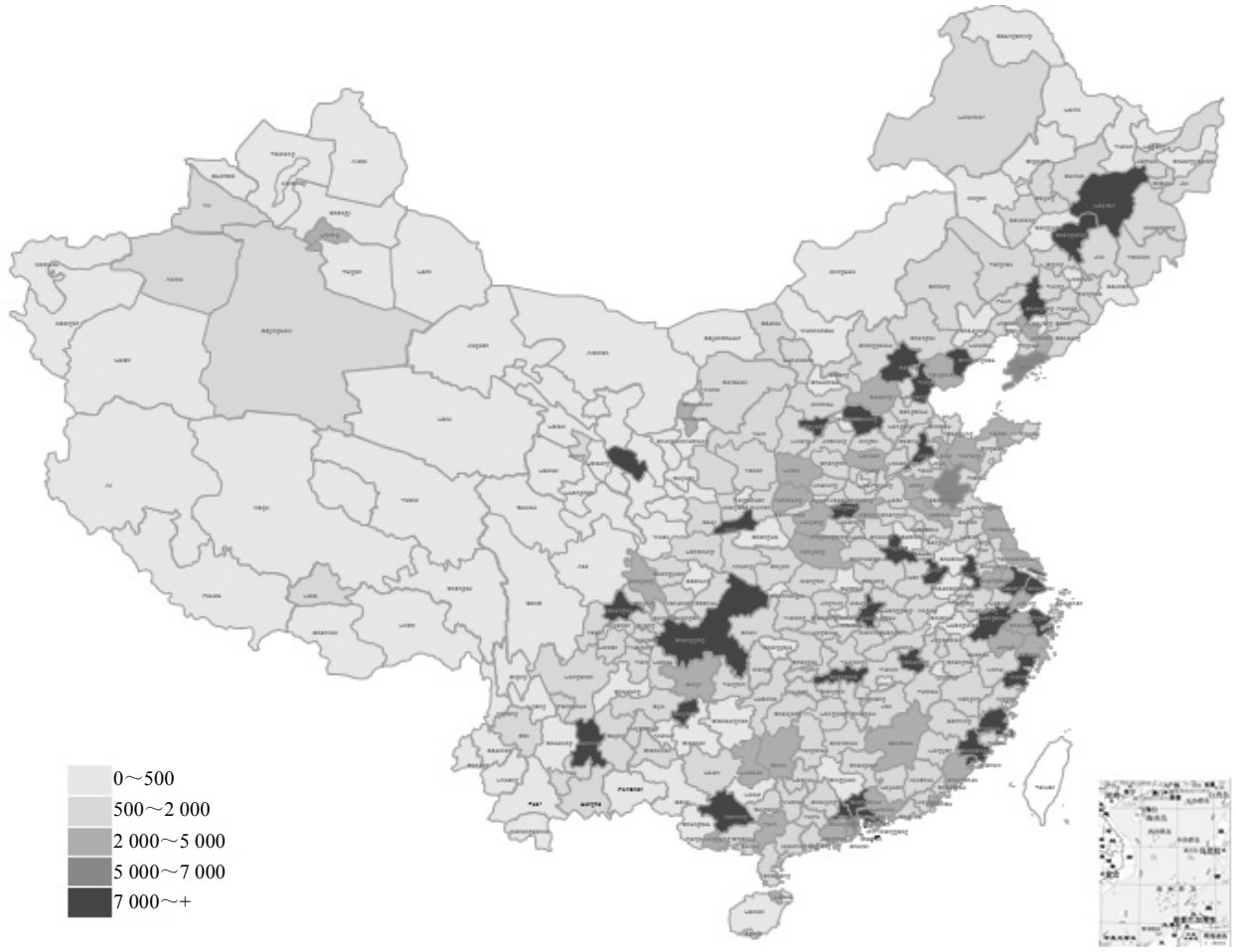
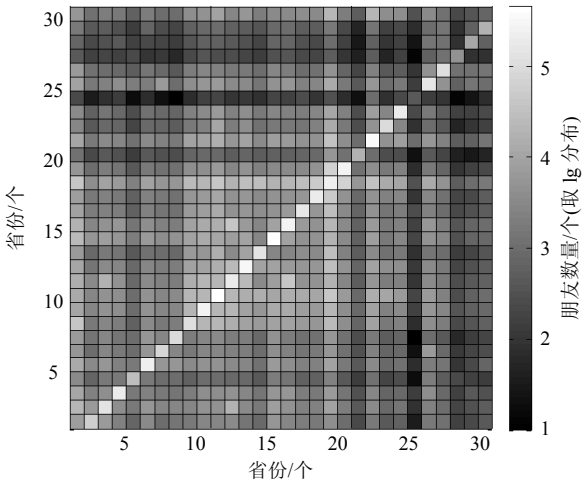
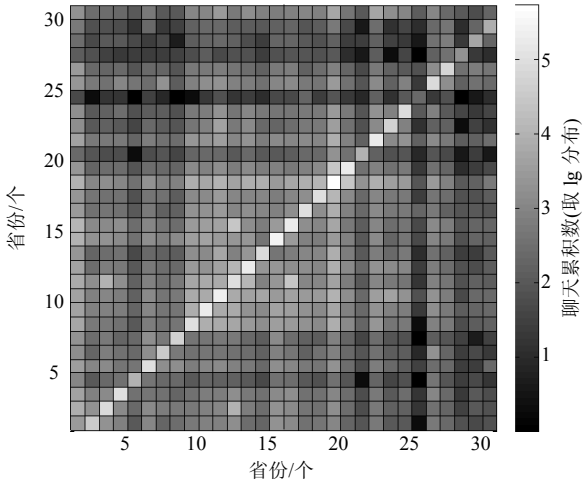


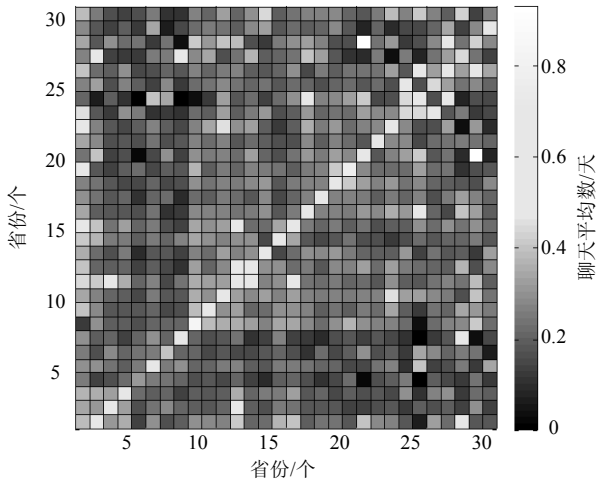
图5　QQ社交网络中各城市用户分布图



a. 不同省份之间朋友数量



b. 不同省份之间30天内朋友聊天累积天数



c.不同省份之间30天内朋友聊天平均天数

图6 QQ数据集中不同省份之间交友与聊天信息

5 在线社交网络结构与区域经济、交通、通信等基础设施发展状况的关系

中国是一个幅员辽阔的国家，对各省市进行经济、交通、人口等普查往往需要耗费大量物力、人力、财力。能不能从其他角度对地方的经济发展程度进行描述是一个非常有意义的课题。为此，本文分析了QQ社交网络中的用户信息和《2013年中国城市统计年鉴》中的数据，试图通过社交网络来推测区域人口、经济、交通、通信等的发展状况。图7~图10是每个城市的QQ用户数量和城市的人口、GDP、公共汽(电)车数量、移动电话用户数等能反映一个城市建设水平的指标之间的关系。在这些图中，QQ用户数与各个指标的皮尔森相关系数分别是0.77、0.84、0.80与0.87。可见，每个城市的QQ用户数与这些主要指标有着非常高的正相关性。不仅是QQ用户数，每个城市的累积QQ用户聊天天数也与这些指标有着比较高的相关性。表1详细统计了每个城市QQ用户数、城市的累积社交网络聊天天数与各类城市发展指标之间的相关关系。QQ用户数、城市累积QQ聊天天数都与固定电话、移动电话、互联网安装数等能体现城市发展状况的指标呈现出最强的相关性，皮尔森相关系数都在0.84以上。在交通运输方面，令人惊讶的是社交网络中的用户数量和聊天累积天数都与民航客运量最为相关，皮尔森相关系数分别为0.75和0.70，要高出与公路运输客运量和铁路运输客运量。本文从网络上爬取到不同省份之间民航客运的运输关系，如图11所示。可以看出，图11与图6社交网络中不同省份用户的互动频率非常吻合。

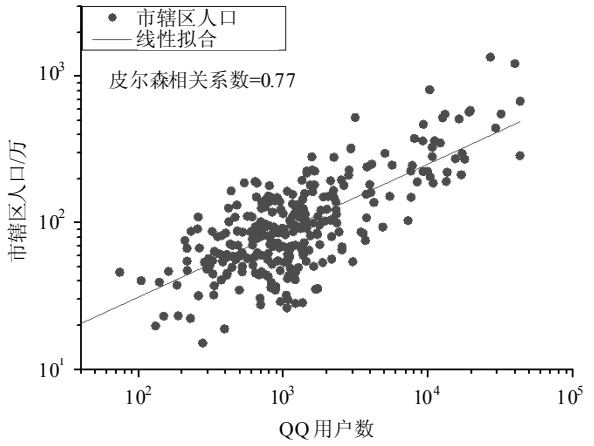
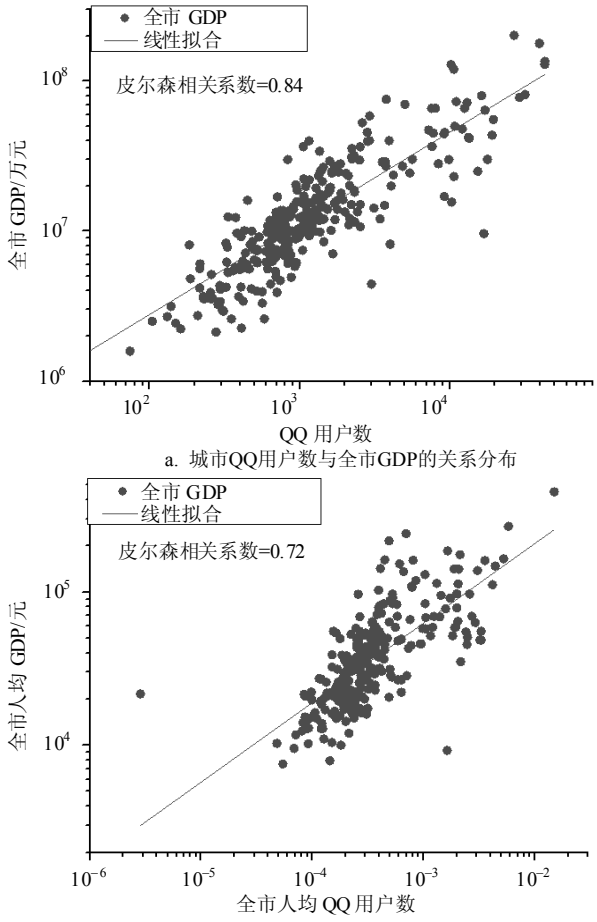


图7 城市QQ用户数与市辖区人口的关系分布

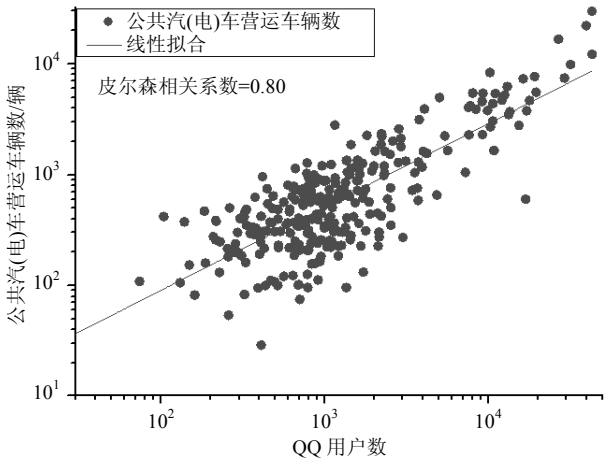
与人口普查、经济普查不同，在线社交网络上的数据获取成本更低，分析更加方便快捷。同时，在线社交网络越来越成为人们生活中不可缺少的一部分，其中隐含了人们在真实生活中的经济、交通、通信、教育等信息，并成为现实经济社会在网络空间的一种映射。从在线社交网络的角度对国民经济建设进行分析是一个新的途径的视角，本文仅给出了最初步的讨论，还有很多更加多样的分析形式、更加多元的分析方法有待进一步探讨。



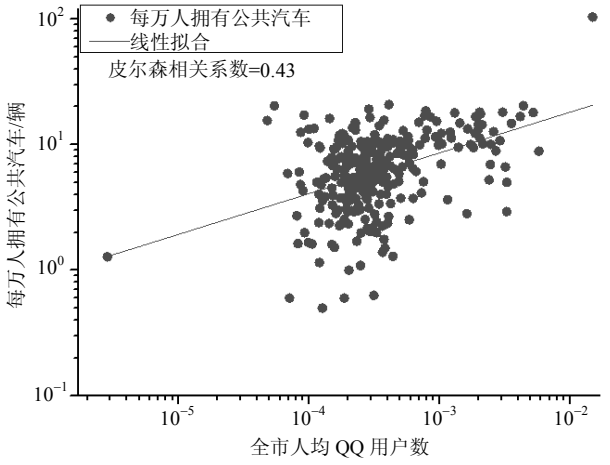
a. 城市QQ用户数与全市GDP的关系分布

b. 城市人均QQ用户数与人均GDP关系分布

图8 城市QQ用户数与城市GDP的关系分布

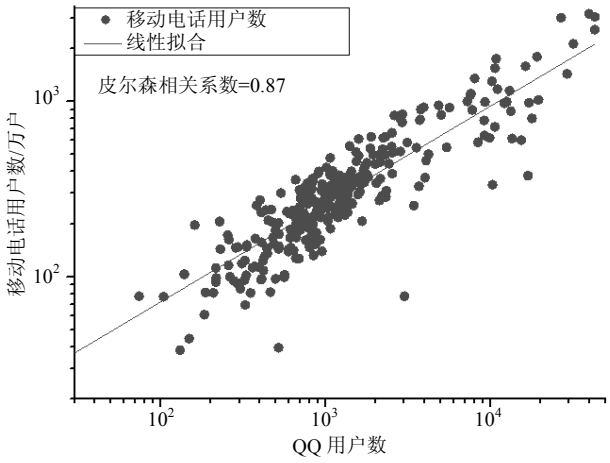


a. 城市QQ用户数与公共汽(电)车量的关系分布

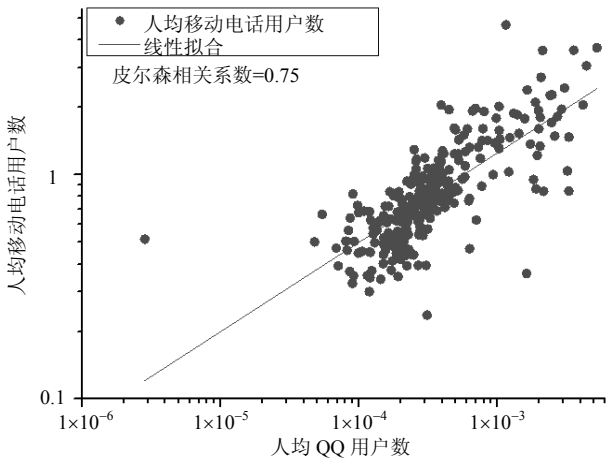


b. 城市人均QQ用户数与每万人公车关系分布

图9　城市QQ用户数与公共车辆的关系分布

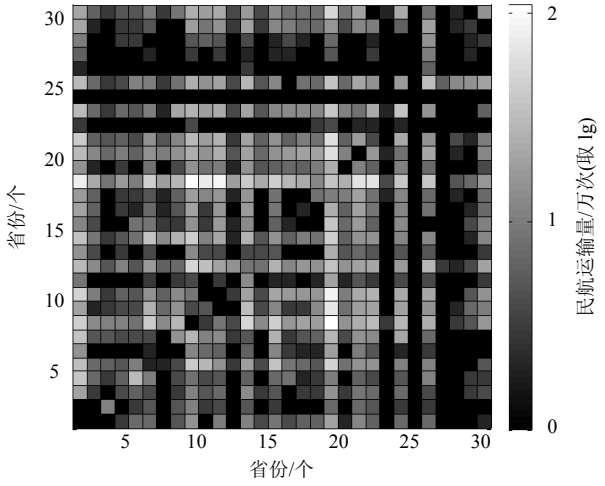


a. QQ用户数与移动电话用户数的关系分布

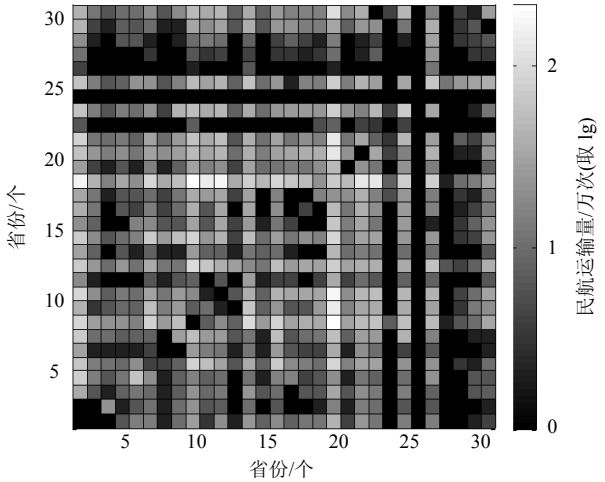


b. 人均QQ用户数与人均移动电话用户数的关系分布

图10　QQ用户数与城市各种发展指标的关系分布



a. 省份之间的民航运输量



b. 不考虑民航运输的方向时的运输量

图11　不同省份之间民航运输图

表1 QQ网络数据与众多城市发展指标的相关关系(用皮尔森相关系数衡量)

城市发展指标	与QQ用户数量的关系	与QQ聊天累积时间的关系
市辖区人口	0.77	0.70
客运总量	0.66	0.66
铁路旅客客运量	0.67	0.62
公路旅客客运量	0.61	0.62
民航客运量	0.75	0.70
城市道路面积	0.74	0.79
固定电话用户数	0.88	0.85
移动电话用户数	0.87	0.84
互联网宽带接入用户数	0.88	0.86

6 结论与展望

腾讯QQ社交网络是中国具有代表性的在线社交网络之一。本文详细分析了QQ在线社交网络的基本统计特征，用户的性别与年龄构成、交友与聊天信息的地理分布、不同省份之间交友与互动之间的地域差别，以及按不同行政单位划分社团结构时的模块度等。最后，通过比较在线社交网络结构与中国城市的发展与建设状况的指标，本文发现这两者存在较强的关联。这表明，人们有可能通过社交网络结构推断城市发展状况，并通过社交网络的演化趋势来预测区域经济的发展。

社交网络分析是一个长盛不衰的话题，从不同的角度在不同的数据集上经常能发现各种各样有意思的结论。希望在接下来的工作中，更加深入地分析在线社交网络上朋友之间的互动信息与不同省市之间的经贸来往之间的关系，并尝试用在线社交网络的结构变化预测未来经济发展状况。相信该方向的研究将给人们对社交网络的演化带来新的理解与思考。

本文的研究工作得到杭州师范大学科研启动经费(PE13002004039)及杭州师范大学阿里巴巴复杂科学研究中心开放基金(PD12001003002003)的资助，在此表示感谢。

参 考 文 献

[1] WASSERMAN S. Social network analysis: Methods and applications[M]. Cambridge: Cambridge University Press, 1994.

[2] KUMAR R, NOVAK J, TOMKINS A. Structure and evolution of online social networks[C]//Link mining: Models, algorithms, and applications. New York: Springer, 2010: 337-357.

[3] DOREIAN P, STOKMAN F. Evolution of social networks[M]. New York: Routledge, 2013.

[4] GARTON L, HAYTHORNTHWAITE C, WELLMAN B. Studying online social networks[J]. Journal of Computer-Mediated Communication, 1997(3): 0-31.

[5] MISLOVE A, MARCON M, GUMMADI K P, et al. Measurement and analysis of online social networks[C]// Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. New York: ACM, 2007: 29-42.

[6] 周涛, 韩筱璞, 闫小勇, 等. 人类行为时空特性的统计力学[J]. 电子科技大学学报, 2013, 42(4): 481-540.
ZHOU Tao, HAN Xiao-pu, YAN Xiao-yong, et al. Statistical mechanics on temporal and spatial activities of human[J]. Journal of University of Electronic Science and Technology of China, 2013, 42(4): 481-540.

[7] ZHAO Zhi-dan, ZHOU Tao. Empirical analysis of online human dynamics[J]. Physica A, 2012, 391: 3308-3315.

[8] 周涛, 汪秉宏, 韩筱璞, 等. 社会网络分析及其在舆情和疫情防控中的应用[J]. 电子科技大学学报, 2011, 25(6): 742-754.
ZHOU Tao, WANG Bing-hong, HAN Xiao-pu, et al. Social network analysis and its application in the prevention and control of propagation for public opinion and the epidemic [J]. Journal of University of Electronic Science and Technology of China, 2011, 25(6): 742-754.

[9] WALLSTEN S. What are we not doing when we're online[EB/OL]. [2014-06-21]. http://www.nber.org/papers/w19549.

[10] HOLMBERG L. Seeking social connectedness online and offline: Does happiness require real contact?[D]. Orebro: Orebro University, 2014.

[11] HRISTOVA D, MUSOLESI M, MASCOLO C. Keep your friends close and your facebook friends closer: a multiplex network approach to the analysis of offline and online social ties[C]//Proceedings Of the Eighth International Conference on Weblogs and Social Media. California: AAAI Press, 2014: 206-215.

[12] ADAMIC L, BUYUKKOKTEN O, ADAR E. A social network caught in the web[EB/OL]. (2003-06-02). http://firstmonday.org/article/view/1057/977.

[13] POOL S, KOCHEN M. Contacts and influence[J]. Social Networks, 1979, 1(1): 5-51.

[14] LIBEN-NOWELL D, NOVAK J, KUMAR R, et al. Geographic routing in social networks[J]. Proc Natl Acad Sci USA, 2005, 102(33): 11623-11628.

[15] PALCHYKOV V, KASKI K, KERTÉSZ J, et al. Sex differences in intimate relationships[J]. Scientific Reports, 2012(2): 370.1-370.5.

[16] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proc Natl Acad Sci USA, 2002, 99(12): 7821-7826.

[17] BACKSTROM L, HUTTENLOCHER D, KLEINBERG J, et al. Group formation in large social networks: membership, growth, and evolution[C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2006, 44-54.

[18] YOU Zhi-qiang, HAN Xiao-pu, LÜ Lin-yuan, et al. Empirical studies on the network of social groups: the case of Tencent QQ[EB/OL]. [2014-8-24]. http://arxiv.org/

pdf/1408.5558.pdf.

[19] VISWANATH B, MISLOVE A, CHA M, et al. On the evolution of user interaction in facebook[C]// Proceedings of the 2nd ACM workshop on Online social networks. New York: ACM, 2009: 37-42.

[20] CHIERICHETTI F, LATTANZI S, PANCONESI A. Rumor spreading in social networks[C]//Automata, Languages and Programming. New York: Springer, 2009: 375-386.

[21] LIBEN D, KELEINBERG J. The link-prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology, 2007, 58: 1019-1031.

[22] ZHANG Qian-ming, LÜ Lin-yuan, WANG Wen-qiang, et al. Potential theory for directed networks[J]. PLoS One, 2013(8): e55437.

[23] 任晓龙, 吕琳媛. 网络重要节点排序方法综述[J]. 科学通报, 2014, 59: 1175-1197.
REN Xiao-long, LÜ Lin-yuan. Review of ranking nodes in complex networks[J]. Chin Sci Bull (Chin Ver), 2014, 59: 1175-1197.

[24] XIA Ying-jie, REN Xiao-long, PENG Zheng-chao, et al. Effectively identifying the influential spreaders in large-scale social networks[C]//Multimedia Tools and Applications. New York: Springer, 2014: 1-13.

[25] LÜ Lin-yuan, ZHANG Yi-cheng, YEUNG C H, et al. Leaders in social networks, the delicious case[J]. PLoS One, 2011(6): e21202.

[26] WENG J, LIM E P, JIANG J, et al. Twitterrank: Finding topic-sensitive influential twitterers[C]//Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York: ACM, 2010: 261-270.

[27] BASTIAN M, HEYMANN S, JACOMY M. Gephi: an open source software for exploring and manipulating networks[C]//Proceedings of the Third International ICWSM, 2009(8): 361-362.

[28] KITSAK M, GALLOS L K, HAVLIN S, et al. Identification of influential spreaders in complex networks[J]. Nature Phys, 2010(6): 888-893.

[29] ONNELA J-P, SARAMÄKI J, HYVÖNEN J, et al. Structure and tie strengths in mobile communication networks[J]. Proc Natl Acad Sci USA, 2007, 104: 7332-7336.

编 辑 蒋 晓