在线社会网络的测量与分析

徐 恪[1)2)]，张 赛[1)]，陈 昊[1)]，李海涛[3)]

[1)](清华大学 计算机科学与技术系，北京 中国 100084)

[2)](清华信息科学与技术国家实验室（筹），北京 中国 100084)

[3)](西蒙弗雷泽大学 计算科学学院，温哥华 加拿大 V5A1S6)

摘 要 Facebook、Twitter、人人网和新浪微博等社交网站逐渐成为互联网上用户数量最多、最受欢迎的网站。近年来，国内外已有大量研究工作深入考察在线社会网络的拓扑结构和用户行为，这对理解人类的社会行为、改进现有的网站系统和设计新的在线社会网络应用具有重要意义。本文从测量角度对在线社会网络的拓扑结构、用户行为和网络演化等方面进行了综述，总结了常见的测量方法和典型的网络拓扑参数，着重介绍了用户行为特征、用户行为对网络拓扑的影响以及网络的演化。可以看出，随着研究的深入，在线社会网络的新特征逐渐被大家认识和理解，包括：好友少的用户的交流范围集中在小部分好友，而好友多的用户联系的好友更均匀；用户之间的交互减小了在线社会网络的聚类系数，使网络结构更松散；边的生成受优先连接和临近偏倚的共同影响；小社团倾向于和大社团合并，大社团倾向于分裂为两个规模相当的小社团等。

关键词 在线社会网络；测量；网络拓扑；用户行为；演化

Measurement and Analysis of Online Social Networks

XU Ke[1) 2)], ZHANG Sai[1)], CHEN Hao[1)], LI Hai-Tao[3)]

[1)](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

[2)] (Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China)

[3)](School of Computing Science, Simon Fraser University, Vancouver V5A1S6, Canada)

Abstract Social network sites, like Facebook, Twitter, Renren and Sina Weibo, are now becoming increasingly popular on the Internet. For the past few years, numerous research have been made to investigate the topological structure and user behaviors of online social networks, which is quite important for the understanding of human social behaviors, the improvement of current Website systems and the design of online social networks' new applications. This paper provides an overview of online social networks' topology, user behaviors and network evolution. It also summarizes several common measuring methods and typical topological features; highlights user behavior characteristics and their impacts on network topology, and the network evolution. The conclusion can been drawn that as research progresses, the new characteristics of online social networks are gradually recognized and understood: users with a smaller number of correspondents tend to interact more with a subset of correspondents, while users with a very large number of correspondents actually spread their activity evenly across all of the correspondents; users' interactions decrease the clustering coefficient and loose the connections between neighbors; edge creation is influenced by both preferential attachment and proximity bias; small communities tend to merge with large ones which tend to split into two comparable size communities.

Key words online social networks; measurement; network structure; user behavior; evolution

---

本课题得到国家支撑计划 (2011BAK08B05-02)、国家科技重大专项 (2012ZX03005001)、国家自然科学基金 (61170292) 、国家“863 计划” (2013AA013302)和国家“973 计划” (2009CB320501, 2012CB315803)资助. **徐 恪**，男，1974 年生，博士，教授，博士生导师，计算机学会高级会员，主要研究领域为新一代互联网体系结构、高性能路由器、网络科学与社会网络、物联网，E-mail: xuke@mail.tsinghua.edu.cn. **张 赛**，男，1988 年生，硕士研究生，主要研究领域为在线社会网络，E-mail: zs11235@gmail.com. **陈 昊**，男，1979 年生，硕士研究生，主要研究领域为在线社会网络，E-mail: jasonchenhao@yahoo.com.cn. **李海涛**，男，1983 年生，博士研究生，主要研究领域为在线社会网络、云计算、P2P，E-mail: lhtao0607@gmail.com.

1 前 言

互联网的出现极大地改变了人们的生活方式，增进了人与人之间的交流。用户通过各类互联网应用连接成一个庞大的网络。在这个网络中，用户是节点，用户之间的关系或交流是有向边，这是我们观察互联网应用拓扑的一般形式。当我们的观察角度或场景不同时，就可以形成不同的网络。比如在wiki中，当两个用户A和B编辑了同一个页面时，他们之间就存在一条边，这是因为同时编辑一个页面的用户常常会直接进行交流[1]。又如在万维网（World Wide Web，WWW）中，数百亿的网页就是节点，而网页之间通过嵌入的超链接相互连接。

在现实生活中，与他人交流是个人的主要行为，人们总是会通过某种关系相互连接形成一个**社会网络**（Social Network），也称为**社交网络**。社会网络分析是社会学的重要分支，关于它的工作最早可以追溯到Auguste Comte（1798-1857）和Georg Simmel（1858-1918）。与通常把社会看成个人的集合不同，他们把社会看成关系的集合。这就打破了以个人心理和行为为研究重点的传统，把人与人之间的关系提升到与个人属性同等的地位。现代社会网络分析的概念和方法来源于1934年Jacob Moreno的工作[2]，之后人们对Moreno的工作进行完善和系统化，逐渐形成了今天的社会网络分析理论[3, 4]。

互联网的兴起和普及为社会网络分析提供了难得的机遇。为了区别一般的社会网络，我们把互联网上的社会网络统称为**社会媒体网络**（Social Media Network，SMN）。例如，Leskovec和Horvitz计算了微软Messenger服务的用户网络参数[5]，他们发现，每个用户都有一些自己经常联系的密友（buddies），这些密友通常会列出自己的地理位置，这就形成了一个从真实世界到虚拟网络的社会关系的映射。Adamic和Adar[6]发现政治类博客可以清晰地分为两类，分别代表了普通人群不同的政治观点。除了这些由用户驱动而成的SMN，人们还研究了超链接网络（网页和超链接构成的网络）的结构特征（如Park和Thelwall的工作[7]）。

伴随着互联网的发展[8]，万维网上SMN的研究趋于成熟。随着Web2.0的普及而逐渐流行的Facebook、Twitter、新浪微博、人人网等社交网站（Social Network Site，SNS）使SMN的研究进入了另一个热潮。最新统计[1]表明，Facebook用户每天要观看约500年时长的YouTube视频，Twitter上每分钟要分享700个YouTube视频。截至2012年12月底，我国使用社交网站的用户规模为2.75亿，较上年底提升了12.6%，占网民比例近5成[2]。社交网站强调用户的直接参与，用户可以建立单向或者双向的好友关系，通过SNS与他人进行交流、获取消息、发布消息、上传照片和视频等。他们不再是被动接受信息的媒体受众，而积极地参与到网络活动中来，成为信息的制作者、分享者和传播者，他们更具自主性和互动性。根据SNS不同的功能和定位，基于SNS而形成的用户网络或是真实社会网络的虚拟映射，或是互联网用户自发形成的在线网络。这个网络不再像wiki一样需要通过进一步抽象而构成用户之间的连接，而是存在用户之间的直接交互。为了区别和强调这种差异，我们把社交网站用户构成的社会媒体网络称为**在线社会网络**，或**在线社交网络**（Online Social Network，OSN）。

在线社会网络的测量与分析是指通过采集、整理OSN的原始数据，利用复杂网络、社会网络和数据挖掘的理论方法和技术，挖掘和提取OSN的结构特征和用户行为特性。由于OSN的快速发展，无论从科学角度还是实用角度都迫切需要对其进行分析研究，以便更好地加以改进利用。另外，OSN大量的用户、丰富的数据、低成本的测量是传统社会学测量和采样所无法比拟的，把经典的社会网络分析理论和技术应用到OSN中我们可以得到关于社会网络更准确的结果。因此在线社会网络的测量与分析有助于人们理解和认识OSN的结构特征、演化和用户的社交行为，对研究开发新的基于OSN的应用具有很好的指导意义。

近年来学术界对OSN的研究日益深入，研究内容也多种多样，包括网络拓扑、用户行为、用户隐私和安全、系统架构、社团挖掘、信息传播等。无论是系统设计还是理论分析，对OSN的测量与分析是认识和研究OSN的第一步，因此本文主要从测量的角度讨论有关OSN拓扑结构及其演化、用户行为分析等方面的研究工作。OSN作为一类复杂网络，对它的研究延续了传统复杂网络的研究方法；同时，作为社会网络的特例，社会网络分析理

1 YouTube[EB/OL]. http://www.youtube.com/t/press_statistics/

2 中国互联网络信息中心（CNNIC）第31次《中国互联网络发展状况统计报告》[EB/OL]. http://www.cnnic.net.cn/gywm/xwzx/rdxw/2012nrd/201301/t20130115_38507.htm/

论也适用于 OSN。已有一些从复杂网络[9, 10]和社会网络[11]的角度论述 OSN 的专著，本文则从 OSN 自身出发，更关注其与传统社会网络的差别、拓扑结构的动态特征、用户行为特征和它们之间的关系。本文对近年来的相关工作进行了梳理，以测量结果和分析为重点，相关的理论研究较少涉及，感兴趣的读者可以参考[12-14]。

一个典型的 OSN 测量与分析的工作流程如下：首先利用某些测量方法采集 OSN 的原始数据，然后针对所关注的问题使用软件工具或者设计新的算法挖掘数据（统计）特征，最后分析所得结果的意义及其所体现的 OSN 特征。

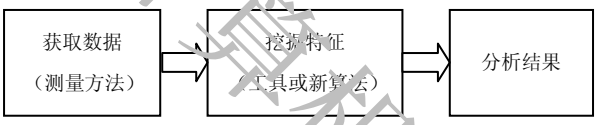


图 1 OSN 测量与分析流程

基于上述流程，本文组织如下：第 2 节简要介绍一些典型的数据集，总结 3 种常用的测量方法；第 3 节简单介绍常见的社会网络分析工具。第 4 到 6 节阐述现有主要的测量分析结果，包括常见的分析方法。第 4 节总结典型的网络拓扑参数和 OSN 在这些参数上的表现，并介绍用户之间的交互对 OSN 拓扑结构的影响；第 5 节从不同角度分析 OSN 上的用户行为特征，包括 SNS 应用上的用户行为和 OSN 中的信息传播；第 6 节考察 OSN 的演化问题，分析不同拓扑参数的动态特征，分析边的生成和社团演化的规律。最后一节对全文进行总结和展望。

2 在线社会网络的测量方法

数据是 OSN 研究的前提和基础。在未获得社交网站服务提供商支持的情况下，自主测量是开展 OSN 研究的第一步。在 OSN 测量方面已经出现了大量的论文，这些工作以测量为基础，较深入地统计、分析了 OSN 的很多特征，使我们对 OSN 有一个较直观的认识，同时也是进一步理论工作的基础。OSN 的测量往往与研究目的相辅相成，测量工作不可能一劳永逸，需要针对特定的研究目的应该选择适当的测量方式。

2.1 典型的社会网络数据集

很多研究者把自己测量的数据共享出来供其他研究者使用，比如加州大学欧文分校（University of California，Irvine，UCI）网络组共享了他们有关Facebook的测量数据[1]，Meeyoung Cha等人共享了他们对Twitter的测量数据[2]。我们选取斯坦福大学网络分析项目组（Stanford Network Analysis Project，SNAP）[3]收集的较丰富的数据集做简单介绍，供读者参考。

SNAP 收集了大约 50 个大型网络的数据，这些网络的规模（节点数和边数）从数万到几千万不等，种类各异，包括社会网络（social networks）、Web 图（web graph）、公路网络（road networks）、互联网网络（Internet networks）、论文引用网络（citation networks）、合作网络（collaboration networks）、通信网络（communication networks）。表 1 列出了一些社交网站的数据集，点代表用户，边（有向或无向）代表用户之间的好友关系。

还有其他一些网络数据集，比如含有社团的社会网络、基于地理位置的社会网络、论文作者的合作网络、Amazon 购物关系网等，因篇幅限制这里不再一一介绍。

2.2 常见的测量方法

虽然很多研究者共享了自己的测量数据，但我们不可避免地会受到测量数据本身的限制，比如如果我们关心消息在社会网络中的传播，我们就需要知道转发消息的来源，而很多已公开的数据集并不包含此类信息。因此主动测量是 OSN 研究中最常用的数据获取方法。OSN 的测量方法主要有以下几种：网络爬虫、抓取数据包、采集日志。

(1) 网络爬虫

网络爬虫是一种最常用的抓取网页数据的方法，当然也可以用来测量OSN。通过下载和分析网页源码，我们可以得到所需的数据。网络爬虫是典型的图遍历过程，例如以用户为中心，以好友关系为线索遍历好友关系网。虽然这种方法的逻辑结构简单，但其工作量大、数据形式单一，局限性较大。通过网络爬虫采集的数据往往是静态的，无法用来分析OSN的动态行为，所以这种方法常出现在早期对OSN拓扑结构的研究当中[15]。当然也有研究者基于SNS（如人人网）自身的功能利用网络爬虫抓取动态数据[16]，但此类方法并不通用。由于很多SNS（如Facebook[4]、Twitter[1]、人人网[2]、新浪微博[3]等）

1 Networking Group[EB/OL]. http://odysseas.calit2.uci.edu/wiki/doku.php/public:online_social_networks/
2 The Twitter Project Page at MPI-SWS[EB/OL]. http://twitter.mpi-sws.org/
3 Stanford Network Analysis Project(SNAP)[EB/OL]. http://snap.stanford.edu/
4 Facebook Developers[EB/OL].

提供了开放的API以方便开发者开发应用程序，我们也可以通过API采集（调用）用户数据[17]。这些数据较简洁，也会包含一些动态信息，但数据获取范围仍受API本身的限制。

(2) 抓取数据包

通过在网关抓取（HTTP）数据包，解析包（包头），我们可以获得丰富的实时信息。一些常用的包分析工具有Wireshark[4]、Sniffer Pro[5]、Bro IDS[18]等。这种网络层的测量方式非常有利于研究OSN用户的行为特征[19, 20]，这些信息恰恰是网络爬虫得不到的。但是为了获得较完整的数据，测量者必须得到ISP或网络运营机构的支持，解析包的工作也非常繁琐，对于认识OSN的拓扑特征较局限。可以看出网络爬虫和抓包是互补的测量方法，它们分别呈现了OSN的静态和动态特征。

(3) 采集日志

采集日志是一种较便捷的测量方法，它不需要分析网页源码、包头等复杂信息，而直接定位到一些关键信息，有助于我们把主要精力放在所要研究的问题和研究方法上。获得日志通常需要得到SNS服务商的协助，由他们直接提供原始日志[21]。

OSN的测量方法并不限于上述三种。为了获得对OSN的全面认识，很多工作[17, 19]同时采用了多种测量方法。鉴于很多SNS提供了应用开发平台，我们也可以通过开发SNS应用来获取丰富的用户数据[22]。文献[17, 19]均在被称为社交网络聚合器（social network aggregator）[6]的服务网站同时采集多个SNS的用户数据，该网站相当于一个“网关”，用户可在该网站同时登陆多个SNS帐号，这种混合数据便于对不同OSN进行比较。

3 常见的社会网络分析工具

除测量方法外，对OSN拓扑结构的研究与传统社会网络分析理论有很多共同之处，因此延续了较多社会学的研究方法。对社会网络的分析由来已久，目前已出现了很多社会网络分析（Social Network Analysis，SNA）软件[7]。对于一些简单的网络拓扑参数，如聚类系数、度分布等，我们可以直接利用适当的SNA软件对OSN进行分析。

以邻接表或邻接矩阵的方式输入原始网络，SNA 软件可以自动计算多种网络拓扑参数，也可以直观地展示网络结构。一些 SNA 软件可以做预测分析，包括基于连接等网络现象预测个人层面的输出（称为同伴影响或传染建模），基于个人层面的现象预测连接或三元组的形成等网络输出（称为同质性模型），基于网络现象预测其他网络现象，比如根据 0 时刻某一个三元组的形成预测 1 时刻某个连接的形成。

SNA 软件通常包括基于图形用户接口（Graphical User Interfaces，GUIs）的包或为编程语言设计的 API。常见的 GUI 包有 NetMiner、UCINet、Pajek、GUESS、ORA 和 Cytoscape。定位商业用户的 GUI 包有 Orgnet、Keyhubs 和 KXEN。其他 SNA 平台，比如 Idiro SNA Plus，定位于电信和在线游戏等有大规模数据分析需求的企业。

常见的 SNA 脚本工具有：NetMiner（Python）、statnet 组件包（R）、igraph（R 与 Python）、NetworkX 库（Python）以及 C++中大规模网络分析的 SNAP 包。虽然对于初学者来说，这些工具包上手有一定难度，但相比于私人软件，其更新速度更快，功能也更完备，并且有详细的说明文档供使用者阅读学习，更重要的是它们都是开源的。特别地，微软开发的 NodeXL[23]是一款基于 Microsoft Office Excel 的社会网络分析工具，容易上手，功能较强大，具体见表 2。

可视化对于理解和分析网络性质特别重要。可视化有助于直观地认识网络特征，因此很多网络分析工具都带有网络可视化功能，包括网络布局、节点颜色和大小等。NetMiner、igraph、Cytoscape、NetworkX 和 NodeXL 都能生成高质量的网络图。

表 2 列举了一些常见的 SNA 软件。

4 在线社会网络的拓扑结构

对在线社会网络拓扑结构的研究具有较强的应用背景：由于OSN的连边在一定程度上代表了用户之间的信任关系，因此基于OSN的拓扑信息可以增强系统的安全性和保护用户隐私，例如可靠电子邮件系统（Reliable E-mail，RE）[8]通过测量邮

http://developers.facebook.com/docs/reference/apis/

1 Build with Twitter[EB/OL]. https://dev.twitter.com/docs/streaming-apis/

2 人人网开放平台[EB/OL]. http://wiki.dev.renren.com/wiki/API/

3 新浪微博开放平台[EB/OL]. http://open.weibo.com/wiki/API 文档_V2/

4 Wireshark[EB/OL]. http://en.wikipedia.org/wiki/Wireshark/

5 Sniffer Pro[EB/OL]. http://www.sniffer.net.cn/

6 Social network aggregation[EB/OL]. http://en.wikipedia.org/wiki/Social_network_aggregation/

7 Social network analysis software[EB/OL]. http://en.wikipedia.org/wiki/Social_network_analysis_software/

8 Reliable E-mail[EB/OL]. http://www.pittsburgh.intel-research.net/projects/completed/reliable-email.html/

表 1　SNAP 收集的社交网站数据集举例

名 称	类 型	节点数	边 数	描 述
ego-Facebook	无向	4,039	88,234	Facebook（已匿名）的社交圈
ego-Gplus	有向	107,614	13,673,453	Google+的社交圈
ego-Twitter	有向	81,306	1,768,149	Twitter 的社交圈
soc-Epinions1	有向	75,879	508,837	Epinions.com 的信任网络（who-trusts-whom network）
soc-LiveJournal1	有向	4,847,571	68,993,773	LiveJournal 在线社会网络
soc-Slashdot0811	有向	77,360	905,468	2008 年 11 月开始的 Slashdot 社会网络
soc-Slashdot0922	有向	82,168	948,464	2009 年 2 月开始的 Slashdot 社会网络
wiki-Vote	有向	7,115	103,689	Wikipedia 的投票网络（who-votes-on-whom network）

表 2　社会网络分析工具举例

名 称	主要功能	输入格式	输出格式	平 台	许可证和付费	注 释
CFinder	探测并可视化社会网络中的社团	.txt	.txt, .pfd, .ps, .svg, .svg, .emf, .gif, .raw, .ppm, .bmp, .jpg, .png, .wbmp	Linux, Mac OS X, Windows	非商用、免费	基于派系过滤方法发现并可视化网络中的社团，允许可定制的可视化和在社团之间切换，包含命令行版本。
Graph-tool	图分析和可视化的 Python 组件	GraphViz(.dot), GraphML	GraphViz(.dot), .bmp, GraphML, .cmap, .eps, .fig, .gif, .gif, .gtk, .ico, .jpeg, .cmapx, .jpeg, .pdf, .plain, .png, .ps, .ps2	GNU/Linux, Mac	免费（GPL3）	高效率的图分析 Python 组件。其核心数据结构和算法用 C++实现，着重使用了基于 Boost Graph Library 的模板元编程。
igraph	大规模网络的分析和可视化	.txt (edge list), .lgl, .gml, .ncol, .graphml, .net	.txt (edge list), .lgl, .net, .dot, .gml, .ncol, .graphml	Windows, Linux, Mac OS X	开源（GNU GPL）	大规模网络分析的 C 库。包括经典图算法的快速实现和一些网络分析技术，如社团结构搜索、粘性阻塞、结构洞和模体计数估计。拥有适用于 R、Python 和 Ruby 的高层接口。
Jerarca	社会网络分析，社团结构，网络分层聚类	.txt (List of links)	Text, output to MEGA, output to Cytoscape, hierarchical tree in Newick format	Linux, Windows	开源（GNU GPLv3）	分层聚类算法套件。可以根据循环分层聚类有效地把无向无权图转换为层次树，还可以探测社团结构。
NodeXL	社会网络分析和可视化	email, .csv (text), .txt, .xls (Excel), .xslt (Excel 2007), .net (Pajek), .dl (UCINet), GraphML	.csv (text), .txt, .xls (Excel), .xslt (Excel 2007), GraphML, .dl (UCINet)	Windows XP/Vista/7	免费（Ms-PL）	Excel 2007/2010 插件和 C#/.Net 库。与 Excel 整合，以图表形式输入有向图，可以计算各种网络拓扑参数，支持从 email、Twitter、YouTube、Facebook、WWW 和 Flickr 提取社会网络，支持多种网络可视化布局。
UCINET	社会网络分析工具	Excel, DL, text, Pajek .net, Negopy, proprietary (##.d & ##.h), Krackplot	Excel, DL, Krackplot, text, Pajek .net, Mage, Metis, proprietary (##.d & ##.h)	Windows	共享软件	可处理 32,767 个节点，包括中心度、子群识别、角色分析和基本图理论等社会网络分析方法。拥有很强的矩阵分析能力，如矩阵代数和多变量统计。

件发送方和接收方的社会网络距离来协助垃圾邮件检测；网络连边也意味着用户之间具有相似的兴趣，例如Google Co-op[1]和PeerSpective[2]等即利用用户社会网络成员的点击对其感兴趣的内容进行排名和预测。人们对网络的拓扑参数已达成共识，各种拓扑参数从不同方面体现网络的结构特征。这一节我们首先介绍一些常见的拓扑参数，然后总结OSN在这些参数上的表现，最后考察用户的交互行为对OSN拓扑结构的影响。

4.1 静态拓扑结构

表 3 总结了几种常见的网络拓扑参数。下面我们举例介绍 OSN 的拓扑特征。

(1) 平均路径长度

一个网络称为具有小世界性质（small-world properties）[10]，如果对于固定的节点平均度$\langle k\rangle$，平均路径长度L的增加速度至多与网络规模N的对数成正比，即$L \propto \ln N$。相比于 Web，OSN 的平均路径长度和直径都很短[15]；相比 20 世纪 60 年代 Stanley Milgram 提出的六度分离（six degree of separation）假设[24]，OSN 缩小了人们之间的距离。但这并不代表 OSN 的平均路径长度最短，事实上很多现实网络（如蛋白质网络、铁路网、淡水食物网等）的平均路径长度要比 OSN 短得多，具体可参考文献[25]。

(2) 聚类系数

在好友关系网中，一个人的两个好友很可能彼此也是好友，这就是网络的聚类特征。好友关系网是无权网络，对于带权网络也有一个相应的聚类系数的计算公式，具体见[26]。把 OSN 的聚类系数与用不同算法（Erdös-Rényi 随机图[27]和幂律随机图[28]）生成的网络聚类系数进行对比，发现 OSN 的聚类系数要远大于理论模型的聚类系数[15]。这说明在 OSN 中，人们更喜欢通过共同好友相互认识。另外小出度的节点其聚类系数更高，如图 2 所示，说明拥有较少好友的用户紧密地连接在一起。

一个网络称为小世界网络（small-world network）[10]，如果它有较短的平均路径长度和较高的聚类系数。由上述讨论可以看出 OSN 是典型的小世界网络。1998 年，Watts 和 Strogtz[29]提出了 WS 小世界模型来刻画网络的这一结构特征，我们利用 WS 模型可以生成小世界网络。

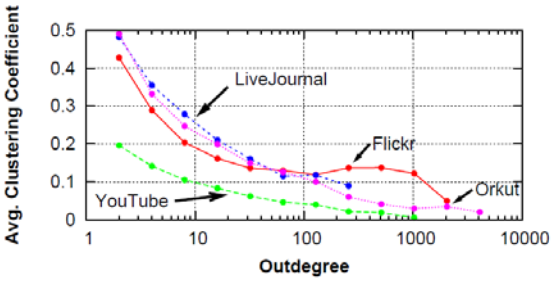


图 2 不同出度节点的聚类系数[15]

(3) 对称性与度分布

OSN 的拓扑结构具有对称性（symmetry）。所谓对称性，是指一个有向图中，边的方向是双向的。比如在 YouTube 中，有 79.1%的好友关系是双向的[15]，这说明用户之间倾向于相互关注。这种对称性与线下社会网络中的情形类似[30]。

完全随机网络的度分布近似为 Poisson 分布，其形状在远离峰值$\langle k\rangle$处呈指数下降，这类网络也称为均匀网络（homogeneous network）。OSN 的一个重要的结构特征是度分布呈幂律形式，也称为无标度（scale-free）分布，即$P(k) \propto k^{-\gamma}$。这类网络称为非均匀网络（inhomogeneous network）。但 OSN 节点度的幂律分布与其他网络不同。由 Barabási 和 Albert[28]提出的 BA 无标度网络模型可以很好地刻画网络的无标度特征，BA 网络的度分布函数可由幂指数为 3 的幂律函数近似描述，Web 的幂指数约为 2.5[31]，而 OSN 约为 1.5[15]，说明 OSN 在度分布方面更加不均匀。

(4) 同配性

大多数情况下，OSN的同配系数r都大于 0，比如人人网的同配系数为 0.15[16]，Facebook为 0.17[32]，Flickr[3]为 0.202[15]，LiveJournal[4]为 0.179[15]，这一特特征把OSN从其他幂律网络中区别开来。Web和Internet的同配系数都小于 0，分别为-0.067 和-0.189[33]。但这并不是说所有OSN都是同配的，我们会在下一节解释这一点。无标度性质和同配性说明OSN中有一些紧密连接的度较大的核心，它们把整个网络连接起来，度较小的节点分布在网络的边缘。

与同配系数相似的另一个参数是平均邻居连接数k_{nn}。图 3 显示出当节点度大于 100 时，人人网用户的k_{nn}与节点度之间呈正相关性，即度较大的节点倾向于与度较大的节点连接，这与 OSN 的同配性一致。

1 Google Co-op[EB/OL]. http://www.google.com/coop/
2 PeerSpective[EB/OL]. http://peerspective.mpi-sws.org/
3 Flickr[EB/OL]. http://www.flickr.com/
4 LiveJournal[EB/OL]. http://www.livejournal.com/

表 3　常见的社会网络拓扑参数

名 称	公 式	注 释	描 述
直径（diameter）	$D=\max\limits_{i,j} d_{ij}$	d_{ij}为连接节点i和j之间的最短路径上的边数	两个节点之间距离的最大值
平均路径长度（average path length）	$L=\dfrac{1}{\frac{1}{2}N(N+1)}\sum\limits_{i\geq j} d_{ij}$	N为网络节点数	包含了节点到自身的距离（为零）
节点i的聚类系数（clustering coefficient）	$C_i=\dfrac{2E_i}{k_i(k_i-1)}$	k_i为节点i的邻居节点的个数，E_i为这些邻居节点之间的实际边数；$0\leq C_i\leq 1$	刻画节点i的邻居的连接程度
聚类系数（clustering coefficient）	$C=\dfrac{1}{N}\sum\limits_{i=1}^{N} C_i$	N为网络节点数；$0\leq C\leq 1$	所有节点聚类系数的平均值，刻画网络的聚类特性；$C=0$当且仅当所有节点都为孤立节点，$C=1$当且仅当网络是全互连的
度分布（degree distribution）	$P_k=\sum\limits_{k'=k}^{\infty} P(k')$	累积度分布函数（cumulative degree distribution function），$P(k')$表示一个随机选定的节点度恰好为k'的概率	表示度不小于k的节点的概率分布；如果$P(k)\propto k^{-\gamma}$，则符合幂指数为$\gamma-1$的幂律：$P_k\propto\sum\limits_{k'=k}^{\infty} k'^{-\gamma}\propto k^{-(\gamma-1)}$
同配系数（assortativity coefficient）	$r=\dfrac{M^{-1}\sum_i j_i k_i-\left[M^{-1}\sum_i \frac{1}{2}(j_i+k_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2+k_i^2)-\left[M^{-1}\sum_i \frac{1}{2}(j_i+k_i)\right]^2}$	j_i和k_i分别为第i条边的两个端点的度，M为网络边数；$-1<r<1$	描述网络中的节点和与其度相同的节点连接的倾向性；若$r>0$，网络是同配的（assortative），表示节点倾向于和与其度相同的节点连接；若$r<0$，网络是异配的（disassortative），表示节点倾向于和与其度相异的节点连接
平均邻居连接数（average neighbor connectivity）	$k_{nn}(k)=\sum\limits_{k'=1}^{k_{\max}} k'P(k'\|k)$	$P(k'\|k)$表示度为k的节点连接到度为k'的节点的概率	给定度节点的邻居节点的平均度
k-核（k-core）	——	反复去掉度小于或等于k的节点后，所剩余的子图	——
核数（coreness）	——	若一个节点存在于k-核，而在$(k+1)$-核中被移去，那么此节点的核数为k	核数表明节点在核中的深度，即便一个节点的度很高，它的核数也可以很小，比如N个节点的星形网络

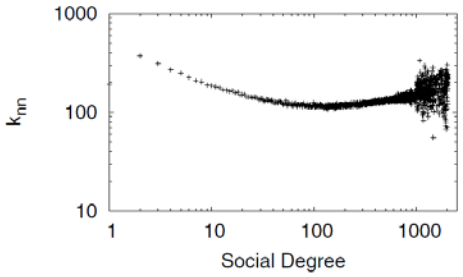


图 3 人人网的k_{nn}分布[16]

(5) k-核与核数

同配性使 OSN 形成了一个“核心”：核心在网络连接中起重要作用，移去核心，网络会变得支离破碎；核心的直径很小。描述核心要用到k-核，具体定义见表 3。Mislove 等人[15]用在 Web 使用过的图分析方法[31]挖掘 OSN 中的核心，结果见图 4，其中 SCC 指网络中的最大连通子图。从图 4 看出，当移去 10%度较大的节点时，网络会分裂成上百万个非常小的 SCC。所以整个网络是由这 10%的核心节点连接而成。此结论的另一个证据是随着移去度较大的节点，网络的平均路径长度逐渐增加。

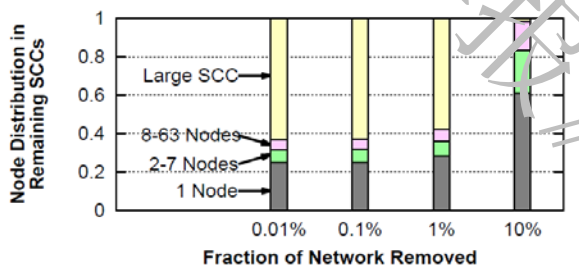


图 4 移去度较大的节点后网络的分离情况[15]

4.2 用户之间的动态交互对拓扑结构的影响

我们在 4.1 节之所以使用“静态拓扑”这个词是因为有一个问题：用户之间建立好友关系之后，是否经常联系呢？之前我们只关注 OSN 中好友关系的声明，而忽略了真实情况下用户之间的交互；很有可能 A 与 B 建立好友关系之后少有联系，但 C 和 D 虽然不是好友却经常交流。因此有必要研究用户的交互行为对网络拓扑结构的影响。

研究者针对由不同的用户交互行为而构成的网络提出了多种描述形式，比如活动网络（activity network）[34]、活动图（activity graph）[22]、隐式交互图（latent interaction graph）[16]等。为了统一概念，也为了区别之前的好友关系网，我们把所有基于用户交互行为而非好友关系建立起来的网络统称为**活动网络**，其拓扑结构称为**活动拓扑结构**。

按如下方式定义活动网络：每个用户为一个节点，节点i到节点j的一条有向边$i \to j$表示i对j有一个行为（比如i向j留言）；有向边$i \to j$的权值ω表示i对j的行为的强度。这样定义的活动网络并不是好友关系网的子集，因为非好友的用户之间仍可以交互。因此活动网络是一个比好友关系网更真实的网络，对研究信息在 OSN 中的传播有着重要意义。

4.2.1 活动拓扑 vs 静态拓扑

Chun 等人[34]以用户之间的留言为实例研究活动网络，他们的研究表明：虽然活动网络是一个有向、加权的网络，但其与好友关系网的结构特征相似，比如活动网络中节点的入度和出度分布都呈幂律形式，且二者分布非常相似，这与 Mislove 等人[15]的结论相同。

当仅考虑对称边时，活动网络比好友关系网更具同配性。以k_{nn}衡量同配性，文献[26]提供了带权网络k_{nn}的计算公式，以Cyworld[1]为例，研究发现[34]：当节点度$k \geq 500$时，k_{nn}呈发散趋势，且好友关系网更发散；当$k < 30$时，网络呈异配性；当$30 \leq k \leq 500$时，网络呈同配性。这说明网络中度较小的节点与度较大节点之间是非对称的，大量节点都去关注“名人”节点，产生所谓的“名人效应”（celebrity effect）[2]。特别地，诸如微博[3]等以弱关系为主社会网络，其名人效应更明显[35]。

分别考察活动网络和好友关系网的k-核，可以明显地看出二者的差别。如图 5 所示，活动网络和好友关系网的k_{nn}曲线在$k < 34$时变化趋势相同，之后活动网络的k-核节点数下降得比好友关系网快。这一点与 Leskovec 等人[5]的结论相同，说明好友关系网中含有不活跃（不联系）的边，而这些边把网络核心连接起来

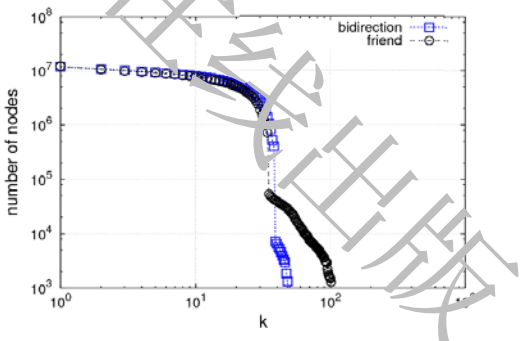


图 5 k-核分析[34]

如果说好友之间的相互留言是一种显式的交互，那么浏览页面则是隐式的交互方式。Jiang 等人

1 Cyworld[EB/OL]. http://www.nate.com/cymain/
2 Special: Micro blog's macro impact[EB/OL]. http://www.chinadaily.com.cn/china/2011-03/02/content_12099500.htm/
3 Microblogging in China[EB/OL]. http://en.wikipedia.org/wiki/Microblogging_in_China/

通过考察人人网用户的访问记录详细考察了这一问题[16]。如果定义用户访问他人主页的次数为该用户的消费，则1%最流行的用户与1%消费最多的用户有 9%的重叠，说明在人人网中有一部分非常流行且活跃的用户。一个有意思的现象是，陌生用户与好友访问个人主页的情况（累积分布）类似，这说明了隐式交互的重要性，因此仅考察好友关系网是不够的，陌生用户在OSN的构成中也非常重要。

以用户为节点，用户之间的浏览关系为有向边同样可以构造活动网络（隐式交互图）。下面总结这个活动网络与好友关系网的异同。

(1) 度分布

隐式交互图的边数（240,408）大于显式交互图（27,347）（以用户之间的评论为有向边），隐式和显式交互图的边数都明显小于好友关系网（753,297）。这说明人人网中有大量的不活跃用户，他们不浏览其他人的页面，也不与他人交互。尽管有这些不同，隐式交互图的度分布仍然呈幂律分布（入向$\gamma = 3.5$，出向$\gamma = 3.39$）。

(2) 聚类系数

同样因为人人网中有很多静默边存在，隐式（0.03）和显式（0.05）交互图的聚类系数都小于好友关系网（0.18），移去这些边导致邻居节点之间的连接松弛。因为很多陌生用户的访问使隐式交互图更加松散，其聚类系数最小。

(3) 同配性

由于很多用户访问流行用户，所以隐式交互图是异配的（$r = -0.06$），而好友关系网（$r = 0.23$）和显式交互图（$r = 0.05$）都具有同配性，这反驳了先前关于活动网络比好友关系网更具同配性的说法[32]。

(4) 平均路径长度

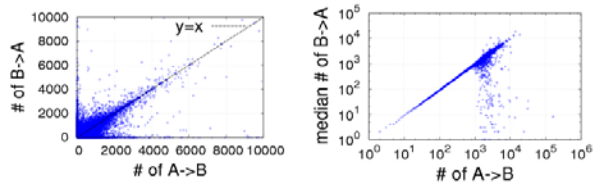
隐式交互图的平均路径长度（4.02）介于显式交互图（5.43）和好友关系网（3.64）之间。由于平均节点度和度较大的节点数的减小，整个网络的连接性能变差，这导致隐式和显式交互图的平均路径长度增加。

4.2.2 活动拓扑特征

OSN 用户之间的交互可以归结为相互作用（reciprocal interaction）。相互作用是社会合作发展的主要机制[36-38]，刻画了关系的强度[30]。虽然性别、宗教、年龄或者文化差别会影响人们之间的亲密感[39]，但相互作用遍及原始生活[40]和社会系统[41]的每一个关系。基于相互作用，我们可以进一步考察网络的活动拓扑特征。

(1) 相互性（reciprocity）

为了考察用户之间的相互性 [1]，可以把一对用户之间的交互次数画成点状图。图 6(a)描述了Cyworld用户之间发消息与接收消息的情况，取中值和对数之后得图 6(b)。从图中可以看出明显的对称性，图 6(b)可以很好地用$x = y$拟合。Hemelrijk[42]把相互性分为三类：相对的（relative），绝对的（absolute）和定性的（qualitative）。因此活动网络中的相互性接近于绝对的相互性，Hemelrijk推测绝对的相互性源于个体之间无差别的交互能力，在Cyworld中体现为用户可以无障碍地浏览他人页面。



(a) no. of received messages vs. no. of written messages between a pair of users　(b) median no. of received messages vs. no.of written message in log scale

图 6 用户之间发消息 vs 接收消息[34]

我们可以用互易系数（reciprocity coefficient）[34]量化相互性。如果用户i访问了用户j，则$v_{ij} = 1$，否则$v_{ij} = 0$。我们定义互易系数ρ为

$$\rho = \frac{\sum_{i \neq j}(v_{ij} - \bar{v})(v_{ji} - \bar{v})}{\sum_{i \neq j}(v_{ij} - \bar{v})^2}, \quad (1)$$

其中$\bar{v} = \frac{\sum_{i \neq j} v_{ij}}{N(N-1)}$，$N$为节点数。显然，$-1 \leq \rho \leq 1$，$\rho$值越大表示相互性越强。隐式浏览意义下人人网的互易系数为 0.23[16]，显式评论意义下为 0.49[16]，而Cyworld活动网络的互易系数为 0.78[34]，WWW为 0.5165[43]，电子邮件网络为 0.231[43]，Slashdot[2]为 0.28[44]，Twitter为 0.58[45]，Wikipedia[3]为 0.32[46]。

(2) 差异性（disparity）

目前为止我们只是衡量用户之间的交互强度，但没有考察一个用户的交互对象在其好友上的分布。直观上讲，一个用户的好友数越少，他对待好友越平等；反之受其精力所限，他无法均匀地与所有好友交互。一个出度为k、入度为k_{in}的节点i的差异（disparity）$Y(k,i)$度量了用户i的活动在其好友

1 这里需要与 4.1 节的对称性区分开，对称性强调好友关系，是静态的，相互性强调用户之间的交互，是动态的。
2 Slashdot[EB/OL]. http://www.slashdot.org/
3 Wikipedia[EB/OL]. http://en.wikipedia.org/

上的分布[47, 48]，定义为

$$Y(k,i)=\sum_{j=1}^{k}\left\{\frac{w_{ij}}{\sum_{l=1}^{k_{in}}w_{li}}\right\}^2, \quad (2)$$

其中，$Y(k)$为所有出度为k的节点的$Y(k,i)$的平均值。如果一个节点的出向边的权值与其入向边的权值是可比的，则$kY(k)\sim 1$；如果一个节点大部分的交互都位于一条出向边或入向边上，则$kY(k)\sim k$。图 7 显示了 Cyworld 活动网络$kY(k)$与k的关系：当$k\le 500$时，$kY(k)\sim k$；当$k>500$时，$kY(k)$开始发散，特别地，当k达到 1000 时，$kY(k)$降为 1，这说明$kY(k)\sim 1$。图 7 说明，一个好友数量少的用户倾向于与少部分好友交互，而好友数量多（大于 1000）的用户联系的好友更多而且更均匀。这一点与我们的直观猜测：好友多的用户似乎并没有受精力所限而有差别地对待好友

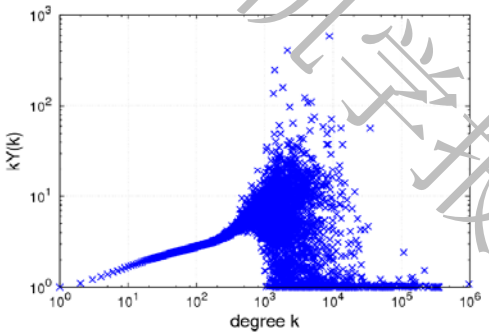


图 7 差异性[34]

(3) 网络模体（network motifs）

由前面的讨论我们可以看出，相互作用使网络拓扑呈对称性，但交互行为在好友中却不是平均分布的。用户之间具体的交互模式如何呢？网络模体（network motifs）是指三个人之间 13 种可能的交互模式，或称为三元组重要性剖面（tiad significance profile, TSP），Milo 等人[49]提出了一种基于网络模体的网络分类方法。基本思想是，计算网络中 13 种网络模体的比例，并与其对应的随机化网络相比较，根据比较结果对网络进行分类。模体i的Z-分数（Z-score）表示它在网络中所占的比例，定义为

$$Z_i=\frac{N_{real,i}-m(N_{random,i})}{\sigma_{random,i}}, \quad (3)$$

其中，$N_{real,i}$表示实际网络中模体i的数量，$m(N_{random,i})$和$\sigma_{random,i}$分别表示对应随机化网络中模体i的平均值和标准差。标准化Z-分数[50]为

$$Z_i/(\sum Z_i^2)^{0.5}。 \quad (4)$$

Chun 等人[34]使用模体探测工具 FANMOD[51]对 Cyworld 好友关系网进行模体分析。图 8 显示了模体分析的结果：传递模体（模体 9、10、12、13）所占的比例很大，而非传递模体（模体 4、5、6）很少出现在 Cyworld 中。这些结果与其他社会网络的结果一致[49, 50]。模体 1 和 2 的正规化Z-分数与我们对社会网络的期望不同：模体 1 是广播型（非传递），对应垃圾信息传播者（spammer），一个用户给两个互不相识的人发消息而没有得到回复；模体 2 是汇聚型（非传递），对应权威或名人，两个互不相识的用户给同一个用户发消息而没有得到回复。

我们可以根据 TSP 对网络进行分类，具有相似 TSP 的网络组成一个网络超家族（superfamily）。Milo 等人[50]把 19 个不同类型的网络划分成了 4 个超家族。这些研究表明，相同类型的网络不仅具有相同的网络模体，而且各个模体在网络中也具有相似的相对重要性；另外，一个网络超家族中可能包含规模差异很大、功能极其不同的网络。

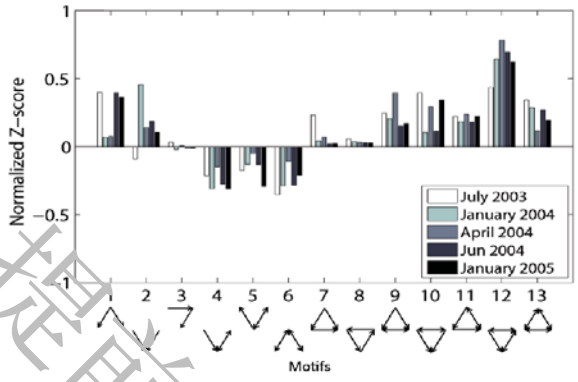


图 8 5 个不同时间活动网络的正规化Z-分数[34]

4.3 小结

本节主要总结 OSN 的拓扑结构特征和用户行为对网络拓扑的影响：

(1) OSN 是小世界网络，具有较短的平均路径长度和较大的聚类系数。

(2) OSN 是无标度网络，度分布呈幂律形式。

(3) 大部分 OSN 是同配网络，度较大的节点倾向于与度较大的节点相连。OSN 中有一些紧密连接的度较大的节点，它们把整个网络连接起来，度较小的节点分布在网络边缘。

(4) OSN 用户的好友关系并不代表他们之间活跃的交互，好友关系网中存在大量的静默边和非活跃用户。用户的不可见行为（如浏览页面）减小了网络聚类系数，使网络结构变得松散，也增加了平均路径长度，使网络的连接性能变差。

(5) OSN 用户的交互行为具有相互性。好友少

的用户其交流对象集中在少部分好友，而好友多的用户联系的好友更多更均匀。

(6) OSN 中的传递模体占很大比例，广播型和汇聚型模体也占较大比例。

5 在线社会网络的用户行为分析

有了对 OSN 拓扑结构的认识，人们还希望进一步考察用户的行为，即用户在 SNS 上做些什么？有些什么规律？对用户行为特征的研究可以帮助改善网络营销机制，改善广告投放模式，添加更为准确和友好的个性化因素，特别是对面向用户的推荐系统的设计有重要意义。由于用户行为的复杂性，这方面的研究往往针对某一问题展开。

在本节中我们首先总结 OSN 的基本用户行为和研究方法，然后分别从社交网站应用和信息传播两个角度深入认识用户的行为特征。

5.1 在线社会网络的基本用户行为

OSN 的活动拓扑是对用户行为的反映，用户行为改变了 OSN 的拓扑结构，后者反映了前者的特征。图 9 显示了 Cyworld 用户好友数与用户发消息的数量中值之间的关系[34]：当好友数小于 200，二者呈正相关，Pearson 相关系数为 0.6235；当好友数大于 200 时图像开始发散，Pearson 相关系数为 0.00913。这说明，200 个好友是 OSN 用户的交互上限，这是一个很有趣的结论。对于人类社交能力的上限，Dunbar 给出的解释是 150，即基于人类大脑新皮质的进化，一个人可维持的人际关系数是 150[52]。Bialik[53]指出，技术依赖型社会网络对社会性修饰提出挑战。图 9 告诉我们 OSN 用户最多可以维持 200 人的社交关系，大于 Dunbar 的结论。

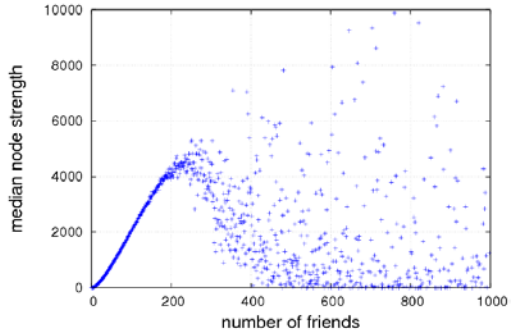


图 9 Cyworld 用户交互能力上限[34]

5.1.1 会话

一个典型的社交网站包括个人主页、日志、相册、应用等，我们首先关注的是用户针对网站不同功能的使用情况和行为特征。一般情况下，为了考察细致的用户行为，我们需要通过抓取和解析 HTTP 请求和应答包以获得点击流（clickstream）信息。这方面的代表工作是[19]和[20]。

获得点击流数据之后的第一个工作是定义会话（session）。可以根据两点来定义一个会话：(1) 用户关闭了浏览器或者注销登录；(2) 用户在 20 分钟内没有任何页面活动。用户会话持续时间的累积分布函数（cumulative distribution function，CDF）均呈重尾分布（heavy-tailed distribution）。

对于在线用户数量的时间特性，Benevenut 等人[19]发现：在线用户数量呈白昼模式（diurnal pattern），下午 3 点左右达到峰值；每个时间段，至少有 50 个用户在使用社交网站，峰值为 700；周末的用户使用量小于工作日。白昼模式也发现于 Facebook[54]、Facebook 应用[22, 55]和其他用户生成内容（user-generated content，UGC）网站中[56, 57]。

5.1.2 会话的统计特征

为了从系统的角度进一步认识用户到达和离开的动态特征，我们可以从不同方面考察用户会话和HTTP请求的统计特征。首先定义一个时间序列$t(i), i = 1, 2, 3, \cdots, t(i)$表示第$i$个会话到达的时间，$a(i)$定义为第$i$个和第$i+1$个会话的间隔时间，即$a(i) = t(i+1) - t(i)$。以Orkut[1]为例[19]，图 10(a)显示了$a(i)$的互补累积分布函数（Complementary Cumulative Distribution Function，CCDF），它逼近于如下函数：

$$f(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{-(\log(x)-\mu)^2}/2\sigma^2, \qquad (5)$$

其中，$\mu = 2.245$，$\sigma = 1.133$。

定义$l(i)$为一个会话期间 HTTP 请求的数量，图 10(b)显示了$l(i)$的频率边缘分布。这也是一种重尾分布，即大多数的会话有很少的 HTTP 请求，而较少的会话有大量的 HTTP 请求。这种分布逼近于幂律分布$\beta x^{-\alpha}$，其中$\alpha = 1.765$，$\beta = 4.888$。幂律分布说明了 OSN 用户的会话时间具有很大的差异性，这与 Huberman 等人在 Web 上的发现类似[58]。

图 10(c)显示了一个会话内请求的间隔时间的统计特征，它同样符合式(5)的函数形式，其中$\mu = 1.789$，$\sigma = 2.366$。Benevenut 等人发现，会话持续时间的长短与会话的时间无相关性[19]，这说明会话的持续时间并不遵守白昼模式，它只是 OSN 用户行为的一个基本性质。这种函数拟合的工作可以用来模拟用户的会话过程。如果细分会话的种

1 Orkut[EB/OL]. http://www.orkut.com

类，不同种类的会话其统计特征也保持一致[20]。

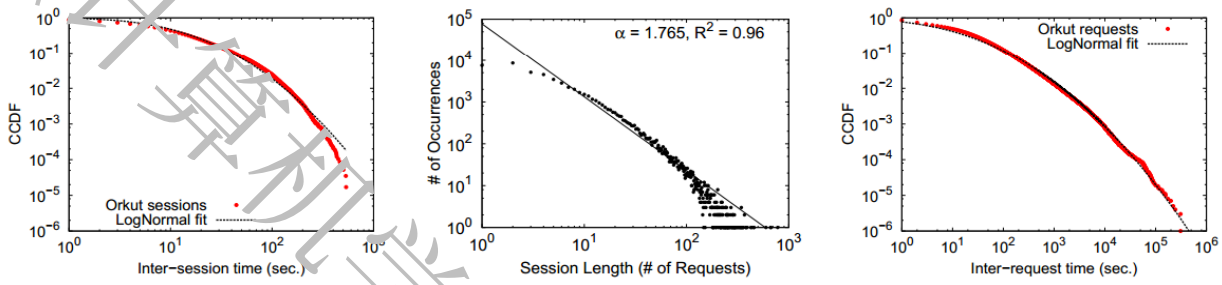
5.1.3 点击流模型

可以通过建立点击流模型（clickstream model）[19]来严格刻画 SNS 中用户的典型行为。建立点击流模型主要分为两步。一是从用户的点击流中区分出典型行为。Benevenut 等人[19]首先从 HTTP 请求中区分出 41 种用户活动，并把这些活动分为以下几类：搜索、剪贴簿、消息、奖状、视频、照片、用户资料和好友、社团和其他。

建立点击流模型的第二步是建立一个一阶 Markov 链，状态代表不同类别的行为，转移概率为用户下一步要做某件事情的概率，具体见图 11。图 11 可以给出很多信息，比如一个用户刚登录 Orkut 帐号，他要做的第一件事情很可能（64%的可能性）是浏览别人的个人资料和好友列表。依靠这个模型可以合理地模拟 Orkut 的用户行为。

5.1.4 用户行为与网络距离

从 4.2 节我们已经知道，用户行为会影响网络的拓扑结构。同样地，网络结构也会影响用户行为。图 12 显示了 Orkut 中不同的用户行为与用户距离之间的关系，比如发生在两跳之外的用户身上的行为



(a) Inter-session time (b) Session length (c) Inter-request time

图 10 Orkut 会话特征及其最佳匹配函数[19]

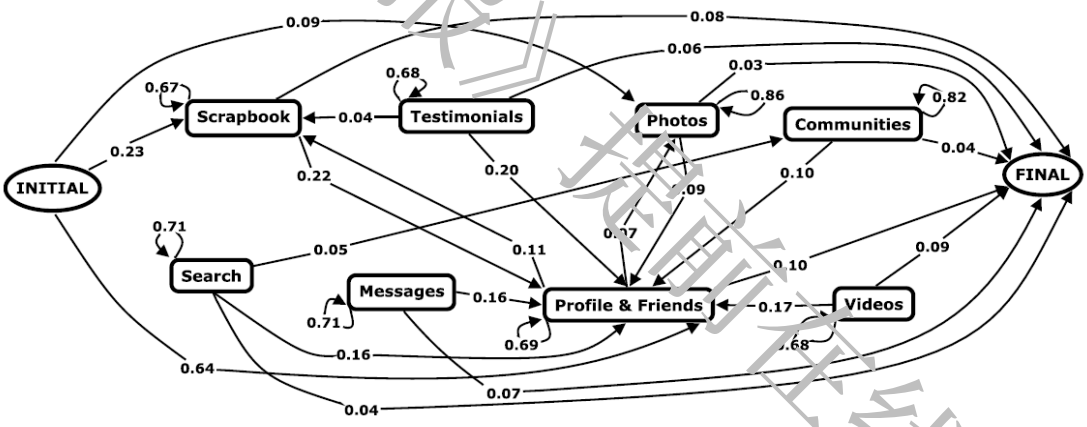


图 11 Orkut 用户行为的点击流模型[19]

数量很少，说明用户一般只与自己的好友交互，而用户与非好友交流的主要方式是写消息。这一点可以与 Jiang 等人[16]的结论作对比：人人网中，个人页面的访问者中大多数也是 2 跳之内的用户，但好友数小于 100 的用户其主要访客是陌生用户。

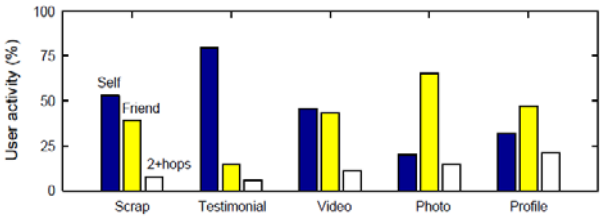


图 12 行为类型与网络距离[19]

通过统计一个用户浏览某一页面之前的所在位置，可以考察是何种行为导致用户浏览此页面[19]：浏览自己的主页最有可能导致用户浏览其直接或间接好友的页面，这与 Orkut 个人主页上包含直接或间接好友的信息有关；大量的访问行为与浏览直接好友的页面有关，这说明用户通常从他的直接好友那里获取信息[59-61]。

5.1.5 隐性行为的比重

同 Jiang 等人[16]的工作一样，点击流的重要优势在于它可以捕捉网页上不可见的行为，比如浏览行为。那么这些不可见行为占多大比重呢？Benevenut 等人[19]通过区分可见和不可见行为，统

计了 Orkut 用户在 12 天内与他人交流的人数（包括浏览页面），他们发现：(1) 12 天内用户平均只联系了 3.2 个用户，如果只考虑可见行为仅为 0.2。Wilson 等人[32]发现 Facebook 中 60%的用户一年内几乎没有和其他人联系，考虑到这里只考察了 12 天的数据，得到如此小的数值是合理的。(2) 所有行为与可见行为的统计结果差别很大，说明用户的主要行为是浏览页面，占所有用户行为的 92%。

5.2 社交网站应用上的用户行为

应用（application）是社交网站的重要组成部分，它拓展了用户体验，增强了用户之间的交流，也增加了用户黏性。Facebook、人人网等 SNS 都有开放的应用开发平台，大量第三方的应用增加了网站流量。Facebook 的网站流量在其开放开发平台后的一周内就增加了 30%[22]，Twitter 的流量在其开放 API 之后增加了 20 倍[62]。这些数据足以说明 SNS 应用的流行程度。SNS 应用也增强了非好友之间的交互[22]，因此研究应用相关的用户行为很有意义。根据我们的调研，基于应用的用户行为的研究主要来自于 Atif Nazir 等人[22, 55, 63]的工作。

5.2.1 Facebook 应用的统计特征

定义日活跃使用量（daily active usage，DAU）为一天内至少一次访问某一应用的用户数。Nazir 等人[22]发现，Facebook 应用的使用情况满足 Pareto 原则或 80-20 规则：20%的应用占有了约 69%的日访问数。

图 13 显示出DAU更细致的统计特征：如果以DAU衡量某一应用的流行度，我们发现，应用流行度开始呈幂律分布，以指数分布截尾（truncated tail），Cha等人[64]指出YouTube和Daum[1]等网站的UGC视频流行度呈相似的分布。

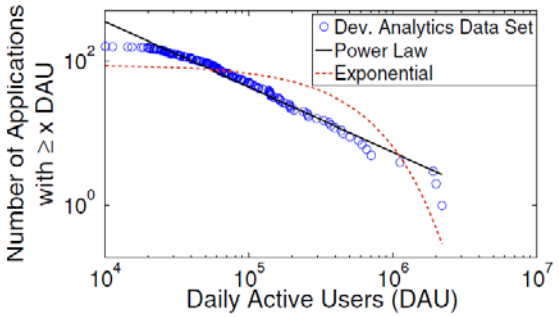


图 13 DAU 分布[22]

幂律流行度在线上以及现实的社会网络中经常可以见到[25, 65]。这里我们对幂律分布做一些解释。一个最直接的解释是优先连接过程（preferential attachment process）产生幂律分布。以 SNS 应用流行度的幂律分布为例，优先连接预示着一个新用户使用某一应用的概率与这个应用现有的用户数成比例。例如，Facebook 维护了一个公告栏实时更新好友动态，这其实是为应用做广告：比如好友 A 订阅了应用 B，该动态就会显示在公告栏上，这就为应用 B 做了广告，这个广告与 B 现有的订阅者（A）产生了联系。正是利用这一点（优先连接），Gjoka 等人[66]设计了一个简单的仿真模型来模拟用户安装应用的过程，生成用户与应用之间的二分图，边表示用户安装应用。该模型的输入为应用列表、每个应用的安装数和用户数，输出为二分图。

幂律分布的指数分布的截尾也是一个老话题[25]，我们这里也做一些解释。文献[67, 68]说明了老化（aging）或生育力（fertility）和优先连接是如何导致以指数分布截尾的幂律分布。生育力指排除优先连接的影响，应用所拥有的最少的初始订阅者（subscriber）。在 Facebook 中，生育力与社会网络自身有关，因此为了挖掘应用的用户潜力，必须要求一定的法定人数（阈值）。老化是指应用在某一时间之后就会变得过时，对用户的吸引力下降。Mossa 等人[69]给出了幂律分布指数截尾的另一个解释——信息过滤（information filtering）。给定一个有限空间，比如 YouTube 主页或 Facebook 公告栏，较少使用的应用信息被过滤掉了，因此一个典型的幂律分布是不会达到的。他们也把信息过滤作为 YouTube 和 Daum 流行度分布的可能解释。

5.2.2 应用的流行度变化

SNS 应用和视频网站的视频都有流行度的概念，因此我们可以考察流行度的变化情况。SNS 应用第 X 天的排名漂移（ranking drift）定义为 $|(RankonDay0) - (RankonDay X)|$。Nazir 等人[22]首先把应用按流行度分为 4 组，然后分别计算每天排名漂移的平均值，具体见图 14。我们可以看出，最流行（top 5%）的应用其排名漂移最小，流行度越低，漂移越大，说明应用随流行度的增加其排名趋于稳定。

1 Daum[EB/OL]. http://www.daum.net/

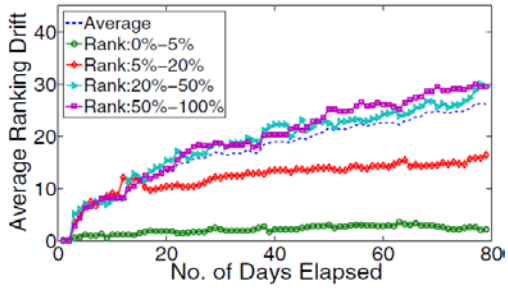


图 14 Facebook 应用的流行度变化[22]

如果我们缩小时间尺度，比如考察每 5 分钟的流行度变化，如上方法可能无法捕捉到更精细的变化，这时我们可以利用 Spearman 等级相关性（Spearman's rank correlation）[70]。Kossinets 和 Watts[71]利用 Spearman 相关性计算社会网络里节点度的稳定性，Cha 等人[72]利用它评估 IPTV 频道流行度的动态变化，他们发现，虽然不同流行度的频道所占有的观看人数的分布具有稳定性，但频道的流行度变化却非常剧烈，并且以天为单位呈现周期性。这说明有一些热门频道的流行度与时间有关，比如儿童频道的观看人数与学校的作息时间密切相关。

5.2.3 应用的活动网络与社团结构

用户在应用上的交互也构成一种关系，这样我们得到另一种 OSN 活动网络，继而可以和第 4 节一样考察该网络的各种拓扑参数。网络中的社团（community）[73]是指这样一组节点，这组节点构成一个连通子图，它们之间的连接要密于它们与外界节点的连接。设e_{ij}是连接社团i中节点与社团j中节点的边，令$a_i = \sum e_{ij}$，网络的结构系数（structure coefficient）定义为$\sum(e_{ij} - a_i^2)$。社团结构称为强的（strong）如果结构系数大于 0.3[73]。

Nazir 等人[22]计算了 Facebook 三个应用 Fighters' Club（FC）、Got Love（GL）和 Hugged 所构成的活动网络的众多参数，得出以下结论。

(1) 大多数用户处在单一的连通分支中。最大连通分支中的用户数与应用 trace 的总用户数成比例，说明随着数据集的增加，越来越多的节点落入最大连通分支中。

(2) 除了 FC，Hugged 和 GL 都具有强的社团结构。FC 最大社团所拥有的用户数占了总数的 72.6%，而 GL 和 Hugged 仅占不到 10%。另外，FC 社团规模的分布与 GL 和 Hugged 差异较大。

(3) 把同一社团中不同地理位置的数目称为社团的地理多样性（geographical diversity of communities）。Nazir 等人发现社团具有较强的地理多样性，社团规模与社团中国家的数目没有关系。

(4) 应用构成的活动网络具有较高的聚类系数和较短的平均路径长度，即是小世界网络。其度分布呈幂律分布，且 FC 的幂律分布比 GL 和 Hugged 更明显。这些都与 OSN 拓扑特征相同。

由以上结论可以看出，SNS 应用有一些通有的性质，比如地理多样性、连通分支的单一性，也与应用的种类有关，比如游戏类应用明显不同于其他应用。进一步地，Nazir 等人[63]比较了基于 SNS 应用的用户活动图（user activity graph，UAG）与好友关系网拓扑结构的不同，发现现有的图模型算法[74, 75]无法准确生成 UAG。最后根据对馈赠（gifting）应用的分析，提出了新的算法，对基于 SNS 应用的 UAG 进行建模。

5.3 在线社会网络中的信息传播

信息传播是用户行为活动的结果和表现。研究信息扩散和传播的机制和特征有助于改善网站系统设计，提高搜索算法和推荐算法的效率，对消息推广、病毒式营销和舆情控制有重要的实际意义。在线社会网络中的信息传播是一个较大的课题，有很多理论上的工作（比如Leskovec等人一系列的工作 [1]），这方面的研究综述可以独立成文，限于篇幅和本文主题，我们仅简要介绍这方面的相关研究工作，感兴趣的读者可具体参阅下面提到的文献。

OSN 是一个由信息所有可能的传播路径组成的复杂网络，信息传播反过来又促进 OSN 结构的变化。无论是研究拓扑结构还是用户行为，我们最关心的还是信息的传播和扩散。复杂网络中的信息传播机制是一个古老而又富有挑战的课题。流行病学的研究已有较长的历史，流行病传播模型[76]也是少有的较完备的传播理论之一。在此基础上有了复杂网络的流行病临界值理论和传播动力学分析。然而这并不意味着我们可以把流行病传播模型照搬到 OSN 中来，因为二者的传播机制和形式都很不相同。因此需要研究新的传播模型刻画 OSN 中的信息传播。

已有相当丰富的理论工作研究信息流与社会网络结构之间的作用。Granovetter[77]提出了一个线性阈值模型，只有当某一用户有足够多（超过阈值）的邻居节点采用了某一创新，他才采用该创新。Watts[78]基于稀疏Erdös-Rényi随机图提出了一个全局相继模型，发现即使在很少的初始接受者的情况

1 Leskovec' publications[EB/OL]. http://cs.stanford.edu/people/jure/pubs/

下也会发生全局相继现象。Watts考察了在用户敏感性具有均匀阈值的情况下这种相继过程发生的条件。Karsai等人[79]考察了小世界网络中信息传播的时间演化，发现权重-拓扑相关和个人的突发活动模式是限制扩散的主要因素。Steeg等人[80]分析了Diggs[1]上的信息扩散，发现Diggs网络的高聚类结构限制了信息的传播范围。

随着OSN数据的逐渐丰富，人们开始了数据驱动的分析，测量社会网络连接上信息传播的模式。Gruhl等人[81]基于博客关键字研究了博客空间中的信息扩散，他们正是利用了传统的流行病模型刻画博客里的信息传播。Adar和Adamic[82]进一步扩展了流行病模型在线上信息传播的应用。Bakshy等人[83]研究了Second Life[2]（一个多人虚拟游戏）社会网络里的内容传播，发现社会网络对内容的接受起着重要作用。Sun等人[84]通过研究Facebook页面的传播发现传播链都很长，且都始于相当多的用户。而Rodrigues等人[85]发现在传播树的形状和发布者与订阅者的影响上，口口相传（word-of-mouth）的信息传播有很多不同，相比于深度，传播树的形状更宽。Gomez-Rodriguez等人[86]考察了扩散路径的问题，提出了一个算法计算近似最优的有向边以最大化影响。Ghosh和Lerman[87]比较了一系列对影响力的度量，声称一个基于中心度的测量是对影响力的最优预测。Scellato等人[88]研究了如何利用从社会级联（social cascade）提取出的地理信息来提高内容传输网络中多媒体文件的缓存。Wang等人[89]发现社会和机构内容对人们传播信息的速率和传播对象影响很大。Galuba等人[90]提出了一个传播模型用来预测Twitter中哪个用户可能提到哪个URL。还有一些工作区分了垃圾信息（spam）[91, 92]和网络钓鱼（phishing）[93]在Twitter上的扩散形式。

对 OSN 信息传播的建模最实际的应用是预测信息的扩散趋势。Cha[64]等人发现 YouTube 视频的早期与晚期的浏览量存在线性相关（相关系数为0.84）。Szabo 等人[94]也发现了类似的规律，并借此提出了 3 种模型来预测内容的流行度。这些模型依据 10 天的历史浏览可以提前 30 天预测视频的流行度，误差为 10%。

基于这些工作，Li 等人[65, 95, 96]系统研究了外部视频分享网站（video sharing sites，VSSes）视频在 OSN 中的传播情况，他们发现：

(1) OSN 极大地放大了视频流行度的偏度（skewness），人人网中 0.31%最流行的视频占了80%的浏览量[95]，如图 15 所示，而 YouTube 中 10%的最流行的视频占了 80%的浏览量[64, 97]；很多 VSS 中流行的视频不一定在 OSN 中流行[95]。

(2) OSN 中的视频流行度分布呈标准的幂律形式，而 YouTube 视频流行度的分布呈幂律形式而以指数分布截尾[64]；OSN 中的视频在达到浏览量峰值前通常需要 2 到 3 天的静默期，而 YouTube 视频在发布到网上之后很快达到浏览峰值[65]。

(3) 与单纯 VSS（比如 YouTube）的情况不同，经典的时间序列（time series）预测模型，如自回归滑动平均模型（Autoregressive Integrated Moving Average，ARIMA）[98, 99]、多元线性回归（Multiple Linear Regression，MLR）[100]和 k-近邻回归（k-Nearest Neighbors Regression，kNN）[101]等，预测 VSS 视频在 OSN 中的流行度的效果很差[96]，如图 16 所示，特别是对于预测早期的浏览峰值和之后的浏览突发（burst），而这都是 OSN 视频传播很常见的现象。

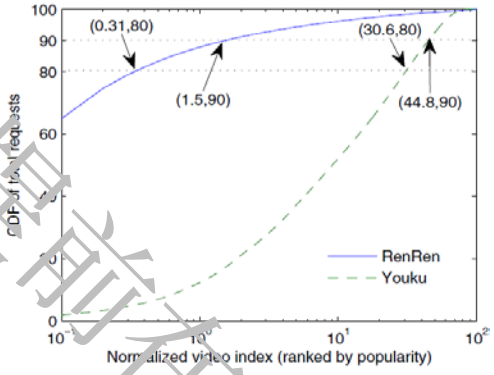


图 15 视频流行度的累积分布[95]

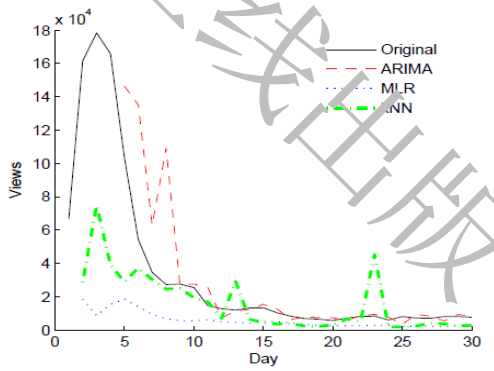


图 16 三种模型对视频浏览时间序列的预测[96]

为了解决这个问题，Li 等人[96]基于视频的固有吸引力和底层传播结构提出了新的视频流行度预测方案——SoVP[96]。他们基于人人网真实的 trace 数据对视频流行度进行预测，结果见表 4。从表 4

1 Diggs[EB/OL]. http://www.digg.com/
2 Second Life[EB/OL]. http://secondlife.com/

可以看出，SoVP 的预测效果明显好于kNN 和 MLR。

表 4 某类型视频的预测结果[96]

	day 2	day 3	day 4	day 5	day 6
Real	161502	178356	165886	103254	54181
kNN	28589	74841	38917	28957	37157
MLR	18349	8533	15487	18628	13956
SoVP	199132	134302	196740	124730	45674

5.4 小结

本节主要介绍了 SNS 的基本用户行为、应用上的用户行为和 OSN 中的信息传播，总结如下：

(1) SNS 的用户行为呈白昼模式。

(2) SNS 用户会话时间具有很大的差异性，且不遵守白昼模式，它是 OSN 用户行为的基本性质。

(3) SNS 用户的平均联系人数很少，他们一般只与自己的好友联系，人人网中好友数小于 100 的用户的主要访客是陌生用户。

(4) 浏览页面是 SNS 的主要用户行为，Orkut 中浏览行为占用户行为的 92%。

(5) 越流行的 SNS 应用其排名漂移越小，排名越稳定。基于应用的活动网络拓扑特征与 OSN 拓扑特征相似。

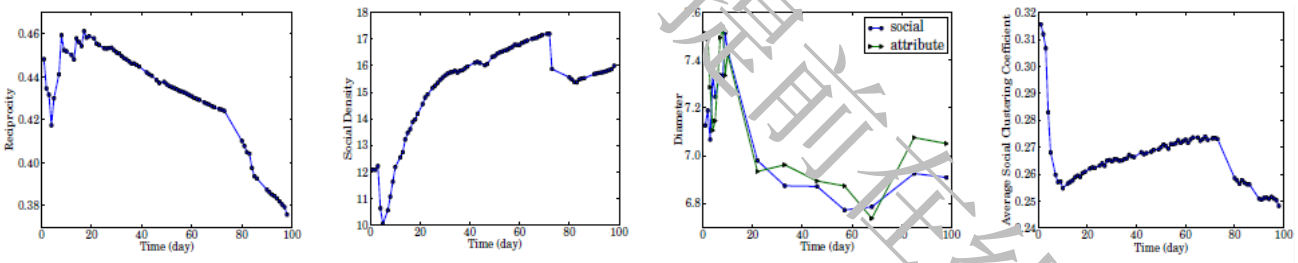
(6) OSN 放大了视频流行度的偏度，视频流行度分布不符合 80-20 原则。

6 在线社会网络的演化

我们已经讨论过在线社会网络的静态拓扑结构和用户行为对拓扑结构的影响，这些仍然都是静态意义上的考察，研究对象为某一时刻网络的快照（snapshot）。那么这些性质是如何随时间变化的呢？这就是 OSN 的演化。对 OSN 演化的研究工作考察网络基本拓扑参数的变化，也有自己关心的问题：节点和边的加入有什么规律？这为网络建模提供了实际依据。我们将会看到，随机图、BA 网络等现有的网络模型无法精确地模拟 OSN 的演化，虽然它们都能展现 OSN 某些特有的拓扑特征。

6.1 基本拓扑参数的变化

平均路径长度和聚类系数等拓扑参数随时间的变化并没有一致的规律，与具体的社交网站有关。以 Google+为例，图 17 显示了若干拓扑参数的变化，可以看出，每种参数的变化基本分为三个阶段，分别对应 Google+的早期阶段、稳定阶段和开放阶段。对于密度，在论文引用网络、作者从属网络[74]和 Facebook[102]中密度随时间增长，Flickr[103]的密度变化呈增—减—增的趋势，而电子邮件网络[71]的密度趋于稳定。对于直径，论文引用网络、自治网络和作者从属网络[74]，Flickr 和 Yahoo! 360[1][103]，Cyworld[105]的直径随时间减小。对于聚类系数，电子邮件网络[71]和人人网[106]的聚类系数趋于稳定。对于同配性，我们知道传统的OSN都是同配的，例如Flickr、LiveJournal和Orkut的同配系数均为正[15]，而Google+的同配系数却不总是正的[104]，如图 18 所示。Google+的同配系数大部分处在 0（大于 0）附近，而逐渐减小。Gong等人[104]解释到：Google+是两种网络的混合体，传统OSN和发布-订阅网络（如Twitter、新浪微博等），前者的同配系数通常为正，而后者为负。开始时，传统的OSN结构在Google+中占优势，随后二者逐渐融合，最后发布-订阅网络占主导，这说明Google+与Twitter的网络结构越来越相似。



(a) Reciprocity (b) Social density (c) Diameter (d) Social clustering coefficient

图 17 Google+四种拓扑参数的变化[104]

从上面的陈述可以发现，不同网络拓扑参数的变化有很大不同，这往往与网站类型和网络事件有关。Zhao 等人[106]考察了两个 SNS 合并对网络结构的影响，他们发现虽然网站合并会对 OSN 的拓扑结构造成干扰，但各种参数或者很快收敛或者很快保持稳定的变化趋势。

1 Yahoo! 360[EB/OL]. http://360.yahoo.com/

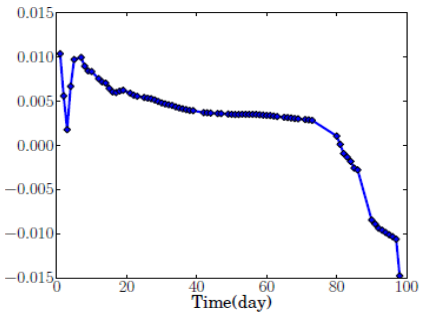
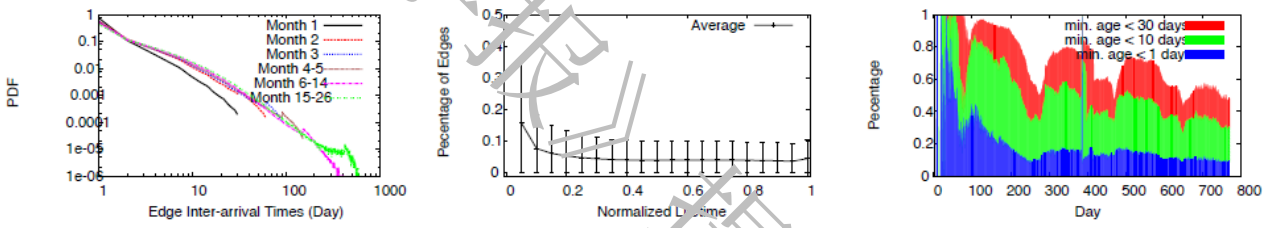


图 18 Google+同配系数的变化[104]

6.2 边生成的时间特征

类似于 5.1 节的点击流，两个节点之间生成一条边可以归结为两个节点之间有一个事件发生。对该事件有一些常见的研究角度，比如事件发生的时间间隔、事件持续的时间、事件发生与节点属性的关系等。按照这个思路，OSN 节点之间边的生成有什么统计规律呢？Zhao 等人[106]考察了人人网节点之间边的生成的若干时间特征。

图 19(a)说明生成边的时间间隔呈幂律分布，Leskovec 等人[107]发现生成边的时间间隔呈以指数分布截尾的幂律分布。Kumar 等人[103]发现在 Flickr 和 Yahoo！360 中绝大多数边生成的时间间隔小于一天。边的生成与用户的生命期有什么关系呢？从图 19(b)可以看出，大多数用户都是在早期建立好友关系，随着节点生命期的增加，边的生成速率趋于稳定。而 Leskovec 等人[107]发现了几种不同的情况：Delicious 和 Answers 边的生成速率具有和上面相似的情况，但 Flickr 和 LinkedIn 边的生成速率随节点生命期的增加而逐渐增加。Zhao 等人[106]进一步考察了边的生成与节点年龄之间的关系，引入网络层面的时间，即在不同时刻不同年龄的节点生成边的比例。从图 19(c)中可以看出网络时间对边的生成有很大的影响，不同时间不同年龄的节点生成边的比例不同。现有的网络模型（如 WS 小世界模型、BA 网络等）都假设只有新加入的节点生成边，但这种假设仅在网络初期成立。



(a) Distribution of Edge Inter-arrival Times　(b) Edge Creation vs. Normalized Life-time　(c) Node Age and Edge Creation

图 19 边的生成的时间特征[106]

6.3 优先连接

我们已经知道，优先连接是指新节点倾向于连接到度较大的节点[28, 108]。优先连接模型要求某一节点选择连接到度为k_i的节点i的概率p_i与k_i^α成比例，其中α为常数。当$\alpha \approx 1$时为线性优先连接，当$\alpha > 1$时为超线性优先连接[108]。线性优先连接是无标度网络的基本特征[28, 108]，很多现实的OSN都表现出线性优先连接，例如Delicious[1]、Flickr、Yahoo! Answers[2][107, 109]，但LinkedIn[3][107]除外。

优先连接是度分布呈幂律形式的直接解释，也是很多网络模型（如 BA 网络）的基本思想。现实 OSN 的优先连接如何随时间变化呢？Mahanti 等人[17]把节点按年龄分为年轻和年老两组（以 50 天为界），分别考察了不同组合的优先连接情况，他们发现：(1) 节点的年龄对优先连接有较大影响，年老节点的线性优先连接较明显；(2) 年轻节点选择年老节点的线性优先连接最为明显。

为了进一步地考察优先连接的动态特性，即α的变化，定义一条边选择度为d的节点作为目标节点的概率$p_e(d)$为[107]

$$p_e(d) = \frac{\sum_t \{e_t(u,v) \wedge d_{t-1}(v) = d\}}{\sum_t |v : d_{t-1}(v) = d|}, \quad (6)$$

其中$\{e_t(u,v) \wedge d_{t-1}(v) = d\} = 1$，当边$e_t(u,v)$的目标$v$的度为$d$，否则$\{e_t(u,v) \wedge d_{t-1}(v) = d\} = 0$。线性优先连接意味着$p_e(d) \propto d$，而一般的 PA 模型有$p_e(d) \propto d^\alpha$。我们想知道$p_e(d) \propto d^{\alpha(t)}$是否是一个好的拟合。Zhao 等人[106]利用均方误差（Mean Square Error，MSE）拟合$p_e(d)$和$d^{\alpha(t)}$，他们发现：(1) 网络快照情况下$p_e(d)$和d^α的拟合情况很好，说明 OSN 边的生成符合优先连接；(2) 选择度较大的节点作为目标节点时，

1 Dlicious[EB/OL]. http://del.icio.us/
2 Yahoo! Answers[EB/OL]. http://answers.yahoo.com/
3 LinkedIn[EB/OL]. http:///www.linkedin.com/

$\alpha(t)$总是大于随机选择目标节点的情况，且差值为常数 0.2；(3) $\alpha(t)$随时间减小，两年内从 1.25 降到 0.65，说明在OSN早期，优先连接较明显，随着网络规模的增加，边的生成对目标节点的度的依赖逐渐减小。

6.4 临近偏倚

优先连接关注的是OSN的整体性质，边的生成有什么局部性质呢？Kossinets等人[71]发现临近偏倚（proximity bias）影响着生成边时目标节点的选择。简单来说，临近偏倚是指节点倾向于与它附近的点建立连接。三元闭包（triadic closure）声称如果两个节点有两跳远，那么它们之间很可能建立连接，这是临近偏倚的一种特殊情况。临近偏倚也出现在现实的OSN中[17, 107, 109]。优先连接与临近偏倚的关系如何呢？以FriendFeed[1]（FriendFeed网络的平均路径长度为 4，度较大的节点与很多节点都相距较近）为例[17]，首先利用线性优先连接模拟边的生成，如图 20 所示，从图中可以发现三元闭包贡献了 82%的新生成的边，这说明优先连接在某种程度上造成了临近偏倚，而实际曲线与优先连接模拟的曲线的不同也说明了临近偏倚对优先连接有补充。Mahanti等人[17]设计了一个简单的边的生成算法模拟临近偏倚的作用：首先根据线性优先连接选择目标节点的度d，然后在度为d的节点中选择最近的节点作为目标节点。从图 20 可以看出这个模拟算法生成的曲线与真实情况更接近。因此在设计网络模型时，边的生成需要同时考虑优先连接和临近偏倚。

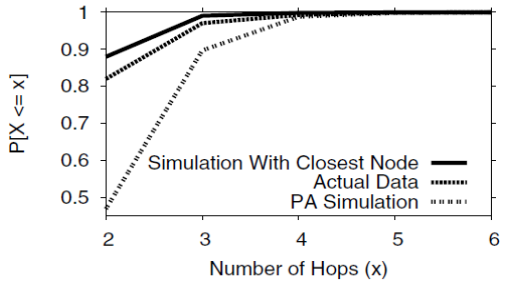


图 20 临近偏倚与优先连接[17]

6.5 组关系

当用户之间存在共同的交互焦点时，他们更有可能建立联系[71, 110]。我们把有共同交互焦点的节点视为在一个组（group）中。Kossinets 和 Watts[71]发现在一个由学生、教员和职工构成的社会网络中，那些参加同一讲座的学生更可能建立联系。Mahanti 等人[17]考察了 FriendFeed 网络中用户订阅相同服务对边的生成造成的影响，他们发现：(1) 对于年龄小于 50 天的节点来说，订阅相同服务的节点与度为 15 的节点建立连接的概率 3 倍于订阅不同服务的节点；(2) 随着节点度的增加，二者之间的差距减小，说明优先连接的影响覆盖了组关系的影响；(3) 对于年龄超过 50 的节点来说，组关系的影响减小；(4) 对于刚加入网络的节点，组关系对生成边的影响强于临近偏倚。

两个在同一个组中的用户往往具有共同的属性（attribute），可以把属性在网络中表示出来。社会属性网络（social attribute network，SAN）[111]是指这样一个网络，网络包含社会节点、属性节点，社会边（用户之间的好友关系）和属性边（用户具有某种属性）。Gong 等人[104]定义了与传统网络拓扑参数相似的属性结构参数，比如直径、聚类系数等，他们详细考察了 Google+社会属性网络的社会拓扑参数和属性拓扑参数，继而研究了 SAN 的演化和节点属性对网络结构的影响。

6.6 社团的演化

社团与组的涵义不同：社团是结构意义上的，组是功能意义上的，共同订阅了某一 SNS 应用的用户构成一个组，而并不一定在一个社团里。社团强调 OSN 中“邻居”的概念，因此社团是邻居对用户行为影响的较好的抽象。复杂网络的社团挖掘也是一个较大的课题，有很多理论工作[112-118]，我们这里仅介绍一些有关社团演化的测量工作。

6.6.1 社团的动态统计特征

(1) 社团的规模

社团规模体现了网络结构的聚类程度。社团规模呈幂律分布[106]。Zhao 等人[106]选取了 3 个平均分布的时刻，考察这 3 个时刻社团规模的分布情况，如图 21(a)所示。我们发现，3 个时刻都有大量的小社团，和少量长尾的大社团，不同时刻社团规模的分布形式较一致。另外，随着时间的增加，小社团数量减少，大社团数量增加。

图 21(b)显示了前 5 个大社团所占节点的比例随时间的变化。我们发现比例随时间逐渐增加，从在 100 天左右的不到 30%增加到 60%。这种趋势——小社团减小，大社团增大——说明随着网络的成熟，网络里主要连通分支的连通性增强，而不同社团之间的差异减小，网络更加均匀。

(2) 社团的生命期

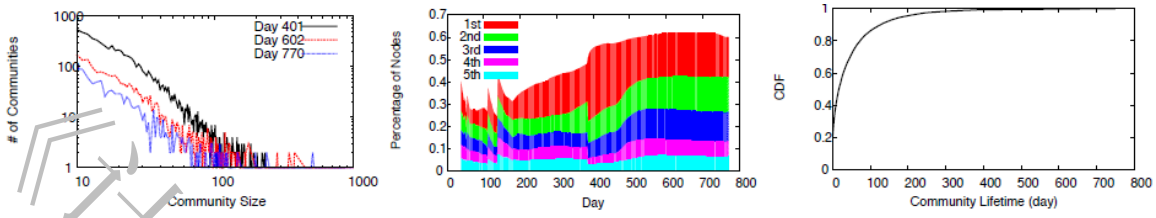
社团的生命期是社团的另一个重要属性。从图

1 FriendFeed[EB/OL]. http://www.friendfeed.com/

21(c)可以看出，大部分社团的存活时间很短，20%的社团的生命期小于 1 天，60%的社团的生命期小于 30 天，这说明社团的活跃性很强。

6.6.2 社团的合并与分裂

社团的合并与分裂是社团生命期开始与结束的主要原因。我们首先关心的问题是，社团规模怎样影响社团的合并与分裂？考察分裂而成的两个最大社团与合并的两个最大社团，考虑较小者与较大者的规模比，比值越小说明二者的规模差别越大。Zhao 等人发现[106]：(1) 78%的合并社团对的规模比小于 0.005，说明对于大多数合并的社团，它



(a) Community Size Distribution　(b) % of Nodes Covered by Top 5 Communities　(c) CDF of Community Lifetime

图 21 OSN 社团的动态特征[106]

们之间的规模差距较大，即小社团不断消失，大社团逐渐增大；(2) 社团分裂的规律恰恰相反，68%的社团分裂对的规模比大于 0.5，即社团倾向于分裂为两个规模相当的小社团，这个现象可用人们维护社会关系能力的局限性（Dunbar 数[52]）解释[106]：由于人们处理社交关系的能力有限，当社团规模超出了社团内某些用户的社交能力时，这些用户就无法继续增加新的好友，导致新用户生成边的分布（在社团内）不一致，这样社团内会产生某些具有更强连通性的“口袋”（pockets），从而使社团分裂。

(1) 预测社团的合并

我们已经了解了社团合并和分裂的某些规律，是否能找到某些结构特征来预测合并呢？Zhao 等人[106]考虑了 3 种结构参数：社团大小（community size），入度比（in-degree ratio，社团内的边数与社团内所有节点度的和的比），相似度（similarity，两个社团相同节点数与不同节点总数的比）。并定义它们的一阶变化指示器（first order change indicator）为：如果社团参数小于之前一个快照时刻的参数，一阶变化指示器的值为-1；否则为 1。同样可以考虑二阶变化指示器（second order change indicator）：如果从快照时刻$i-1$到i社团参数的变化大于从$i-2$到$i-1$的变化，则二阶变化指示器的值为 1，反之为-1。至此我们有 9 种特征参数，分别为社团大小、入度比、相似度和它们相应的一阶参数指示器与二阶参数指示器。利用这些参数及其变化，Zhao 等人[106]采用支持向量机对社团合并进行预测，预测合并的平均准确率为 75%，预测分裂的平均准确率为 77%。这说明我们可以根据短期内社团的变化预测社团的演化。

我们还可以预测一个社团与具体哪个社团合并。Zhao 等人[106]得出了一个较强的结论：一个社团i有 99%的可能性与另外一个与它连接最多的社团j合并。这个结果具有时间稳定性，因此一个社团与其它社团之间的边数是预测其合并对象的可靠依据。

6.6.3 社团对用户的影响

Zhao 等人的研究[106]表明，社团促进用户行为，即社团用户要比非社团用户活跃。他们发现：(1) 社团用户和非社团用户生成边的间隔时间的统计曲线相似，但社团用户的时间间隔较小，说明边的生成较频繁；(2) 社团的规模越大，其中用户的生命期越长；(3) 定义用户的入度比为他在社团内的边数与他总入度的比，社团的规模越大，用户的入度比越大，说明大社团的边较凝聚；(4) 11%到 30%的节点仅与它所属社团内的用户交互，说明社团抽象出了用户的局部行为。

除了组、社团之外，Kumar 等人[103]还考察了 OSN 连通成分的结构特征和演化，此处不再赘述。

6.7 小结

本节总结了 OSN 演化的测量和分析工作，介绍了 OSN 演化的特征和规律：

(1) OSN 拓扑参数并没有统一的（随时间）变化规律，与具体的 SNS 有关。

(2) 多数用户常在早期建立好友关系，不同网络时间不同年龄的节点生成边的比例不同。因此假设只有新节点生成边不够准确。

(3) 边的生成受优先连接和临近偏倚的共同影响。随网络规模的增加，优先连接的影响减弱，优

先连接在某种程度上造成了临近偏倚。

(4) 有共同焦点的用户较易建立连接，新节点受组关系的影响较明显。

(5) OSN 中大社团数量增加，小社团数量减少，大社团逐渐增大，小社团逐渐减小。随着网络的成熟，网络主要连通分支的连通性增强，网络变均匀。

(6) 社团的生命期较短，活跃性很强。小社团倾向于与大社团合并，大社团倾向于分裂成两个规模相当的小社团。一个社团有 99%的可能与另外一个与它连接最多的社团合并。

为了更深入地认识 OSN 的拓扑结构，我们可以利用前面得到的结论（网络中边的生成有何种规律等）设计各种新的网络模型[103, 104, 107, 119]，以期生成更准确、更真实的社会网络。

7 总结及展望

本文从测量的角度综述了现今 OSN 的研究工作。测量方法是首要关注的问题，我们可以通过一些传统的网络测量方法，如抓包、网络爬虫等，来获取数据，也可以借助其他研究人员共享的数据集。由于对 OSN 的研究工作部分延续了复杂网络和传统社会网络的分析方法，特别是针对网络拓扑结构的研究，因此我们可以利用现有的分析软件和工具对采集到的 OSN 数据进行分析。当然更多情况下我们更关注 OSN 有别于其他社会网络的特征，而针对具体问题开发新的算法以挖掘和提取特征是 OSN 测量与分析工作的重点。

基于已有的工作，我们总结了 OSN 的若干结构性质（网络层面）和用户行为特征（用户层面）。总的来说，在网络结构方面，OSN 是一个具有小世界特性的无标度网络，它有一些由度较大的、紧密连接的节点组成的核心，它们把整个网络连接起来。SNS 中用户之间的实际交互会影响 OSN 的拓扑结构，它减小了 OSN 的聚类系数，使网络结构更加松散，并且降低了 OSN 的同配性，增加了平均路径长度。

在用户行为方面，SNS 用户之间的交互具有相互性，他们能维持的社交关系上限是 200 人，大于线下社会网络的 150 人。特别地，SNS 用户的不可见行为（如浏览页面）占了很大比重，而用户平均联系的人数是很少的。

在网络演化方面，边的生成受优先连接和临近偏倚的共同影响，随着时间增加，优先连接对生成边的影响逐渐减小，而优先连接在某种程度上造成了临近偏倚。随着 OSN 年龄的增加，小社团逐渐减少，大社团逐渐增加，网络变得更加均匀。OSN 中社团的存活时间普遍很短，60%的社团的生命期小于 30 天，社团活跃性很强。特别地，OSN 中小社团倾向于与大社团合并，大社团倾向于分裂成两个规模相当的小社团。

还有很多其它方面关于 OSN 的研究工作，比如用户隐私与安全、SNS 系统架构、社团挖掘、地理位置对用户行为的影响、网络博弈等等，限于篇幅，本文无法涵盖所有的研究内容。我们需要认识到，虽然本文总结了很多 OSN 的特征和结论，但有些结论并不是通用的，与具体的 SNS 有关。与实际的研究对象相关是 OSN 的研究难点，亦是各种研究工作百花齐放的原因。比如我们多次强调 OSN 的度分布呈幂律形式，这似乎是 OSN 最基本的性质，但有人[104]发现 Google+的度分布并不服从幂律分布，而是呈对数正态分布。这告诉我们不能盲信 OSN 的研究结论，而要就事论事，适当总结。

当今互联网上各种面向用户的网络应用层出不穷，为研究者们提供了丰富的素材和多变的视角。在线社会网络的研究正在从静态到动态、整体到局部过渡，还存在许多挑战需要进一步解决：

(1) 我们已经多次提到OSN研究中所得结论的“片面性”和“不一致性”。由于在线社会网络的复杂性，得到某种统一的结论或者挖掘不同结论背后的一致规律是最困难也是研究者最为关心的问题。这要求研究者不仅要关心局部，而且不能被局部局限，从而从整体的角度观察和研究问题。

(2) 传统的数据处理方式往往已无法满足 OSN 的研究需求，需要开发新的算法和工具高效地处理庞大的数据。

(3) 在线社会网络是一个交叉领域，这个领域有很多来自不同背景（计算机、物理、数学等）的研究人员，他们都致力于从不同角度研究和考察 OSN 的特征。因此交叉学科之间的碰撞和冲突是该领域的极大挑战，也是灵感和智慧的源泉。

OSN 作为一个复杂巨系统，还有无数有趣的未知信息有待研究者去探索。我们相信会有更多的好的工作出现，从而为我们揭示出 OSN 的本质特征。

致 谢 感谢李刚博士的意见和建议。

参 考 文 献

[1] Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S. Feedback effects between similarity and social influence in online communities//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA, 2008: 160-168.

[2] Moreno J L, Jennings H H. Who shall survive? : a new approach to the problem of human interrelations. Washington: Nervous and Mental Disease Publishing Cooperation, 1934.

[3] Scott J P. Social network analysis: a handbook. 2. London: SAGE Publications, 2000.

[4] Freeman L C. The development of social network analysis: a study in the sociology of science. Vancouver: Empirical Press, 2004.

[5] Leskovec J, Horvitz E. Planetary-scale views on a large instant-messaging network//Proceedings of the 17th international conference on World Wide Web. New York, USA, 2008:915-924.

[6] Adamic L, Adar E. Friends and neighbors on the Web. Social Networks, 2003, 25(3): 211-230.

[7] Park H W, Thelwall M. Hyperlink analyses of the World Wide Web: a review. Journal of Computer-Mediated Communication, 2003, 8(4): 0.

[8] Xu K, Zhu M, Lin C. Internet architecture evaluation models，mechanisms，and methods. Chinese Journal of computers, 2012, 35(10): 1985-2006 (in Chinese).
(徐恪, 朱敏, 林闯. 互联网体系结构评估模型、机制及方法研究综述. 计算机学报, 2012, 35(10): 1985-2006.)

[9] Wang X F, Li X, Chen G R. Network science: an introduction. Beijing: Higher Education Press, 2012 (in Chinese)
(汪小帆, 李翔, 陈关荣. 网络科学导论. 北京: 高等教育出版社, 2012.)

[10] Wang X F, Li X, Chen G R. Complex networks: theory and its applications. Beijing: Tsinghua University Press, 2006 (in Chinese)
(汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用. 北京: 清华大学出版社, 2006.)

[11] Easley D, Kleinberg J. Networks, crowds, and markets: reasoning about a highly connected world. New York: Cambridge University Press, 2010.

[12] Leskovec J, Lang K J, Dasgupta A, Mahoney M W. Statistical properties of community structure in large social and information networks//Proceedings of the 17th international conference on World Wide Web. New York, USA, 2008: 695-704.

[13] Leskovec J, Chakrabarti D, Kleinberg J, Faloutsos C, Ghahramani Z. Kronecker graphs: an approach to modeling networks. The Journal of Machine Learning Research, 2010, 11(3): 985-1042.

[14] K Myunghwan, Leskovec J. The network completion problem: inferring missing nodes and edges in networks//Proceedings of the Eleventh SIAM International Conference on Data Mining. Mesa, USA, 2011: 47-58.

[15] Mislove A, Marcon M, Gummadi K P, Druschel P, Bhattacharjee B. Measurement and analysis of online social networks//Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. New York, USA, 2007: 29-42.

[16] Jiang J, Wilson C, Wang X, Huang P, Sha W P, Dai Y F, Zhao B Y. Understanding latent interactions in online social networks//Proceedings of the 10th annual conference on Internet measurement. New York, 2010: 369-382.

[17] Garg S, Gupta T, Carlsson N, GMahanti A. Evolution of an online social aggregation network: an empirical study//Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. New York, USA, 2009: 315-321.

[18] Paxson V. Bro: a system for detecting network intruders in real-time. Computer Networks: The International Journal of Computer and Telecommunications Networking, 1999, 31(23-24): 2435-2463.

[19] Benevenuto F, Rodrigues T, Cha M, Almeida V. Characterizing user behavior in online social networks//Proceedings of the 9th ACM Internet Measurement Conference. New York, USA, 2009: 49-62.

[20] Schneider F, Feldmann A, Krishnamurthy B, Willinger W. Understanding online social network usage from a network perspective//Proceedings of the 9th ACM Internet Measurement Conference. New York, USA, 2009: 35-48.

[21] Qiu T, Ge Z, Lee S, Wang J, Xu J, Zhao Q. Modeling user activities in a large IPTV system//Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. New York, USA, 2009: 430-441.

[22] Nazir A, Raza S, Chuah C N. Unveiling facebook: a

measurement study of social network based applications//Proceedings of the 8th ACM SIGCOMM conference on Internet measurement. New York, USA, 2008: 43-56.

[23] Hansen D, Shneiderman B, Smith M A. Analyzing social media networks with NodeXL: insights from a connected world. San Francisco: Morgan Kaufmann Publishers, 2010.

[24] Milgram S. The Small World Problem. Psychology Today, 1967, 1(1): 61-67.

[25] Newman M E J. The structure and function of complex networks. Society for Industrial and Applied Mathematics Review, 2003, 45(2): 167-256.

[26] Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A. The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(11): 3747-3752.

[27] Erdös P, Rényi A. On random graphs, I. Publicationes Mathematicae, 1959, 6: 290-297.

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

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

[30] Granovetter M S. The strength of weak ties. American Journal of Sociology, 1973, 78(6): 1360-1380.

[31] Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J. Graph structure in the Web. Computer Networks, 2000, 33(1-6): 309-320.

[32] Wilson C, Boe B, Sala A, Puttaswamy K P N, Zhao B Y. User interactions in social networks and their implications//Proceedings of the 4th ACM European conference on Computer systems. New York, USA, 2009: 205-218.

[33] Newman M E J. Assortative mixing in networks. Physical Review Letters, 2002, 89(20): 208701+.

[34] Chun H, Kwak H, Eom Y H, Ahn Y Y, Moon S, Jeong H. Comparison of online social relations in volume vs interaction: a case study of cyworld//Proceedings of the 8th ACM SIGCOMM conference on Internet measurement. New York, USA, 2008: 57-70.

[35] Zhang S, Xu K, Li H T. Measurement and analysis of information propagation in online social networks like microblog. Journal of Xi'an Jiaotong University, 2013, 47(2): 124-130 (in Chinese).
(张赛, 徐恪, 李海涛. 微博类社交网络中信息传播的测量与分析. 西安交通大学学报, 2013, 47(2): 124-130.)

[36] Axelrod R, Hamilton W D. The evolution of cooperation. Science, 1981, 211(4489): 1390-1396.

[37] Gouldner A W. The norm of reciprocity: a preliminary statement. American Sociological Review, 1960, 25(2): 161-178.

[38] Nowak M A. Five rules for the evolution of cooperation. Science, 2006, 314(5805): 1560-1563.

[39] Aukett R, Ritchie J, Mill K. Gender differences in friendship patterns. Sex Roles, 1988, 19(1): 57-66.

[40] Thurnwald R. Economics in primitive communities. London: International institute of African languages and cultures, 1932.

[41] Simmel G, Wolff K. The sociology of georg simmel. New York: Free Press of Glencoe, 1950.

[42] Hemelrijk C K. Models of, and tests for, reciprocity, unidirectionality and other social interaction patterns at a group level. Animal Behaviour, 1990, 39(6): 1013-1029.

[43] Garlaschelli D, Loffredo M I. Patterns of link reciprocity in directed networks. Physical Review Letters, 2004, 93(26): 268701+.

[44] Gómez V, Kaltenbrunner A, López V. Statistical analysis of the social network and discussion threads in Slashdot//Proceedings of the 17th international conference on World Wide Web. New York, USA, 2008: 645-654.

[45] Java A, Song X, Finin T, Tseng B. Why we twitter: understanding microblogging usage and communities//Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. New York, USA, 2007: 56-65.

[46] Zlatic V, Bozícevic M, Stefancic H, Domazet M. Wikipedias: collaborative web-based encyclopedias as complex networks. Physical Review E, 2006, 74(1): 016115+.

[47] Almaas E, Kovacs B, Vicsek T, Oltvai Z N, Barabási A L. Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature, 2004, 427(6977): 839-843.

[48] Derrida B, Flyvbjerg H. Statistical properties of randomly broken objects and of multivalley structures in disordered systems. Journal of Physics A: Mathematical and General, 1987, 20(15): 5273-5288.

[49] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D,

Alon U. Network motifs: simple building blocks of complex networks. Science, 2002, 298(5594): 824-827.

[50] Milo R. Superfamilies of evolved and designed networks. Science, 2004, 303(5663): 1538-1542.

[51] Wernicke S, Rasche F. FANMOD: a tool for fast network motif detection. Bioinformatics, 2006, 22(9): 1152-1153.

[52] Dunbar R. Grooming, gossip, and the evolution of language. Cambridge: Harvard University Press, 1998.

[53] Bialik C. Sorry, you may have gone over your limit of network friends. The Wall street journal online, 2007, 11.

[54] Golder S, Wilkinson D M, Huberman B A. Rhythms of social interaction: messaging within a massive online network// Proceedings of the Third Communities and Technologies Conference. East Lansing, USA, 2007: 41-66.

[55] Nazir A, Raza S, Gupta D, Chuah C N, Krishnamurthy B. Network level footprints of facebook applications//Proceedings of the 9th ACM Internet Measurement Conference. New York, USA, 2009: 63-75.

[56] Guo L, Tan E, Chen S, Zhang X D, Zhao Y H. Analyzing patterns of user content generation in online social networks//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA, 2009: 369-378.

[57] Duarte F, Mattos B, Bestavros A, Almeida V, Almeida J. Traffic Characteristics and Communication Patterns in Blogosphere. Boston: Boston University, Technical Report: BUCS-TR-2006-033, 2006.

[58] Huberman B A, Pirolli P L T, Pitkow J E, Rajan M L. Strong regularities in world wide web surfing. Science, 1998, 280(5360): 95-97.

[59] Cha M, Mislove A, Adams B, Gummadi K P. Characterizing social cascades in flickr//Proceedings of the first workshop on Online social networks. New York, USA, 2008: 13-18.

[60] Cha M, Mislove A, Gummadi K P. A measurement-driven analysis of information propagation in the flickr social network//Proceedings of the 18th international conference on World wide web. New York, USA, 2009: 721-730.

[61] Sastry N, Yoneki E, Crowcroft J. Buzztraq: predicting geographical access patterns of social cascades using social networks//Proceedings of the Second ACM EuroSys Workshop on Social Network Systems. New York, USA, 2009: 39-45.

[62] Krishnamurthy B, Gill P, Arlitt M. A few chirps about twitter//Proceedings of the first workshop on Online social networks. New York, USA, 2008: 19-24.

[63] Nazir A, Waagen A, Vijayaraghavan V S, Chuah C N, D'souza R M, Krishnamurthy B. Beyond friendship: modeling user activity graphs on social network-based gifting applications//Proceedings of the 2012 ACM conference on Internet measurement conference. New York, USA, 2012: 467-480.

[64] Cha M, Kwak H, Rodriguez P, Ahn Y Y, Moon S. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system//Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. New York, USA, 2007: 1-14.

[65] Li H T, Liu J C, Wang H Y, Xu K. Video sharing in online social network: measurement and analysis//Proceedings of the 22nd ACM Workshop on Network and Operating Systems Support for Digital Audio and Video. Toronto, Canada, 2012.

[66] Gjoka M, Sirivianos M, Markopoulou A, Yang X W. Poking facebook: characterization of osn applications//Proceedings of the first workshop on Online social networks. New York, USA, 2008: 31-36.

[67] Berger N, Borgs C, Chayes J T, D'souza R M, Kleinberg R D. Competition-induced preferential attachment//Proceedings of the 31st International Colloquium on Automata, Languages and Programming. Turku, Finland, 2004: 697-721.

[68] D'souza R M, Borgs C, Chayes J T, Noam B, Kleinberg R D. Emergence of tempered preferential attachment from optimization. Proceedings of the National Academy of Sciences, 2007, 104(15): 6112-6117.

[69] Mossa S, Barthélémy M, Stanley H E, Amaral N. Truncation of power law behavior in "scale-free" network models due to information filtering. Physical Review Letters, 2002, 88(13): 138701+.

[70] Spearman C. The proof and measurement of association between two things. The American Journal of Psychology, 1987, 100(3/4): 441-471.

[71] Kossinets G, Watts D J. Empirical analysis of an evolving social network. Science, 2006, 311(5757): 88-90.

[72] Cha M, Rodriguez P, Crowcroft J, Moon S, Amatriain X. Watching television over an IP network// Proceedings of the 8th ACM SIGCOMM conference on Internet

measurement. New York, USA, 2008: 71-84.

[73] Newman M E J. Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 2006, 74(3): 036104+.

[74] Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: densification laws, shrinking diameters and possible explanations//Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. New York, USA, 2005: 177-187.

[75] Sala A, Cao L, Wilson C, Zablit R, Zheng H T, Zhao B Y. Measurement-calibrated graph models for social network experiments//Proceedings of the 19th international conference on World wide web. New York, USA, 2010: 861-870.

[76] Bailey N. The mathematical theory of infectious deseases and its applications. London: Griffin, 1975.

[77] Granovetter M. Threshold models of collective behavior. American Journal of Sociology, 1978, 83(6): 1420-1443.

[78] Watts D J. A simple model of global cascades on random networks. Proceedings of the National Academy of Sciences, 2002, 99(9): 5766-5771.

[79] Karsai M, Kivelä M, Pan R K, Kaski K, Kertész J, Barabási A L, J Saramäki. Small but slow world: how network topology and burstiness slow down spreading. Physical Review E, 2011, 83(2): 025102+.

[80] Steeg G V, Ghosh R, Lerman K. What stops social epidemics?//Proceedings of the 5th International Conference on Weblogs and Social Media. Barcelona, Spain, 2011: 377-384.

[81] Gruhl D, Guha R, Nowell D L, Tomkins A. Information diffusion through blogspace//Proceedings of the 13th international conference on World Wide Web. New York, USA, 2004: 491-501.

[82] Adar E, Adamic L A. Tracking information epidemics in blogspace//Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence. Washington, USA, 2005: 207-214.

[83] Bakshy E, Karrer B, Adamic L A. Social influence and the diffusion of user-created content//Proceedings of the 10th ACM conference on Electronic commerce. New York, USA, 2009: 325-334.

[84] Sun E, Rosenn I, Marlow C, Thomas L. Gesundheit! Modeling Contagion through Facebook News Feed//Proceedings of the Third International Conference on Weblogs and Social Media. San Jose, USA, 2009: 146-153.

[85] Rodrigues T, Benevenuto F, Cha M, Gummadi K, Almeida V. On word-of-mouth based discovery of the web//Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. New York, USA, 2011: 381-396.

[86] Gomez R M, Leskovec J, Krause A. Inferring networks of diffusion and influence//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA, 2010: 1019-1028.

[87] Ghosh R, Lerman K. Predicting Influential Users in Online Social Networks//Proceedings of the fourth KDD workshop on social network mining and analysis. Washington, USA, 2010.

[88] Scellato S, Mascolo C, Musolesi M, Crowcroft J. Track globally, deliver locally: improving content delivery networks by tracking geographic social cascades//Proceedings of the 20th international conference on World Wide Web. New York, USA, 2011: 457-466.

[89] Wang D, Wen Z, Tong H, Lin C Y, Song C M, Barabási A L. Information spreading in context//Proceedings of the 20th international conference on World wide web. New York, USA, 2011: 735-744.

[90] Galuba W, Aberer K, Chakraborty D, Despotovic Z, Kellerer W. Outtweeting the twitterers - predicting information cascades in microblogs//Proceedings of the 3rd conference on Online social networks. Berkeley, USA, 2010: 3-3.

[91] Benevenuto F, Magno G, Rodrigues T, Almeida V. Detecting Spammers on Twitter//Seventh annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference. Washington, USA, 2010.

[92] Grier C, Thomas K, Paxson V, Zhang M. @spam: the underground on 140 characters or less//Proceedings of the 17th ACM conference on Computer and communications security. New York, USA, 2010: 27-37.

[93] Chhabra S, Aggarwal A, Benevenuto F, Kumaraguru P. Phi.sh/$oCiaL: the phishing landscape through short URLs//Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference. New York, USA, 2011: 92-101.

[94] Szabo G, Huberman B A. Predicting the Popularity of

Online Content. Communications of the ACM, 2010, 53(8): 80-88.

[95] Li H T, Liu J C, Xu K, Wen S. Understanding video propagation in online social networks//Proceedings of the 2012 IEEE 20th International Workshop on Quality of Service. Piscataway, USA, 2012: 21-21.

[96] Li H T, Ma X Q, Wang F, Liu J C, Xu K. On popularity prediction of videos shared in online social networks// Proceedings of the 2013 IEEE 21th International Workshop on Quality of Service. Submitted.

[97] Cha M, Kwak H, Rodriguez P, Ahn Y Y, Moon S. Analyzing the video popularity characteristics of large-scale user generated content systems. IEEE/ACM Transactions on Networking, 2009, 17(5): 1357-1370.

[98] Niu D, Liu Z, Li B C, Zhao S Q. Demand forecast and performance prediction in peer-assisted on-demand streaming systems//Proceedings of the 30th IEEE International Conference on Computer Communications. Shanghai, China, 2011: 421-425.

[99] Gürsun G, Crovella M, Matta I. Describing and forecasting video access patterns//Proceedings of the 30th IEEE International Conference on Computer Communications. Shanghai, China, 2011: 16-20.

[100] Kutner M H, Nachtsheim C J, Neter J. Applied Linear Regression Models. 4. New York: McGraw-Hill/Irwin, 2004.

[101] Navot A, Shpigelman L, Tishby N, Vaadia E. Nearest neighbor based feature selection for regression and its application to neural activity//Advances in Neural Information Processing Systems 18. Vancouver, Canada, 2006: 995-1002.

[102] Backstrom L, Boldi p, Rosa M, Ugander J, Vigna S. Four Degrees of Separation//Proceedings 4th ACM International Conference on Web Science. Evanston, USA, 2012: 45-54

[103] Kumar R, Novak J, Tomkins A. Structure and evolution of online social networks//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA, 2006: 611-617.

[104]Gong N Z, Xu W, Huang L, Mittal P, Stefanov E, Sekar V, Song D. Evolution of social-attribute networks: measurements, modeling, and implications using google+//Proceedings of the 2012 ACM conference on Internet measurement conference. New York, USA, 2012: 131-144.

[105] Ahn Y Y, Han S, Kwak H, Moon S, Jeong H. Analysis of topological characteristics of huge online social networking services//Proceedings of the 16th international conference on World Wide Web. New York, USA, 2007: 835-844.

[106] Zhao X, Sala A, Wilson C, Wang X, Gaito S, Zheng H T, Zhao B Y. Multi-scale dynamics in a massive online social network//Proceedings of the 2012 ACM conference on Internet measurement conference. New York, USA, 2012: 171-184.

[107] Leskovec J, Backstrom L, Kumar R, Tomkins A. Microscopic evolution of social networks//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA, 2008: 462-470.

[108] Watts D J. The "New" Science of Networks. Annual Review of Sociology, 2004, 30(1): 243-270.

[109] Mislove A, Koppula H S, Gummadi K P, Druschel P, Bhattacharjee B. Growth of the flickr social network//Proceedings of the first workshop on Online social networks. New York, USA, 2008: 25-30.

[110] Newman M E J. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences, 2001, 98(2): 404-409.

[111]Gong N Z, Talwalkar A, Mackey L, Huang L, Shin E C R, Stefanov E, Shi E, Dawn S. Predicting Links and Inferring Attributes using a Social-Attribute Network (SAN). The Computing Research Repository, 2011, abs/1112.3265.

[112] Lin Y R, Chi Y, Zhu S H, Sundaram H, Tseng B L. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks//Proceeding of the 17th international conference on World Wide Web. Beijing, China, 2008: 685-694.

[113] Tantipathananandh C, Berger-Wolf T. Constant-factor approximation algorithms for identifying dynamic communities//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA, 2009: 827-836.

[114] Tantipathananandh C, Berger-Wolf T, Kempe D. A framework for community identification in dynamic social networks//Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data

mining. New York, USA, 2007: 717-726.

[115]Sun J, Faloutsos C, Papadimitriou S, Yu P S. GraphScope: parameter-free mining of large time-evolving graphs//Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA, 2007: 687-696.

[116] Kim M S, Han J. A particle-and-density based evolutionary clustering method for dynamic networks. Proceedings of the VLDB Endowment, 2009, 2(1): 622-633.

[117] Greene D, Doyle D, Cunningham P. Tracking the Evolution of Communities in Dynamic Social Networks//Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining. Washington, USA, 2010: 176-183.

[118] Asur S, Parthasarathy S, Ucar D. An event-based framework for characterizing the evolutionary behavior of interaction graphs. ACM Transactions on Knowledge Discovery from Data, 2009, 3(4): 16:1-16:36.

[119] Allamanis M, Scellato S, Mascolo C. Evolution of a location-based online social network: analysis and models//Proceedings of the 2012 ACM conference on Internet measurement conference. New York, USA, 2012: 145-158.

XU K, born in 1974, Ph. D. , professor, Ph. D. supervisor. His research interests mainly include architecture of next-generation Internet, high performance router, networking science and social networks, Internet of Things.

ZHANG Sai, born in 1988, master candidate. His main research interests include measurement and modeling of online social networks.

CHEN Hao, born in 1979, master candidate. His main research interests include measurement and analysis of online social networks.

LI Hai-Tao, born in 1983, Ph. D. candidate. His main research interests include online social networks, Cloud Computing, P2P.

Background

Social network sites are becoming increasingly popular these days. This stimulates researchers to study online social networks more deeply and systematically, for a better understanding of human behaviors and social networks. There is a lot of work on OSNs in respect of measurement, privacy, system design, modeling etc. Focusing on the measurement and analysis of ONSs, this paper provides a thorough and comprehensive overview of the current study on network topology, users' behaviors and the evolution of network structure; it sums up abundant related research progresses in OSNs and is of great referential significance for researchers.

This work is supported by the National Science and Technology Major Project (NO. 2012ZX03005001), the National Natural Science Foundation (NO. 61170292, NO. 60970104), 863 Program (NO. 2013AA013302) and the National Basic Research Program (973 Program) of China (NO. 2009CB320501, NO. 2012CB315802). These projects aim to make advances to the Next Generation Internet and the evolution of the Internet. This paper summarizes the measurement and analysis of online social networks.