网络科学的发展新动力：大数据与众包

许小可[1]，刘肖凡[2]

(1. 大连民族学院信息与通信工程学院 辽宁 大连 116600；2. 东南大学计算机科学与工程学院 南京 211189)

【摘要】大数据时代的来临给网络科学带来了新的发展机遇，但如何处理海量数据也成为网络科学领域面临的严峻挑战。与大数据时代同时到来的，是近年来兴起的众包项目模式。公开竞赛和数据公开等众包形式，已成为解决数据领域问题非常流行的方法。该文概述了海量数据和众包模式在多个方面对网络科学发展的促进作用，并详细介绍了2013年首届阿里数据平台创新大赛的竞赛流程和本团队的获奖成果。在众包模式的驱动下，人们期待以大数据处理为中心的数据科学和网络科学相辅相成、共同发展。

关 键 词 大数据; 复杂网络; 众包; 网络科学
中图分类号 N94 文献标志码 A doi:10.3969/j.issn.1001-0548.2013.06.001

New Driving Forces of Network Science: Big Data and Crowd Sourcing

XU Xiao-ke[1] and LIU Xiao-fan[2]

(1. College of Information and Communication Engineering, Dalian Nationalities University Dalian Liaoning 116600;
2. School of Computer Science and Engineering, Southeast University Nanjing 211189)

Abstract The era of big data has brought big opportunities to network science, yet the study of network science is also facing big challenges of processing huge amount of data. Meanwhile, a new project outsourcing model, the crowd-sourcing model, has been widely applied recently to solving data related problems. This paper presents a brief introduction to the new driving forces of network science: big data and crowd-sourcing model. We also review the 2013 Alibaba data platform innovation competition as well as our award-winning work. We expect that driven by the crowd-sourcing model, data science, which deals with big data processing techniques, shall evolve together with network science into a prosperous future.

Key words big data; complex networks; crowd sourcing; network science

近年来网络科学发展十分迅速，每年都有一批影响力较大的成果涌现出来。2013年度，该领域的发展得到进一步深化和提高，出现了两种不同于往年的独特现象。

一是著名大学和科学家协会新办了多个以网络科学理论及应用为主旨的学术刊物。如世界著名大学牛津大学和剑桥大学分别创办了网络科学方面新刊物：Journal of Complex Networks 和 Network Science杂志。全球最大的非营利性专业技术学会美国电气和电子工程师协会(IEEE)也成立了两份新的会刊 IEEE Transactions on Network Science and Engineering 和 IEEE Transactions on Control of Network Systems，主旨在收录网络科学方面的成果。毫无疑问，这些新刊物的诞生将对网络科学今后的深化发展起到促进作用。

二是大数据时代的到来给网络科学提供了新的发展机遇和挑战，“众包”成为驱动网络科学理论与工程发展的新动力。“大数据”这个词是2013年最火热的IT行业词汇，专家学者们都十分关注如何在大数据的背景下推进网络科学研究。本文拟就参加的由杭州师范大学信息经济研究所主办的“大数据时代下复杂网络的机遇与挑战”研讨会及第九届复杂网络大会组委会和阿里研究中心承办的“阿里数据平台创新大赛”的参赛经历谈谈这方面的感想和体会。

1 网络科学发展和大数据分析相辅相成

1.1 数据是网络科学发展的原动力

众所周知，复杂网络研究领域最原始的发展动力是新数据的获得。20世纪90年代以前，在无法获

收稿日期：2013– 10 – 15；修回日期：2013 – 11 – 11
基金项目：国家自然科学基金(61004104，61374170，61304167)；中央高校基本科研业务费专项基金(DC120101132，DC13010215)
作者简介：许小可(1979 –)，男，教授，主要从事非线性时间序列分析和复杂网络方面的研究.

取实证数据的情况下，长期以来随机网络(或称随机图)是网络科学领域最主要的网络拓扑模型。随着科技的进步，数据获取越来越容易，网络科学也因此开始蓬勃发展。如1998年Watts和Strogatz为了拟合很多实际网络中出现的较短平均路径长度和较高的聚类系数，提出了小世界模型。1999年，Barabasi与Albert为了重现多个实证网络中的度分布规律而提出了无标度网络模型，从此掀起了网络科学研究的热潮。由此可见，数据是网络科学发展的原动力。

回顾十几年网络科学的发展历程，能拿到契合研究课题的实证数据的研究小组往往能占领网络科学研究的最前沿。以该领域最享有盛誉的Barabasi小组为例，他们的诸多研究成果都是基于自身拥有的独一无二的实证数据完成的，有些数据甚至他们发表论文数年以后其他研究者还无法获得。在2013年，华东理工大学周炜星小组的PNAS论文也是基于中国移动通信公司的590万用户手机通话记录的分析完成的[1]。尽管拥有数据的研究者不一定能做出好的研究，但是优秀的研究成果中往往包含了大量的实证数据分析。因此拥有什么样的数据做研究已经成为网络科学领域学者们关心的重要问题，鉴于很多时候都是好数据对应好成果、大数据对应大发展，所以广大学者们都期待着大数据时代的来临。

1.2 网络科学是数据科学研究的重要工具

“数据科学”是近年出现的热门词汇，数据科学是关于数据的科学或者研究数据的科学，是大数据时代产生的新概念。尽管大数据时代给网络科学研究带来了重要机遇，但如何处理海量数据也是网络科学研究领域面临的严峻挑战。2013年5月，杭州师范大学商学院信息经济研究所、阿里研究中心、阿里复杂科学中心共同组织研讨会分析了“大数据时代下复杂网络的机遇与挑战”。由于数据量级和维度的增加大大增加了很多问题的复杂性，因此传统复杂网络中的基本理论问题、实际应用以及各种经典算法都将在大数据时代面临严峻的挑战。

目前，在线社交网络、脑科学和基因组分析等是大数据研究的重要领域，而网络科学在其中大有用武之地[2]。以在线社交网络研究为例，近年来Facebook、Twitter、Google等互联网公司都提供了在线社交网络服务，数据科学领域的专家们都为如何来分析网络上的海量数据而发愁。网络科学理论可以很自然的将社交网络中的每个用户当作一个节点，用户之间互动情况当作连边来加以分析，因此网络科学是数据科学研究领域的重要工具。

综上，网络科学和大数据研究是相辅相成、相互促进的关系。大数据给网络科学提供了重要的发展机遇和挑战，数据是网络科学发展的动力，网络科学的研究者应重视数据的重要作用。同时，以大数据处理为中心的数据科学也需要网络科学的相关理论作为支撑，网络科学已成为数据科学领域的重要手段和工具。

2 众包模式驱动网络科学发展

2.1 为什么网络科学领域偏爱众包

“众包”是一种分布式的问题解决模式，指的是一个公司或机构把过去由特定人员执行的工作任务，以自由自愿的形式外包给(通常是网络上的)非特定大众的做法[3]。众包和通常意义上外包的不同之处在于，前者的任务和问题是外派给不确定的群体，而后者是外派给确定的个体。众包的好处在于：企业和结构可以充分利用网络资源，借助外部的智慧，节约大量的研发成本。

值得注意的是，众包模式尤其适合大数据中无法界定学科界限的问题研究。以互联网为平台的发布模式将以往架设在不同学科之间的藩篱打破，集结了不同领域的学者协同合作解决难题。网络科学是一门典型的交叉学科，从事统计物理、系统科学和计算机科学等领域研究的人员均对网络科学的发展做出了巨大贡献[4]。这些研究人员零散分布于各个研究机构、各个学科之中，意味着企业如遇到网络科学相关问题，将很难在传统的学科门类培养体系中直接招聘到合适人才，也很难通过外包方式找到直接相关的研究人员。因此，公开竞赛、数据开放等众包模式就成为了解决网络科学领域问题非常有效的方式。

2.2 国内外网络科学领域众包项目介绍

据粗略统计，2012年到2013年国内外的各种网络科学领域的众包任务超过了20项，限于本文篇幅仅选取几个有代表性的众包来介绍。在世界著名的科研众包网站Kaggle上，Facebook等大公司举行了多次网络科学领域的竞赛，并将该项比赛作为招聘数据科学家的一条非常重要的途径[5]。在2012年，Facebook在首次在Kaggle举行的竞赛是有关社交网络中链路预测研究的题目，这也是近年来网络科学领域中的热点问题[6]。由于该竞赛取得了非常好的效果，Facebook不久之后又举办了关于Internet网络拓扑结构动态分析和演化预测研究的第二次竞赛。

此外，具有举办竞赛传统的计算机顶级会议如

国际知识发现和数据挖掘年会(KDD)去年也将目标关注于在线社交网络研究，题目为预测微博用户对系统推荐对象的关注。甚至像美国大学生数学建模竞赛(MCM/ICM)这种非常经典的初级交叉学科竞赛中也出现了分析恐怖分子社交网络分析这类和网络科学强相关的题目，由此可见网络科学方面的研究已经成为众包中的热门题目。

在国内，网络科学方面的众包研究也如雨后春笋般发展起来。早在2011年，首届全国大学生数据挖掘邀请赛的题目就是以某大型婚恋网站交友数据为依托设计推荐算法。2012年，相关的网络科学题目众包开始增多，而在2013年几乎所有的主流互联网公司都举办了这方面的众包竞赛。百度举办了“机器知我心——推荐引擎创意大赛”，阿里巴巴举办了“阿里数据平台创新大赛”，腾讯联合中国计算机协会发起了以支持在线社交网络研究为主的“犀牛鸟基金”。此外，华为也通过“全国大学生智能设计竞赛”和“创新研究计划”的方式参与到网络科学领域的众包中。在近期中国计算机学会大数据专委会举办的“第一届大数据技术创新与创业大赛”命题竞赛中，中国移动通信研究院也提供了和网络科学有关的题目:“移动用户交往圈构建和特定类型用户识别”。

2.3 众包对网络科学发展的促进作用

众包对网络科学发展的促进是多方面的。首先，众包项目中的众多网络科学相关的研究课题拓展了网络科学在学术界和工业界的影响，让不同领域的研究者意识到网络科学的学术魅力和发展潜力。其次，网络科学科学家们也通过众包项目开阔了视野，了解了工业界的需求，进一步明确了这一领域亟待解决的基本科学问题。同时，众包项目也吸引了其他领域的研究者投入到网络科学领域，增加了网络科学发展的有生力量。最后，传统上科学家们交流的渠道一般是参加学术会议和阅读学术论文，众包模式让参与者在互联网平台上相互学习、相互竞争，客观上促进了网络科学学者之间的交流，提高了研究者的科研水平和解决实际问题的能力。

3 参加阿里数据平台创新大赛的感想和体会

3.1 阿里数据平台创新大赛简介

以上章节总体阐述了大数据和众包对网络科学发展的影响和促进作用，接下来以阿里数据平台创新大赛这个具体例子以管窥豹，了解一下融合大数据和众包这两个特征的数据科学竞赛的方方面面。

“首届阿里数据平台创新大赛”暨第三期阿里巴巴青年学者支持计划，是由阿里数据创新平台大赛组委会发起并主办，由第九届复杂网络大会组委会和阿里研究中心承办的全国性互联网数据平台创新大赛活动[7]。该项大赛围绕国家互联网及电子商务产业发展战略，以基于电子商务平台中在线用户交易行为数据的数据分析技术及应用模式创新为竞赛内容。参赛选手团队需要在规定时间内给出指定数据集的命题和处理方法，在“阿里云”平台进行分析和计算，并在给定时间内自主命题完成，并参与数据创新大赛评奖。

在本次阿里数据平台创新大赛中，主办方提供了一个真实的大规模阿里旺旺数据集，具体为2011年11月份10%的阿里旺旺通信记录。数据抽样的方式是随机记录抽样，文件大小为解压缩前为2.8 GB，文档类型是CSV格式，总的记录数为1.3亿条。数据的结构如表1所示，第一列为通信发生的时间，精确到天，第二列是通信的发起人匿名化后的编号，第三列是通信的接收人匿名化后的编号。

表1 阿里旺旺通信记录格式

时间/d	通信发起人	通信接收人
1	16	743 772
1	26	93 302
1	26	905 410
1	27	1 999
1	40	47 662
1	47	100
1	62	9 393
1	77	173 043
1	77	299 133

3.2 阿里数据创新大赛中本团队的研究成果

在竞赛中本团队主要取得了以下3个方面的研究成果，并获得了该项赛事唯一的一等奖。首先，基于抽样理论分析了阿里旺旺即时通系统的各种宏观统计特性，如总用户数量、每天通信量的动态波动以及每天的活跃用户数量等，这些信息是衡量阿里旺旺在整个即时通软件领域市场地位和商业份额的重要指标，对于研究天猫和淘宝商城用户的商业集群行为也有重要的辅助作用。

其次，阿里旺旺系统中最有商业价值的就是那些高活跃性用户，基于随机记录的抽样方法也使数据集中这部分用户的信息相对更准确，因此研究中使用社交网络理论[8]对阿里旺旺的高活跃性用户进行了分类。根据出权(发短消息数量)和入权(收消息数量)之间关系将高活跃性用户分为：客服中心型、广告传播型和高影响力型。这三类用户不仅在社交网络的出入权上有不同的特点，他们和其他用户信

息交互的人类动力学特征也明显不同，这些特征对于挖掘垃圾广告传播者、寻找优质商家等实际应用具有重要的参考价值。

最后，发现用户交互的动力学特征和其商业行为息息相关，因此使用加权社交网络分析中的友谊关系传递性理论来区分高影响力用户的社交和商务两类活动，发现高影响力用户具有迅速衰减的边权重分布和极弱的友谊传递性，自我中心网具有和其他社交通信系统明显不同的星型结构，说明阿里旺旺中的商务交流远大于社会交流，因此可以确定阿里旺旺属于典型的商务即时通工具。

3.3 阿里数据大赛的双赢局面及其局限性探讨

在本次阿里数据平台创新大赛中，研究学者得到了非常有价值的数据集，并通过阿里巴巴云计算平台对数据集进行了分析和交流。该项竞赛产生了双赢的局面：网络科学的研究者了解了阿里巴巴这个全球领先的电子商务公司的技术需求和面临的理论难题，拓展了网络科学理论在实际问题中的应用；而阿里巴巴获得了研究人员所提出的解决具体问题的理论基础和新思路，同时也扩大了企业在网络科学研究社区的影响力。阿里集团将通过“阿里青年学者支持计划”进一步加深双方的合作。

尽管在阿里数据平台创新大赛中多方共赢，但在参加的过程中也发现企业众包模式中大数据项目的一些局限性。众包参与者和大数据提供的企业之间往往存在矛盾，相互之间必须不断进行沟通和妥协，这主要体现在以下两个方面：一方面，企业往往比较关心大数据众包出去后能否获得一些商业利益，而科学家们更注重自己研究的学术影响，因此企业和学者之间的目标往往是不同的；另一方面，出于保护用户隐私和商业秘密的原因，企业提供的大数据往往都匿名化并刻意采用一些抽样方法，这样的数据往往不能满足研究者的需求。对于阿里数据平台创新大赛来说，目前的数据中短消息记录的时间信息仅仅精确到天，如果能提供精确到秒级的时间信息，那么使用人类动力学理论[9]来细致分析用户的行为特征，进一步挖掘 VIP 用户的使用习惯便成为可能。此外，如果能提供短消息相关的商品联动数据、以及信息发送-接受方的用户属性(如用户是买方还是卖方)，则能从微观上揭示阿里旺旺中社交网络特征形成的内在驱动力。我们期待着在今后的阿里数据平台创新大赛中，阿里巴巴开放更多更好的数据给研究者，使该竞赛无论在商业价值上，还是学术价值上都更上一层楼。

4 总结与展望

网络科学是一门新兴的交叉学科，企业界和一些研究机构遇到的网络科学问题很难找到直接对口的研究人员通过外包方式解决，因此近年来通过公开竞赛或公开招标的众包模式来解决网络科学领域问题就成为非常流行的方式。大数据为网络科学的发展提供了新的机遇，成为网络科学发展的新动力。同时，以大数据处理为中心的数据科学也需要网络科学的相关理论作为支撑，网络科学将成为数据科学领域的重要手段和工具。我们期待着今后更多的众包模式大数据项目能够推动网络科学与数据科学相辅相成，协同发展。

参 考 文 献

[1] JIANG Zhi-qiang, XIE Wen-jie, LI Ming-xia, et al. Calling patterns in human communication dynamics[J]. Proc Natl Acad Sci, 2013, 110(5): 1600-1605.

[2] 方锦清. 大数据浪潮冲击下网络科学与工程面临的挑战与机遇[J]. 自然杂志, 2013, 35(5): 345-354.
FANG Jin-qing. Network science and engineering faced with a new challenge and developing opportunity under the wave impact of big data[J]. Chinese Journal of Nature, 2013, 35(5): 345-354.

[3] DOAN A, RAMAKRISHNAN R, HALEVY A Y. Crowdsourcing systems on the world-wide web[J]. Communications of the ACM, 2010, 54(4): 86-96.

[4] 汪小帆, 李翔, 陈关荣. 网络科学导论[M]. 北京: 高等教育出版社, 2012.
WANG Xiao-fan, LI Xiang, CHEN Guan-rong. Network science: an introduction[M]. Beijing: High Education Press, 2012.

[5] NARAYANAN A, SHI E, RUBINSTEIN B I. Link prediction by de-anonymization: How we won the kaggle social network challenge[C]//The 2011 International Joint Conference on Neural Networks (IJCNN). [S.l.]: IEEE, 2011: 1825-1834.

[6] 吕琳媛, 周涛. 链路预测[M]. 北京: 高等教育出版社, 2013.
LÜ Lin-yuan, ZHOU Tao. Link prediction[M]. Beijing: High Education Press, 2013.

[7] 阿里数据创新平台大赛组委会. 阿里数据平台创新大赛[EB/OL]. [2013-11-01]. http://cccn.eugene.cn/EnableCE/t.php?tid=29.
Ali data innovation platform competition organizing committee. Ali data platform for innovation contest[EB/OL]. [2013-11-01]. http://cccn.eugene.cn/EnableCE/t.php?tid=29.

[8] WASSERMAN S, FAUST K. Social network analysis: methods and applications[M]. Oxford: Cambridge University Press, 1994.

[9] BARABASI A L. The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005, 435(7039): 207-211.

编 辑 蒋 晓