社交网络用户的社交关系和签到行为分析

梁霄,赵吉昌,许可

北京航空航天大学软件开发环境国家重点实验室,北京 100191

摘要　理解社交关系和移动行为的相关性,对于研究社交网络演化及建模人类移动是非常重要的。分析了两个基于位置社交网络网站用户的社交关系和签到行为,以量化社交关系与移动性的相关性。结果表明,社交或签到排名的概率分布反比于其排名,这意味着社交关系和移动性间存在着隐含的联系。通过对不同粒度下用户社交关系和签到行为的比较,以及用户皮尔逊相关系数的计算,证明社交关系和人类移动性存在较强的相关性。

关键词　社交关系;移动模式;相关性;签到

中图分类号　N94;TP393　　**文献标志码**　A　　**doi**　10.3981/j.issn.1000-7857.2014.11.006

Analysis of Social Ties and Checkins in Location-based Social Networks

LIANG Xiao, ZHAO Jichang, XU Ke

State Key Laboratary of Software Development Environment, Beihang University, Beijing 100191, China

Abstract　Understanding the relationship between friendship and mobility is crucial to studying the evolution of social networks and modeling human movements. This paper aims to quantify the correlation between social ties and checkins by investigating two location-based social networking websites. It is discovered that the probability of social or checkin's rank is inversely proportional to the corresponding rank, which implies the potential connection between friendship and mobility. Therefore, the fractions of friends and checkins in the same county at different scales are compared and the Pearson correlation coefficients of users are computed. These results all demonstrate that social friendship correlates closely with human movements.

Keywords　social ties; mobility pattern; correlation; checkins

近年来,各种社交媒体(如Facebook、Twitter、Foursquare和Flickr等)正变得日益流行,极大地便利了人们的日常生活。基于这些社交媒体提供的服务,人们能够更容易认识新朋友、与朋友沟通交流、分享照片和当前位置信息等。同时,这些海量的社交行为数据也被记录下来,为相关研究提供了重要的基础。

关于社交网络的研究开展较早,可以追溯到社会心理学家Milgram著名的"六度分隔"(six degrees of separation)实验[1]。近来,通过利用因特网发送电子信件的方式,类似的实验被重复,也得到了相似的结果[2]。这些实验揭示了一个惊人的"小世界"现象,即在现实如此庞大的社会网络中,人与人之间的联系路径是如此之短。从这些实验中发现所谓的"小世界"现象具有两层含义:1)在拓扑方面,两个陌生人能够通过较少的中间人联系起来。Watts和Strogatz提出了一个模型,并解释了"小世界"现象的产生机制[3]。目前,在亿级用户规模的在线社交网络MSN[4]和Facebook[5]中,拓扑的"小世界"属性也得到了进一步的证明。2)在算法方面,人们能够仅利用网络的局部知识找到两人之间一条较短的连接路径。

收稿日期:2014-03-05;修回日期:2014-03-11

基金项目:高等学校博士学科点专项科研基金项目(20111102110019);北航博士研究生创新基金项目(YWF-12-RBYJ-036)

作者简介:梁霄,博士研究生,研究方向为人类动力学,电子信箱:liangxiao@nlsde.buaa.edu.cn;许可(通信作者),教授,研究方向为算法、数据挖掘和网络等,电子信箱:kexu@nlsde.buaa.edu.cn

引用格式:梁霄,赵吉昌,许可. 社交网络用户的社交关系和签到行为分析[J]. 科技导报, 2014, 32(11): 43-48.

Kleinberg扩展了Watts-Strogatz模型[3],发现针对二维网格上分布的用户,只有当他们的长程连接概率与地理距离的平方成反比时,这个网络才具有算法"小世界"特性[6]。实际中,人口在空间上的分布是非均匀的。通过分析LiveJournal站点的博主居住地和他们组成的社交网络,Liben-Nowell等[7]发现个体v与个体u成为朋友的概率反比于居住距离u更近的人口数目,并且证明这种连接构成的网络符合算法的"小世界"特征。此外,这种社交网络结构和用户的地理位置间的类似关系也被发现存在于社交网络Facebook中[8]。

如今,随着手机基站、GPS和基于位置的在线社交服务等各种定位技术的普及,位置数据变的越来越容易获得。因而,关于人类移动性的研究正逐渐成为热点,并且已经引起各学科学者的广泛关注。通过分别从美国钞票的流通[9]及欧洲的手机基站[10]提取人类的移动轨迹,研究者发现人们的出行距离能够很好的符合幂率或截断幂率分布。这说明人类移动具有与Lévy游走相似的特点。但是通过对提取的城市出行轨迹分析,人们发现行程分布更符合指数分布[11~14]。对于人类移动性的建模,主要分为建模个体和群体两类。对于个体的移动,许多基于主体(agent-based)的模型[15~18]被提出来用于模拟对整个群体聚合的统计特征。然而,人口移动模式中个体移动的多样性不能被忽视[19]。对于群体移动,早期的引力模型(gravity model)[20]可以用来建模和预测地点间的流量。最近,创新性地提出无参数的辐射模型(radiation model)[21],能够预测任何两地点间的有向流量。这个模型能够准确地预测通勤和人口迁移的流量,但是它似乎不能刻画城市内更频繁的日常移动[22,23]。因此,为更好地模拟城市内的出行,一些群体移动模型被提出[23]。

尽管在社交网络和人类移动性方面已分别有大量的研究,但是对于两者之间关系的理解仍然十分有限。事实上,社交网络处于地理空间中,它们的形成与人们的位置和出行紧密相关;相反地,人们的出行也依赖于社交网络的拓扑结构。这些都说明移动性和社交性之间应该存在某种潜在的联系。近来的一些工作已经注意到两者潜在的关联性:一些研究通过挖掘用户的历史位置信息推断社交关系或推荐好友[24,25];另一些研究将社交关系看做一个重要的因素来预测用户可能出现的位置[26,27]。这些研究结果都强调了社交关系和移动性的关联是确实存在的。但是据我们所知,目前还没有相关的研究明确地量化两者之间的相关性,也不清楚它们在多大程度上相互影响。

本文通过对两个基于位置的社交媒体网站的社交网络和用户历史出行记录进行研究,量化这种社交关系和移动性的相关性。首先,基于社交排名的交友现象[7]也出现在这两个社交网络中,这说明这两个网络同样具有"小世界"的特性。其次,发现这两个网络中签到的排名分布也类似于交友的排名分布,这表明社交关系和人们的出行之间存在某些内在的联系。最后,在不同的粒度下,通过对社交关系和移动性进行比较,它们之间的相关性被量化。

1 数据集描述

Brightkite和Gowalla是国外两个基于位置的社交网络站点。这两个网站都提供了类似的签到服务,使得用户能够分享自己当前位置周边的相关信息。这类网站除了记录用户的历史签到信息,也记录了用户之间的社交关系。因而,利用它们的相关数据,能够对用户社交关系和移动性之间的相关性进行研究。

本文使用了公开的数据集(http://snap.stanford.edu/data/#locnet)[22],它包含了Brightkite从2008年4月至2010年10月以及Gowalla从2009年2月至2010年10月的历史签到信息和社交网络结构。研究画出了美国(夏威夷和阿拉斯加除外)的Brightkite用户在本国内的所有签到位置,如图1所示。从图中可以观察到美国路网的大致结构,而且签到点往往集中分布在纽约、洛杉矶、西雅图等少数大城市中。同时,Gowalla用户的签到点也具有相似的分布,在此省略。这意味着人们日常的大部分活动仍然集中在城市内。

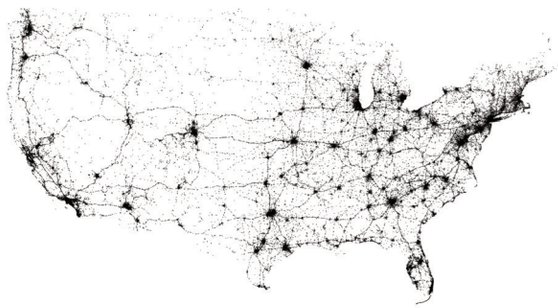


图1 美国Brightkite用户的签到分布

Fig. 1 Checkins of Brightkite´s users in the US

2 数据分析

2.1 推断用户的居住地

最近一些研究尝试从用户的历史签到记录中推测他们的居住地点[26,28,29],其主要思想是用户更倾向于在家的附近频繁地从事购物、就餐和娱乐等活动。其中,Cho等[26]将地理空间划分成25 km×25 km的网格,推断用户的居住地点为访问次数最多的网格中所有签到地点的平均位置。Scellato等[29]将用户的居住地点近似地看作访问次数最多的兴趣点(point of interests,POI)。Cheng等[28]考虑了在划分网格时,网格间的边界可能会对推断结果造成影响。他们首先将地理空间按照经纬度各1°划分网格,选择用户访问次数最多的网格及其相邻的8个网格作为候选区域;然后对这个候选区域再按照经纬度各0.1°划分网格,并重复上述过程;最后,直到网格的经纬度为0.001°,这时选择访问次数最多的网格的中心位置作为用户的居住地点。Cheng方法的思想是迭代细分用户居

住地的候选区域,直到实现要求的推断精度。因此,这里利用Cheng方法推断用户的居住位置,所推断的美国Brightkite用户的居住地分布如图2所示。从用户居住点分布图可以看出,用户主要集中在一些热门的大城市,这在一定程度上能够反映美国人口的地理分布情况。事实上,本研究也检查了Cho和Scellato的推断方法,得到了大致相似的分布结果。这说明3种推断方法在城市或县级精度下能够给出较为一致的结果。

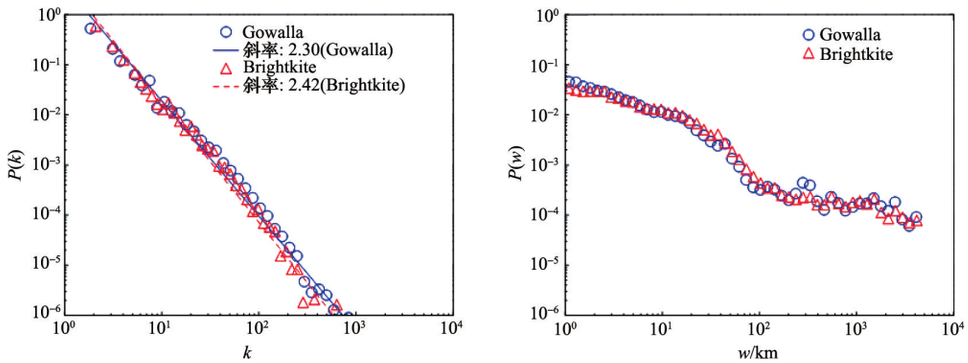
图2　美国Brightkite用户居住地的分布

Fig. 2　Home locations of Brightkite´s users in the US

因为这两个社交网站中,美国的用户较为活跃,所以本研究主要关注美国的用户以及他们在美国的签到信息。从这个公开数据集中,可提取出Brightkite网站的30156个用户和2810364次签到记录,以及Gowalla网站的50042个用户和3507539次签到记录。

2.2　社交关系

从数据集中可提取出美国用户间的社交网络。其中,网络的节点表示用户,边表示连接节点间存在社交关系,而边的权值定义为节点所代表的用户间的地理距离。最终,从Brightkite和Gowalla中可得到两个社交网络,分别有30156个节点、104085条边和50042个节点、211630条边。这两个网络的节点度k和边权值w分布如图3所示。从图3可以发现,这两个网络有非常接近的节点度和边权值分布。图3(a)显示出这两个网络度分布如其他复杂网络一样符合幂率分布,幂指数介于2和3之间。此外,如图3(b)所示,朋友间地理距离的概率随着距离的增加从均匀衰减到加速衰减,而当距离大于约100 km时衰减速度明显变缓。还分析了两个社交网络中任意两个用户居住地间的距离,其分布如图4所示。从图4可以看出,当居住地距离在大约100 km范围内时,其概率值也呈现出快速衰减的趋势,但随后缓慢上升。值得注意的是,从图3(b)和图4可以看出,朋友间距离的分布和用户居住地间距离的分布有基本一致的拐点(约100 km),这说明朋友间距离分布(边权值分布)衰减减缓的趋势是由于用户居住地的不均匀分布引起的。



(a) 节点度分布　　(b) 边权值分布

图3　社交网络的节点度和边权值分布

Fig. 3　Distributions of degree and weight for social networks

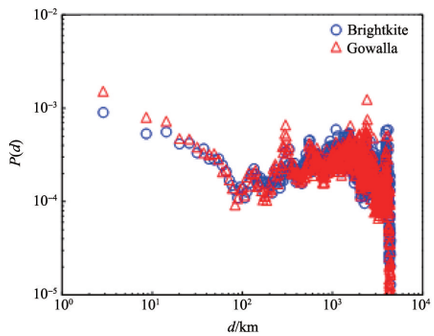


图4　用户居住地间的距离分布

Fig. 4　Distance distributions of users´ home locations

此外,还研究了用户间建立社交关系的内在特征。下面分别从距离和排名两个方面研究用户的交友概率。在距离方面,计算用户在某特定距离d建立社交关系的概率,可表示为相距d的朋友二元组数目除以所有相距d的用户二元组数目,其数学形式为

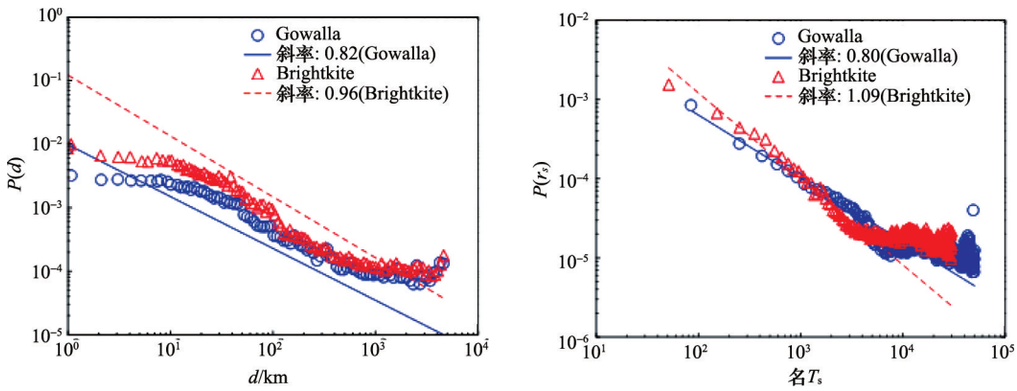
$$P(d)=\frac{|\{(u,v)|d(u,v)=d,\text{任意用户}u\text{和}v,\text{其中}u\text{和}v\text{是朋友}\}|}{|\{(u,v)|d(u,v)=d,\text{任意用户}u\text{和}v\}|}$$

式中,$d(\cdot,\cdot)$为用户居住地之间的距离。这个概率表示距离因素对交友的影响,并且排除了用户居住地点非均匀分布所造成的影响,其分布如图5(a)所示。从图5中可看出,关于Brightkite和Gowalla的社交网络的概率分布曲线有非常相似

的趋势,即交友概率是关于距离 d 的幂函数 $P(d)\propto d^{-\alpha}$。其中,幂指数 α 分别是0.96和0.82,都非常接近1。这也能够说明两个社交网络同样具有算法"小世界"性质[6]。在排名方面,根据文献[7]中的定义,用户 v 相对于用户 u 的排名(这里称作社交排名)被定义为居住地离用户 u 比用户 v 离 u 更接近的用户数目(包含用户 u 本身),可形式化为

$$rank_u(v)=\left|\{w|d(u,w)<d(u,v),\text{任意用户}w\}\right|$$

图5(b)为基于排名的交友概率。同样,可以观察到两个分布都能很好地被幂函数 $P(r_s)\propto r_s^{-\beta}$ 拟合,幂率指数都很接近1。这个结果也和以前的工作[7]十分吻合。总之,这两个社交网络基于距离或排名的交友概率都反比于距离或排名。事实上,这两者之间是等价的,并且可以相互导出[30]。它们是社交网络具有算法"小世界"特征的必然结果,即社交网络中,个体总是能够仅凭借局部信息,找到它们之间的较短路径。



(a) 根据距离　　(b) 根据排名

图5　建立社交关系的概率分布

Fig.5　Probability distribution of building social relationship

2.3　用户签到

研究主要关注这两个基于位置的社交网络站点的用户签到行为。图6为用户签到地点与其居住地间的距离分布。从图6可以看出,这两个分布概率在距离小于约100 km时下降得较快,然后衰减开始变得缓慢。需要说明的是,与社交网络边权值分布(图3(b))和用户居住地间的距离分布(图4)类似,曲线趋势转折点同样发生在约100 km处。这并不是偶然的,因为在实际生活中,用户的出行会受到人口在地理空间上的分布的限制。

类似于社交排名,这里定义签到排名(checkin's rank)。用户 u 的某次签到 c 的排名定义为居住地点离用户 u 比签到 c 的地点离用户 u 更接近的用户数目,其数学形式化为

$$rank_u(v)=\left|\{w|d(u,w)<d_u(c),\text{任意用户}w\}\right|$$

式中,$d_u(c)$ 表示用户 u 的居住地和签到 c 的地点之间的距离。因而,图7描述了所有签到的排名分布。有趣的是签到排名的分布也较好地符合幂函数形式 $P(r_c)\propto r_c^{-\beta}$,其中,对于Brightkite和Gowalla的分布指数β分别为1.19和1.10,都非常接近1。这一结果类似于社交排名的分布(图5(b)),揭示出社交关系的产生和签到行为之间存在某些内在的联系。

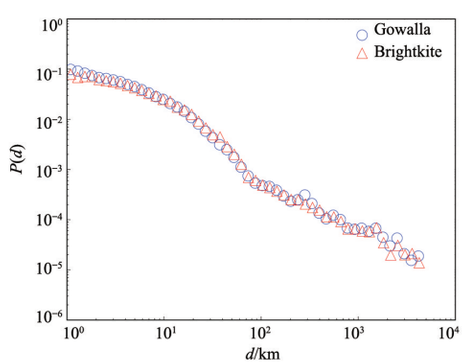


图6　用户居住地与签到地点间的距离分布

Fig.6　Probability distribution of distance from checkins to users´ homes

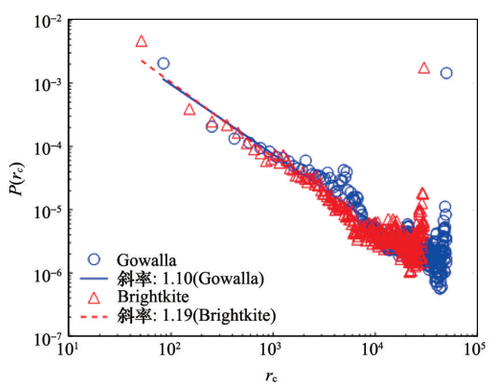


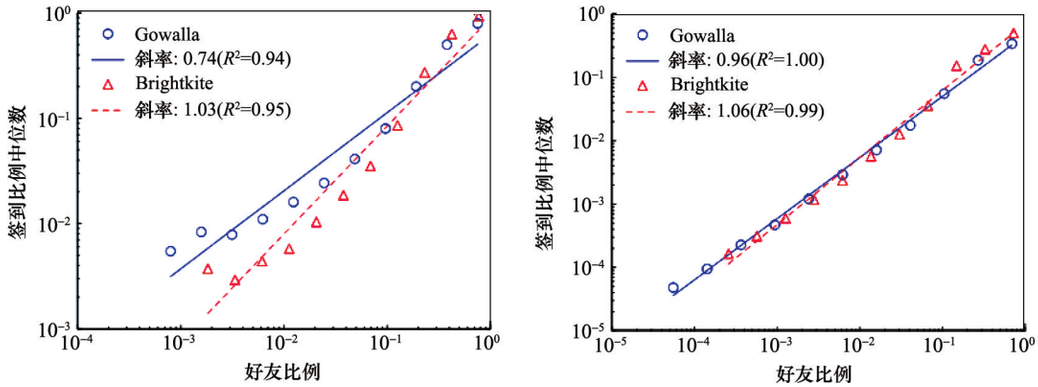
图7　签到排名的概率分布

Fig.7　Probability distribution of checkin´s rank

2.4　社交关系和签到行为的相关性

直观地说,人们一方面出于社交的目的常会拜访朋友;另一方面,当去不熟悉的地方时,也可能会在那里结识新的朋友。通过比较前两节中关于社交排名和签到排名的概率分布,发现两者都有非常相似的规律,即分别反比于社交和签到排名。这意味着社交关系和移动性可能有一些潜在的相关性,也符合直观感觉。

下面从两种不同的粒度来考虑社交关系和签到行为的相关性。在个体粒度下,计算了每个用户在每个县的好友数目和签到数目,并将每个用户在每个县的好友比例和签到比例看作一个有序点对。它们之间的比较结果如图8(a)所示,其中横坐标表示好友比例,纵坐标表示同一好友比例下,所有签到比例值的中位数。从图8可以看出,好友比例的增加会相应的导致签到比例的增加,并且好友比例和签到比例的中位值之间存在较严格的幂函数关系,其中Gowalla和Brightkite的幂指数分别为0.74和1.03。这说明在个体粒度下,社交关系和移动性间存在较强的相关性。同样,在县的粒度下,聚合出同一个县的所有用户在其他各县的好友数目和签到数目,它们之间的比较结果如图8(b)所示。相比于个体粒度(图8(a))而言,图8(b)显示出更为严格的幂函数关系,并且幂指数都非常接近1,揭示出社交关系和移动性间更强的相关性。



(a) 个体粒度　　(b) 县级粒度

图8　社交关系和移动性的相关性。

Fig. 8　Correlation between social friendship and mobility

由于对每个用户来说,都可以计算出在各个县的朋友数和签到数,从而形成两个相对应的序列。对每个用户计算得出关于这两个序列的皮尔逊相关系数(Pearson correlation coefficient),其累积分布如图9所示。从图9可以看出,对于这两个社交网站,有超过67%的用户相关系数大于0.6,50%的用户相关系数大于0.8。这也进一步证明了社交关系和移动性存在很强的相关性。

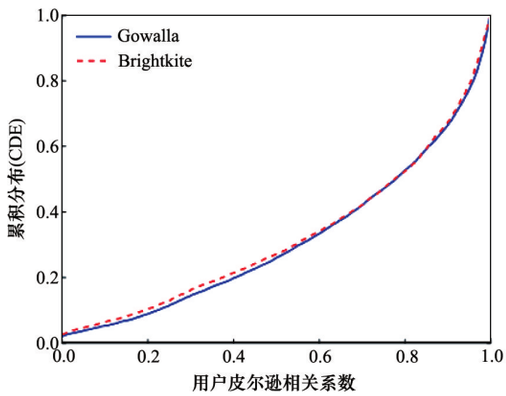


图9　用户皮尔逊相关系数的累积分布

Fig. 9　Cumulative distribution function of Pearson correlation coefficients for users

3　结论

研究了两个基于位置的社交网络站点的社交关系、签到行为以及它们之间的相关性。首先,通过对美国用户的社交网络进行分析,证明两个社交网络都具有"小世界"的性质。具体来说,用户间建立社交关系的概率反比于他们之间的距离或社交排名。这意味着仅利用局部知识,用户可以自发地找到他们之间较短的路径。其次,通过对用户签到行为进行分析发现,签到排名的分布概率也反比于签到排名。最后,通过比较在个体粒度和县级粒度下好友数和签到数的比例,以及用户社交关系和签到间的皮尔逊相关系数,量化了社交关系和移动性,并且发现它们之间存在着非常强的相关性。

此外,尽管社交关系和移动性间的强相关性被发现,但在多大程度上基于社交关系能够预测移动性以及如何根据社交关系来建模人类移动仍然是还未解决的问题,值得我们进一步深入研究。

参考文献(References)

[1] Milgram S. The small world problem[J]. Psychology Today, 1967, 1(1): 60-67.

[2] Dodds P S, Muhamad R, Watts D J. An experimental study of search in global socialnetworks[J]. Science, 2003, 301(5634): 827-829.

[3] Watts D J, Strogatz S H. Collective dynamics of ´small-world´ networks [J]. Nature, 1998, 393(6684): 440-442.
[4] Leskovec J, Horvitz E. Planetary-scale views on a large instant-messaging network[C]//Proceeding of the 17th International Conference on World Wide Web. NewYork: ACM Press, 2008: 915-924.
[5] Backstrom L, Boldi P, Rosa M, et al. Four degrees of separation[C]// Proceedings of the 3rd Annual ACM Web Science Conference. New York: ACM Press, 2012: 33-42.
[6] Kleinberg J. Navigation in a small world[J]. Nature, 2000, 406(6798): 845.
[7] Liben-Nowell D, Novak J, Kumar R, et al. Geographic routing in social networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(33): 11623-11628.
[8] Backstrom L, Sun E, Marlow C. Find me if you can: Improving geographical predictionwith social and spatial proximity[C]//Proceedings of the 19th International Conference on World Wide Web. Raleigh: ACM Press, 2010: 61-70.
[9] Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel [J]. Nature, 2006, 439(7075): 462-465.
[10] González M, Hidalgo C, Barabási A-L. Understanding individual human mobility patterns[J]. Nature, 2008, 453(7196): 779-782.
[11] Kang C, Ma X, Tong D, et al. Intra-urban human mobility patterns: An urban morphology perspective[J]. Physica A, 2012, 391(4): 1702-1717.
[12] Bazzani A, Giorgini B, Rambaldi S, et al. Statistical laws in urban mobility from microscopic GPS data in the area of Florence[J]. Journal of Statistical Mechanics: Theory and Experiment, 2010, 2010(5): P05001.
[13] Liang X, Zheng X, Lü W, et al. The scaling of human mobility by taxis is exponential[J]. Physica A, 2012, 391(5): 2135-2144.
[14] Roth C, Kang S M, Batty M, et al. Structure of urban movements: Polycentric activityand entangled hierarchical flows[J]. PLOS ONE, 2011, 6(1): e15923.
[15] Song C, Koren T, Wang P, et al. Modellingthe scaling properties of human mobility[J]. Nature Physics, 2010, 6(10): 818-823.
[16] Han X, Hao Q, Wang B, et al. Origin of the scaling law in human mobility: Hierarchy of traffic systems[J]. Physical Review E, 2011, 83 (3):2-6.
[17] Hu Y, Zhang J, Huan D. Toward a general understanding of the scaling laws in human and animal mobility[J]. Europhysics Letters, 2011, 96 (3): 38006.
[18] Jia T, Jiang B, Carling K, et al. An empirical study on human mobility and its agent-based modeling[J]. Journal of Statistical Mechanics: Theory and Experiment, 2012, 2012(11): P11024.
[19] Yan X Y, Han X P, Wang B H, et al. Diversity of individual mobility patterns andemergence of aggregated scaling laws[J]. Scientific Reports, 2013, 3: 2678.
[20] Barthélemy M. Spatial networks[J]. Physics Reports, 2011, 499(1-3): 1-101.
[21] Simini F, González M C, Maritan A, et al. A universal model for mobility and migrationpatterns[J]. Nature, 2012, 484(7392): 96-100.
[22] Liang X, Zhao J, Dong L, et al. Unraveling the origin of exponential law in intra-urbanhuman mobility[J]. Scientific Reports, 2013, 3: 2983.
[23] Yan X Y, Zhao C, Fan Y, et al. Universal predictability of mobility patterns in cities[J]. Physics and Society, arXiv:1307.7502v1: 1-19.
[24] Crandall D J, Backstrom L, Cosley D, et al. Inferring social ties from geographic coincidences[J]. Proceedings of the National Academy of Sciences of the Uniited States of America, 2010, 107(52): 22436-22441.
[25] Cranshaw J, Toch E, Hong J, et al. Bridging the gap between physical location and onlinesocial networks[C]//Proceedings of the 12th ACM International Conference on Ubiquitous Computing. Copenhagen: ACM Press, 2010: 119-128.
[26] Cho E, Myers S, Leskovec J. Friendship and mobility: User movement in location-based social networks[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego: ACM Press, 2011: 1082-1090.
[27] Sadilek A, Kautz H, Bigham J P. Finding your friends and following them to where youare[C]//Proceedings of the fifth ACM International Conference on Web Search and Data Mining. Seattle: ACM Press, 2012: 723-732.
[28] Cheng Z, Caverlee J, Lee K. Exploring millions of footprints in location sharing services[C]//Proceedings of the Fifth International Conference on Weblogs and Social Media, Menlo Park: AAAI Press, 2011.
[29] Scellato S, Noulas A, Mascolo C. Exploiting place features in link prediction on location-based social networks[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego: ACM Press, 2011: 1046-1054.
[30] Hu Y, Wang Y, Li D, et al. Possible origin of efficient navigation in small worlds[J]. Physical Review Letters, 2011, 106(10): 1-4.

（责任编辑 吴晓丽）