第 24 卷第 2 期 中南大学学报(社会科学版) Vol.24 No.2
2018 年 3 月 J. CENT. SOUTH UNIV. (SOCIAL SCIENCE) Mar. 2018
DOI: 10.11817/j.issn. 1672-3104. 2018. 02. 019

“计算”的边界：互联网大数据与社会研究

郝龙

(武汉大学社会学系，湖北武汉，430072)

摘要：互联网大数据计算，是当前社会研究方法创新的主要方向之一。部分纯数据驱动型学者认为，大数据独立于研究之外生成，不仅能记录下人们的真实态度与自然行为信息，又可以摆脱研究者与研究本身的干扰，由此形成了“总体性”“真实−自然性”与“客观性”三大认识假定。然而，无论是由数字鸿沟造就的年龄与阶层边界和由差异化生产划定的群体与主题边界，还是由数据操纵和数据引导带来的虚假(非真实)与偏态(非自然)状况，以及潜藏在整个数据生产−挖掘−分析过程中的人为干扰，都证明上述假定在很多情况下并不成立。认清互联网大数据的可“计算”边界，对于推动数据计算在社会研究中的应用有着重要的理论与方法意义。

关键词：互联网；大数据；计算范式；数据缺失；数据偏态；数据操纵

中图分类号：C91 **文献标识码**：A **文章编号**：1672-3104(2018)02−0148−11

社会学的量化研究以数据资料为基础，大数据时代的到来，使运用海量数据和新的数据处理技术，对人类行为、群体互动乃至社会复杂适应系统进行研究成为可能。可用于社会研究的大数据，依其生成方式大体可分为三类：第一类是基于人机互动在互联网和移动互联网平台上生成采集的互联网大数据①，包括社交关系数据、网络文本数据、电子踪迹数据等；第二类是通过各种传感器采集而来的物联网大数据，手机位置信息是其典型类型；第三类则是通过数字化与数据化手段由既有信息资料转制而成的大数据，例如谷歌图书语料库(Google Books Corpus)[1]。在三类数据中，互联网大数据由于承载着大规模、长时段、连续关系性和意义性信息，被认为将赋予社会学“改变我们对生活、组织和社会的理解”的潜力[2]。

单从名称上看，“大数据”好像是在强调与传统量化数据相比所具有的更大个案数量或信息规模。然而实际上，两种数据无论是在数据性质还是生产逻辑上都存在着质的差异：传统计量方法分析的是数值型数据(numerical data)，这些数据是出于特定研究目的而运用实验、问卷调查等方法有计划地观测的结果，即数据生产本身就构成了研究的一项重要组成部分。新型计算方法所处理的则是计算机代码型数据(code data) ——“作为数据的可解释代码和作为代码的数据”[3]，这些数据独立于社会研究之外。数据生产的独立性，也决定了其在社会研究中的边界。在计算范式下，数据分析的焦点不再是能测量到什么，而是“已经生产出什么”；不再是“能否有效且稳定地测量”，而是“是否真实且准确地生产”。[4]在由“可观测性”议题转向“可获得性”议题的过程中，围绕着大数据计算形成了一系列认识假定，其中对社会研究最为重要的有“总体性”“真实−自然性”“客观性”三大假定。“总体性”假定指大数据时代的到来，开启了“样本=总体”的全数据模式，数据代表性问题将不复存在；“真实−自然性”假定指互联网上记录的是人们行为互动的真实踪迹和“自然状态”下的表达；“客观性”假定指基于大数据的研究可以避免研究者个人因素的影响，能够获得传统研究方法无法企及的、带有真理性、客观性和准确性的见解。然而，将大数据运用于社会研究，就会发现实际情况并没有预想的那么乐观。

一、缺失与分隔：互联网大数据的代表性边界

“总体性”假定来自于迈尔-舍恩伯格和库克耶的《大数据时代：生活、工作与思维的大变革》，他们将大数据理解为不同于抽样数据的全体数据，称大数据

收稿日期：2017−12−10；**修回日期**：2018−02−23
基金项目：国家社科基金重大项目“大数据时代计算社会科学的产生、现状与发展前景研究”(16ZDA086)
作者简介：郝龙(1988—)，男，山东新泰人，武汉大学社会学系博士研究生，主要研究方向：数字社会学与计算社会学

是指不用随机分析法这样的捷径，而采用所有数据的方法"[5](56)；并且认为"社会科学是被'样本=总体'撼动得最厉害的学科"[5](41)。这样笼统地宣称"采用所有数据"的潜台词似乎是——在大数据时代，一切社会科学研究都能够用总体数据来分析。这一观点对传统定量研究者而言无疑有着巨大的吸引力，因为如果真的可以获得"全样本"，就意味着不存在数据代表性问题，社会研究结论的准确性和适用范围将得到显著提升。国内有些学者直接接受了"总体性"假定，认为"抽样误差曾经是长期困扰社会科学研究的重要难题，而全样本作为大数据最重要的特征，甚至可以将抽样误差降为零"[6]。然而，"总体性"假定在表述上是含混不清的，在社会科学研究中，"总体"是相对于研究对象和研究问题而言的，在没有明确研究对象的情况下谈论总体，其实是毫无意义的。舍恩伯格等研究者未能对数据的"可计算性"和"可获得性"之间的差异作出清晰的分辨，他认为随着计算能力的日益强大和数据处理技术的日益进步，对获取到的所有数据已有能力进行有效的分析，无需再因计算条件(能力、成本、时效等)的限制而采取随机抽样方法压缩数据体量。然而，在数据生产与科学研究相分离的背景下，可获得的所有数据不一定等同于研究对象的所有数据。这是不能脱离具体研究问题来下结论的。正因如此，国内有些学者对此问题的论述陷入自相矛盾，他们一方面沿袭舍恩伯格的观点，强调大数据的全样本特性，另一方面又承认很多时候并不能获得总体数据[7−8]。鉴于此，有必要对"总体性"假定进行细致的分析，以矫正相关认知偏差。

(一) 数据缺失："数字鸿沟"下的年龄与阶层边界

在现实生活中，计算设备的获得和使用会直接受到支付成本、技能学习、生活需求等社会因素的影响，从而使互联网大数据生产过程本身具有明显的社会属性[9]。对这种社会属性最直接的考察，便是检视网民群体结构与总体人口结构的对应程度。

据第 41 次"中国互联网络发展状况统计报告"显示，2017 年中国网民规模达到 7.72 亿，而按照当年总人口数计算，中国互联网普及率只有 55.8%，仍有近一半的中国人口未能成为互联网大数据的生产主体。当然，如果这种缺失只是群体比例上的随机缺失，可以通过统计手段加以修正[10](186−187)。但现实情况却不尽然，仅从年龄结构来看，2017 年，中国网民群体以 40 岁以下人口为主，40 岁以上网民只占总网民数的 23.6%，不到 1.82 亿人；而同年龄段的实际人口，占总人口数的比例却接近五成[11−12]。以往的研究表明，中国互联网的使用不仅会受到使用者年龄因素的影响，更与其收入、受教育程度和城乡差异等因素紧密相关[13−14]。即便只是对使用者的年龄、收入、受教育程度与城乡结构四个因素的交叉列联也会发现，仅凭 40 岁以上的网民群体规模是无法实现对同年龄段总人口变异性的整体覆盖，尤其是覆盖那些年龄较大、收入较低、学历不高、居住在农村的群体，其中大部分人的日常行为和态度意见都没有被记录在互联网大数据之中。例如，新浪微博发布的"2016 微博用户发展报告"显示，82%的微博用户年龄在 30 岁以下，40 岁以上用户不足 7%；77.8%的用户受教育程度为大学及以上层次，初中及以下层次用户同样不足 7%[15]。

"数字鸿沟"(Digital Divided)的一系列研究对数据缺失背后所隐含的社会意义有所揭示。数字鸿沟概念，最初被用于描述因网络设备接入的不均衡所引发的信息分配的不平等现象[16]。对互联网大数据而言，"数字鸿沟"现象的存在意味着部分社会成员作为数据生产主体的缺场，其态度与行为信息无法在网络中获取。"数字鸿沟"不仅出现在网民与非网民群体之间，同样也出现在网民群体内部。随着研究的不断深入，社会学家们普遍意识到，由互联网的接入与否所引发的区隔问题，不过是"数字鸿沟"的表现形式之一。社会的结构性不平等因素，同样会在网民群体之间制造出使用频率、需求程度、技能水平和信息素养等方面的显著差异，由此引发数据生产上的"次级数字鸿沟"问题[17]。

"数字鸿沟"理论表明，受个人技术能力、经济条件和社会需要等因素的限制，社会大龄群体和底层群体在成为互联网大数据生产主体问题上普遍面临着更多的障碍。这些群体中只有少数成员成为了网民，他们无论是在行为方式还是态度意见方面都不足以代表全部成员，其所生产出的数据信息也无法涵盖群体内的所有变异性特征[18]。可以说，"数字鸿沟"现象的存在，使互联网大数据不可避免地存在数据缺失问题。在以往量化研究方法中，数据缺失是指所要观测的变量取值未能被测量到，或测量结果的信度太低而无法使用。然而在大数据研究领域，"缺失"的内涵发生了变化，用以描述受成本支付和主体偏好等因素的影响，社会研究所需要的数据未能在互联网络中生产或储存下来，因而研究者无法获取关于特定社会群体或研究主题的全部必要信息。由此类数据的绝对缺失所带来的信息恒定缺损，以至难以甚至无法以统计学方式来加以弥补或矫正。

(二) 数据分隔：差异化生产下的群体与主题边界

在网络经济时代，作为纯粹人工制品的互联网服务，多数情况下都是以一种商品化的形象呈现在世人面前。由这些服务平台所生成的各类互联网大数据，也就变相成为对互联网商品/服务消费过程或消费结果的一系列记录资料的集合。因此，在互联网大数据所蕴含的全部意义中，最首要也最基本的便是数字化媒介的消费意涵。各种信息商品，在具备使用价值的同时，被赋予了远比以往更为丰富的符号属性，指向着“通过区别符号来生产价值社会编码的目标”[19](69)。隐含在数字消费背后的种种“社会性功能”，会在互联网大数据的生产过程中画出一条不甚清晰的群体与主题边界，进而制造出一种“数据分隔”现象，即不同数据源所承载的信息在生成主体、内容主题等方面存在明显的差异[20]。

互联网络数据分隔现象的出现主要有两大动力，一是发生在不同专业领域之间的专门化过程，二是发生在同一领域或服务类型内部的差异化过程。伴随着现实社会中的领域分化与职能分工，互联网中出现了诸多面向特定领域的专门化网络服务平台(例如评论股市行情的炒股论坛、探讨病情的病友贴吧、分享技术知识的科技论坛)。与综合性信息网站相比，这些专门化网站通常具有专业信息覆盖面宽、信息内容规范且体系完整、原创信息丰富和信息增长率快等优势，并由此成为相关专业群体获取信息与沟通交流的重要平台。由此，专业/领域边界便转化为数据信息中的群体与主题边界。另外，在自由竞争的互联网市场中，尽管的确存在着一些占优势甚至主导地位的平台或服务，但多个服务商竞相满足同一需求的状况仍是市场中的主流。现代消费社会的一大特征是，商品的消费取代生产成为经济活动组织的主导形式。网络服务商们普遍有针对性地依据各种必要的或建构的细分市场需求来组织商品的生产与营销，从而带来同一类型互联网服务之间的差异化。这种差异化又会因消费者的选择偏好——既有统计意义上的个体偏好，亦有数字认同(涉及数字媒介中对不同象征符号的组织、串联、赋值与解读)作用下的群体偏好[21]——而持续得以巩固，并由此引发用户群体的进一步分化。

由中国互联网信息中心(CNNIC)开发的中国互联网数据平台，提供了各种类型的网络站点与软件客户端总覆盖人数及其比例的重合分析计算功能。从新闻、购物、浏览器与视频客户端四种网站与软件类型中各选取三个代表性服务商，对覆盖人数的重合比例(见表 1)进行分析后发现，数据分隔现象的确存在。表 1 显示，各网站或软件尽管存在着覆盖人数比例上的差异，但均拥有多则千万少则百万的独占用户规模(即只访问或使用特定网站或软件的用户)。当这些独占用户不具备独特的群体属性时，其存在并不会对社会学互联网大数据计算造成严重影响。然而，通过对用户群体的基本属性分析，可以发现不同类型网站或软件的独占用户之间存在着明显的群体特征差异(见表 2)。例如，由表 2 可知，在三种视频客户端平台的独占用户中，(1)爱奇艺与搜狐的男性用户多于女性，后者的男性用户占比甚至超过七成，而优酷的女性用户则要远高于男性用户；(2)四成以上的爱奇艺独占用户年龄在 20~29 岁之间，近五成的搜狐用户为 30~49 岁群体，优酷用户中 19 岁以下年龄段群体占比则接近四成；(3)爱奇艺独占用户集中于华东与中南地区，搜狐用户集中在华北和西南地区，而优酷用户则集中在华东和西南地区，其中西南地区比例占比超过四成；(4)爱奇艺和搜狐均有七成左右的独占用户分布在二、三级城市之中，后者有着更高的四级城市用户比例，而优酷则

表 1 2017 年第 1 季度四种类型网站与软件覆盖人数的重合比例分析(%)

购物网站访问	比例	新闻网站访问	比例	网络浏览器	比例	视频客户端	比例
1-淘宝	87	1-网易新闻	56	1-IE 浏览器	57	1-爱奇艺	62
2-京东	60	2-搜狐新闻	74	2-360 浏览器	63	2-优酷	68
3-天猫	72	3-新浪新闻	63	3-搜狗浏览器	7	3-搜狐视频	40
只访问 1	13	只访问 1	11	只使用 1	31	只使用 1	18
只访问 2	8	只访问 2	18	只使用 2	40	只使用 2	25
只访问 3	3	只访问 3	9	只使用 3	2	只使用 3	10
1-2 重叠	50	1-2 重叠	39	1-2 重叠	22	1-2 重叠	39
1-3 重叠	67	1-3 重叠	36	1-3 重叠	4	1-3 重叠	27
2-3 重叠	45	2-3 重叠	48	2-3 重叠	1	2-3 重叠	26
1-2-3 重叠	43	1-2-3 重叠	31	1-2-3 重叠	0	1-2-3 重叠	23

注：本表数据采集自中国互联网数据平台；四类网站与软件的总覆盖人数分别为 25 235.2、287 89.4、28 033.8 和 18 056.2 万人；网址：http://www.cnidp.cn/

表 2　四类网站与软件代表性服务商独占用户群体特征的比例分布(%)

群体属性		购物网站			新闻网站			网络浏览器			视频客户端		
		淘宝	京东	天猫	新浪	搜狐	网易	IE	360	搜狗	爱奇艺	搜狐	优酷
性别	男	58.33	53.66	58.07	63.86	53.81	58.84	47.36	58.57	42.29	54.21	72.69	40.65
	女	41.67	46.34	41.93	36.14	46.19	41.16	52.64	41.43	57.71	45.79	27.31	59.35
年龄	19 岁以下	12.41	−	−	0.94	0.38	16.16	13.66	15.44	0.58	−	11.92	39.41
	20~29 岁	22.72	36.05	60.9	33.48	36.88	23.45	26.7	26.99	24.81	44.69	19.82	23.22
	30~39 岁	27.77	26.56	20.14	25.82	36.09	32.36	29.46	26.53	36.46	23.88	26.21	17.65
	40~49 岁	16.24	25.51	11.40	19.48	14.23	19.98	19.9	15.25	14.98	19.41	22.25	8.73
	50~59 岁	10.01	7.12	6.72	8.44	6.39	8.01	6.27	4.30	19.00	3.60	4.04	2.40
	60 岁以上	10.84	4.76	0.85	11.83	6.03	0.04	4.01	11.49	4.17	8.42	15.76	8.60
地域分布	华北	12.03	9.77	7.24	7.63	11.30	13.96	9.9	15.37	15.65	7.94	24.54	5.65
	华东	41.01	19.22	15.12	33.14	27.04	42.64	31.49	35.8	36.06	31.65	17.65	26.33
	中南	29.23	34.42	39.39	26.67	26.86	23.21	31.42	24.3	19.44	23.85	20.3	17.21
	东北	7.22	12.50	6.53	5.72	11.90	4.61	8.89	4.73	14.14	3.76	9.94	5.84
	西北	2.91	4.09	3.27	2.48	8.18	7.23	5.12	3.32	9.70	7.67	5.85	2.02
	西南	7.56	19.99	28.45	24.35	14.71	8.35	13.19	16.47	5.01	20.61	21.72	42.95
城市分布	一级城市	8.12	20.45	8.05	14.52	14.82	6.76	18.28	15.1	14.03	14.5	13.92	7.12
	二级城市	55.64	51.21	52.07	45.07	46.61	56.07	42.28	41.77	31.46	46.44	26.48	67.49
	三级城市	26.71	15.66	24.79	33.32	22.67	28.02	24.3	29.96	36.66	27.61	34.22	21.6
	四级城市	8.22	12.57	15.09	6.47	14.73	8.60	14.51	12.69	17.84	10.08	22.64	3.39
	五级城市	1.30	0.11	−	0.29	1.16	0.55	0.63	0.48	−	1.38	2.73	0.40
独占用户数(万人)		3 281	2 019	757	3 167	5 182	2 951	8 690	11 214	561	3 250	4 514	1 806

注：本表数据采集自中国互联网数据平台；时间为 2017 年第一季度；网址: http://www.cnidp.cn/

有七成以上的用户居住在一、二级城市之中。独占用户之间的类似群体差异，同样出现在生成购物大数据的电商网站平台和生成电子踪迹数据的浏览器与视频客户端软件平台之上。

性别、年龄、地域分布与城市分布等变量对亚文化群体的形成具有直接影响，这几乎是社会学、人类学等学科的基本共识。由此可知，在专业化与差异化过程的双重作用下，数据信息中不同专业群体之间、不同社会阶层之间、不同主题内容之间的边界会被隐秘地树立起来。数据分隔现象的存在意味着社会研究难以藉由单一或少数几个数据源获取关于研究对象的全部可用数据，不同的数据源同研究主题操作化定义的匹配程度(即数据效度)也会有明显的差异。因此，在社会学互联网大数据研究，特别是那些与专业化议题或特殊群体紧密相关的研究中，预先明确数据源的群体属性与核心主题范围，就成为考察数据代表性与数据效度的重要指标。

二、虚假与偏态：互联网大数据的准确性边界

互联网大数据的“真实−自然性”假定指向两个问题，一是数据信息本身的真伪，二是数据生成过程的“自然”与否。在 IBM 公司早期的 5Vs 模型中，真实性(veracity)曾一度被视为大数据的基本特征之一[22]。这种认识的建立，主要基于以下理由：一方面，“机器不会说谎”，人们在互联网上所呈现的任何心态与行为信息都会被计算机直接保留下来，不存在对信息的选择性记录与存储；另一方面，他们也相信摆脱了调查/实验情境和研究者面对面的影响，大众将普遍缺乏造假或说谎的直接动机而会表露出自身的真实意图。有学者在同传统测量方法比较后提出，“许多大数据是人们活动行为的实时和真实的记录，鲜受人类记忆、偏好和情感的干扰，这将会在很大程度上排除人

们因主观性以及对概念的误解等因素对调查内容的误填和烂填”[23]。然而面对各种谣言、刷单等网络虚假信息层出不穷的现实，这一特征已开始受到越来越多的质疑。与对“真实性”的广泛存疑不同，“自然发生性”(naturally occurring)仍被许多研究者视为互联网大数据的基本属性之一[24]。他们相信，人们更多是根据个人兴趣与需要来有选择地获取与筛选不同主题的信息，并基于自身知识与态度对信息做出判断和反馈；其中，信息的发布、搜索、阅读、转发与评论，作为大众个体意识的自然表达或“平常状态”[5](42)被记录在互联网大数据之中，成为“描绘复杂的人类感官世界，展现个体真实的内心世界，体察驿动的心理动态，挖掘丰富的人际交互，探索人类社会总体价值走向”[25]的可靠资料。对于上述假设，产生如下疑问：互联网大数据都是真实可靠的吗？数据生产果真都是自然发生的吗？数据所承载的信息是否会存在着某种形式的缺损？

(一) 数据操纵：互联网大数据生产中的“非真实”

如前所述，绝大多数互联网服务平台都带有浓厚的商业色彩，互联网大数据的生成在一定意义上可以被理解为销售行为与消费行为交互作用的结果。在市场逻辑之下，围绕着信息的生产、分配与交换形成了一种“数据商业”。所谓“数据商业”，在此指的不是某种纯粹的数据或信息买卖业务，而是一种数据生产的基本逻辑：商业化的盈利导向为互联网大数据的生成提供了一种“例行程序”，一方面，指导和限定着数据的整体生成框架；另一方面，“还要尽力掩盖渗透其中的经济逻辑”[26](129)。正是这种商业化逻辑的存在，将大量人为操纵因素注入到互联网大数据中。例如在网络购物数据中，考虑到销售量和既有评价是潜在消费者购物的评判依据，部分网络销售商便采用恶意刷单、雇佣“水军”等方式人为篡改销售数量和好评度，甚至出现了许多专业“刷单”公司和“水军”公司；部分消费者也可能出于获取返利等目的刻意编造好评，将大量虚假信息注入互联网大数据之中。网络打车平台上，也曾出现过大量为骗取平台补贴的“恶意刷单”现象。再如，网络搜索引擎服务商作为“信息把关人”，在数据商业逻辑的影响和缺乏外部监管的情况下，难免会进行“权力寻租”——由于搜索结果的排名先后会直接影响其点击率，搜索服务商出于增加广告营利目的，普遍对搜索结果中的优先位置进行计价销售。2008 年的“屏蔽百度抓取”事件[27]和 2016 年的“魏则西”事件[28]，便充分暴露出百度搜索引擎通过竞价排名对搜索结果排序的人为操纵以及由此所产生的社会后果。

除了商业利益驱使外，政治利益也是数据造假的重要动机。大数据时代带来的政治–文化后果便是“大数据政治”，即一种技术“殖民”社会的权力结构体系。随着互联网对政治活动影响的不断加深，以往身体化的参与行为日渐让位于虚拟的鼠标点击行动(clicktivism)，支持或反对的程度被认为可以通过点击、阅读和转发的数量来衡量[29]。由此，大数据使政府与企业决策过程中的公众角色不断弱化，取而代之的则是数据化的“幻影公众”[30]，此后果不仅使数据生产者的主体性受到侵蚀——这正是哈贝马斯对“技术官僚统治”的忧虑所在[31]，而且也使数据信息本身的真实性遭遇严峻的考验。例如，英美等国媒体就曾曝光过美国政府通过开发网络机器人和注册虚假社交媒体账号等方式伪造民意的新闻[32]。对于互联网大数据研究而言，这些人为操纵之下形成的虚假数据，如果不能被有效甄别与剔除，就意味着数据可能存在巨大的系统性偏差，势必导致研究结果出现严重错误。然而，虚假数据的甄别与剔除，目前仍是有待深入解决的技术难题和社会难题。

(二) 数据偏态：互联网大数据生产中的“非自然”

有学者认为，无论社交媒体中的聊天信息、电子邮件，还是各类服务平台上的购物记录和电子踪迹等，都是在未受研究者干预条件下自然发生的，反映着行动者的客观真实状态。该观点的缺陷在于，过分关注互联网大数据生成过程的技术维度而忽略其社会维度。一方面，互联网大数据的生成平台能够通过程序设计与议程设置等方式，对数据生成过程产生直接引导作用，影响着所能生成的数据形式与信息内容。另一方面，社会大众的数据生产过程实质上就是其在互联网空间开展社会行动与互动的过程，这一过程除了受到群体环境的影响，其本身会带有明显的现实情境特征。正因如此，即使能摆脱研究者与研究本身的影响，也改变不了互联网大数据中存在着诸多其他社会因素影响的现实。

1. 数据引导

数据引导，即通过人为设计与限制等方式影响信息生产过程与结果的行为。除了权力监管(如网络删帖、敏感词屏蔽)这种显性形式之外，数据引导还会以“数据算法”的隐蔽形式潜藏在互联网大数据的生成过程。有学者指出，那些看似“自然”的互联网大数据，其实在生成过程中就已经掺杂进了大量人为的设

计因素。Facebook 和 Twitter 等社交网站通过不停地调试，将友谊、受欢迎程度等转换成某种算法，同时把这种算法宣称为某种"社会共享"的价值观念。点"赞"和"热门话题"这样的网站按钮虽然可能被认为是自然的在线社交活动，但并不能掩盖构成这些按钮的算法，本质上是被精心调制出来用于引导人们点击响应的[33]。

除了基于算法设计的技术引导，数据引导还可以通过直接的人为干预方式发生。社会注意力研究也证明，网络时代的大众媒介和部分精英群体同样也能够通过议程设置和框架建构，对受众的注意力分配发挥明显的引导与形塑作用。与心理学关注神经性活动不同，社会学和经济学将"注意力"视为特定结构与情境条件下，与信息处理相关联的可组织配置的一种社会性资源[34]。在信息爆炸与信息过载的网络时代，"注意力"开始取代信息成为社会中的稀缺资源②，其分配——注意力资源在不同信息对象之间的配置结构——会对信息获取的方向、主题及其处理方式与效率产生重要影响[35]。传播学认为，大众媒介对信息受众的注意力分配发挥着引导与建构作用。议程设置与框架理论指出，在以往信息与信源匮乏的时代，人们为获取信息必须紧紧依附于有限的大众媒介，并在媒介的影响下配置自身的注意力。"随着时间的推移，媒介议程中报道对象的显著性会转移到公众议程上，媒介不仅能成功地告诉我们去想什么，而且能成功地告诉我们如何去想"[36]。进入网络时代之后，这种议程设置现象并未随着信源数量的迅猛增长和信息议题的多元化而消失。表 3 呈现了由 2017 年 3 月 26 日至 5 月 20 日 8 周时间内，中国传统新闻媒介在新浪微博中每周影响力的排名。由表中数据可知，尽管存在着许多网络意见领袖(微博 VIP 会员)和自媒体账号，但微博中传统新闻媒介依然拥有巨大的舆论影响力；与印刷形式上的多样性相比，大量受众的注意力逐渐集中到少数传统新闻媒介上[37]。议程设置现象的存在，将会造成网络舆情大数据在主题分布上的极度不均衡，那些未被纳入议程的主题很可能面临数据量过小或样本代表性不足等潜在问题。此外，各类自媒体的出现，很多情况下也不过是将议程设置的主体由大众媒介拓展至部分网络精英群体[38−39]。"2016 微博用户发展报告"就指出，新浪微博的 VIP 会员的发文量超过普通用户近四倍[15]。一项关于 Twitter 网中信息生产主体的研究也显示，精英用户群体尽管只占全部用户的极少部分，却生产出该平台近 50%的信息[40]。自媒体的精英属性，也使议程的选择难免带有偏见与人为谋划色彩。

表 3　部分传统新闻媒体微博账号影响力每周排名

(2017.3.26—2017.5.20)

微博账号	主办单位	周数							
		一	二	三	四	五	六	七	八
人民日报	人民日报社	2	2	3	2	11	2	1	5
央视新闻	中央电视台	6	4	6	6	−	7	7	11
环球时报	人民日报社	7	5	7	8	−	13	10	24
人民网	人民日报社	8	8	8	7	−	17	9	19
中国新闻网	中国新闻社	13	12	11	17	−	−	18	−
新京报	光明日报社与南方日报社	20	24	23	−	−	20	−	−
澎湃新闻	上海报业集团	−	19	−	−	−	−	16	−
新华网	新华社	−	−	−	−	−	−	25	−

注：本表数据采集自清博指数微博榜单数据平台；该榜单为全部微博账号的每周总排名，各微博账号影响力由发博数、转发数、评论数、点赞数等活跃度和传播度指标计算而来；受篇幅限制，只选取排名前 25 位的传统新闻媒介账号，如果周未出现在该范围内则用"−"号标示

2. 环境塑造

在社会学互联网大数据的生产主体中，政府、媒介与商业公司等专业内容生产者只构成了其中的一小部分，绝大多数是那些普通的互联网用户。这些用户被认为会以"主动自我报告"或"自我曝光"形式在互联网上持续生产各种类型的心态与行为信息，将自己的真实面记录在数据之中。由此，可以引申出如下问题，即在互联网上大众是否真的所言/行如所想？关于网络舆论中从众行为、传染行为与"沉默的螺旋"现象的研究均显示，许多情况下人们并非会按照自己所想的那样去行事。首先，互联网中的数字认同与社交互动中的同质性偏好，造成了网络结构上的不均匀，信息受众普遍分散在内部关系紧密而外部关系稀疏的各个子网络之中。受子网络群体的影响和压力，网络

成员有可能倾向于隐匿自身的想法而试图与其他群体成员保持一致[41−42]。其次，由从众行为衍生而来的社会传染研究，更进一步揭示出网络中特定成员会将关系邻接者的行为作为现实的情境因素加以解读，并可能受其传染而出现行为上的主动趋同化[43−44]。最后，在被动从众与主动趋同之外，还存在着一种受众保持“沉默”的可能。“沉默的螺旋”理论认为，人们会出于害怕孤立的心理，预先评估特定议题下的意见分布状况，并判断不同意见之间的优劣地位。当他们估计优势意见与其个人意见相去甚远，且不愿改变自身立场时，便倾向于保持沉默[45]。沉默的直接后果是使优势意见的强者地位得到进一步强化，劣势意见则更趋于沉默，这种循环往复的作用会严重损害网络舆情大数据中的态度多样性信息[46]。

实际上，除了群体压力以外，权力监控下的自我隐私保护同样会带来互联网用户的主动沉默。福柯曾指出现代权力体系的两大特征，即从统治权向生命权力的拓展以及与之相配合的“全景敞视主义”。当数字化技术成为人们身体的延伸，大数据计算在一定意义上便成为强化生命权力的工具；而时时刻刻的“数据监测”，则进一步提升了对社会的“全景监控”能力。在部分学者看来，数据规模愈大，数据生成主体就会变得愈加“透明”，这与现代社会所强调的隐私权利保护背道而驰[47]。尽管存在着各种数据的匿名与脱敏技术，但对性别、年龄、族群或亚文化群体信息的披露，仍会涉及到对群体隐私权的侵犯[48]。所谓群体隐私，是一个群体以其整体的名义而非群体内各成员的个人名义所享有的社会权利[49]。以往关于“群体污名”的研究已经表明，在一个不平等社会中，任何群体间的明显差异都有可能成为建构群体污名甚至社会区隔的意义基础。这种对个人/群体隐私的潜在侵扰，势必会给数据生产制造障碍，对个体或群体隐私保护意识的强化，会窒息数据生产主体的创造意愿，使其对重要信息进行刻意隐瞒甚至主动篡改，并由此对数据质量带来严重损害[50]。

三、价值有涉：互联网大数据的客观性边界

自从孔德创立“社会物理学”并将其视为“标志着实证主义的最终胜利”开始[51]，实证主义方法论就在社会研究中划分出一条科学与非科学的边界。它在“将社会科学放置到了低于自然科学的位置上”的同时，也“将人文学科贬低到了一个虚幻的主观性领域”[52]。然而长久以来，无论是质性研究中“萨摩亚”之争和“墨西哥特波茨兰村”之争背后隐含的研究者主观价值分歧[53]，还是量化资料收集过程中调查者对被调查者的外在干扰[54]，都持续证明着绝对“价值中立”与完全“客观”在社会研究中的难以企及[55]，甚至引发出一场关于社会学方法论危机的讨论[56−57]。

大数据时代的到来，将社会学的实证主义情结再次唤醒。二进制的计算机代码将大数据时代描绘成一个摆脱人为干涉的纯数字时代，“客观性”被视为互联网大数据的基本特征之一。由于数据生成与社会研究相分离，加之可用数据集中信息的极度丰富及其多维属性，研究者个人因素所造成的“观念先行”“材料拼凑”和“以偏概全”等问题被认为可以有效避免。在此基础上，一种称为“大数据神话”的观点被提出，它认为“大数据集提供了一种智力和知识的更高级形式，可以生产出以往无法企及的、带有真理性、客观性和准确性的见解”[58]；“从海量的客观数据中得出的结论要比传统抽样统计分析得出的结论更为可靠”[59]。然而也有学者提醒我们，大数据其实“并没有看起来的那么简单”[60]。事实上，数据生产与社会研究的相互独立，不但未能将人为干扰从互联网大数据中排除出去，反而会招致比传统研究方法更多的干扰因素。各种或显性或隐性的人为干扰，潜藏在由数据生成到数据挖掘再到数据分析的整个处理链条之中，持续威胁着计算结果的客观中立性。

首先，数据生成环节上的算法设计及其变更，构成了互联网大数据研究的第一层人为干扰。网络数据的生产通常都需要依靠特定的程序设计才能得以实现，但这些程序作为一种人工制品本身远非尽善尽美，需要经过不断的调试、更新与升级。在此过程中，程序本身的明显调整(如搜索推荐算法变更)，会在基于该程序所生成的数据中制造出某种不甚清晰的断裂。拉泽尔等就曾指出，谷歌出于商业目的对网络搜索推荐算法的变更，是造成谷歌疾病预测走向失败的一个重要原因[61]。与此同时，不同平台间算法设计上的差异，也为数据的匹配和关联制造了障碍。一个典型的例子是，中国几大主要门户网站(如网易、搜狐)在互联网平台和移动互联网平台上普遍采用了两种不同的推荐算法——前者以人工审核和操作为主，后者则使用着基于机器学习的智能推荐算法，其结果将造成无法对两大平台的数据直接关联或相加。

其次，即便数据生成本身足以客观中立，但对数

据的选择与获取仍无法逃脱人为因素的干扰。一方面，作为有价值的商品和有意义的信息，数据公布或隐藏背后普遍隐含着商业利益或权力斗争方面的考量；另一方面，数据采集同样是一个主观操作的过程，无论是主题的选择还是信息的取舍，背后都难免涉及各种差异性甚至矛盾性的价值取向。有学者提出大数据作为一个庞大的原始信息集合，本身并不是不言自明的，对数据的解释必然要向各种哲学辩论开放[62]13。换句话说，存在着社会学互联网大数据研究的价值边界，这要求研究者应当始终持有一种反思意识，即"基于谁的利益，出于什么目的"进行数据的收集与计算[33]。

最后，人为干扰因素同样充斥在数据分析阶段。由于数据与其生产者之间的关系并非是不证自明的，对任何数据的分析都难免掺杂进分析者的个人解读。人们绝不仅仅是制度规范的机械执行者，社会行为及其意义必须被置于其所发生的特定情境中才能被更好地理解。然而，现实情况是许多情境因素都未能被记录在大数据之中；即使那些被记录下来的情境因素，一部分也难免会在数据清洗与抽取过程中被排除出去。脱离了具体情境下的意义结构，我们即使能够发现人们在"抽动眼皮"，却永远也难以直接明确分辨出这一行为的意义所在[63](7-8)；而对于那些被解读出来的意义，也应当不断反问，这是数据本身的意义，还是分析者个人所赋予其的意义？清醒认识到这一点，有助于在社会研究中避免对数据意义的过度解读。

四、结语：互联网大数据的社会研究边界

当下热门的"大数据"一词，最早是由著名计算机企业的专业技术人员提出的。由于这一概念很快就夹杂了大量商业宣传的声音，许多有市场炒作之嫌的观点也逐渐大行其道[64]。部分社会学研究者囿于自身的知识结构，一时难以对以计算机科学为基础的各种大数据观点进行准确的判断，故而出现了一些误解或认知偏差。对于社会研究来说，大数据的总体性、真实-自然性、客观性假设，在很多数情况下其实并不成立。首先，无论是由"数字鸿沟"现象划分出的年龄与阶层边界，还是由"数据分隔"现象所筑起的主题与群体边界，都说明大数据尽管体量庞大、类型多样，但只能完成对现实社会信息的部分记录，在大多数情况下并非什么"总体性"数据，故而对数据中可能存在的信息缺失必须加以考量，数据代表性问题依然需要检视。其次，在那些已经被生产出来的互联网大数据中，还或多或少存在着信息造假与数据偏态的问题，有些数据的真实性和准确性值得怀疑。最后，在互联网大数据的生成与计算过程中，还暗含着大量显性或隐性的人为干扰因素。从源数据生成阶段的信息造假与意义建构，到数据处理阶段的算法设计和变量选取，再到数据分析阶段的意义解读，大数据的生成、采集和计算中很可能存在着比传统研究方法更多的人为操纵和干扰因素。

受线上/线下诸多因素的影响，大数据中普遍存在数据缺失、数据偏态与人为干扰等问题，这决定了将其运用于社会研究会存在一定的范围限制。这些限制部分源于当前信息社会发展的不成熟，随着互联网与移动互联网的持续普及，越来越多的社会成员将有机会成为数据生产的主体，从而使大数据的代表性问题得到不断优化；另一些导源于互联网大数据结构性问题——如专业化和偏好因素作用下的数据分隔现象、经济利益驱使下的数据造假现象、个体性格与社会心态影响下的信息偏态现象、变量操作化与挖掘分析中的价值有涉现象——的限制将会始终存在。

本文反思三大假定的目的，在于认清当前互联网大数据的研究边界，这对运用大量新型数据和新兴工具与方法开展计算范式下的社会研究而言至关重要。因为，对于市场营销而言，大数据获取与计算的及时性、高效性和低成本，能够服务于企业的快速决策和市场细分，对数据的代表性、真实性和客观性的要求有时并不严格[65]。但对于学术研究而言，上述三大假定的成立与否，直接决定着计算结果是否真正具有学术价值。尤其是在将研究结果应用于社会治理时，更应注意数据中潜藏的不平等、刻意隐瞒与人为操纵等现象，防止缺失、偏态与强加意义的数据分析结果成为政策制定的错误依据。

对研究边界的讨论，一定程度上也是在回答有关大数据计算与传统研究方法之间关系的问题。部分纯数据驱动型学者认为，传统的研究方法、研究逻辑和认识路径(理论假设——数学模型——统计检验)基本上已经过时，借助于丰富的海量数据和强大的复杂算法，可以在不需要理论的前提下做出准确理解和精确预测[66]。在拉泽尔等人看来，以为大数据计算可以取代传统方法的观点是一种"大数据狂妄"(Big Data Hubris)[61]。如今，互联网早已不再是与现实社会相平行的"虚拟社会"，其本身已成为现实社会的一部分。大数据计算尽管首先表现为一种技术和方法层面上的

意义，但同时也越来越带有着浓厚的社会属性。因而基于大数据计算所得出的结论并不一定是绝对客观的真理，可能还需要通过传统研究方法加以补充和验证。

当然，边界的存在决不意味着没有价值。在研究边界内，大数据计算能够从远比以往更为丰富的数据资料(其中大部分过去并不存在)中挖掘出有价值的信息，可能会帮助研究者发现一些以往未被认识或未能深入了解的社会规律[67]。尤其是在转向复杂自适应系统理论的过程中，大数据计算对社会复杂性、社会适应性、微观行为的宏观涌现性等问题的处理要比传统方法更具优势[68]。此外，大数据计算的相关技术，也为传统质性与量化资料的处理提供了新型的、更高效的分析方法和手段。当前，基于大数据的社会研究尚处在方兴未艾的时期，关于其基本概念、理论、方法和技术的研究仍有待进一步深入。将传统质性与量化研究方法同大数据计算，特别是人工智能技术相结合，将会为社会研究的未来发展助益良多。

注释：

① 根据生产主体的不同，互联网大数据可以细分为两种类型，即专业生成内容(professional generated content，PGC)和用户生成内容(user generated content，UGC)。本文讨论的互联网大数据主要着眼于后一类型，即非专业机构用户在互联网上生产和累积的各类数据化信息。

② 受人脑智力条件和社会时间资源等因素的限制，人类被认为只具备获取与处理信息的有限能力。尽管各类先进的数字化设备与技术极大地拓展了这一能力，但其本质上仍存在着一定的上限。

参考文献：

[1] 郝龙, 李凤翔. 社会科学大数据计算——大数据时代计算社会科学的核心议题[J]. 图书馆学研究, 2017(22): 20−29.

[2] Lazer D, Pentland A, Adamic L, et al. Social science. computational social science[J]. Science, 2009, 323(5915): 721−723.

[3] Wing J M. Computational thinking[J]. Visual Languages and Human-Centric Computing, 2006, 49: 33−35.

[4] 段伟文. 大数据知识发现的本体论追问[J]. 哲学研究, 2015(11): 114−119.

[5] 迈尔-舍恩伯格，库克耶. 大数据时代: 生活、工作与思维的大变革[M]. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版社, 2014.

[6] 唐皇凤, 谢德胜. 大数据时代中国政治学的机遇与挑战[J]. 新疆师范大学学报(哲学社会科学版), 2016(1): 95−104.

[7] 唐文方. 大数据与小数据: 社会科学研究方法的探讨[J]. 中山大学学报(社会科学版), 2015(6): 141−146.

[8] 刘涛雄, 尹德才. 大数据时代与社会科学研究范式变革[J]. 理论探索, 2017(6): 27−32.

[9] 何其聪, 喻国明. 我国城市互联网用户使用社会化媒体的现况考察——使用时长、类型偏好、认知-使用率及在线社交活动的若干特征[J]. 当代传播(汉文版), 2015(3): 29−32.

[10] 阿利森. 缺失数据[M]. 林毓玲, 译. 上海: 上海格致出版社, 2012.

[11] 中国互联网信息中心. 第 41 次“中国互联网络发展状况统计报告”[EB/OL]. (2018-01-31) [2018-02-23]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201801/P020180131509544165973.pdf.

[12] 中华人民共和国国家统计局. 中国统计年鉴 2017[EB/OL]. (2017-06-26) [2018-02-23]. http://www.stats.gov.cn/tjsj/ndsj/2017/ indexch.htm.

[13] 李冠强, 陈雅, 李强. 中国互联网用户网络使用行为分析[J]. 中国图书馆学报, 2004, 30(5): 43−46.

[14] 汪明峰. 互联网使用与中国城市化——“数字鸿沟”的空间层面[J]. 社会学研究, 2005(6): 112−135.

[15] 新浪微博数据中心. 2016 微博用户发展报告[EB/OL]. (2017-01-11) [2018-02-23]. http://data.weibo.com/report/reportDetail?id=346.

[16] 曹荣湘. 数字鸿沟引论: 信息不平等与数字机遇[J]. 马克思主义与现实, 2001(6): 20−25.

[17] Van Dijk, J. Digital divide research, achievements and shortcomings[J]. Poetics, 2006, 34(4): 221−235.

[18] 罗俊, 罗教讲. 数据密集型知识发现的边界与陷阱——以美国大选预测为例[J]. 学术论坛, 2017, 40(3): 1−7.

[19] 布希亚. 消费社会[M]. 刘成富, 全志刚, 译. 南京: 南京大学出版社, 2001.

[20] Liu J, Li J, Li W, et al. Rethinking big data: A review on the data quality and usage issues[J]. Isprs Journal of Photogrammetry & Remote Sensing, 2016, 115: 134−142.

[21] Akerlof G A, Kranton R E. Economics and identity[J]. Quarterly Journal of Economics, 2000, 115(3): 715−753.

[22] Bendler J, Wagner S, Brandt T, et al. Taming uncertainty in big data: Evidence from social media in urban areas[J]. Business & Information Systems Engineering, 2014, 6(5): 279−288.

[23] 丁小浩. 大数据时代的教育研究[J]. 清华大学教育研究, 2017, 38(5): 8−14.

[24] Shah D V, Cappella J N, Neuman W R. Big data, digital media, and computational social science: Possibilities and perils[J]. Annals of the American Academy of Political & Social Science, 2015, 659(1): 6−13.

[25] 陈潭, 刘成. 大数据驱动社会科学研究的实践向度[J]. 学术界, 2017(7): 130−140.

[26] 麦克马那斯. 市场新闻业: 公民自行小心？[M]. 张磊, 译. 北京: 新华出版社, 2004.

[27] 杜骏飞. 百度“屏蔽门”事件: 网络社会的敌人[J]. 传媒,

2008(10): 14−17.
[28] 凌永辉, 张月友. 市场结构、搜索引擎与竞价排名: 以魏则西事件为例[J]. 广东财经大学学报, 2017 (2): 109−116.
[29] Halupka M. Clicktivism: A systematic heuristic[J]. Policy & Internet, 2014, 6(2): 115−132.
[30] 袁光锋. 政治算法、“幻影公众”与大数据的政治逻辑[J]. 学海, 2015(4): 49−54.
[31] 曹卫东. 开放社会及其数据敌人[J]. 读书, 2014(11): 73−78.
[32] Fielding, N & Cobain, I. Revealed: US spy operation that manipulates social media [EB/OL]. (2011-3-11) [2018-02-23]. https://www.theguardian. com/technology/2011/mar/17/us-spy-operation-social-networks.
[33] Van Dijck, J. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology[J]. Surveillance & Society, 2014, 12(2): 197.
[34] 练宏. 注意力分配——基于跨学科视角的理论述评[J]. 社会学研究, 2015(4): 215−241.
[35] 汪丁丁. “注意力”的经济学描述[J]. 经济研究, 2000(10): 67−72.
[36] 麦克斯韦尔-麦考姆斯. 议程设置理论概论: 过去、现在与未来[J]. 郭镇之, 邓里峰, 译. 新闻大学, 2007, 93(3): 55−67.
[37] Hamilton J. All the news that's fit to sell: How the market transforms information into news[M]. Princeton: Princeton University Press, 2004: 197.
[38] 王平, 谢耘耕. 突发公共事件中微博意见领袖的实证研究——以“温州动车事故”为例[J]. 现代传播-中国传媒大学学报, 2012, 34(3): 82−88.
[39] 禹建强, 李艳芳. 对微博信息流中意见领袖的实证分析: 以“厦门BRT公交爆炸案”为个案[J]. 国际新闻界, 2014, 36(3): 23−36.
[40] Wu S, Hofman J M, Mason W.A., et al. Who says what to whom on twitter[C]. In Proceedings of the 20th international conference on World Wide Web, 2011: 705−714.
[41] 朱琳, 汪蕾, 陈长, 等. 网络信息传播的从众行为研究——以微博为例[J]. 现代情报, 2014, 34(12): 17−22.
[42] 刘锦德, 王国平. 网络舆情传播的从众效应[J]. 江西社会科学, 2015(5): 234−239.
[43] 王世龙, 谢光明. 社会网络中的行为传染研究述评[J]. 人民论坛, 2016(8): 164−166.
[44] Cheng J, Danescu-Niculescu-Mizil C, et al. Anyone can become a troll[J]. American Scientist, 2017, 105(3): 152.
[45] Noelle-Neumann E. The spiral of silence: A theory of public opinion[J]. Journal of Communication, 1974, 24(2): 43−51.
[46] 罗俊, 罗教讲. 互联网舆情偏态传播与引导[J]. 人民论坛, 2015(36): 25−27.
[47] Rubinstein I S. Big data: The end of privacy or a new beginning?[J]. Social Science Electronic Publishing, 2013, 3(2): 74−87.
[48] Zwitter A. Big data ethics[J]. Big Data & Society, 2014, 1(2): 1−6.
[49] Floridi L. Open data, data protection, and group privacy[J]. Philosophy & Technology, 2014, 27(1): 1−3.
[50] Tene O, Polonetsky J. Privacy in the age of big data: A time for big decisions[J]. Stanford Law Review Online, 2012, 64(63): 63−69.
[51] Giddens A. Positivism and its critics[C]// In Bottomore, T., Nisbet, R. eds. A History of Sociological Analysis. New York: Basic Books, 1978: 237−286.
[52] 约翰·扎米托. 科学哲学: 从实证主义到后实证主义[J]. 刘鹏译, 淮阴师范学院学报(哲学社会科学版), 2013(1): 28−35.
[53] 朱炳祥. 反思与重构: 论“主体民族志”[J]. 民族研究, 2011(3): 12−24.
[54] 郭淑华. 现代社会调查真实性所面临的挑战[J]. 社会, 2003(5): 22−24.
[55] 苏国勋. 社会学与社会建构论[J]. 国外社会科学, 2002(1): 4−13.
[56] 张小山. 实证主义社会学面临挑战[J]. 社会学研究, 1991(5): 114−126.
[57] 吴小英. 社会学危机的涵义[J]. 社会学研究, 1999(1): 52−58.
[58] Boyd D, Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon[J]. Information, Communication & Society, 2012, 15(5): 662−679.
[59] 罗小燕, 黄欣荣. 社会科学研究的大数据方法[J]. 系统科学学报, 2017(4): 9−12.
[60] Goldston D. Data wrangling[J]. Nature, 2008, 455(7209): 15.
[61] Lazer D, Kennedy R, King G, et al. The parable of Google Flu: Traps in big data analysis[J]. Science, 2014, 343(6176): 1203−1205.
[62] Bollier D. The promise and peril of big data. Washington[R]. Washington DC: Aspen Institute, Communications and Society Program , 2010.
[63] 格尔茨. 文化的解释[M]. 韩莉, 译. 南京: 译林出版社, 1999.
[64] Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics[J]. International Journal of Information Management, 2015, 35(2): 137−144.
[65] 冯仕政. 大数据时代的社会治理与社会研究: 现状、问题与前景[J]. 大数据, 2016, 2(2): 3−16.
[66] Anderson C. The end of theory: The data deluge makes the scien tific method obsolete [EB/OL]. (2008-06-23) [2018-02-23] https: //www.wired.com/2008/06/pb-theory.
[67] Watts D. Computational social science: Exciting progress and future challenges[J]. The Bridge, 2013(4): 5−10.
[68] Conte R, Gilbert N, Bonelli G, et al. Manifesto of computational social science[J]. European Physical Journal Special Topics, 2012, 214(1): 325−346..

The boundary of computation: Internet big data and social survey

HAO Long

(Department of Sociology, Wuhan University, Wuhan 430072, China)

Abstract: The computation of internet big data is one of the main directions of innovation in current social survey methods. Some purely data-driven scholars believe that independently generated big data can not only record information about people's real attitudes and natural behaviors, but also get rid of the interference by researchers and the research itself. Therefore, big data is considered to be “general” “real-natural” and “objective”. However, whether the age-class boundaries created by the digital divide and the boundaries of groups and topics delineated by differentiated production, or the false (non-real) and skewed (unnatural) conditions brought about by data manipulation and data guidance, or human interference in the process of data mining-mining-analysis, they all prove that the above assumption does not hold in many cases. Therefore, recognizing the computable boundary of internet big data has important theoretical and methodological significance for promoting the application of data computing in social survey.

Key Words: internet; big data; computational paradigm; missing data; skewed data; manipulated data

[编辑：谭晓萍，游玉佩]

(上接第 95 页)

Transportation, information accessibility and regional ecological efficiency: Considering the study of space spillover effects

ZOU Xuan, HUANG Meng, YU Yantuan

(School of Economics and Trade, Hunan University, Changsha 410079, China)

Abstract: Based on the development of green, coordination, sharing ideas，the present essay analyzes the impact of transportation and information accessibility on regional ecological efficiency from both aspects of theory and empirical study. Empirical study shows that if the space factors are not taken into consideration, the traffic density is u-shaped relationship with regional ecological efficiency, the threshold value reaches 1.048 miles per square kilometer, transportation accessibility promotes the ecological efficiency mainly by improving the transmission mechanism of pure efficiency, and internet communications penetration has significant linear positive correlation with ecological efficiency. But when the space factors are taken into consideration, the interregional ecological efficiency lag factor is significantly negative, the traffic and information accessibility of neighboring provinces has a negative influence on the ecological efficiency of the province, “adjacent competition” effect is stronger than the “coordinated development” effect, and the development idea of the green, the collaborative, the shared needs to be deepened.

Key Words: ecological efficiency; transportation accessibility; information accessibility; adjacent competition; green Shared; space spillover

[编辑：谭晓萍]