ISSN 1000-1239/CN 11-1777/TP
49(4): 691-699, 2012

微博网络测量研究

樊鹏翼　王　晖　姜志宏　李　沛
(国防科学技术大学信息系统与管理学院　长沙　410073)
(fanpengyi@gmail.com)

Measurement of Microblogging Network

Fan Pengyi, Wang Hui, Jiang Zhihong, and Li Pei
(*College of Information Systems and Management, National University of Defense Technology, Changsha* 410073)

Abstract　With the breakthrough of mobile communications and Web technology, online social network led by microblog has developed widely in China. More and more people begin to share information and opinions through microblogging system. In order to gain insights into the topological characteristics of microblogging network and online user behavior characteristics, we launch an active measurement of Sina microblog which is the biggest microblogging system in China. This paper analyzes the results from our measurement and investigates on topological characteristics and user behavior patterns in Sina microblog, compared with existing results via measuring other online social networks. Our major findings include: 1) Sina microblog has apparent smalll-world effect; 2) While the indegree of Sina microblog follows power-law distribution, the outdegree distribution appears to have multiple separate power-law regimes with different exponents; 3) Unlike other online social networks, Sina microblog has weak correlation of indegree and outdegree; 4) The overlay graph of Sina microblog appears assortative mixing; 5) Tweeting time of users exhibits daily and weekly patterns; 6) The number of tweets in Sina microblog approaches Weibull distribution; 7) Actions of retweeting and replying have strong correlation in Sina microblog, and the probability of retweeting is higher than that of replying. These research and findings will be helpful not only for designing mathematical or computational models which are coincident with actual microblogging characteristics in China, but also for monitoring, directing and dominating of microblogging system.

Key words　microblog; social network; network measurement; topological characteristics; user behavior

摘　要　随着移动通信和 Web 技术的不断突破,以微博为代表的在线社会网络在中国广泛发展起来,越来越多的人开始使用微博进行信息分发和舆论传播.为了了解中国微博网络中的拓扑结构特征和用户行为特征等内在信息,对国内最大的微博系统——新浪微博——开展了主动测量,并结合已有的在线社会网络测量结果,对新浪微博的网络拓扑和用户行为特征进行了分析和比较.主要发现包括:1)新浪微博网络具有小世界特性;2)新浪微博网络的入度分布属于幂次分布,而出度分布表现为某种分段幂率函数;3)与类似社会网络相比,新浪微博网络的出入度不具有相关性;4)新浪微博网络属于同配网络;5)新浪微博用户发博时间具有明显的日分布和周分布模式;6)新浪微博用户博文数目分布表现为威布尔分

收稿日期:2010-09-07;**修回日期**:2011-03-07
基金项目:国家"八六三"高技术研究发展计划基金项目(2008AA01Z407);国家自然科学基金项目(60872053,61105124)

布;7)新浪微博用户博文的转发和评价行为具有很强的相关性,且博文转发概率要高于评价概率.这些测量研究和发现不仅有助于设计出符合中国微博网络结构特征的数学模型和计算模型,也是实现对微博舆论的监测、引导、控制等方面的重要依据和基础.

关键词 微博;社会网络;网络测量;拓扑特征;用户行为

中图法分类号 TP393

随着移动通信网络和 Web 技术的不断发展,微博逐渐成为人们日常交流、通信、娱乐的基本工具.与传统博客不同的是,微博对博文信息进行了字数限制,其发布方式也不再局限于传统的计算机,大量移动设备的引入使得微博成为一种流动的信息传播平台,应用十分广泛.Twitter 作为全球最早的微博平台,在短短的两三年时间内便发展成为全世界最受欢迎的网络服务之一,充分展现了微博潜在的影响力.2009 年,随着中国各主流门户推出了各自的微博产品,微博在中国也得到了迅猛发展,用户数目与日剧增.微博用户之间相互联系组成了一个关系紧密、结构复杂的社会网络,信息以最快的速度在微博中广泛传播,其已发展成为一种新型的信息传播媒介,具有巨大的应用前景和商业价值.同时,由于微博在中国还是一个处于上升期的新生事物,管理机制仍不完善,虚假信息和违法信息有可能得以滋生和蔓延,这将对我国的社会稳定和经济发展产生一定的负面影响.因此,针对微博网络开展测量研究,了解微博系统的拓扑结构特征、用户行为特征等信息,不仅能加强对复杂网络与社会网络的理论探索,而且有助于微博网络信息传播模型和计算模型的建立,同时也对实现微博舆论的监测、引导、控制等提供了重要依据和基础,具有十分重要的理论价值和实践意义.

网络测量是对实际网络进行指标测量,评估网络运行状况,建立网络行为分析模型,是感知、优化和控制网络的有效途径,一般采用的方法主要包括基于爬行器的主动测量技术与利用节点部署获取流量的被动测量技术.随着互联网规模的持续扩大和应用的不断更新,以博客、播客、知识共享、图片共享等为代表的在线社会网络(online social network)服务受到了用户的青睐,其主要包括博客网络 LiveJournal、微博网络 Twitter、视频共享网络 YouTube、图片共享网络 Flickr 等.同时对于这些社会网络的测量研究也引起了越来越多学者的关注,其中研究的重点集中于拓扑结构测量、用户行为模式挖掘、社区结构发现[1]等相关领域.Mislove 等人[2]采集了 4 种不同的社会网络数据:Flickr,YouTube,LiveJournal 和 Orkut,对比分析这 4 种社会网络的拓扑特性,验证了社会网络的幂率特性、小世界效应以及无标度特性,并观测到社会网络中存在一个高度节点互连、聚集度很大的核心子网,其他低度节点连接在这个子网边缘组成了复杂的社会网络.Guo 等人[3]针对 3 种不同类型的知识共享网络(博客网络、书签共享网络、知识问答网络),对其用户行为进行了测量研究,发现用户时长并不服从指数分布,用户对网络的贡献服从广延指数(stretched exponential)分布而非幂率分布,并证明了这样的网络并非由少部分核心节点所能支配的.Cha 等人[4]则在 Mislove 的基础上对 Flickr 网络中的图片拓扑分布、时间演化分布以及信息传播过程进行了分析,发现 Flickr 中的信息传播主要依靠社会网络中的节点连接关系,其传播过程需要经历较长的时间.而 Cha 等人[5]和 Cheng 等人[6]则主要针对 YouTube 网络进行了测量研究,分别从用户的行为特征、视频内容的属性(类别、长度、大小等)特征、社会网络的拓扑结构和动态演化特征等多个角度进行了测量分析,得到了 YouTube 独有的统计行为模式.Twitter 作为全球最大的微博系统,对于它的测量研究也主要集中于拓扑和用户行为模式的挖掘,Java 等人[7]主要对 Twitter 的网络拓扑、地理分布进行了测量,并利用文本处理技术分析了用户发博的兴趣和动机.Huberman 等人[8]则对 Twitter 中潜在的朋友关系进行了挖掘,发现 Twitter 网络是由密度很大的相互关注网络与稀疏的真实朋友网络组成的.Kwak 等人[9]利用爬行技术采集了比先前研究更为精确的 Twitter 数据,从 Twitter 网络的拓扑结构、用户排序方法、热点话题传播模式等多个角度展开研究,实现了 Twitter 网络及其信息分发的量化分析.

虽然在线社会网络在中国取得了巨大的成功,但对于这些网络的测量工作却很少看到,由于不同的文化背景和社会习惯,国内外在线社会网络可能存在一定的差异,这也是本文研究的出发点.本文选取国内最大的微博平台——新浪微博[10]——作为研究对象,采用爬行器技术获取微博网络数据,对微博覆盖网络的拓扑特征和用户行为特征进行了研究,并与国外相关的社会网络进行了对比分析.从目前已知的研究文献看,本文是第 1 个对中国的微博网络展开测量研究工作的.

1 新浪微博数据采集

1.1 新浪微博爬行器

作为国内最大的微博系统,新浪微博凭借其优质的服务和新颖的运营模式,已发展成为当下最热门的互联网服务.在新浪 2010 年 9 月公布的《中国微博元年市场白皮书》中显示,新浪微博的月覆盖人数约 4 400 万,而总微博条数为 9 000 万,这足以说明新浪微博的巨大影响力.新浪微博规定用户每次更新的信息不超过 140 个字,并提供图片、视频、音频等多媒体嵌入功能,用户可以通过网页、WAP、手机短信与彩信、手机客户端、MSN 等多种方式更新自己的微博,做到"随时、随地、随性"地信息共享与分发.同时新浪微博还引入"明星"认证机制,依靠"明星微博"的影响力来不断吸引"草根用户",从而推动微博的发展.新浪微博中,用户还可以对自己喜欢的用户进行关注,成为这个用户的关注者,通过用户之间的关注行为,整个微博系统将抽象成为一个有向网络,即微博系统的拓扑网络 $G=(V,E)$.其中节点集 V 表示微博用户集合,而有向边集 E 则表示这些用户的关注行为.

在采集新浪微博的网络拓扑结构和用户博文数据时,本文利用 Python 语言编写了一个面向新浪微博的网络爬行器——SMCrawler.该爬行器主要采用了"滚雪球"(snowball)的爬行算法[11],即以一部分节点集合作为初始节点,利用节点之间的连接关系进行广度优先搜索,并抽取出每个节点的相关信息,实现新浪微博网络的数据获取.由于新浪微博中存有大量的"明星微博",他们一般具有较大的连接度,本文选取这些节点作为初始节点,从而保证了爬行的连续性.同时,由于新浪微博对节点的"粉丝"页面进行了限制(最多显示 2000 个"粉丝"),我们无法获取到节点完整的 Inlink 信息,因此,在爬行的过程中应重点考虑节点之间的 Outlink 关系,在保证每个节点 Outlink 列表完整的基础上,利用节点的 Inlink 不断拓宽爬行的广度,本文提出了一种"出度优先、广度获取"的爬行策略(如图 1 所示),当访问一个用户页面时,首先对该节点的链接信息 Link_Info 和用户信息 User_Info 进行信息抽取并存储入库,同时提取该节点的所有 Inlink URL 与 Outlink URL,进行相应的优先级设置以实现出度优先,过滤重复的 URL 加入至访问列表,最后按照优先级对访问列表重新排列,这样既能保证每个节点出度信息的完备性,又可以拓展节点爬行的广度.同时,为了更为快速地获取数据,爬行器采用了并发访问机制,可以同时向多个用户发送请求,从而优化了爬行器的爬行效率.

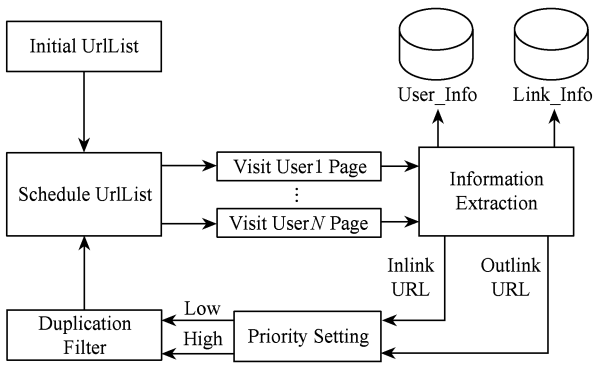


Fig. 1 Flow chart of SMCrawler.

图 1 SMCrawler 流程图

新浪微博网络主要存在两类页面:用户页面与链接页面.其中用户页面由包含出入度、博文数等信息的用户首页与该用户发表的博文页面组成;而链接页面则主要包括用户的"关注"列表页面和"粉丝"列表页面.针对不同的页面形式,本文采用 XPath[12] 语言对新浪微博的网络拓扑信息与用户博文信息进行实时抽取与存储,其具体的处理信息如表 1 所示:

Table 1 Information Lists Extracted by SMCrawler

表 1 SMCrawler 抽取信息列表

Information	Content	Page	Use
Network topology	Outlink list	"follows" page	For computing Smalll-world effect, correlation of indegree and outdegree, mixing pattern
	Inlink list	"Fans" page	
	Outdegree	User front page	For computing power-law characteristic
	Indegree		
User behavior	Tweeting number	User front page	For computing the pattern of tweeting time
	Tweeting time	Tweet page	For computing the pattern of tweeting, retweeting and replying
	Retweeting number		
	Replying number		

1.2 微博数据采集与描述

在本文的测量实验中,将 SMCrawler 部署于一台服务器上,测量节点的互联网出口宽度为 100 Mbps. 测量开始时间为2010-04-01,采集时长持续了 50 d,最终得到的数据集如表 2 所示:

Table 2 Dataset Measured by SMCrawler

表 2 SMCrawler 测量的结果数据集

# Nodes	# Links	# Tweets
829 222	27 608 171	41 833 647

在社会网络测量中,由于网络规模十分庞大且处于动态变化中,获取整个网络的拓扑数据是十分困难的,因此在测量实验中,都会做一定的采样以实现网络的简化. 文献[2]与文献[4]均采用"滚雪球"算法采集在线社会网络数据,这是因为"滚雪球"采样算法能够有效地获取一个很大的连通团,在统计指标计算中虽然会造成社会网络的幂率偏差,但对于其他测量指标(全局聚集系数等)的估计却十分有效[2]. 本文对新浪微博网络节点出入度的统计,并不是通过邻居节点求和的方法来计算,而是直接利用用户页面显示的真实数据进行统计,这样可以保证节点出入度的有效性,从而减小"滚雪球"采样对网络幂率估计造成的误差. 如图 2 所示,我们可以看出,随着采集数据的增加,微博网络的出度与入度分布十分相似,并且幂率均维持在较小的变化范围内,因此,我们认为本文采样获取的微博网络具有自相似性,能够有效地描述新浪微博网络的拓扑特性.

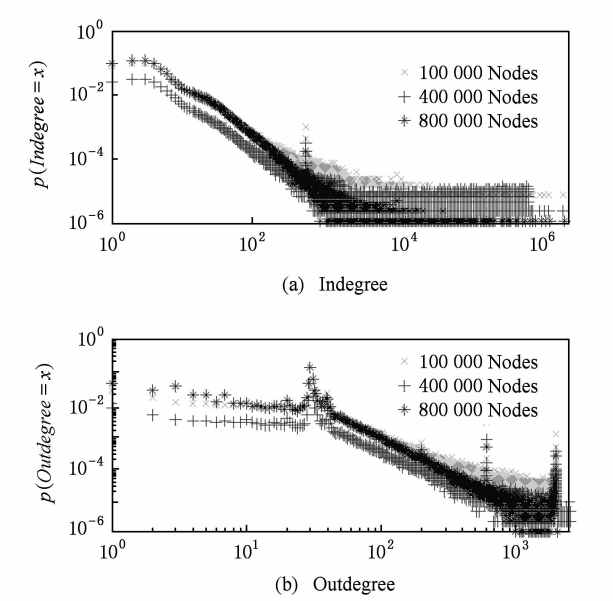


Fig. 2 Degree distribution varying in data capturing process.

图 2 数据采集过程中度分布变化

2 新浪微博网络拓扑测量

新浪微博网络拓扑测量主要是从微博网络的拓扑结构出发,研究微博网络的平均路径长度、聚集系数、节点度分布、出入度相关性、网络混合模式等物理性质,从而得到微博网络的基本属性,为微博网络的拓扑生成与相关应用研究奠定基础.

2.1 小世界特性

Watts 等人[13]定义了小世界网络应同时具有以下两类属性:1)网络中的大多数节点以较短的路径相连;2)网络具有较高的聚集度. 大量的测量实验研究表明,真实网络几乎都具有小世界效应,尤其在社交网络中更是普遍存在. 小世界理论表明,不论网络规模有多大,只要经过有限的几条边,就可以从网络的一节点到达其他任意节点,该理论也被称为"六度分离"理论. 新浪微博网络是否同样具有小世界特性? 它与其他在线社会网络之间有何区别? 为了解答以上问题,本文主要采用了以下测量指标:

1) 网络平均路径长度(average path length)是指网络中所有节点对之间最短路径的平均值,而网络直径(diameter)则是指网络中任意节点对之间最短路径的最大值. 网络平均路径长度和网络直径与网络的连通性、可达性以及传输延迟等特征密切相关,主要用于研判网络小世界特性的第 1 类属性.

2) 聚集系数(clustering coefficient)用于描述一个节点邻居之间的相互连接的紧密程度,即网络的集团化程度,主要用于研判网络小世界特性的第 2 类属性. 聚集系数的计算方法包括局部聚集系数、平均聚集系数与全局聚集系数,由于全局聚集系数能够最准确地从全局刻画网络的聚集特性,因此,本文采用与文献[14]相同的全局聚集系数来描述网络聚集性能,其计算公式如下:

$$C=\frac{3\times\text{number of triangles in the network}}{\text{number of connected triples of vertices}}. \quad (1)$$

由于本文采集的数据十分庞大,很难对所有的节点和边进行指标计算,因此本文采用随机采样求均值的方法. 每次采样的节点数目为 300 000,在一台配置为 4 核 2.33 GHz,内存为 8 GB 的服务器进行计算,平均计算时长为 130 h,采样 3 次最终得到新浪微博网络的测量结果如表 3 所示:

Table3 Small World Characteristic of Sina Microblogging

表 3 新浪微博小世界特性

Social Network	Average Path Length	Diameter	Clustering Coefficient
Sina microblogging	4.0	12	0.2137
Random network	3.88		0.00004
Twitter[7]		6	0.106
YouTube[2]	5.1	21	0.136
Flickr[2]	5.67	27	0.313

从表 3 可以看出,新浪微博与其同等规模的随机网络相比,网络平均路径长度 $l>\approx l_{\text{Random}}$ 并且聚集系数 $C\gg C_{\text{Random}}$,因此,我们可以认为新浪微博网络具有小世界特性. 相比于其他社会网络,新浪微博的平均路径长度和网络直径都较小,聚集系数明显高于 Twitter 和 YouTube,这说明新浪微博网络中节点之间的联系更为紧密,这十分有利于信息的传播. 同时,新浪微博的小世界特性也会加速网络中的谣言、病毒的传播,从而为社会舆论和互联网的监管带来困难,这也是微博网络发展亟待解决的一个重要问题.

2.2 节点度分布

在新浪微博中,节点之间的连接关系可表示为两类:"粉丝"和"关注",其对应的网络关系则为节点的"链入"和"链出"关系. 在测量节点度分布时,我们应分别对节点的出度和入度概率分布进行计算,从而得到新浪微博节点度分布的特性.

从图 3 中可以看出,新浪微博的节点入度和出度概率分布具有明显的幂率(power-law)分布特征,并且节点入度在较低的节点度范围内出现类似小变量饱和(saturation for small variables)的现象,考虑我们在测量过程中以高度节点为初始节点实施"滚雪球"采样,丢失了大量低入度节点信息,这可能也是造成这种现象的原因之一. 通过最小二乘法进行曲线拟合,可以得到节点入度概率分布的幂指数为 −1.435,而节点出度概率分布具有明显的幂率分段特性,其拟合曲线表达式为

$$p_{\text{out}}(k)=\begin{cases}0.02692\times k^{(-0.4449)}, & k\leqslant 27,\\ 5.822\times k^{(-1.851)}, & k\geqslant 40.\end{cases}\quad(2)$$

同时,当节点出度 k 满足 $k\in(27,40)$ 时,其分布出现了一个较明显的"尖刺"现象,考虑到该分布区间较小以及拟合的复杂性,本文并未对其进行曲线拟合.

通过与其他社会网络的节点度分布特性进行对比(如表 4 所示),我们可以看出新浪微博的出度分布具有明显的差异,其具有分段幂率特性,这与新浪微博的节点加入机制相关. 对于新加入的节点,新浪微博系统会自动推荐 30 个已有节点作为"链出"节点,用户可以自主选择是否接受推荐,从而导致节点出度初始值为 0 或 30,因此,在这两个点附近将产生不同的幂率分布. 此外,我们可以观测到出度为 2 000 附近出现了"尾部提升"现象,这主要由新浪微博的出度上限机制所造成(出度最大值为 2 000),将原有的重尾分布压缩至较小的区间内,从而形成了尾部提升.

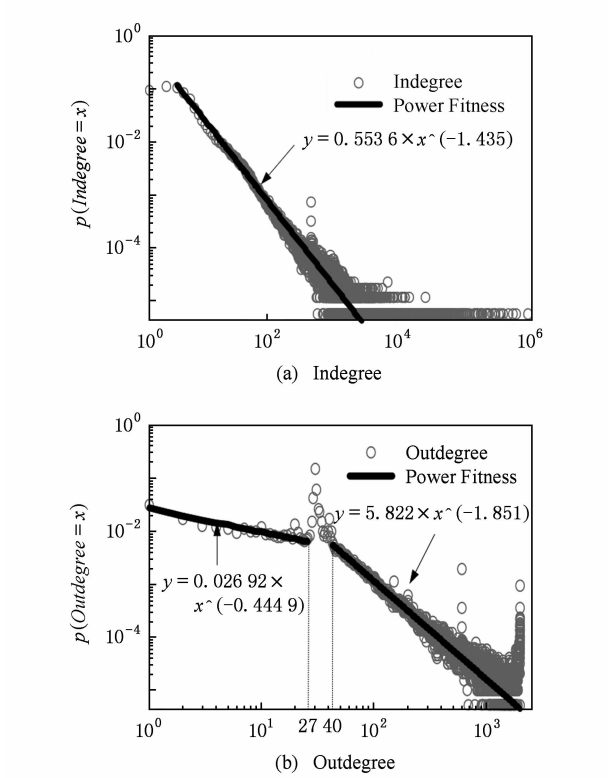


Fig. 3 Degree distribution for nodes in Sina microblogging.

图 3 节点度分布

Table 4 Power-Law Coefficient Compared with Different Social Networks

表 4 不同社会网络节点度分布幂率指数对比

Social Network	Slope of Indegree	Slope of Outdegree
Sina microblogging	−1.435	$k\leqslant 27$: −0.4449 $k\geqslant 40$: −1.851
Twitter[7]	−2.4	−2.4
YouTube[2]	−1.63	−1.99
Flickr[2]	−1.74	−1.78

2.3 出入度相关性

已有的测量研究表明,现有的社会网络中,节点的出度与入度之间具有明显的相关性. 一般而言,高

出度节点应同时具有相似的高入度,而低出度的节点其入度也相对较小[2],这与社会网络中节点的互惠性和链接的对称性有着密切的联系.本文对新浪微博节点的出入度相关性进行了统计,发现新浪微博不同于 Twitter,YouTube 以及 Flickr,其出入度相关系数仅为 0.033,而 Twitter 测量得到的相关系数则为 0.59[2],这表明新浪微博的出入度之间并无明显的相关性.

图 4 是新浪微博节点出入度相关系数随节点出度和入度变化的曲线,从图 4(a)可以看出,节点出度的变化对相关系数的影响不是很大,其波动范围均维持在较低的数值;而从图 4(b)可以看出,当节点入度较小时(小于 2 500 时),其相关系数与 Twitter 的值基本保持一致,此时的节点一般为"草根节点",这些节点的出入度具有明显的相关性.但随着节点入度的增加,相关系数呈幂次指数下降,最终到达较低的均值.由此,我们可以看出造成新浪微博节点出入度分布不相关的原因主要包括:1)新浪微博对节点出度进行了限制,这直接造成了节点出入度的绝对不平衡,从测量数据来看其极大值差距可高达 10^3 量级;2)由于存在大量的"明星节点"和推荐机制,使得新浪微博用户更倾向于连接这些入度很大的"明星节点",而这些节点的出度却很小,进一步增大了节点出度和入度的差距,影响了整个网络节点出入度的相关性.同时从"明星节点"的高入度特性,我们也能看出新浪微博具有很强的节点从众性和名人效应,这也是新浪微博能够在短时间内迅速成长的一个重要因素.

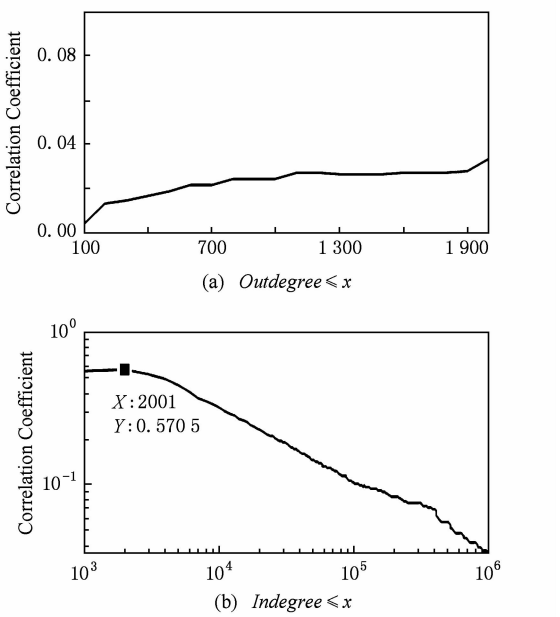


Fig. 4 Plot of correlation coefficient versus degree of nodes.

图 4 节点出入度相关系数随节点度变化曲线

2.4 网络混合模式

在多数网络中,两节点之间存在连接边的概率,通常依赖于这两个节点的类型,这种现象通常称之为网络的混合模式(mixing pattern).在真实的网络中,节点之间的连接通常会表现出某种倾向性,我们将高度节点倾向于连接其他高度节点的网络称为同配网络,而高度节点倾向于连接低度节点的网络则为异配网络.这种节点之间连接的相关性,对网络可靠性、传播动力学等都有着不同于度分布特征的重要影响,受到很多研究者的关注.

节点度联合分布(joint degree distribution)能够直观地表示网络的混合模式,它可以通过节点度相关函数 k_{nn} 近似定义.对于无向网络,k_{nn} 是指网络中度为 k 的节点与其邻接节点的平均度函数,可以表示为节点度的条件概率形式:$K_{nn}(k)=\sum_{k'} k'P(k'|k)$,当 K_{nn} 是关于节点度 k 的递增函数,则该网络为同配网络,反之为异配网络[15].对于有向网络,文献[2]对 K_{nn} 进行了相似定义:节点出度 k 与其邻居(链出)节点的平均入度之间的映射,并且利用该方法证明了 Flickr,LiveJournal 和 Orkut 网络均为同配网络,而 YouTube 网络则为异配网络.但由于在新浪微博中,不同节点的入度差异很大,如果利用求和取平均的方法来计算入度均值则可能会产生较大的误差,因此,本文采用中位值的方法计算节点的平均入度.从图 5 可以看出,随着节点出度的增加,均值 K_{nn} 曲线基本保持稳定且略有上升,而中位值 K_{nn} 曲线明显呈上升趋势,这说明新浪微博网络中各节点一般都会连接度很大的"明星节点",并且高度节点更倾向于连接其他高度节点,具有同配特性,而这些高度节点相连组成了新浪微博网络的核心子网.新浪微博网络的同配性也说明了该网络具有更低

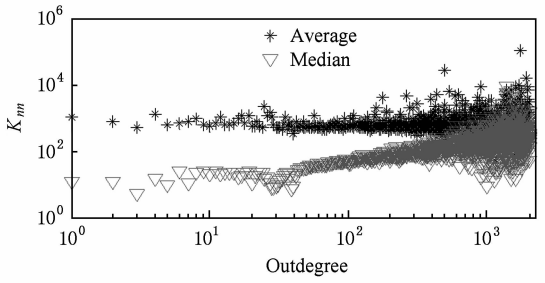


Fig. 5 Log-log plot of the outdegree versus the average indegree of friends.

图 5 节点出度与邻居节点平均入度对数变化曲线

的传播临界值,十分有利于网络舆论与信息的快速传播.

3 新浪微博用户行为测量

新浪微博用户行为测量主要是从微博用户的行为模式出发,研究微博用户所涌现出来的行为特征,主要包括用户时间分布、发博行为分布、转发与评价行为分布等,可以为个体行为建模、信息传播模式的研究奠定相应的基础.

3.1 用户时间分布

新浪微博用户时间分布是指用户发表博文数目在时间轴上的统计分布,该指标体现了大量用户节点所涌现出来的一种"在线"模式和行为特点,这对于新浪微博用户行为的建模研究具有一定的指导意义.为了直观地表示用户发博的时间分布特点,本文采用了两种不同的时间单元:小时和日.

图 6 是在不同时间尺度下,对用户在某时间段内发表博文的数目进行归一化的结果.其中,图6(a)表示以小时为时间单元、以日为时间周期对新浪微博的博文数目进行统计,发现新浪微博用户发博时间与用户的作息时间十分吻合,其具体的时间分布为:临晨 1 点到上午 8 点为用户发博的低潮期(≤2%),上午 9 点到晚上 12 点为用户发博的高峰期(4%~6%),而晚上 10 点到 11 点则为用户发博的最高峰(7%).图 6(b)表示以日为时间单元、以周为时间周期对新浪微博的博文数目进行统计,可以看出博文数目的变化在该时间单元上具有一定的自重复性.同时,新浪微博用户不同于 P2P IPTV 用户[16],不存有明显的周末效应,周末的博文数目较工作日略有下降,具体来说主要是由白天高峰时期的博文数目减小造成的,这说明了微博用户更倾向于工作时间和每日晚间发表微博,并具有一定的周期行为(periodic behavior).

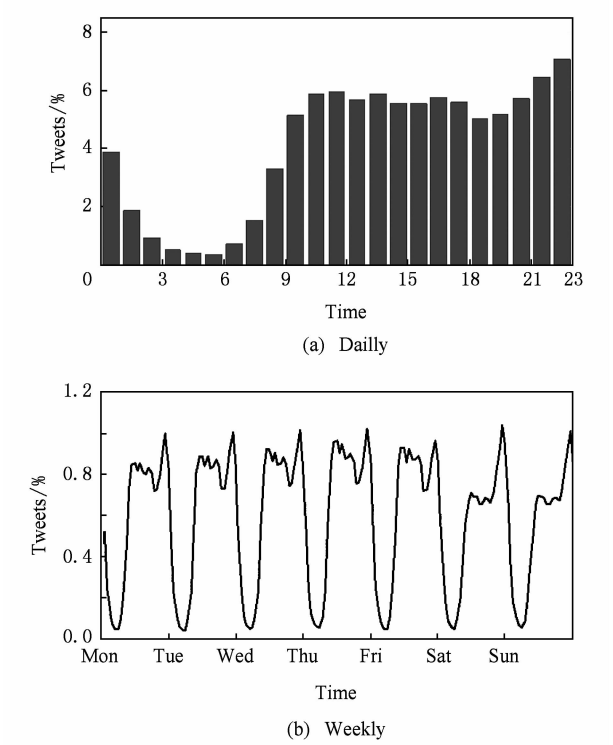


Fig. 6 Daily/weekly pattern of tweets in Sina microblogging.

图 6 博文日/周分布模式

3.2 用户博文分布

新浪微博用户博文分布是指用户发表、转载以及回复博文所呈现出的分布规律,该指标描述了博文信息产生、传播等行为的统计特征,用定量分析的方式描述了微博用户的基本操作行为.

本文首先对用户博文数目的分布进行了统计分析(如图 7 所示),可以发现用户博文数目存在重尾特性,近似服从威布尔分布.这主要是因为新浪微博中存在许多"哑"用户,这些用户虽然注册了微博帐号,但长期处于"潜水"或"未登陆"状态,从而使得博文数目较小的用户分布概率较大,而在信息传播过程中,这些用户节点不具有传播性,一般均对其忽略不计;同时由于各用户注册时间和发博频率的不同,使得博文数目在较大的范围内(10^3~10^5)呈现出重尾特性,且分布概率较小.

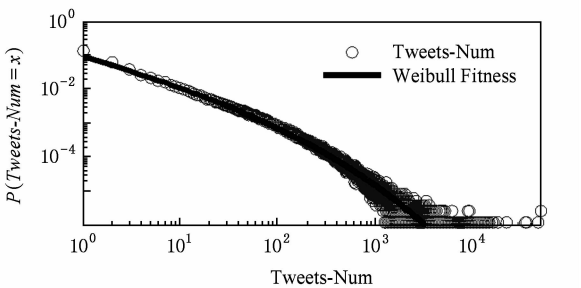


Fig. 7 The distribution plot of Tweets number.

图 7 用户博文分布曲线

其次,本文还对用户的博文平均转发数和平均评价数分布进行了统计分析,其中博文平均转发数表示用户平均每篇博文被转发的数量,平均评价数表示用户平均每篇博文被评价的数量,这两项指标不仅能够反映不同用户博文受欢迎的程度,而且还能反映微博网络的基本行为模式.其数学描述如下:

$$Ave\_Retweets_i=\frac{\#Retweets_i}{\#Post_i},$$
$$Ave\_Replys_i=\frac{\#Replys_i}{\#Post_i}. \tag{3}$$

图 8 表示博文平均转发数和平均评价数的概率分布函数与互补累积分布函数(complementary cumulative distribution function),可以看出转发和评价这两种行为基本一致,均近似服从幂率分布,而且两者的相关系数高达 0.73,说明转发和评价行为具有很强的相关性. 其中,大约有 50% 的用户博文平均转发数和平均评价数小于 1,这说明在新浪微博中,约有一半的用户对整个网络的信息传播贡献为零. 而且,在相同的转发(评价)数目上,转发概率要明显高于评价概率,这说明新浪微博用户更倾向于微博的转发而非评价,这也同时反映了微博的传播功能要优于用户之间的评价功能.

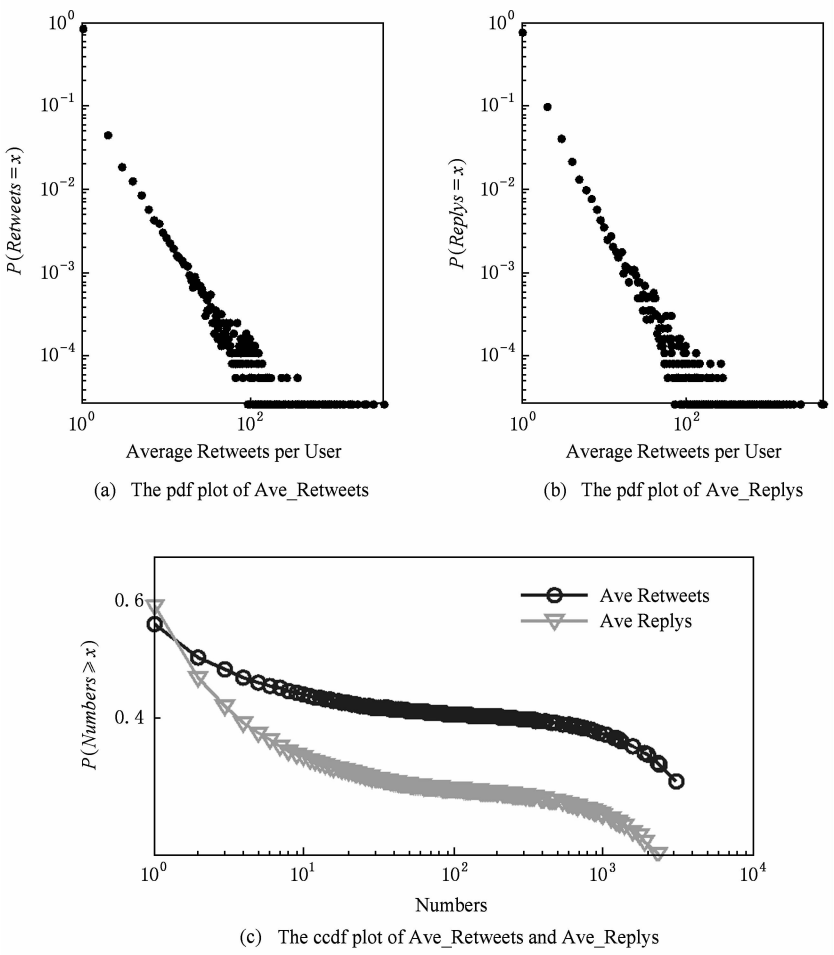


Fig. 8 The distribution plot of Ave_Retweets and Ave_Replys.

图 8 平均转发与平均评价数的分布曲线

4 结 论

随着移动通信网络和 Web 技术的发展,微博已成为人们进行信息共享和舆论传播的重要媒介. 为了实现微博系统的有效监测、引导、控制以及相应传播模型的建立,需要对已有的微博系统展开测量研究,了解其网络拓扑特征与用户行为特征等基本属性. 本文选取中国最大的微博系统——新浪微博——作为研究平台,采用自行设计的新浪微博爬行器 SMCrawler 对新浪微博实施数据采集,并结合已有的在线社会网络测量结果,对新浪微博的网络拓扑和用户行为特征进行了分析和比较. 主要发现包括:1)新浪微博网络具有小世界特性,其网络平均路径长度低于 Flickr 与 YouTube,而聚集系数高于 Twitter 和 YouTube,说明新浪微博具有更为紧密的网络结构;2)新浪微博网络的入度分布属于幂次分布,由于初始用户链出节点推荐机制的存在,其出度分布表现为某种分段幂率函数;3)与类似社会网络相比,新浪微博网络的出入度不具有相关性,这主要与新浪微博的出度设置上限和大量“明星节点”的推荐机制相关;4)新浪微博网络属于同配网络,即高度节点倾向于连接高度节点;5)新浪微博用户发博时间具有明显的日分布和周分布模式;6)新浪微博

用户博文数目分布表现为威布尔分布;7)新浪微博用户博文的转发和评价行为具有很强的相关性,且博文转发概率要高于评价概率,这说明相对于用户之间的通信,微博用户更倾向于信息的传播.

下一步的研究工作包括两个主要部分:1)对微博中的话题传播模式进行统计测量,建立微博网络影响力传播模型,实现对微博网络中话题传播的预测与控制;2)对微博用户发博、转发等行为特征进行时间序列分析,挖掘微博用户个体的行为模式,并通过聚类的方法实现微博用户的分类,最终实现基于用户行为的微博网络计算模型.

参 考 文 献

[1] Yang Nan, Gong Danzhi, Li Xian, et al. Survey of Web communities identification [J]. Journal of Computer Research and Development. 2005, 42(3): 439-447 (in Chinese)
(杨楠, 弓丹志, 李忻, 等. Web 社区发现技术综述[J]. 计算机研究与发展, 2005, 42(3): 439-447)

[2] Mislove A, Marcon M, Gummadi K P, et al. Measurement and analysis of online social networks [C] //Proc of the 7th ACM SIGCOMM Conf on Internet Measurement. New York: ACM, 2007: 29-42

[3] Guo L, Tan E, Chen S, et al. Analyzing patterns of user content generation in online social networks [C] //Proc of the 15th ACM SIGKDD Int Conf on Knowledge discovery. New York: ACM, 2009: 369-378

[4] Cha M, Mislove A, Gummadi K P. A measurement-driven analysis of information propagation in Flickr social network [C] //Proc of the 18th Int Conf on World Wide Web. New York: ACM, 2009: 721-730

[5] Cha M, Kwak H, Rodriguz P, et al. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system [C] //Proc of the 7th ACM SIGCOMM Conf on Internet Measurement. New York: ACM, 2007: 1-14

[6] Cheng Xu, Dale C, Liu Jiangchuan. Statistics and social network of YouTube videos [C] //Proc of the 16th Int Workshop on Quality of Service. Piscataway, NJ: IEEE, 2008: 229-238

[7] Java A, Song Xiaodan, Finin T, et al. Why we twitter: Understanding microblogging usage and communities [C] //Proc of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. New York: ACM, 2007: 56-65

[8] Huberman B A, Romero D M, Wu Fang. Social networks that matter: twitter under microsope [J]. First Monday, 2009, 14(1): 1-5

[9] Kwak H, Lee C, Park H, et al. What is twitter, a social network or a news media [C] //Proc of the 19th Int Conf on World Wide Web. New York: ACM, 2010: 591-600

[10] Sina Microblogging [OL]. [2011-03-07]. http://t.sina.com

[11] Lee S H, Kim P J, Jeong H. Statistical properties of sampled networks [J]. Physical Review E, 2006, 73(1): 016102

[12] XPath [OL]. [2011-03-07]. http://www.w3.org/TR/xpath, 2011

[13] Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks [J]. Nature, 1998, 393(6684): 440-442

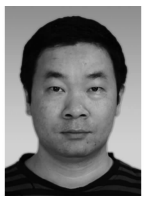
[14] Newman M E J. Random graphs with clustering [J]. Physical Review Letters, 2009, 103(5): 058701

[15] Boccaletti S, Latora V, Moreno Y, et al. Complex network: structure and dynamics [J]. Physics Reports, 2006, 424(4): 175-308

[16] Jiang Zhihong, Wang Hui, Fan Pengyi. Research on crawler-based measurement of large scale P2P IPTV systems [J]. Journal of Software, 2011, 22(6): 1373-1388 (in Chinese)
(姜志宏, 王晖, 樊鹏翼. 基于爬行器的大规模 P2P IPTV 测量研究[J]. 软件学报, 2011, 22(6): 1373-1388)

Fan Pengyi, born in 1984. PhD candidate of the National University of Defense Technology. His current research interests include social computing and network measurement.

Wang Hui, born in 1968. PhD, professor and PhD supervisor of the National University of Defense Technology. Member of China Computer Federation. His current research interests include social computing and information system engineering.



Jiang Zhihong, born in 1975. PhD and lecturer of the National University of Defense Technology. His current research interests include social computing and network measurement.

Li Pei, born in 1981. PhD of the University of Science and Technology of China, lecturer of National University of Defense Technology. His current research interests include social computing and network measurement.