流动的信息网络：反思社交媒体大数据研究的迷思与问题

◎王喆[1，2]

（1.浙江传媒学院 文化创意学院，浙江 杭州 310018；2.台湾政治大学 传播学院，台北）

【摘要】社交媒体上的人类活动形成了大数据资料库，吸引了社会科学研究的注意力，造就了大量的研究和讨论。本文针对社交媒体大数据分析在方法上和概念上所面对的迷思和挑战，提出大数据具有其技术存有论之基础，参与了现实的构造，进而造就了当下的社会人类。文章进一步提出可探讨的关键议题，如“数据需要海量的大吗”“数据分析是量化研究方法吗”以及“传播研究应该如何用数据来说故事”等等，进一步厘清究竟应该如何通过社交媒体上的流动数据来进行社会科学研究，以实际行动来增进研究者在这一领域的分析能力。

【关键词】大数据　社交媒体　研究方法　复杂网络　效度

【中图分类号】G201　【文献标识码】A　【文章编号】1003-6687（2017）1-075-05

【DOI】10.13786/j.cnki.cn14-1066/g2.2017.01.014

一、信息社会的研究挑战

互联网技术已经渗透到人们的思想与社会空间，信息增长的速度非常之快。Stevenson指陈：从各种网络所负载的信息和它们对各种社会组织所具有的更加广泛的影响角度来看，传播网络是最重要的。[1]网络的重要性在日常生活中已经显而易见，譬如，社交媒体已经在当代人们日常生活中扮演起重要角色。人们都有这样的共同经验：在捷运或公交车上、下课十分钟，人们几乎一有空闲，就低头滑手机或平板。他们使用微博、Facebook或Twitter，通过点赞、分享、聊天或打卡以阅读讯息和分享讯息，表达当下心情或传递自己的实时动态，这些活动已经成为人们每日生活的例行实践。

社交媒体中的信息现在变成了最关键商品，在个人与团体的松散联结中流通于时与空之中，形成并再度形成（form and re-form）某些流动身份／液态身份（fluid identities）与联结；这些身份与联结在过往是受制于地点、时间、阶级、性别与种族等因素的。此外，社交媒体往往又并非只具有单一的社交性。网络平台由于逐利性，会在平台基础上叠加很多诸如新闻传播、资金金融以及营销推广等其他层面的可能性，虚拟和现实世界的互动日趋复杂。以此角度来看，如Urry所说，网络不是反映出社会，而更是形塑甚至生产了社会。[2]Latour则提出，社会在日常实践中被聚合，并构成多重、异质、不平均、动态的“全球流”（global fluids）的全球网络（global networks）；[3]这网络中，人、信息、物、金钱、影像与危机以惊人速度和难以预测之形态混乱并跨地域地移动。Urry说，这些“流”（fluids）没有清晰的出发与到达，没有一定必要的结束状态，并且其特色为浮现（emergent）、没有意图（unintended）与非线性结果（non-linear consequences）。

如是一个流动网络对社会科学研究者提出了很大的挑战。对于那些关注网络上的人们所留下来的踪迹以

【基金项目】浙江传媒学院2015年度校级科研项目（ZJ15XJ023）

【作者信息】王喆（1985— ），女，湖南长沙人，浙江传媒学院文化创意学院媒体创意系讲师，台湾政治大学传播学院博士在读，主要研究方向：社会化媒体与数据挖掘。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

及人们在互动中所形成的社群的研究者而言，必须找到更好的探究切入方式和观察对象。在社交媒体中，研究者既要看到个人，也需要观察小团体，另外这些个人/小团体处于一个网状结构（即多点对多点），信息实时更新、水平流动，造就了大数据，这些数据不仅仅是数字、符码，更是文字和图像（参见表1）。

表1　大众媒体与社交媒体的特质比较

特质	大众媒体	社交媒体
消息源头	社会机构	个人/小团体
发布结构	星状结构（一点对多点）	网状结构（多点对多点）
更新周期	固定的周期更新	实时更新
讯息流向	讯息垂直流动	讯息水平流动
发言主体	专业组织内的人员	实名/匿名的大众
样本数量	有限	大数据

那么，在这样一个流动网络中，研究对象和传统的研究对象的异同之处为何？如果进行研究的话，因果关系应如何确认？可以肯定的是，在这样的流动网络中，能即时将流动踪迹捕捉到的，必然是“量化一切的”数据化，以数据来测量、记录和分析世界。数据化并不是数字化，文字可以变成数据，方位/空间可以变为数据，沟通也可以成为数据。当我们这样去看待流动网络中的数据时，必然导向大数据研究的更深层面。

二、网络大数据的技术存有论

由于通过计算机运算来收集与分析网络数据可以让研究者捕捉到线性/非线性历程中的记号，在Google、Yahoo等大型网络公司，或美国政府组织之中，可计算的社会科学变成一个很重要的领域，通过收集与分析大量资料，可以揭露个体与团体的行为模式。这促成了一个很热的新词：“大数据”。坊间对其最流行的定义是四个V：海量（volume），快速（velocity），庞杂（variety）和价值（value）。而就学术研究而言，boyd与Crawford的界定更加值得人们注意，他们将“大数据”定义为奠基在科技（计算机运算技术、大量链接、分析能力）、分析（找出经济、社会、科技的趋势）与神话（大众普遍认为“大数据/海量数据”能提供进阶智能或知识，解决以往悬而未决的问题，且更为客观真实）交互作用下的现象。[4]“大数据”现象的出现，并不仅仅是人们表面上所看到的，采用了大量的数据、分析工具和操作程序，它其实重新形塑了知识建构与研究流程，以及人们应该如何处理和分类信息，这一趋势被称为“向计算转”（computational turn），并逐渐扩及人文社会研究领域。根据祝建华近年对SCI/SSCI期刊上已发表的270篇有关大数据的研究论文的统计，大部分都是2011年后出现的。其中，来自计算器科学和工程技术的最多，分别占了27%；其次是医学、生物、化学（20%）和基础研究的数学、物理等（11%）；最少的是工商管理（8%）和社会科学（7%）。[5]

在“大数据”这个概念备受关注的同时，也携带而生了很多关于这一研究技术的迷思，而且这些迷思往往纠缠于“大数据”的4个V特质，认为拥有“大数据”就能“多快好省”地洞悉一切，认为其将客观、精确地预测未来、建立模型、洞察世事，认为其是“取之不尽，用之不竭”的数据创新。科技决定的自动化研究也许会让我们产生全知者的幻觉，误以为自己只要分析数据的模式就能对复杂、多元面向的真实了如指掌。

这种对“大数据”的信仰是一种技术为中心的迷思。社会科学是建立在工具文明上的，欲求采用可分析/测量及有分析/测量的办法。早期哲学和社科经典都写作于实证几乎不大可能的时代。比如，在二战之前，基本上不存在大规模的民调、完整的社会数据和科学上严谨的统计技术等等，所以大多数社科经典的写作方式是逻辑式的，从概念到概念，从推断到推断，创造出了很多漂亮的理论框架，但却没有检测、证伪的可能性。待到测量技术发展起来，才从经验判断过渡到了数据检验。

根据海德格尔的看法，技术并不只是人类达成某种目的的工具，而是参与了我们所面对的现实之构成；仅仅将技术视为人类的工具，并不能捕捉到技术的本质。我们造就了现代科技（计算器、网络、社交媒体、数据分析工具），但同时现代科技也参与了现实的构造，进而造就了我们。人与技术之间的关系并非一种工具性的关系，而是具有存有论意义的关系。比如，“大数据”的运算转向强调其快速性，而黄厚铭也提出技术发展的核心是速度，有了更快速的科技，在一定的时间内可以完成更多的事情，因此在和过去相同的时间间隔里，事情却也可以有更巨大的变化。[6]速度与加速不仅是科技发展的目标，也是科技发展对社会文化所产生的影响。但是，这一“大数据”的迷思会带来研究上的风险，亟待研究者进行反思，也值得进一步厘清究竟应该如何通过社交媒体上的流动数据来进行社会科学研究。

三、数据分析的争议性问题

把运用数据分析流动的信息网络作为研究方法与当下虚实夹杂的网络社会实存互为映像，然而，对数据的想象，不同典范/学门的研究者恐怕会有多重的方向以及断裂。数据分析的起点是EAD（Exploratory Data

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Analysis，探索性数据分析），它由John W. Tukey在1977年提出，指出统计建模应该结合数据的真实分布情况，对数据进行分析，而不应该从理论分布假设出发去构建模型。EDA是对调查、观测所得到的一些初步的、杂乱无章的数据，在尽量少的先验假定下进行处理，通过作图、制表等形式和方程拟合、计算某些特征量等手段，探索数据的结构和规律的一种数据分析方法。在这步骤之后，恐怕数据分析可以有两个分岔的取径，一个是统计学的取径，计算变项之间的关系；另外一个则可以是诠释学的取径，讨论深层意义的浮现。可以说，对于社会科学而言，数据的诠释不应只止于统计的诠释，而更应该有关于人的洞察。

1. 数据需要海量的大吗？

正如前文所言，大数据四个V的第一个就是海量，而“大数据”的神话之一就是过分夸大其无与伦比的“大”。在此先举一例，自2008—2009年起，美国国家研究计划（NEH）开始运用高效计算机运算进行数字人文研究，包括对于十八世界思想家、地图、文件、影像的分析，及其与十九世界美国铁路、犯罪数据的关联，数据量共约1.27亿字。然而，即使是在这个架构下最大型的计划，它的数据量实际上也远小于计算机科学家所言的“大数据”。这些实践让学者们看到，在探讨数据分析和研究时，更应该看到的是数据是否能够有效地回答研究者所提出的研究问题，而不是一味求“大”。

“大数据”所带来的去抽样的可能性在目前环境下是没法实现的。由于技术、商业、机密和其他原因，除了少数大数据的原始拥有者（一般都是商业权力所有者，如淘宝、Facebook，或政治权力所有者，如国家电网、教育网等），真正的大数据恐怕是难以拿到的，有更多变项的、跨数据库的大数据更难近用。祝建华将大数据分为小规模、中型规模和巨型规模三种。小规模的数据非常多，免费就可以得到。中规模在大部分情况下也可以是低成本的。但真正巨型规模的大数据其实是得不到的。

“大数据”并不意味着它是“更好的数据”。举例而言，很多海量数据的研究是从Twitter之类的网站找出人们在特定事件发生时的行为。但是Twitter使用者不代表人本身——账号与人之间的联结为何、有无断裂？此外，社交媒体上所获取的大数据也不等于全部的数据。Twitter会过滤掉问题字眼的推文（像是色情暴力等），因此能捞到的数据样本无法呈现完整的数据。最后，研究者在捞资料的时候，只能捞到公开文章，且不确定手上的样本与全部文章的比例关系。

最后，“大数据”仍然会有效度的问题。郑宇君整理了批评者对于“大数据”的总体性方面的批评。[7]这些批评者认为，大数据分析中，研究者多半只从一两个平台进行研究，可得性偏误（availability bias）、模式生物（model organism）的问题仍然存在。根据在线资料所得的研究成果未必能延伸到不同平台，而且还需要将数字落差纳入考虑。[8]即使是经由Google、Twitter等API取得大数据，继而从中推论人类行为，这些数据的量也都还太小，比如相较于大学程度与中产阶级生活形态的人，低收入与低教育程度的人较少有机会近用网络或是创造数字内容。因此，需要从方法论的角度来探讨这些捕捉而来的踪迹如何有可能代表社会整体，而不只是那些上线的人，研究者必须认知到这些大数据比原先所想象的还要小。

2. 数据分析是量化研究方法吗？

如前文所言，“大数据”可以将一切量化，但是在此基础上的分析方法是量化研究方法吗？首先可以探讨的是，量化研究应该处理什么问题？更重要的是，它能够处理什么样的问题？

过度迷信数据即量化，仍然是一种将社会科学与自然科学视为相同的研究途径，他们认为社会科学与自然科学一样，有其不变的、固定的本质存在，因此社会研究者要做的便是去搜集与世界相关的资料，如此便有可能完全了解这世界的存在以及事件的因果关系。然而，事实是社会与自然有其差异，社会存在着许多偶然性，事件与事件之间并不存在着强而有力的决定性关系，其本质并非固定的，因而我们不可能百分之百地掌握到世界运作的模式。

当将数据分析纯粹视为量化研究方法时，所存在的问题是可能会被量化研究方法牵着鼻子走，抛弃了以问题意识出发，进而使用相应的研究方法这一正确做法，反倒以研究方法、标准出发而去设定问题、观察事件。如果再深入去探讨数据分析，特别是“大数据”的测量问题，会发现它其实只是相对浅层次的测量：只采用可捞取的数据、可实用的测量方式，但未必是最有效的方法，仅仅是试图以低层次的数据（具有变异量的数据较易在低层次找到）去解释高层次的问题。

此外，一般认为“大数据”是计算器自动化分析方法的主导研究（尤指机器编码），一般人也会认为计算机的自动分析方法比较可靠，不易发生偏误。不过这一论点仅在讯息量相当庞大（数以千计）的情况下有效。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

在实际研究过程中，研究者处理数据时仍是加入了主观因素。研究者是数据的诠释者，他们对于数据的想象、选择或排除的变项，以及诠释方式都未必中立客观。

综上所述，恐怕无法断言数据分析，特别是“大数据”是所谓严谨的量化研究。数据现象所造成的结果很有可能只是类似的实际现象，未必能进行进一步的推论。比如，即便两个人通话记录频繁、通话时间又长，这一数据也不能代表这两人感情好。如果要进一步推论，需要更多的信息才行。从这个层面上而言，数据分析重视的是“关系”，而非“因果”。运用“大数据”进行数据分析来切入社交媒体的研究可以不是量化研究。

如何运用“大数据”来进行综合分析？Tufekei提出“质性抽取”（qualitative pull-outs）作为辅助，[9]研究者可以随机抽取样本子集询问一些问题来检验行为的多样性，这些质性抽取并不一定要大量，但可以帮助研究者进行诠释。Manovich则强调，通过计算机运算与人文研究所收集到的深层数据，是两种截然不同的数据内容，并没有任何一方比另一方优越，也没有一方能取代另一方，研究者必须要懂得如何问出新形态的问题，结合两种不同的数据内容进行分析探索，才能扩展社会科学领域的研究方向。[10]

3. 传播学应该如何用数据做研究？

在“大数据”相关研究兴起之后，由于社交媒体数据之取得往往掌控在商业经营者手上，大部分的研究只有这些社交媒体公司才能接近真正大量的社群数据，特别是传输数据（用户上传与下载数据的时间与地点），其实其他学术社群则无法近用。然而，这些商业经营者在获取“大数据”之后，往往将其视为具有“商业意义”的预测工具，用来“校准”或“实验”使用者的行为和意图，以便更好地吸引广告商对社交媒体进行投入，获取更多的利益。而很多时候，这一切研究都只是表层的描述，而非解释。

社交媒体数据运算和数据追踪分析研究中的关键问题是“数据的社会意义”。数字痕迹（digital traces）无处不在不代表它们具备研究价值，它们的真正价值在于“是谁产生它们”。在方法论层次上，应该把持的立场是，这些数据不是不证自明，而需要更精细的概念语汇与理论架构来进行探讨。在此，笔者希望在今后的研究中，进一步从浮现的社交网络中看到浮现的讯息复制/再制网络，并从复杂理论切入，而并非固执于“预测”“监控”和“实验”。Ovortrup认为可将复杂理论用于新数字媒介，认为浮现的媒介增加了社会复杂性，同时也成为管理混沌的一个工具。[11]基本的做法是透过社交媒体的数据收集、整理与分析，首先提取社交媒体上某些静态特性，包括节点（nodes，也就是使用者）和边（edge，也就是一个使用者与另外一个使用者之间的直接传递）及其他重要特征，在此基础之上，加入空间和时间的面向，检验社交媒体信息网络的动态特性，突显其浮现的复杂网络。

在这网络之中，意义是关于某人，某事，且在特定范围之内的。更重要的是，数据分析后所看到的浮现网络指向了类似于“何谓有意义的行为”这样的问题。Winch对于社会科学的理念中强调的是：[12]社会科学研究的对象应是他所谓的“有意义的行为”。何谓有意义的行为呢？在Winch看来，也就是社会成员的“遵守规则的行为”，而影响行动者行动的规则并非普世皆同的，宗教有其规则、科学有其规则，更细致地说，不同社群、不同阶层的人都有着不尽相同的规则对其行动产生着影响力。因此，研究者若是想要理解社会上的行动者，首先必须理解影响着行动者行动的背后规则为何，由被研究者的角度去赋予行动的意义。这样的观点与自然科学寻求的一律性以及一般性通则相差甚巨。

数据跟踪与分析研究尚处于起步阶段，其复杂性无疑将随时间增加。传播和社交运算这两个领域无疑会保持自己的独特方式，因为他们有各自的基础性任务。但是，这两个领域彼此间有很多地方可以相互学习，每个人都应该继续借鉴对方的长处。

结语：从数据到知识

马克思：“一切科学都是多余的，如果事物的外表和本质全部吻合。”存有论的深度——在存在于表层的东西之外，在我们能够看到的东西之外，它们还有别的存在层次。这些“底层”有其特殊的意义，因为它们能够解释我们所看到的。如果数据分析只让我们看到浮现的信息网络的现象，那么恐怕研究者需要以其他的研究方法来进行补足。

数据是杂乱无关的，信息是整理之后的，最后知识是经过验证、可以被拿来再讨论的东西。我们的知识是因我们的历史处境（而非个人的处境）而定的：不同历史阶段的不同社会会产生不同形式的知识。“大数据”带来了极大的争议性，随之产生的知识也并非确凿。我们能够在不同的理论之间做选择，并说某一种理论比另一种理论好，这就说明并非任何形式的知识都一样好。然而，这样选择的基础判准是包容性，即某一理论涵盖

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的范围比其他理论大。而这就预设了两件事：第一，称职的知识乃是关于整体的知识，而非关于不同枝节部分的知识；第二，历史有其发展的逻辑，引导我们去获得这样一种称职的知识。本文只是一个开始，还有更多关于“大数据”的方法论讨论可以进行。

参考文献：

[1] Stevenson N. Understanding media culture [M]. London, UK: Sage Publicatiion of London Newdehi, 1995: 182.

[2] Urry J. Mobile sociology[J]. British Journal of Sociology, 2000 (1): 185-203.

[3] Latour B.Reassembling the social [M]. UK, Oxford: OUP, 2006.

[4] boyd D. & Crawford K. Critical questions for big data [J]. Information, Communication & Society, 2012(5): 662-679.

[5] 祝建华．一个文科教授眼中的大数据：多、快、糙、噪？[J]. 大数据中国，2013（1）：10-12.

[6] 黄厚铭．迈向速度存有论——实时性电子媒介时代的风险[J]. 新闻学研究，2009（101）：139-175.

[7] 郑宇君．向运算转：新媒体研究与信息技术结合的契机与挑战 [J]. 传播研究与实践，2014（1）：67-83.

[8] Schradie J. Big Data not big enough? How the digital divide leaves people out[J/OL]. PBS Mediashift, 2013. http://www.pbs.org/mediashift/2013/07/big-data-not-big-enough-how-digital-divide-leaves-people-out.

[9] Tufekci Z. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls [J]. In ICWSM' 14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 2014.

[10] Manovich, L.Trending: The promises and the challenges of big social data[M]. In M. K. Gold (Ed.), Debates in the digital humanities. Minneapolis, MN: The University of Minnesota Press, 2011.

[11] Ovortrup, L. The hypercomplex society[M]. New York, NY: P. Lang, 2013.

[12] Winch, P.The idea of a social science[M]. London, UK: Routledge & Kegan Paul, 1958.

Fluid Information Network: Rethinking the Myth and Problem in Social Media's Big Data Researches

WANG Zhe（School Of Cultural Creativity, Zhejiang University of Media and Communications, Hangzhou 310018, China; National Chengchi University, Taipei 11605 , China）

Abstract: Large-scale databases of human activity in social media have captured social scientific attention, producing a number of researches and discussions. This paper considers methodological and conceptual challenges for this emergent field, with special attention to the social media big data analyses. This paper tends to raise discussions based on technology ontology, which participates in constructing reality, and even affects social human being. Further key issues are proposed, such as "should big data be that big" "data analysis is a quantitative research method" and "how could researches in communication studies narrate via big data" and so on. In order to answer how to do social scientific research through fluid data in social media, and to improve researchers' ability by actual action.

Key words: big data; social media; research method; complex network; validity

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net