科研信息化技术与应用
2014, 5(2): 3–13

· 论坛 / PERSPECTIVE ·

计算社会科学在新闻传播研究中的应用

祝建华[1]，彭泰权[2]，梁海[1]，王成军[1]，秦洁[1]，陈鹤鑫[1]

1. 香港城市大学媒体与传播系互联网挖掘实验室
2. 新加坡南洋理工大学黄金辉传播与信息学院

摘　要：本文旨在回顾和讨论新兴的计算社会科学在新闻传播研究中的应用。按照新闻传播研究中经典的5W 模型，本文分别介绍了计算社会科学在“ 谁 (传播者)，通过什么 (渠道)，对谁 (受众)，说了什么 (内容)，并产生了什么 (效果) ”等五个领域的主要应用案例，并讨论了计算社会科学和网络大数据对这些研究领域的主要贡献和现存问题。

关键词：5W 模型; 传播者; 受众; 内容; 渠道; 效果

doi: 10.11871/j.issn.1674-9480.2014.02.001

Computational Social Science in Communication Research

Jonathan J. H. Zhu[1], Peng Tai-Quan[2], Liang Hai[1], Wang Chengjun[1], Qin Jie[1], Chen Hexin[1]

1. Web Mining Lab, Department of Media and Communication, City University of Hong Kong, HKSAR
2. Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore

Abstract: We review and discuss how computational social science has been employed in communication research. Following the classic 5W model in communication research, we describe key studies of computational social science on the process of “who (communicators), says what (content), to whom (audiences), through what (channel), with what (effects)”. We also discuss major contributions and existing problems of computational social science and online big data in these research areas.

Keywords: 5W model; communicator; audience; content; channel; effects

引言

本文旨在回顾和讨论计算社会科学在新闻传播研究中的应用。什么是“计算社会科学”(Computational Social Science，CSS)？计算社会科学是最近 10 年内兴起的一种采用互联网、大数据、机器学习等计算技

基金项目：本研究由香港城市大学媒体与传播系 Faculty Research Grant (9610249) 资助完成。

术来研究社会科学问题的新思潮和新方法。计算社会科学不是社会科学家们的专利，而是一个涉及科学、技术、医学、社会、人文等各领域的跨学科“群众运动”。例如，David Lazar 等人 2009 年在《科学》上发表了一篇以“计算社会科学”为标题的短文，15 位作者分别来自于计算机、物理学、公共卫生、政治学、社会学、传播学、历史、心理学、认知科学、商学等 10 余个学科[1]。斯坦福大学新近成立的 CSS 中心 (https://css-center.stanford.edu/) 目前有 13 位成员，也分别来自计算机 (包括中国读者比较熟悉的 Jure Leskovec)、经济学、语言学、教育学、政治学、社会学等学科。与计算社会科学相关的还有“e-社会科学”、“社会计算”等概念，三者虽然在概念上各有不同，但是就研究人员、研究方法和研究成果等而言，均有很大程度的交叉重叠。

什么是“新闻传播研究”？按照 Lasswell (1948) 提出的经典的 5W 定义，新闻传播涉及“Who，says what，to whom，through what channel，with what effects”(谁、通过什么渠道、对谁、说了些什么、并产生了什么影响”)[2]。由此可把新闻传播研究分为包括了“传播者”(communicator)、“内容”(content)、“渠道” (channel)、“受众”(audience)、“效果”(effects) 五个分支。当然，在社会化媒体主导的今天，越来越多的受众成为了媒体内容的自创者，他们与传播者之间的界限日益模糊。但是，我们将在下面谈到，受众与传播者还是应该独立存在 (而不是合并)，新闻传播研究仍然包括五个 (而不是四个) 分支。当然，计算社会科学在这五个分支中的发展并不平衡。我们将根据各分支采用计算社会科学的程度而进行详略不同的论述。最后，受篇幅限制，本文主要集中在新闻内容，而不过多涉及娱乐、广告、知识等同等重要并与新闻密切相关的传播内容。

1 传播者研究

传播者研究原先专指针对消息来源 (如政府新闻发言人或企业公关人员) 和媒体专业人士 (如记者、编辑、专栏作家等) 的研究。然而，在 Web 2.0 时代，社会化媒体由广大用户自创内容。因此，传播者与受众 (或用户) 之间的界限已经日益模糊。当然，不是每一个用户都是传播者。为了便于讨论，我们可以将每个社会的成员按其参与社会化媒体的程度分为五类：一、意见领袖；二、活跃用户；三、被动跟随者；四、纯粹旁观者；五、非用户。其中前两类 (意见领袖和活跃用户) 应该划入传播者之列，而后三类属于受众 (见下节)。

用计算社会科学方法研究社会化媒体的传播者有两个技术难点：如何发现传播者 (这在传统媒体研究中根本不是问题)，以及如何描述传播者。机器学习中的两类基本方法 (无监督机器学习和有监督机器学习) 按道理均适用于研究社会化媒体的传播者，但绝大多数现有研究更青睐有监督的方法。Wu 等人 (2011)[3] 对 Twitter 的研究也许最有代表性①。他们使用了两种非随机的抽样方法 (“滚雪球”和“活跃发贴人”)，按事先设定的四组关键词 (“明星”、“媒体”、“机构”和“专业博客”)，找到了 54 万合格用户，并将每组关注度或发贴量最高的 5 000 人定义为“精英用户”(elite users，相当于我们前述的意见领袖②)。而其余的定义为“普通用户”(ordinary users，相当于我们前述的活跃用户)。这些方法显然具有相当程度的主观和人为成分，但好处是容易操作，其结果也容易解读 (用社会科学的话来说是具有一定的 face validity)。例如，他们的研究发现：以人均发贴量计，媒体最活跃 (其发贴量分别为专业博客的近 4 倍、机构的 10 倍、明星的近 40 倍)，但普通用户收到的帖子中只有 15% 是直接来自媒体，46% 由媒体经过其他普通用户朋友的转发，而 35% 来自其它精英用户 (博客、机构和明星)。再结合其它发现，作者认为 Twitter 上的新闻传播过程基本上还是经典的“媒体→意见领袖→普通受众”的“二级传播”(2-step flow of information[4])。当然，这一结论尚

① 有趣的是，该论文的标题就是模仿了 Lasswell[2] 的 5W：“Twitter 上谁对谁说了什么？”

② Wu 等人将“意见领袖”宽泛地定义为任何转发过媒体消息、而又被第三者转发过的“中介用户”(intermediaries)。据此定义，99% 的意见领袖是普通用户而不是精英用户[3]。

有待于后人用无监督机器学习的数据加以验证。

我们实验室近年来也在研究 Twitter 上的传播者。例如，我们将以前研究传统媒体时提出的“议程设置零和游戏”理论 (zero-sum theory of agenda-setting[5]) 扩展到社会化媒体环境中。议程设置指媒体通过新闻报道而塑造或强化公众对某些社会问题的关注。在传统媒体时代，各种媒体上充满了关于各种社会问题的报道，从而使得这些问题背后的“议题”(issues) 相互竞争以获取公众的注意力和政府资源；但是，各种媒体对这些议题的排序 (即 agenda 或“议程”) 却高度相似；因此我们在研究传统媒体的议程设置时，可以将所有媒体的报道视为同一议程，而着重分析各议题之间的竞争关系。到了社会化媒体时代，传播者的议程已不再同质化，而需要区别对待。我们在研究 2012 年美国大选期间的 Twitter 内容时，参照 Wu 等人的方法，将 Twitter 上的传播者分为“媒体”、“政党”、“专业博客”(其中包括了明星作者) 三类，分别计算各自对普通用户在“选举”、“经济”、“国际”等六大类议题上的影响力，并采取“川流式”(river theme) 可视化系统来展示三类传播者在不同议题和不同时间上的竞争关系[6]。

2 受众研究

我们将受众限定为只接收而不贡献内容的用户。很多人认为，在社会化媒体主导的今天，只收不发的受众不存在或者很少见了。这是由时下流行的研究方法而造成的一种错觉。这些方法从社会化媒体的帖子中寻找用户，其结果当然是大多或者全部用户都是传播者 (如发表原始贴的意见领袖或转发别人贴子的活跃分子)，而忽略了相当数量的从不发贴或转发的“围观者”。其实，社会化媒体上的“围观者”并不是一种新现象，例如在线论坛上的“潜水员”已经存在二十多年了。现在的问题是如何找到这些“围观者”，并定量计算其规模和特征。进一步说，用户的参与行为是动态的，曾经活跃的用户也许中断退出，与此同时新人又不断加入，从而形成了社会化媒体上“你方唱罢我登场”、“铁打的营盘、流水的兵”的复杂局面，更增加了准确统计传播者和受众规模及特征的难度。表 1 显示了有关问题的复杂性。

我们实验室对此在多种社会化媒体平台上做了研究，其结果相当一致：社会化媒体的多数用户还是“受众”。为了准确地估算“受众”在社会化媒体用户中的比例，需要以“用户” (而不是“贴子”) 为统计单位；否则会夸大活跃用户的比例。为此，我们参照社会调查中随机抽取电话号码 (Random Digit Dialing，RDD) 的方法，设计了一套“随机数码搜索” (Random Digit Search，RDS) 方法[7]，从新浪博客、新浪微博、Wikipedia、YouTube、Flickr、Maze (P2P 平台) 等社会化媒体的用户总体中抽取随机样本 (即 uniform sample 或“等概率样本”)，并分别统计其不活跃用户和潜水用户的比例及其他特征。我们的发现是：在各种社会化媒体平台上的用户帐号中，一半或更多是“单篇作者” (即只发过一次贴的，其中往往是启动时系统自动发布的“欢迎光临我的页面”)，另有三、四成的用户从活跃发贴/转贴逐渐转型为“潜水员”和“围观者”，剩下的一、两成用户才是真正长期坚持参与自创内容的传播者③。

其他学者例如 Fu 和 Chau[10] 也用类似的随机抽样方法，抽取了 3 万新浪微博用户，发现其中近六成 (57%) 的发帖时间 (timeline) 是空的 (从未发过贴者，即表 1 中的第3和第4组)；其余四成 (43%) 为传播者，但其中的大多数 (87% 的传播者或 37% 的所有用户) 在最近七天内没有发帖任何原创贴。

表1 社会化媒体的传播者与受众之区别

Table 1 Distinctions between communicators and audiences of social media

参与时长	创造内容	消费内容
持续不断	1. 活跃传播者	3. 活跃受众
中断放弃	2. 不活跃传播者	4. 不活跃受众

③ 这种现象也广泛发生在其它社会化媒体的使用上。如在美国著名 MOOC 平台 Coursera 上注册的学生中，只有 10% 是完成所有作业、取得证书的[8]；中国国家精品课程的常用者 (平均每周一次以上) 只占学生的 16%、教师的 8%[9]。

Benevenuto 等人[11]是少数用第一手数据直接比较用户在社会化媒体上创造与消费信息行为的研究④。他们发现如果按用户人数或访问次数来计，消费信息的行为占 92%，而创造信息的行为只占 10% 不到。虽然这项研究涉及的是社交网 (而不是新闻网) 的使用行为，我们估计新闻传播内容的创造与消费之间的比例，概也是在 1：9 到 2：8 之间。当然，这种猜测需要采用类似 Benevenuto 等人的方法，通过对 Twitter、微博等网站的第一手访问日志数据进行分析而加以验证。如果可以做到，那将是对新闻传播受众研究的一个重大贡献。

3 内容研究

在新闻传播研究中，“内容为王 (Content is king)”是最令人耳熟能详的口号之一。在以报纸、杂志等传统媒体为主的时代，优质的新闻内容是媒体的安身立命之本。在互联网时代，新闻传播学界或业界都存在争论，是否仍然要奉“内容为王”为圭臬？信息传播的渠道是否会取代内容，成为媒体的制胜法宝？现在，越来越多的互联网从业者发现，优质的内容是稀缺资源，坚持“内容为王”仍然是成功的互联网媒体的核心竞争力之一。然而，社会化媒体的新闻内容生产确实与传统媒体时代有不同之处。前者的生产主力不再是记者、编辑等专业人士，而是用户。诸如微博、搜索词等用户生产内容 (user-generated content) 极大地改变了传媒生态。我们在下面介绍两个研究成果最为丰硕的领域，即社会化媒介 (social media) 的内容与搜索词 (web search query)。

社会化媒介已经成为信息传播的重要平台。社会化媒介上发布的信息和传统媒体 (比如报纸，广播或电视) 上发布的信息之间有什么异同，是内容研究的一个重要问题。Zhao 等人[12]运用主题建模 (topic modeling) 的方法，对 2009 年 11 月~2010 年 1 月间 Twitter 上的内容和同时期美国《纽约时报》的报道进行比较⑤。他们分别按照两种分类法对 Twitter 和《纽约时报》所涉及到的话题 (topic) 进行了分类。按照话题的主题 (subject area)，Twitter 和《纽约时报》的内容被归属到 10 个类别。首先，他们发现 Twitter 和《纽约时报》的报道重心有所不同。Twitter 上，家庭与生活、艺术以及时尚是发布最多的信息类别，而在《纽约时报》中，艺术、国际新闻、商业新闻则位居前三。此外，他们还发现 Twitter 和《纽约时报》上同一类报道的覆盖面 (breadth) 有差别。比如在《纽约时报》，艺术类新闻会涉及书籍、小说、博物馆、历史等，覆盖面较广；而在 Twitter 上，艺术类别的信息则集中在流行音乐和明星上 (例如 Lady Gaga)，覆盖面较窄。他们又按照话题的本质 (nature) 将 Twitter 和《纽约时报》的内容划归为 3 个类别：事件主导的话题 (event-oriented topics)、人物或组织主导的话题 (entity-oriented topics)，以及持续性话题 (long-standing topics)。他们发现，在 Twitter 上，人物或组织主导的话题要远远多于《纽约时报》，这些话题多半是关于明星或大公司的。而在《纽约时报》，事件主导的话题要远远多于 Twitter。同时，Twitter 和《纽约时报》在报道事件主导的话题上有较高的重合度，而在人物或组织主导的话题上重合度较低。Kwak 等人[13]曾基于 Twitter 用户很少“互粉”的事实而认为 Twitter 是新闻媒体，不是社交媒体；但 Zhao 等人的研究则发现 Twitter 和传统的新闻媒体之间在内容和话题上还是存在一定的差异。

我们实验室的研究则进一步深入分析了 Twitter 和传统新闻媒体对同一事件的不同呈现框架 (frame)。Qin[14] 借鉴了 Hemphill 等人[15] 用机器学习识别话题

④ 他们从一个以拉美地区用户为主的社会化媒体整合网站获得了一批十分罕见的数据。该整合网站与四个社交网站 (包括 Orkut、MySpace、LinkedIn 和 Hi5) 有合作关系，用户登入该网站后，就同时接通了上述四个社交网站的帐号。他们得到的数据中，包含了 3 万 7 千多个帐号 (其中 98% 来自 Orkut) 在 12 天内的所有 460 多万条行为记录，因此可以具体解析出每个用户在这些社交网上所留下的每一步踪迹 (即何时访问了哪些网页，并停留了多少时间)。他们将所有被访问的网页，按其内容分成 41 类。它们分别属于搜索、留言、私信、浏览朋友网页、观看朋友的照片或视频、社区活动等 9 大类。

⑤ 他们的数据中总共包括了 4 916 位 Twitter 用户发布的 122 万条信息以及纽约时报的 11 924 篇新闻报道。

的方法，并结合已有的语义挖掘工具 (如 hashtagify 和 sensebot)，绘制出了“棱镜门”事件在 Twitter 和传统媒体报道中的语义网络 (semantic network)。她发现 Twitter 用户会将“棱镜门”主角斯诺登 (Edward Snowden) 与之前的泄密者、个人隐私、反税运动联系起来，将斯诺登塑造为一个英雄；而传统媒体则套用国土安全及反恐框架，将斯诺登塑造为一个叛徒。这一发现在某种程度上也呼应了 Zhao 等人[12]的结论，即 Twitter 和大众媒体的话题不一定重合，其呈现方式及舆论后果甚至可能南辕北辙。究其原因，Qin 提出了社会化媒介内容生产方式与传统媒体的三点不同，即话题范围 (scope)、人为操纵 (manipulation)、语义组合 (association)。

除了上述常见的社会化媒体之外，我们认为用户在搜索引擎上留下的数字化痕迹 (digital traces)，即“公众注意力”，也是一个重要的社会化媒体内容。“公众注意力”指公众对于某些社会议题进行思考的过程中所投入的时间和认知资源[16]。过去传播学者主要是依靠民意调查的结果 (例如美国盖洛普公司 Most Important Problems 调查系列) 来测量公众注意力，并跟踪其变化。近年来，研究者开始利用网民在搜索引擎提交的搜索词来测量公众注意力。搜索引擎可以说是最传统的社会化媒介之一，因为用户可以通过在搜索引擎自我设定并提交关键词来获取他们感兴趣的信息。用户自我提交 (self-initiated) 的搜索词，可以被归属于传播研究的不同领域。搜索词可以代表用户的使用行为 (behavior)，也可以反映用户使用行为的效果 (outcome)。但是在更多的时候，搜索词被看作是用户贡献的内容 (content)。这方面最具影响力的研究当属 Ginsberg 等人[17] 发表在 Nature 上的文章。他们利用 45 个与流感有关的关键词，来测量公众对流感的关注程度。基于这些关键词搜索趋势的变化，他们准确地预测了美国流感的爆发。自此之后，搜索词被广泛运用于测量现实世界中公众对疾病、商业产品以及社会议题的注意力，比如登革热[18]、股票[19]、就业[20]等。

虽然越来越多的实证研究开始应用搜索词来测量公众注意力，但是有一个简单但非常重要的问题还没有得到系统地回答：作为公众注意力的测量工具，搜索词是否具有其测量效度 (measurement validity)？我们实验室的两项研究都涉及到这个问题[21–22]。Zhu 等人[19] 通过比较住房、交通、治安等话题在搜索引擎与深圳幸福指数调查中的走势，发现搜索词这种新的工具具有一定效度，但是还有诸多因素会影响到搜索词作为公众注意力的测量工具的效度，比如搜索词的选择、议题本身的特点，以及互联网的扩散程度等等[23]。我们的另一项研究则比较了环保及能源议题在谷歌趋势搜索 (Google Trends) 与盖洛普民意调查中的走势，也得出了与 Zhu 等人 (2012) 一致的结论[22]。

4 渠道研究

在新闻传播研究中，渠道是指新闻信息的传播路径。常见的渠道包括媒介系统 (如电视、广播、报纸)、社会网络 (如社区、参考群体)、组织 (如政府、公司) 等。渠道研究系统地连接了新闻传播学研究的各分支 (如传播者、信息内容、传播效果)，并衍生出众多的研究传统 (如新闻扩散、创新的扩散) 和研究视角 (如传播网络分析)。我们这里主要探讨有半个多世纪历史的新闻扩散这一研究传统。渠道研究的一个核心问题是各种传播渠道的优劣比较。传统的新闻扩散研究主要比较媒体和人际网络这两种传播渠道对于信息扩散的影响[24]。早期研究发现，对于一些重大和琐碎的新闻，人际传播渠道是主要消息来源 (如肯尼迪遇刺之后，超过 50% 的人是从人际网络中获悉该新闻)；而对于一些中等程度的新闻事件而言，媒体则成了主要渠道[25–26]。但是，传统的新闻扩散研究的研究方法和理论框架常受批评[27]。例如，重大新闻事件的突发性迫使研究者必须匆忙地开展调查，往往忽略了研究设计；采用调查和访问等方法收集的被访者自我报告的数据 (self-report data) 则会因为事件发生时间和访问时间之间的间隔而产生遗忘问题；大多数新闻扩散研究是个案分析，并且个案数量和抽样规模都很有限[28]。正是由于这些限制，人际传播和新闻媒体对于新闻扩散哪个更重要的问题并没有得到很好的解答。

社会化媒体时代大量丰富易得的数字化痕迹则极大地便利了渠道研究。目前成果主要集中于两个

领域：一个是在数字媒体中信息扩散模式 (diffusion pattern)，另一个是新兴渠道与传统渠道的比较。在扩散模式领域，Kwak 等人[13] 和 Wang 等人[29] 的研究发现，信息 (也包括新闻) 在社会化媒体的扩散是广度优先而非深度优先，换句话说，依赖单一信息源无法有效地在社交媒体上传播新闻信息。在跨渠道比较领域，Petrovic 等人[30] 比较了 Twitter 以及美联社、路透社等传统通讯社 70 多天中对各种新闻事件的贴子和报道。他们采用了人工阅读和机器学习两种方法，从这些文章中寻找“新闻事件”。结果发现：平均而言，Twitter 与传统通讯社在对同一批重大事件的报道时效性上，并没有明显区别；相反，Twitter 的优势在于报道了一大批被传统媒体忽略了的“微事件”。这一研究有助于澄清对 Twitter 等社会化媒体的过高期望。Kim 等人[31] 研究了 284 条新闻在整个互联网的扩散，结果表明新闻扩散渠道与新闻信息类型也有关系。例如，新闻网站可以有效传播艺术和经济新闻，社交媒体和博客可以有效扩散政治和文化新闻。另外，争议性新闻可以跨越多个渠道传播，而娱乐型新闻则主要集中于一个传播渠道。

在众多社会化媒体中，例如 Digg、Reddit 等社会新闻网站 (social news website) 是很好的研究对象⑥。我们实验室对 Digg 的研究发现，通过协同过滤 (collaborative filtering) 的方式产生了“集体把关”(collective gatekeeping) 的现象[32]。在新闻扩散过程中，集体把关主导了超过 59% 的信息渠道，人际网络渠道占 23%，而最新新闻页面渠道占 18%。我们的发现拓展了新闻传播研究中的“把关人”理论在公民新闻 (civic journalism) 研究中的应用。

5 效果研究

效果研究是指新闻传播对受众认知、态度、行为方面的影响。所谓认知，简而言之，是指个体对事物的认识 (包括 knowledge、perception 等)。我们有一项研究通过比较网络论坛参与者语义网络的相似性来测量参与者之间所达成共识 (common ground) 的程度[33]。用户在论坛中的回帖包含了大量的文本数据，可以从每个用户的文本中提取出一个语义网络用来推断用户大脑中知识的结构和阐述问题的框架。比对任意两个参与者的语义网络便可推断出二者在知识结构和阐述框架上的异同。在一个讨论网络中，语义网络的相似性在整体上体现出了人们所具有的共识。即使人们在对某一问题上的态度是不同的，他们仍然可以在更基本的问题上达成一致——讨论同一个问题，而不是自说自话。

也有学者通过自然语言处理的方式来研究个体的情绪或态度[34]，甚至是政治意识形态[35] 等。态度或情绪在新闻传播中是非常重要的因素，计算社会科学可以帮助我们理解情绪和意识形态是如何在网络讨论中形成的，对集体行动又有怎样的影响。而行为则是网络上是最容易被观察到和测量的变量。例如 Himelboim[36] 分析了 35 个新闻组中 20 多万个参与者 6 年时间里的讨论行为。他发现在这个讨论网络中，入度 (in-degree) 遵循幂律分布。这种不平等的分布代表着网络讨论中人们所受到的注意的不平等。他进一步发现这种不平等随着网络参与者的增加而增加。这一现象被解读为一个民主的悖论：政策制定者希望更多的人参与网上的讨论来促进公民社会的发展，但是这种大规模的讨论会导致注意力集中在少数几个人身上，从而不利于民主的发展。

将计算社会科学引入效果研究领域，有两个明显的方法优势。首先，传统的效果研究多使用控制实验法，因为这是判断因果关系最有效的办法。而互联网正是一个自然实验 (natural experiment) 的平台。传统实验只能测试几个、几十个对象，然而互联网上的控制实验则可以触及规模巨大的群体。这一根本性的变革也受到了计算社会科学研究者的重视。其中最为著

⑥ 好处在于：在社会新闻网站 (如 Digg、Reddit) 上，用户可以自由地向网站提交新闻信息，并对提交的信息的重要性投票。如果一条新闻可以在一段时间获得较高的票数，就会被社会新闻网站推荐为流行新闻 (这一点跟微博相似)。此外，用户可以彼此添加好友关系。因此，用户在社会新闻网站可以从至少三个渠道阅读新闻：最新新闻页面、朋友页面、流行新闻页面。

名的是 Salganik 等人[37] 对 14 341 位在线用户下载音乐的研究，以及 Bond 等人[38] 针对 6 100 万 Facebook 用户的实验。其次，计算社会科学研究者往往采用一种无干涉 (unobtrusive) 的方式来观察和计算真实环境中的传播和效果。这一特征是传统社会定量方法，比如问卷调查和实验室实验无法做到的。

但是将计算社会科学应用于效果研究，还存在两个技术难点。第一、如何自动化地判断受众的认知、态度和行为特征。第二、如何判断变量间的因果关系。从数据属性上来说，数字化痕迹在判断因果关系上有其独特的优势⑦。但是在很多计算社会科学的文章中，这种优势并没有得到足够的利用⑧。Liang[33] 提出的通过比较语义网络的方法来判断共识度，并利用了在线讨论中这种面板特性 (panel data)，用网络分析的方法区分出了社会选择和社会影响在政治讨论中的作用，并证明了社会选择的作用要大于社会影响。该研究对于上述两个技术难点都有所贡献。

6 结论与探讨

我们在前文中，分别讨论了新闻传播过程中的传播者、受众、内容、渠道和效果等 5W 要素的计算社会科学研究。总体说来，计算社会科学给新闻传播研究带来了革命性的变化和进步，短短几年之内就已经在 5W 的各方面均涌现了一系列有创意的研究案例，其发现或者验证了悬而未决的猜测、或者挑战或颠覆了长年流行的理念。但是，这五个领域的研究还各自有其局限性。

在传播者研究方面，面向 Twitter 新闻的研究最多、而面向其它平台的研究甚少，因此限制了对社会化媒体上新闻传播多样性的全面了解。而且，这五个领域发展不平衡——传播者研究比较丰富、而受众研究最显不足，从而造成了对自媒体普及程度的过高估计。日后研究中有必要修正这些偏向。

在受众研究方面，非活跃用户及隐性传播行为是盲点之一。虽然我们已经有了切实可行的用户抽样方法，但也仅仅局限于“活跃用户”及“显性行为”。那么，如何挖掘论坛上只看不发言的“潜水员”？如何测量那些包括浏览网页在内的“隐性行为”？学术界对这两个方面的研究显然是不足的。

在内容研究方面，现有的分析还比较粗糙。例如，在以搜索词来测量公众注意力的研究中，一个重要的问题就是选择与相关议题 (issue) 相对应的关键词 (keyword)。例如研究公众对流感议题的关注，哪些关键词代表了公众对流感的关注，哪些与其无关，这是相关研究中的重要一步。目前，大部分研究都依靠研究者本身的主观判断来选择检测。这种主观选择带来的后果就是，针对同一议题会有不同的关键词来测量公众的注意力，造成相关研究之间无法对话，而且结论可能大相径庭。所以，更细致、更基础的工作还是非常必要的。

在渠道研究方面，如何收集和分析数据成为第一道门槛。例如，社交网络中的个体进行抽样至今未能得到较好的解决。其次，逐渐交叉的研究问题使得跨学科的合作开始越来越多。从我们的介绍中可以看到，大量来自计算机、物理等学科的科学家已经积极地投入到社会化媒体的研究中来。如何更有效地展开跨学科的研究和合作成为不可忽略的问题。

在效果研究方面，现有的分析技术还过于局限于整体层面 (aggregate level)，而对于更为微观的个体行为还缺少细致的研究。其次，现有方法过分依赖于数据本身而缺乏通过研究设计来解决重要的问题。许多研究只是“有什么数据做什么研究”而不是“要解决什么问题而收集什么数据”。很多时候研究者

⑦ 优势在于：首先，通过无干涉地追踪人们在互联网上的行为，很容易得到一个面板 (固定样本) 数据 (panel data)，这是传统的一次性问卷访问的方法不容易做到的。面板数据因其具有时间上先后变化的特征而常用于判断因果关系的方向。

⑧ 比如 Wu 等人[39] 的文章就是通过分析整体层面 (aggregate level) 上态度的变化来得出舆论在走向中庸的结论的。这并没有体现出每一个个体 (比如一个帖子或者个人) 是否也同样满足这样的因果关系。在对网络讨论中同质性 (homophily) 的研究中同样存在这样的问题，比如 Yardi 和 boyd[40]、Conover 等人[41] 对 Twitter 上群体激化的研究在同质性的假设下讨论了不同意识形态和观点间人们的讨论。但是却并没有涉及到群体激化的现象是来自于人们更愿意与相似的人讨论还是讨论导致他们更为相似了。

过多地强调数据变量间的相关性，而忽略了理论的重要性。

综上所述，我们认为将计算社会科学引入新闻传播学的优势有四点。第一，数据的价值优势。在社会化媒体时代，诸如阅读、评论、网购等数字化痕迹使得个体行为都变得有迹可循。这类数据的价值在于：首先，它是对个体行为第一手的、客观、细致记录；其次，它往往包含时间变量，为研究行为演化提供了可能。第二，对大数据的处理优势。随着越来越多的报纸杂志开始由纸质印刷转为电子出版，新闻传播学者所要处理的数据必然要从千字节 (kilobytes，例如几十篇新闻报道) 向千兆字节 (gigabytes，例如上万本书籍的规模) 进化。计算社会科学对大数据的处理能力是传统的劳动密集型分析方法所不能企及的。第三，自下而上 (bottom-up) 的归纳优势。由于处理能力的局限，传统的分析大多采取自上而下的演绎推理 (deductive reasoning)。而计算社会科学则是从观察 (observation) 出发，总结模式 (pattern)，继而验证假设并提出理论，其逻辑则是基于归纳推理 (inductive reasoning)。归纳推理更加适用于新事物不断产生的社会化媒体环境。第四，非介入性 (unobtrusive) 的方法优势。关于介入性与非介入性方法的争论在社会学界已经存在许久。学界对介入性方法 (例如采访、问卷等) 的疑虑在于：自我报告 (self-report) 的数据是否可靠？计算社会科学中涌现的数字化痕迹为新闻传播研究提供了自我报告数据之外的另一种选择。当然，它是否真的比自我报告数据更为可靠，还有待进一步检验。

当然，挑战与机遇并存。第一个问题，跨学科对话不足。本文中的研究案例大多数由计算机学者完成，但也有新闻传播学者独立进行[8]，或者由双方合作而成[6–7]。这种局面与整个社会化媒体研究中的学科分布基本一致。例如，根据我们的检索，Web of Science (科学网) 的 SCI 和 SSCI 期刊至 2012 年底共发表了近 1 000 篇关于 Facebook 和 Twitter 的研究，其中近六成 (58%) 由科学技术学科完成，三分之一 (34%) 由社会科学完成，而只有8%由科技和社科合作进行。当然，这种局面并不理想。我们预期，随着社会化媒体的日益普及和计算社会科学的逐步完善，将涌现出越来越多的跨学科合作的研究团队和成果。第二，“名实相怨”的问题。“名实相怨”这一概念出现在春秋战国时期，是百家争鸣的一个核心问题。这一概念实际上包含了三个层次的问题：(1) 新现象不断涌现，我们不知道如何对其进行初步描述；(2) 已经有了初步描述，但是还缺乏进一步的理论化；(3) 对现象的描述与理论化都已经存在，但二者之间存在错位。因此，学界多展开一些关于概念、定义的探讨，也是非常必要的。

参考文献

[1] Lazar, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D. D., Christakis, N. A., Contractor, N., Fowler, J. H., Gutman, M. P., Jebara, T., King, G., Macy, M., & Van Alstyne, M. (2009). Computational social science [J]. *Science*, 323(5915), 721–723.

[2] Lasswell, H. D. (1948). The structure and function of communication in society [M]. In L. Bryson (Ed.), *The communication of ideas* (117-130). Urbana, IL: University of Illinois Press.

[3] Wu, S. M., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on Twitter [A]. *WWW2011*.

[4] Lazarsfeld, P. F.,Berelson, B., & Gaudet, H. (1968). *The people’s choice: How the voter makes up his mind in apresidential campaign* (*3rd edition*) [M]. New York: Columbia University Press.

[5] Zhu, J. J. H. (1992). Issue competition and attention distraction: A zero-sum theory of agenda-setting [J]. *Journalism & Mass Communication Quarterly*, 69(4), 825–836.

[6] Xu, P. P., Wu, Y. C., Wei, E. X., Peng, T. Q., Liu, S. X., Zhu, J. J. H., & Qu, H. M. (2013). Visual analysis of topic competition on social media [J]. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2012–2021.

[7] Zhu, J. J. H., Mo, Q., Wang, F., & Lu, H. (2011). A

random digit search (RDS) method for sampling of blogs and other web content [J]. *Social Science Computer Review*, 29(3), 327–339.

[8] 吴文峻. 美国 MOOC 考察见闻[J]. 中国计算机学会通讯, 2013, 9(10): 46–50.

[9] 孙茂松. 从技术和研究角度看 MOOC[J]. 中国计算机学会通讯, 2013, 9(10): 51–53.

[10] Fu, K. W., & Chau, M. (2013). Reality check for the Chinese microblog space: A random sampling approach [J]. *PLOS ONE*, 8(3), e58356.

[11] Benevenutoy, F.,Rodriguesy, T., Cha, M. Y., &Almeiday, V. (2009).Characterizing user behavior in online social networks [A]. Paper presented at the *2009 Internet Measurement Conference (IMC'09)*, Chicago, USA.

[12] Zhao, X., Jiang, J., Weng, J. S., He, J., Lim, E. P., Yan, H. F., & Li, X. M. (2011). Comparing Twitter and Traditional Media using Topic Models [A]. *Research Collection School of Information Systems (Open Access)*. Retrieved from http://ink.library.smu.edu.sg/sis_research/1375

[13] Kwak, H., Lee, C., Park, H., & Moon, S. (2010). *What is Twitter, a social network or a news media*? [A].Paper presented at the Proceedings of the 19th international conference on World Wide Web, Raleigh, North Carolina, USA.

[14] Qin, J. (2013). Snowden Wins on Twitter but Fails in News: The Mismatch between Social Media Frame and Mass Media Frame [A].In *The 2013 Asian Symposium of Doctoral Students in Communication*, Hong Kong, November 17–19.

[15] Hemphill, L., Culotta, A., & Heston, M. (2013). Framing in Social Media: How the US Congress Uses Twitter Hashtags to Frame Political Issues [A]. *Available at SSRN 2317335*.

[16] Newig, J. (2004). Public Attention, Political Action: the Example of Environmental Regulation [J]. *Rationality and Society*, 16 (2), 149–190.

[17] Ginsberg, J, Mohebbi, M., & Patel, R. (2009). Detecting influenza epidemics using search engine query data [J]. *Nature*, 457, 1012–1014.

[18] Althouse, B. M., Ng, Y. Y., & Cummings, D. A. T. (2011). Prediction of Dengue Incidence Using Search Query Surveillance [J]. *PLoS Neglected Tropical Diseases*, 5(8), 1–7.

[19] Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends [J]. *Scientific reports*, 3, 1684.

[20] Fondeur, Y., & Karamé, F. (2013). Can Google Data Help Predict French Youth Unemployment? [J]. *Economic Modelling*, 30(0), 117–125.

[21] Zhu, J. J. H., Wang, X., Qin, J., & Wu, L. (2012). Assessing Public Opinion Trends based on User Search Queries: Validity, Reliability, and Practicality [A]. In *Annual conference of the World Association for Public Opinion Research*, Hong Kong, June 14–16.

[22] Qin, J., & Peng, T. Q. (2014). Measuring Public Attention on Environment and Energy Issues with Google Trends: A Validity Assessment [A]. In *Annual conference of the International Communication Association (ICA)*, Seattle, Washington, USA, May 22–26.

[23] Mellon, J. (2011). Search Indices and Issue Salience: the Properties of Google Trends as a Measure of Issue Salience [A]. *Sociology Working Papers*, 1.

[24] Larsen, O. N., & Hill, R. J. (1954). Mass media and interpersonal communication in the diffusion of a news event [J]. *American Sociological Review*, 19(4), 426–433.

[25] Miller, D. C. (1945). A research note on mass communication: How our community heard about the death of president Roosevelt [J]. *American Sociological Review*, 10(5), 691–694.

[26] Greenberg, B. S. (1964). Person to person communication in the diffusion of a news event [J]. *Journalism Quarterly*, 41(3), 489–494.

[27] De Fleur, M. L. (1987). The growth and decline of research on the diffusion of the news, 1945-1985 [J]. *Communication Research*, 14(1), 109–130.

[28] Funkhouser, G. R., & McCombs, M. E. (1971). Rise and

fall of news diffusion [J]. *Public Opinion Quarterly*, 35(1), 107–113.

[29] Wang, D., Wen, Z., Tong, H., Lin, C.-Y., Song, C., & Barabási, A.-L. (2011). *Information spreading in context* [A]. Paper presented at the Proceedings of the 20th international conference on World Wide Web.

[30] Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., & Shrimpton, L. (2013). *Can Twitter replace newswire for breaking news*? [A]. Paper presented at the Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM2013), Boston, MA, USA.

[31] Kim, M., Newth, D., & Christen, P. (2013). Modeling dynamics of diffusion across heterogeneous social networks: News diffusion in social media [J]. *Entropy*, 15(10), 4215–4242.

[32] Wang, C. J. (2012). Jumping over the network threshold: How widespread could news diffuse on news sharing websites? [A]. *In the 62nd Annual Conference of International Communication Association*, Phonix, Arizona, May 24–28.

[33] Liang, H. (2014). Coevolution of political discussion and common ground in web discussion forum [J]. *Social Science Computer Review*, 1–15.

[34] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment [A]. ICWSM, 10, 178–185.

[35] Mullen, T., & Malouf, R. (2006).A preliminary investigation into sentiment analysis of informal political discourse [A]. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (pp. 159–162).

[36] Himelboim, I. (2011). Civil society and online political discourse: The network structure of unrestricted discussions [J]. *Communication Research*, 38(5), 634–659.

[37] Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market [J]. *Science*, 311(5762), 854–856.

[38] Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization [J]. *Nature*, 489(7415), 295–298.

[39] Wu, F., & Huberman, B. A. (2010). Opinion formation under costly expression [J]. *ACM Transactions on Intelligent Systems and Technology (TIST)*,1(1), 1–13.

[40] Yardi, S., & boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on Twitter [J]. *Bulletin of Science, Technology and Society*, 30(5), 316–327.

[41] Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Flammini, A., & Menczer, F. (2011). *Political polarization on Twitter* [A]. Paper presented at the 5th International Conference on Weblogs and Social Media.

收稿日期：2013 年 12 月 3 日

祝建华：香港城市大学媒体与传播系，教授，互联网挖掘实验室创办人，美国印第安纳大学博士，主要研究方向为社会化媒体的结构、内容、使用与影响。
E-mail: *j.zhu@cityu.edu.hk*

彭泰权：新加坡南洋理工大学黄金辉传播与信息学院，助理教授，香港城市大学博士，主要研究方向为社会化媒体的结构、内容、使用与影响。
E-mail: *winsonpeng@gmail.com*

梁　海：香港城市大学媒体与传播系，博士研究生，互联网挖掘实验室成员，主要研究方向为数据挖掘与政治传播、互联网活动的时空结构。
E-mail: *hai.liang@my.cityu.edu.hk*

王成军：香港城市大学媒体与传播系，博士研究生，互联网挖掘实验室成员，主要研究方向为社会化媒体、新闻扩散。
E-mail: *wangchj04@gmail.com*

秦　洁：香港城市大学媒体与传播系，博士研究生，互联网挖掘实验室成员，主要研究方向为超链接网络分析、公益组织新媒体使用。
E-mail: *jieqin4-c@my.cityu.edu.hk*

陈鹤鑫：香港城市大学媒体与传播系，博士研究生，互联网挖掘实验室成员，主要研究方向为新媒体使用与影响、互联网用户的时间特质。
E-mail: *hexin.chen@my.cityu.edu.hk*