动，也未能向网络服务商和网民提供相应救济保障。为此，需要尽早完善网络法律救济体系，通过个人利益、公共利益、社会利益的公开意见表达与多元共同协商，形成治网共识与治理决策的社会参与机制。以此为基础，厘清互联网领域内的责、权、利关系。

〔责任编辑：李凌静　责任编审：冯小双〕

互联网与社会学定量研究*

孙秀林　陈华珊

近几十年来，迅速发展的互联网改变了社会生活的各个方面。卡斯特认为，网络构建了一种前所未有的社会形态，全面改变了人类社会的生产、权力和文化进程。网络社会的崛起，已经成为一种历史趋势和一种正在浮现的社会结构。① 互联网的发展，也极大地影响了社会学研究，尤其是定量研究。传统社会学研究的对象是实体社会，互联网兴起以后，“虚拟社会”开始出现，在博客、微博、微信等网络工具影响下，很多议题已超出传统社会学研究范式。不仅如此，“大数据”概念的广泛应用和巨大影响，对社会学研究的冲击更为直接。这种冲击涉及数据来源、研究方法、社会测量等诸多重要领域。众所周知，互联网产生了一个相对开放的空间，促成了网络资源共享观念，人们开始反思数据的公共性问题。当下，数据开放特别是公共数据的开放与共享已渐成共识，不仅一些互联网门户入口（如谷歌、百度等）开始提供开放数据，传统数据管理者（例如金融机构、电信运营商以及政府部门等）也开始加入这一开放的潮流。这些数据不同于传统问卷调查收集而来的数据。这种数据来源与数据结构的变化对传统社会学定量研究方法构成巨大挑战。本文试图就互联网对社会学量化研究方法的影响进行分析。

* 本文是国家社会科学基金资助项目“我国新社会群体研究”（批准号 14BSH026）的阶段性成果。

① 参见曼纽尔·卡斯特：《网络社会的崛起》，夏铸九等译，北京：社会科学文献出版社，2006 年。

一、互联网产生新的数据生成与获取方式

首先，互联网产生新的数据生成方式。

互联网的发展，突破了社会学传统问卷收集数据的方式。人们在网络上的一切活动，都可以被记录下来，成为一种可以分析的数据格式。例如，城市中每天几十万人使用公交卡上下班的信息，构成了研究城市生活的重要数据库。这些数据的产生，为社会学的定量研究提供了新的可能。就互联网对传统数据生成方式的突破而言，一个重要例子来自非社会学领域的“谷歌流感趋势”。2009 年 2 月 19 日，《自然》杂志刊登了一篇关于谷歌预测流感的文章。流感初期人们会在网上搜索某些特定关键词，通过汇集这些关键词搜索数据，可以很好地标示流感疫情，这一预测模型被命名为“谷歌流感趋势”。谷歌将预测结果与美国疾病预防控制中心的监测报告进行了比较，结果发现二者惊人地一致。① 虽然针对谷歌的这一研究争论持续不断，② 但却极大地促进了“用户生成数据”这一新的思路在社会科学界的借鉴利用，对社会学定量研究产生了具有突破性意义的影响。社会学定量研究可以不再囿于传统问卷调查所形成的静态、滞后的数据局限，迈向包括行为、态度、心理和空间等多维度的动态的数据分析时代。

其次，互联网产生新的数据获取方式。

互联网新的数据形成方式使得社会学定量研究获取数据的方式也发生巨大变化，不仅使数据的可获得性大大加强，获取的速度也得到大幅提升。以对上海市社会组织的空间分析为例。上海市社会组织的数据来自上海社团管理部门网站所公示的上海所有在册社会组织信息。这一研究通过“网络爬虫”，获得了上海所有在册社会组织的详细信息，包括组织名称、组织注册代码、注册时间、证书有效时间、组织类型、注册地、主管单位、法人代表、地址、邮编、电话、网址、主要业务内容以及奖惩情况等。③ 这种数据获取方式，相对于传统方式，无疑极大降低了社会科学的研究成本。

互联网的发展，还可以使研究者快速获得以前由政府统计部门掌握的数据。如

① Jeremy Ginsberg et al.，“Detecting Influenza Epidemics Using Search Engine Query Data，” *Nature*，vol.457，no.7232，2009，pp.1012-1014.

② Michael W. Davidson，Dotan A. Haim and Jennifer M. Radin，“Using Networks to Combine ‘Big Data’ and Traditional Surveillance to Improve Influenza Predictions，” *Scientific Reports*，vol.5，no.8154，2015，January 29，2015，http://www.nature.com/articles/srep08154.

③ 孙秀林：《社会科学中的空间分析：概念、技术和应用实例》，《山东社会科学》2015 年第 8 期。

在传统社会学研究中，对于贫困指标的测量数据多来源于社会调查和普查数据。近年来，学者们利用手机通讯记录，发现电话通话记录与其他来源的数据指标高度相关，如人口结构、经济活动、贫困指数、能源使用等。在一项对科特迪瓦的分析中，发现手机通讯数据与贫困指标有很强的相关关系，通过手机通讯数据的收集与分析，可以更加精确、快速地评估经济发展情况。①

可佩带式智能设备的迅速普及，使得研究者可以收集人们之间进行社会互动的实时数据。一个研究小组用一年时间研究了一个大学宿舍内部的人际交往行为。他们为所有参与研究的大学生提供了带有特别软件的智能手机，并跟踪收集他们与朋友之间的互动，包括50万小时的对面互动、电话、短信。② 除了手机，新型传感器技术和数据采集智能设备层出不穷，例如智能眼镜、手环、手表等等，不仅可以随时记录人们的社会活动，同时也能实时记录个人的各种生理信息。这为社会学定量分析的跨学科视野提供了数据基础。

互联网的发展，还可以使研究者获得历史性数据。通过考察谷歌图书中的词频可以发现，“美国”一词从复数到单数的演变是一个漫长的过程，复数形式到单数形式的演变在南北战争后稍有加速，直到美国内战结束15年后的1880年，“美国”一词作为单数名词才开始在美国普及。③ 再如，国内有学者利用谷歌图书的百万书籍大数据，提取计算中国近300座城市在英语书籍中出现的历年频率，在300年的时间跨度上展示和分析了这些城市国际知名度的历史变迁。④ 这显然是传统抽样数据无法达到的。

二、互联网促成新的社会学研究方法

首先，互联网促成新的社会学测量方法。

① Sanja Šćepanović et al.，“Mobile Phone Call Data as a Regional Socio-Economic Proxy Indicator，” *Plos One*，vol. 10，no 4. April 21，2015，http://dx. doi. org/10. 1371/journal. pone. 0124160；Christopher Smith et al.，“Ubiquitous Sensing for Mapping Poverty in Developing Countries，” In 3*rd International Conference on the Analysis of Mobile Phone Datasets*，Boston，MA.，May 2013.

② Anmol Madan et al.，“Sensing the ‘Health State’ of a Community，” *IEEE Pervasive Computing*，vol.11，no.4，2012，pp.36-45.

③ 埃雷兹·艾登、让-巴蒂斯特·米歇尔：《可视化未来：数据透视下的人文大趋势》，王彤彤等译，杭州：浙江人民出版社，2015年，第4页。

④ 陈云松：《大数据中的百年社会学——基于百万书籍的文化影响力研究》，《社会学研究》2015年第1期；陈云松、吴青熹、张翼：《近三百年中国城市的国际知名度——基于大数据的描述与回归》，《社会》2015年第5期。

定量社会学研究过程中的一大障碍是缺乏适当的测量。一是概念的模糊性；二是难以获取个体行动者的互动数据。由于数据和观测手段经过多重转换，某种程度上偏离了概念的原始内涵。因此，定量测量与概念之间的偏差在经典社会学研究中是一个长期存在且被相对忽视的问题。在互联网时代，新的数据记录设备和获取手段的出现在一定程度上解决了第二个问题，但定量研究面临的第一个问题或将更为突出。在互联网及大数据背景下，人们所收集或观测的数据基本可归类为日志数据或行为踪迹数据。与基于问卷调查的测量数据不同，这类观测数据在时间和空间上具有连续性，且随着技术性手段的丰富，数据精度越来越高，数据规模越来越膨胀，因此传统的测量方式，例如量表、复合指数或因子分析等或不再适用，或难以提取足够的信息。基于小样本的方差分析和经典线性回归模型面对高维数据已力不从心。在过去 20 年中统计学家发展出的罚则回归（Regularized Regression）模型已被广泛运用于数据挖掘和生物基因科学研究领域，该模型也适用于大规模社会网络的测量。

除了行为数据，数字化文本也开始进入定量研究的视野。基于社交媒体的网络民意或舆情研究就面临如何处理海量文本信息的问题。对于每天产生的社交媒体信息，以人工阅读搭配手工编码的传统定性研究方式显然无法胜任。在经典的文本分词、分类、相似度计算、词频分析等文本算法基础上，社会科学研究者开始研究如何以定量的方式将日常的文本表达与理论概念相对接。这一领域涉及主题建模（Topic Modeling）等文本语义分析方法。

其次，互联网推进实验方法在社会学中的应用。

在社会学研究过程中，导致无法像自然科学那样做出因果推断的一个重要原因，在于社会学研究对象的特殊性，即难以进行完全的随机实验。为解决这一问题，社会科学家设计了多种方法，力图从数据上实现因果推断，包括倾向值匹配、工具变量、结构方程等等。在这些方法中，近年来备受社会学家关注的是一种“准自然实验”（Quasi Nature Experiment）的方法。在复杂的现实社会中，力图使用“准自然实验”方法进行社会干预，获得更为随机分配的实验组和控制组数据，以进行更为严格的因果推断。但这种技术由于投入的人力物力成本巨大，远远超过了传统的抽样调查技术，因此“准自然实验”方法在现实应用中受到很多限制。

互联网的发展，为大规模的“准自然实验”甚至是“自然实验”方法提供了新的手段。例如在 2010 年美国国会选举过程中，有学者研究了社会网络对于个体行为的影响效果。设计思路非常简单，随机给 6100 万名“脸书”（Facebook）用户发送不同类型的“出去投票”的信息：参照组仅仅收到“出去投票”的信息，实验组不仅收到这一信息，还可以看到已经投票的朋友的头像。实验结果令人吃惊，给用户看到自己朋友的投票情况，显著提高了实验组的政治投票行为，社会网络中亲密朋友对实际投票行为的影响效果是“出去投票”

这个参照组的4倍。① 另外一个著名的网络随机试验来自哈佛大学金加里（Gary King）教授。为检验互联网治理的逻辑与运作实践，他的团队做了一项大规模的网络随机试验：在社交媒体上大量创建帐号，并上传随机分配的文本，然后通过全球计算机网络侦测这些文本是否遭到屏蔽。②

最后，互联网促进了社会学可复制性研究的推广。

可复制性研究对于社会学定量分析具有重要意义。陈云松与吴晓刚提出，③ 建立一个透明和开源的学术机制，让研究数据和模型公开共享，使得研究成果可以得到他人的验证和进一步拓展，从而深化社会学定量研究。互联网为定量分析的可重复提供了便捷载体，定量研究的模型技术处理细节，如样本的删节、数据的编码等等，都被详细记录在计算机程序源代码里，算法与数据的公开，有利于研究者利用自己的方法和视角来复制已有的研究结果，从而可以更加有效地完善已有研究。

三、互联网拓展社会学的研究领域

首先，互联网扩展了社会网络研究。

在传统社会网络研究中，存在两种不同的研究脉络：一是个体网研究，主要考虑个体的社会网络（如网络的规模、质量、异质性等）对于个体的效用与影响（如对于收入、找工作、职场晋升的影响等等）；二是整体网研究，如某组织（如公司、学校）内部的人际网络。这些对于个体网和整体网的研究，目前主要表现为两个特点：一是多集中于小规模的、界限相对清晰、闭合的网络结构，如一个村庄、学校、公司等等；二是单一类型网络关系研究，缺乏对复合网络关系的观照。因为无论个体网还是整体网研究都需要获得个体当下或过往的社会互动信息，需大量采用回溯式的调查方式收集个体信息，并依赖被访者个体回答的网络情况数据。但无论是数据获得的可靠性还是数据的规模以及覆盖面，均受制约。

互联网尤其是各种社交网络平台和智能设备的发展，使得社会网络研究突破了这一限制。获取互联网上的社会网络数据的方法，比现实世界中容易得多；同时，

① Robert M. Bond et al.,“A 61-million-person Experiment in Social Influence and Political Mobilization,” *Nature*, vol. 489, no. 7415, 2012, pp. 295-298.

② Gary King, Jennifer Pan and Margaret E. Roberts, “How Censorship in China Allows Government Criticism but Silences Collective Expression,” *American Political Science Review*, vol. 107, no. 2, 2013, pp. 1-18; Gary King, Jennifer Pan and Margaret E. Roberts, “Reverse-engineering Censorship in China: Randomized Experimentation and Participant Observation,” *Science*, vol. 345, no. 6199, 2014, pp. 1-10.

③ 陈云松、吴晓刚：《走向开源的社会学——定量分析中的复制性研究》，《社会》2012年第3期。

在互联网上，个体之间的互动是实时变化的，属于天然的跟踪性轨迹数据。利用在中国新浪微博上抓取的数据，有学者考察了中国社会化媒体空间中的群体类型。还有人采用网上问卷调查的方式，在国内多个网络社区发布答题链接，邀请网友自愿进行答题，并对答案进行分析。也有研究者从互联网上搜索获得中国千人学者的个人信息、发文信息、引文信息等，建构了千人学者的论文合作网络，计算并分析了千人学者的跨国学术资本转移情况。①

其次，互联网深化了社会参与研究。

线上与线下的互动成为社会学研究的一个新热点。以中国的业主论坛为例，随着互联网的发展，业主论坛在中国社区生活和社区治理中发挥着越来越重要的作用。陈华珊以一个业主论坛为切入点，通过对复合网络关系的分析，区分了不同类型的虚拟社区用户参与在线讨论的特征及其和社区在线参与之间的关系。②

最后，互联网推进了城市社会学研究。

国内社会学对于城市议题的关注，具有明显的人文主义色彩，定量化的实证分析比较缺乏，如西方城市研究中已经非常成熟的“社会区”（social area）分析、因子分析（factor analysis），尚不多见。究其原因，是国内获取含有地理信息系统（GIS）的城市数据比较困难。在互联网时代，随着城市生活中移动通讯、全球定位系统、社会化网络等技术的日益广泛应用，人类活动本身成为分析城市空间结构与城市活动的重要数据来源，极大地扩展了社会学城市研究的视角。通过这些带有地理信息的数据，不仅可以分析城市中活动的行为轨迹，还可以分析城市空间结构对人们活动的影响。例如公交卡刷卡数据，显示了一个城市中较为基本的活动情况，从中可以发现城市活动的模式、规律；把这种新的数据格式与传统的数据格式结合在一起，就可以发现许多以往研究的视角盲区，更好地理解人们在城市中的行为模式与空间特征。

四、互联网时代的社会学研究：一个初步的思考

互联网尤其是大数据的发展为社会学研究扩展了新的领域。但很多情况下，学者们只是用新的数据、新的方法重新验证了旧的问题而已。如对于网络空间的研究，

① 参见桂勇等：《网络极端情绪人群的类型及其政治与社会意涵——基于中国网络社会心态调查数据（2014）的实证研究》，《社会》2015 年第 5 期；马得勇、王丽娜：《中国网民的意识形态立场及其形成——一个实证的分析》，《社会》2015 年第 5 期；杨张博等：《近朱者赤：基于社会网络分析方法的归国者跨国社会资本转移研究》，《社会》2015 年第 4 期。

② 陈华珊：《虚拟社区是否增进社区在线参与？一个基于日常观测数据的社会网络分析案例》，《社会》2015 年第 5 期。

虽然学者们也在不断强调线上线下的互动，但更多学者还仅仅把网络空间作为现实社会关系的一种虚拟映射。那么，互联网带给社会学研究的，仅仅是传统研究议题的一个网络版本，还是其本身就能够生成传统社会学之外的全新的研究题目？在互联网背景下，社会学理论的意义何在？定量研究者如何应对理论指导下的因果判断？

对上述问题的回答，需要学者们做出大量的实证研究并在此基础上进行讨论和理论提升。

首先，互联网时代，由于新的数据来源多样海量、更新迅速，对传统社会学定量研究提出严峻挑战。仅就互联网数据获取来说，涉及编程、数据库、网络传输、文本解析、格式转换甚至分布式计算和云存储等各种技术环节，这些技术已成为获取数据的一种必要手段。伴随着数据规模的膨胀以及数据异构性的增加，在分析阶段，建模方式也不再局限于传统的基于假设检验的概率统计模型，主体建模、文本语义分析、深度学习、复杂网络建模等都已进入社会学研究者视野。

其次，互联网快速发展的时代，对于任何一个研究者来说，面对复杂纷繁的统计模型和算法，想要了解所有的统计分析模型，已成为一种不可能的任务。因此，新时代的社会学研究更需要突破传统单兵作战的思路，应鼓励学者参与不同学科之间的交叉合作。

从技术与方法层面看，互联网尤其是大数据的发展，使社会学的传统方法面临巨大挑战，也为社会学定量研究的方法更新与变革带来不可多得的机遇。不失时机地把握互联网新技术并运用到社会学研究尤其是定量研究中，可以更加深入研究中国社会的独特议题和社会发展脉络，进而发展出具有中国本土意义的研究题目。

〔责任编辑：刘亚秋　责任编审：冯小双〕