计算传播学：国际研究现状与国内教育展望

沈浩[1]，罗晨[2]

（1. 中国传媒大学新闻学院、协同创新中心、媒体融合与传播国家重点实验室，北京 100024；
2. 清华大学新闻与传播学院，北京 100084；加州大学戴维斯分校传播系，美国加利福尼亚州 95616）

摘要：计算传播研究是传播学领域内富有潜力且发展迅速的一脉新兴研究取向。通过对传播学领域内影响因子位居前 20 的 SSCI 期刊中的 252 篇计算传播研究论文全文进行关联主题模型分析，研究者识别出包含社交媒体分析、多元社会议题分析、新闻与新闻工作者研究、社会运动与社会参与研究、政治竞选研究、用户媒介接触研究六大领域在内的研究图景。从高水平研究论文中折射出的计算传播研究现状呼吁理论与方法紧密融合的专业教育实践。其中，理论旨在强调培育跨学科的、混合逻辑的、呼应语境的研究思维；方法旨在强调提升数据获取、统计分析、视觉化呈现等计算传播研究必备能力。

关键词：计算传播；关联主题模型；研究领域；自动化文本分析；网络分析；专业教育

The State-of-the-Art Research and Domestic Education on Computational Communication Science

SHEN Hao[1], LUO Chen[2]

(1. Communication University of China, 100024, Beijing, China;
2. Tsinghua University, 100084, Beijing, China; the University of California, Davis, 95616, California, United States of America)

Abstract: As a relatively new research approach in communication studies, computational communication has considerable potential and experiences a rapid development in recent years. This study analyzes 252 computational communication research articles from the top 20 journals under the communication category (from 2015 to 2019). By adopting the correlated topic modeling, six primary research areas emerged, including social media analysis, investigation of diverse social issues, journalistic studies, social movements and social participation, political campaign, and users' media exposure. This quantitative review also sheds light on the domestic education of computational communication science. In the theoretical aspect, we emphasize multidisciplinary thinking features mixed logic and context sensitiveness. In the methodological part, we advocate the cultivation of necessary research abilities incorporating data acquisition, advanced statistical analysis, and data visualization.

Key words: computational communication science; correlated topic model; research area; automated text analysis; network analysis; professional education

---

基金项目：国家留学基金委 201906210115

作者简介：沈浩（1963-），男（汉族），上海人，中国传媒大学新闻学院教授、协同创新中心博士生导师、媒体融合与传播国家重点实验室高级研究员，shenhao@cuc.edu.cn。

罗晨（1994-），男（汉族），江西吉安人，清华大学与美国加州大学戴维斯分校联合培养博士，luoc18@mails.tsinghua.edu.cn。

1 引言

大数据的兴起和计算能力的提升共同推促了传播学领域的“计算转向”（computational turn）[1-2]，计算传播学（computational communication science）亦受该转向驱动，成为发展迅速且备受关注的一支新兴研究取向。van Atteveldt 等[3]曾归纳了计算传播研究的显著特征：第一，使用大体量甚至是复杂数据集；第二，研究数据通常源自“数字足迹”（digital trace）或其他“自然发生”（naturally occurring）的数据；第三，需要依赖算法来开展分析；第四，能够结合传播理论来探索人类传播行为。前两项特征关注数据构成，第三项特征强调分析路径，第四项特征则聚焦于计算传播导向下方法与理论的对话。无独有偶，Hilbert 等[4]在综述文章中同样论及上述要点，包括大规模“数字足迹”如何关照传播学经典理论、计算机模拟方法如何建立个体行为与集体行动机制之间的桥梁、前沿方法（如：在线田野实验）如何对传统研究方法予以补足、不断更新的机器学习算法如何开辟文本分析的更多可能。

以传统传播学研究为参照，计算传播常被冠以“新兴”前缀。这不仅源于延循该取向的研究论文在近年来密集出现，更在于计算传播内在的“常新”特质。“常新”受益于可用数据的扩展及跨学科视角的深化。以计算传播研究的数据来源为例，社交媒体平台上由用户主动发布的内容已屡见不鲜，不少研究者开始从大型游戏的服务器日志[5-6]、网络流量统计[7-8]等新型“富矿”中抽取数据。跨学科视角则体现在传播学与其他学科的“联姻”，例如率先诞生于政治科学的结构主题模型（structural topic model）[9]被挪用至形象建构研究中[10]；计算机视觉相关知识被运用到党派媒体的报道偏见研究中[11]。计算传播研究的内生驱动力带来丰富潜能，激励传播学研究者围绕更多样的问题采用更科学的方法来对社会现实展开探索，进而对理论作出延展。

与计算传播研究不断向前迈进相同步的是计算传播学建制化水平的持续攀升。在课程设置上，国内外众多高校已将计算传播相关课程纳入新闻传播学科培养体系；在专业社群建设上，一系列计算传播研究团体相继诞生，最典型的莫过于国际传播学会（International Communication Association）中的计算方法兴趣小组（interest group）已晋升为分会（division）；在研究标准确立上，诸如《计算传播研究》（*Computational Communication Research*）等专刊和一系列特刊的发行强化了计算传播领域的专业化水准[12]。纵使如此，不少传播学研究者依然面临着明显的“技能迁移”问题[4]，根植于社会科学和人文学科的传播学相较于 STEM（科学、技术、工程、数学）领域而言，在计算传播研究所需技能的培养上并无优势，这无疑给计算传播研究的长远发展提出挑战。

有鉴于此，本研究拟从两个方面对计算传播学展开讨论：其一，计算传播研究具备不同于传统研究的特质，且时刻处于更新状态。那么，近年来的高质量计算传播研究论文主要在关注什么？换言之，计算传播研究者的关注视野是否存在某些共通之处？他们青睐的数据来源和分析方法有哪些？其二，顺承第一项研究问题，对前沿研究的梳理能够给予国内计算传播人才培养以哪些启迪？我们应该追求怎样的计算传播专业培养路径？

2 研究设计

2.1 数据

参照前人经验[13-14]，我们认为高质量期刊刊载的计算传播研究论文具备足够代表性，可以视为国际计算传播发展的前沿阵地与“风向标”。我们同样从科学网（Web of Science）下的期刊引证报告（Journal Citation Reports）入手，选择传播学分类下 5 年影响因子位列前 20 位的 SSCI 刊物。5 年影响因子相较于单年度影响因子而言更能折射出期刊的长期表现。最新的期刊引证报告统计截止时间为 2019 年，当年度传播学下的 SSCI 期刊共计 92 本，将 20 本作为入选数量接近于选择 SSCI 期刊的一整个分区（92/4=23），且入选的 20 本期刊 5 年影响因子皆在 3.0 以上，具有显著的代表意义。初步浏览后，进一步排除侧重理论研究的刊物 *Communication Theory* 与无访问权限的刊物 *Journal of Advertising Research*，最终，18 本期刊[1]被纳入本研究的分析范围。

研究者招募了 12 名传播学专业的硕士研究生[2]来从入选期刊中挑选符合要求的计算传播研究论文。由于文章总数较多，我们将期刊的时间检索范围统一设定为 2015 年 1 月 1 日至 2019 年 12 月 31 日（对应前述的 5 年影响因子），并秉持“查准”原则将检索词设定为“computation”。计算传播研究论文的判断标准为：1. 必须为量化研究论文，排除诸如围绕“计算传播”展开的质性研究；2. 研究数据来源必须是“数字足迹”（如：推文、微博）或其他数字化档案（如：经数字化处理后的新闻数据库）[3]；3. 分析方法上，必须运用自动化分析方法（如：自动化文本分类）。我们并未限定论文必须使用 Python 或 R 等编程语言，原因在于部分计算传播研究将大体量数据进行清洗或降维后，常用的非开源统计软件也可胜任之后的数据分析任务。此外，我们要求挑选者在面临判别模糊状况时及时向两位作者汇报，由两位作者来执行最终判断。判别环节结束后，共计 252 篇计算传播研究论文（包含标题、摘要、正文部分）成为本研究的分析语料。

2.2 分析方法

从核心研究问题出发，研究者选择在全文本基础上构建关联主题模型（correlated topic model，后文简称 CTM）来发掘入选论文的共通关注领域。Blei 和 Lafferty[15]指出，基于潜狄利克雷分布（latent Dirichlet allocation）的主题模型（后文简称 LDA）忽略了主题间的隐含联系，例如一篇关于遗传的文章可能在同时论述疾病，而不太可能包含天文知识。LDA 之后的变分推断方法无法胜任探索主题关联的任务（即发现“遗传”与“疾病”的高度相关及“遗传”与“天文”的弱相关），相较于主题关联假设，LDA 更倾向认为主题之间是相互独立的。CTM 依托于逻辑斯正态分布（logistic normal distribution），其原理如图 1 所示，历经一系列生成推断环节后，CTM 将习得主题之间的协方差结构。

---

[1] 18 本期刊的刊名和 5 年影响因子为：*Journal of Communication* (7.18), *Journal of Computer-Mediated Communication* (6.27), *Journal of Advertising* (5.69), *Political Communication* (5.07), *New Media & Society* (4.97), *Digital Journalism* (4.96), *Information, Communication & Society* (4.79), *Communication Research* (4.50), *International Journal of Advertising* (4.13), *Human Communication Research* (4.04), *Comunicar* (3.83), *Journal of Public Relations Research* (3.74), *Communication Monographs* (3.70), *Media Psychology* (3.65), *Public Opinion Quarterly* (3.26), *Social Media + Society* (3.20), *The International Journal of Press/Politics* (3.11), *The Information Society* (3.10). 统计信息取自：https://jcr.clarivate.com/JCRJournalHomeAction.action?pg=JRNLHOME&categoryName=COMMUNICATION&categories=EU#

[2] 论文作者向 12 位同学的辛勤工作致以感谢，他们是（按姓名首字母排序）：中国传媒大学新闻学院 2020 级硕士研究生成淑原、刁华莉、顾小妍、洪东方、姜人文、李明辉、李秋萍、李爽、梁梓琦、刘川、吴馥梅、徐杰。

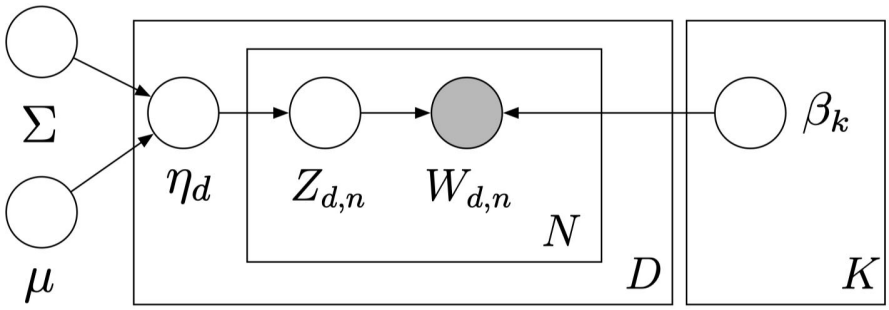

图 1 CTM 原理示意

图 1 中，W 代表词语，D 代表文档，这两项是模型建构环节的可见要素，之后的系列生成过程以此为基石。N 表示文档 D 中的词汇总数，$W_{d,n}$ 表示文档 d 中的第 n 个词，$Z_{d,n}$ 代表主题分配函数。在最右侧矩形框内，K 表示主题个数，β_k 表示第 k 个主题（主题实质上为词汇的分布）。可以看出，每一个词语都由主题分配函数（文档-主题矩阵）和主题函数（主题-词语矩阵）联合决定。进一步地，主题分配函数事关 η_d（与主题比例相关的多项分布表达式），而 η_d 由 $K * K$ 规模的均值矩阵（μ）与协方差矩阵（Σ）联合决定。CTM 与 LDA 的生成过程十分类似，二者的不同之处在于 LDA 的主题比例是从潜狄利克雷分布中计算得出[16]，而 CTM 则是从逻辑斯正态分布中推演出主题比例。

关于 CTM 的实用性，Blei 等[17]曾将 CTM 应用于《科学》（*Science*）期刊刊登的文章（1990 年至 1999 年）之上，发现 CTM 在拟合效果与预测能力上胜过 LDA。作者还提出在探索非结构化数据时，CTM 是一种更为自然的路径[15]。在 Song 等[14]对传播研究知识结构的分析中，CTM 的长处也得到明显体现，尤其是 CTM 相比文献共被引等传统分析路径的全面性与可靠性优势。

3 研究发现

Maier 等[18]在回顾传播学领域使用 LDA 方法的研究后总结出一套提升 LDA 信效度的标准化操作流程。虽然本研究使用 CTM，但 CTM 由 LDA 发展而来，适用于 LDA 的操作策略对 CTM 而言仍具一定的启发意义。该策略包含四个流程：语料预处理、模型参数选择（主题数及超参数确定）、模型信度评估、主题阐释与效度评估[18]。其中，模型参数选择与模型信度评估的目标指向较为接近，可以合二为一。后续分析将按照这三项步骤展开[3]。

第一，语料预处理。研究者首先执行小写转换，接着利用正则表达式匹配去除文内引用及图表相关信息，并执行分词、词形归类、移除标点、保留特定词性（名词、形容词、专有名词等信息量较为丰富的词性）、过滤短词（长度为 1 的词语）、过滤停用词、移除所有数字和非英文字符操作。经过这一步骤，共有来自 252 篇文档的 19432 个独特词汇被保留。

第二，模型参数选择与信度评价。该环节需要搭建一系列竞争模型，进而根据模型拟合指标来确定最优的预先设定参数。理想的参数组合不仅要保证模型可以较好地拟合数据，更要保证模型结果可以得到有效阐释。可解释性（interpretability）与模型信度紧密交织，并共同影响主题模型的效度。常用的模型拟合指标包括主题词概率相似性、独立验证似然（held-out likelihood）、复杂度（perplexity）。本研究选择最常用的复杂度作为模型拟合判断依据。复杂度关乎模型的可靠性，

[3] 该部分后续的分析环节若无特殊交代，皆沿用 Maier 等（2018）的梳理，作者不再添加文内引用。

例如将语料拆分为 10 个子集，以其中 9 个子集为训练语料生成的模型有多大概率适用于最后 1 个验证子集。其次，在迭代次数选择上，本研究依托单个词语对数似然（log likelihood per-word）变化来确定恰当的迭代次数。超参数设置遵循既有成熟机器学习库 scikit-learn[19]和 Gensim[20]中的默认设定，取主题数量的倒数。图 2、3 分别展示了复杂度随主题数量的变化趋势和单个词语对数似然随迭代次数的变化趋势。

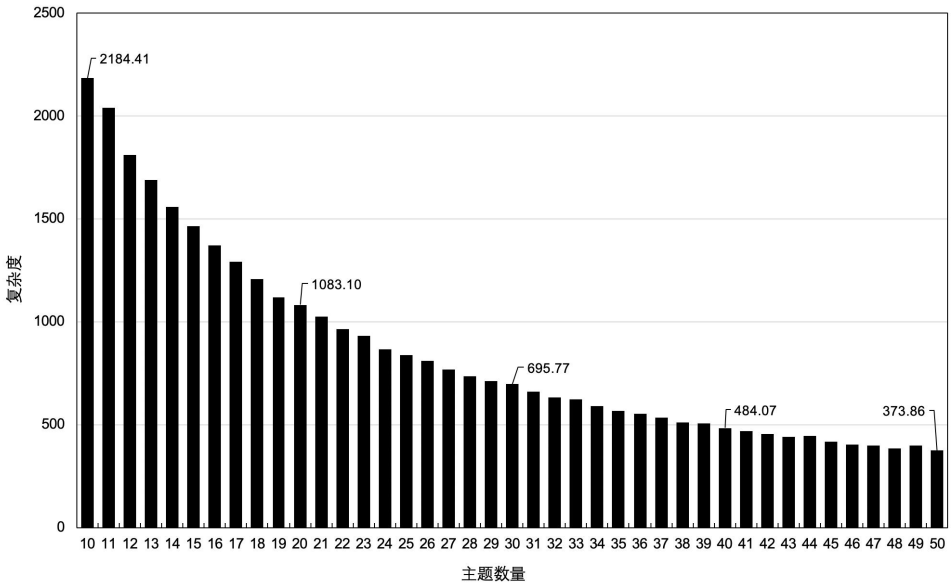

图 2 复杂度随主题数量变化趋势

（注：超参数设置为主题数量倒数，采用逆向文档频率对词语进行加权，迭代次数预设为 1000）

如图 2 所示，我们将主题数量范围设定为 10 至 50。伴随主题数量增长，模型复杂度总体呈下降趋势。图中标注的数据点反映复杂度下降速率逐渐变缓。为保证合理的主题数量与主题阐释清晰性，研究者选定主题数量等于 40 带入后续运算。

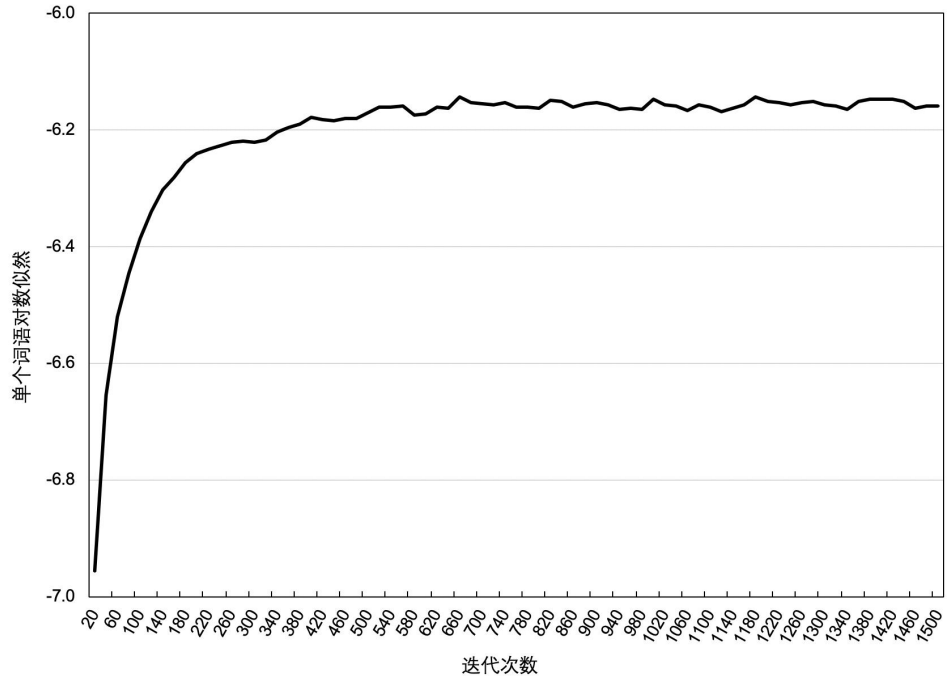

图 3 单个词语对数似然随迭代次数变化趋势

（注：主题数量设置为 40）

图 3 中的对数似然值在迭代 400 次后波动非常微弱，在迭代 660 次时到达最大值（-6.144）。因此，后续模型运算将迭代参数赋值为 660。

第三，主题阐释与效度评估。在 LDA 中，这一环节常用的量化指标包括 Rank-1 度量、连贯性度量、相关性度量、Hirschman-Herfindahl 指数等。由于 CTM 与 LDA 有所差别，且本研究的语料数量偏少，因此我们选择采用人工比对的方法来概括主题并衡量主题的语义覆盖能力。具体来说，挑选者在判别文章、保存文本的同时记录了文章的研究主题和主要分析方法。研究者最后将 CTM 结果与编码环节的记录进行核验。图 4 展示了 40 个主题（从 0 开始编号）之间的关联图谱，每一个节点代表一则主题，节点标签展示该主题下对应的前 5 个关键词。节点之间的连边粗细与对应主题之间的相关系数呈正比。为求清晰简约的展示效果，研究者将 0.10 设定为相关系数截点，相关系数小于 0.10 的主题关系被移除。为更好地把握共通领域，研究者根据主题间相关性执行网络模块分析，进而发掘出 6 个关键领域。

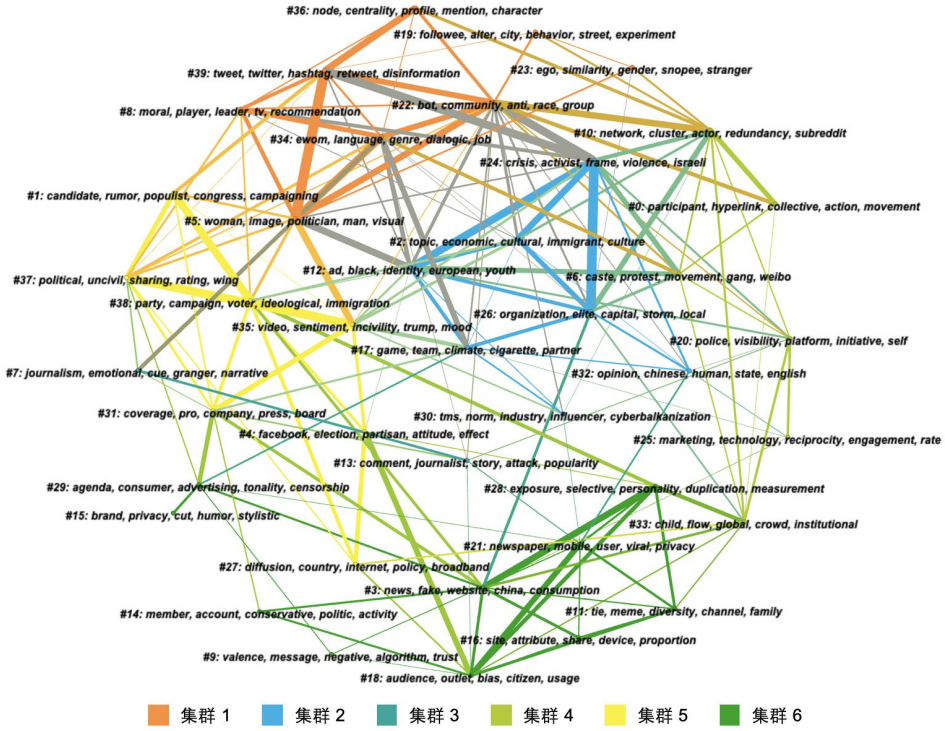

图 4 主题关联与核心领域示意

（注：图中由不同颜色代表的“集群”即为文中提及的“领域”，灰色连边部分为模块分析算法自动排除的部分，其中覆盖的节点皆包含于其他集群中，本研究不单独对灰色部分进行讨论；网络基础指标：节点数 = 40；连边数 = 170；网络密度 = 0.218；模块化指数 = 0.301）

领域 1（占比 20%）主要涉及针对社交网络展开的网络分析（涉及 node, centrality, ego, alter, community, followee, retweet 等关键词）、自动化文本分析（涉及 genre, language, hashtag 等关键词），乃至自动化图像分析（涉及 image, visual, gender, leader 等关键词）。该领域根植于最为典型的“数字痕迹”产生平台（如：推特），并强调利用新兴计算方法分析平台上承载的丰富数据。典型代表包括 Ogan 和 Varol[21]综合使用网络分析和内容分析方法来探究推特在土耳其社会运动当

中扮演的角色，其中涉及到对包含发布推文（tweet）、转发推文（retweet）在内的信息传播行为进行考察，以及利用发布内容来对用户进行聚类；Ji 等[22]融合自动化文本分析与传统内容分析来研究推文中包含的道德判断要素。除却推文之外，社交媒体营销（eWOM）过程中的产品评价也吸引了研究者的关注[23]。一些建基于自动化内容分析和网络分析之上的二次提炼指标同样被纳入计算传播研究者的视野，比如借助社交媒体机器人识别工具 Botometer（基于账号发布内容和账号间互动模式生成二级指数）来探索政治竞选事件中社交媒体机器人产生的影响[24]。

领域 2（占比 17.5%）相较于领域 1 的数据源与方法面向而言，更侧重运用计算传播方法来探究多元化社会议题。包括利用网络分析等方法对社交媒体行动主义中关键行动者（activists）和话语框架（discursive frames）进行识别[25]、利用自动化文本分析方法发掘危机事件（crisis events）时期公众的框架建构（frame building）特征[26]。此外，这一领域还包含了许多群体传播相关论题，比如在交互式存储系统（TMS, transactive memory system）视角关照下研究协作式游戏（game）中的群体（team）特质如何影响游戏结果[6]、社交媒体空间的意见极化（opinion polarization）与“赛博空间巴尔干化”（cyberbalkanization）[27]、媒体内容如何传达关于烟草和电子烟（e-cigarette）的群体规范（population norms）与个体规范（individual norms）[28]。这一领域中涵盖的社会议题还包含身份认同（identity）、移民群体（immigrants）、犯罪集团（criminal gang）等。

领域 3（占比 5%）包含的主题与新闻（journalism）和新闻工作者（journalists）密切相关。典型研究包括对新闻文本进行自动化分析[29-30]、结合特定概念（如：popularity, virality, immediacy）讨论新闻特征与受众新闻消费行为之间的关联[31-32]。传统的新闻研究多采用人工内容分析和相对基础的统计模型，计算传播取向在方法层面提供了更多的创新可能，比如使用事件史分析（event history analysis）、格兰杰因果检验（Granger causality test）来挖掘隐藏的新闻生产及新闻扩散机制。

领域 4（占比 15%）高度聚焦社会运动（social movement）与社会参与（social engagement），这一领域属于鲜明的主题导向，下辖研究方法较为多样。比如结合社会网络分析的相关知识研究在线政治讨论网络的演化[33]，结合话语调适理论展开自动文本分析来研究调解政策如何影响 Reddit 网站的在线讨论[34]。剩余的主题还有网络自组织与志愿行动（voluntary action）、在线抗议（online protest）等。

领域 5（占比 17.5%）同属主题导向，高度关注政治竞选。这一领域和西方政治制度密切衔接，具有一定的语境特殊性。由于该类型研究较为普遍，研究者不在此尽数列举。总体来说，政治竞选中的计算传播研究关注但不限于以下议题：政治讨论的文明（civility）程度、党派媒体（partisan media）对候选人的报道偏见、对政治信息的选择性接触（selective exposure）与意见极化（attitude polarization）、竞选过程中的谣言（rumor）与虚假信息、意识形态极端化（ideological extremity）等。

领域 6（占比 25%）主要指向用户（user/audience）媒介接触（exposure）相关议题。这其中既包括近年来常见的选择性接触、隐私保护（privacy protection）、算法与推荐系统（algorithms and recommender systems）、假新闻接触（fake news consumption）、用户生成内容（user-generated content），也包含从互联网地理（Internet geography）、用户交叠（audience overlap/duplication）等新颖视角出发开展的探索。创新视角的典型代表是 Taneja 等学者的研究[35-36]，他们借助网络流量数据探究了国家或地区互联网使用模式的接近性与差异性，并从区域文化、地理边界、政治经济情境出发对研究结果予以阐释。类似地，Mukerjee 等[37]从网站间的共享用户量出发进行用户网络与媒体网络的构建。诸如此类的新颖研究概念和方法尝试延展了传统受众研究的边界。

4 对国内计算传播学教育的启示

通过对 252 篇计算传播研究论文进行关联主题建模，可以发现计算传播学研究高度倚重社交媒体等平台上的数字痕迹、经过数字化处理的新闻语料；主要利用包含自动化文本分析、社会网络分析在内的一系列方法，来对多元化社会议题作出讨论。正如 Hilbert[4]所揭示的，计算传播取向为传播学研究供以方法上的“催化”。因此，对国际高水平研究的梳理首先为国内计算传播学教育提供方法层面的启迪。这其中包括但不限于：1）获取数据的能力。如 Possler 等[38]所指出的，获取大型数据集有包含与数据持有方合作、网络抓取、调用 API（Application Program Interface，应用程序接口）在内的多种路径，这不仅要求研究者需要掌握相关的数据采集技术，更要具备评估数据质量的能力（如：代表性、维度全面性）和基本的伦理意识；2）分析数据的高阶技巧。计算传播研究中，属性数据和关系数据兼有，国内的专业教育应当有意识地强化对于新兴数据结构分析技能的传授，如开设与网络分析相关的课程。数据结构的转型也意味着传统的以小数据为基础的分析方法有可能难以胜任大型复杂数据分析任务。目前国内传播学定量研究课程中教授的统计模型往往建立在正态分布假设上，然而计算传播研究仰赖的许多数据指标往往呈现出长尾特征，附加季节性、爆发性变动趋势（如：King, Pan, & Roberts 关于在线审查的研究[39]），数据分布的特殊性要求计算传播研究的稳健推断应当依赖于更多的非参数模型与半参数模型，而对于这些统计方法的教导恰是目前国内新闻传播学教育所严重缺失的；3）数据视觉化的能力。相对于可纯粹用统计表格概括研究发现的传统量化研究，计算传播研究的数据异质性和多样性更为明显，这同样呼吁研究结果展现上的变更。诸如选用分布倾向图（参见：Pak 的把关研究[40]）、热力图、网络图谱（参见：Röchert 等的意见同质性研究[41]）等新颖的视觉化形式。

除却方法上的启示，对计算传播研究进行杷梳还赋予国内计算传播教育以研究理念上的参照。传统传播学研究多遵循演绎路径，也即以理论框架为基础推演系列研究假设，进而通过经验数据对假设进行验证，由此检验、改进、扩展理论。计算传播被一些研究者称为科学研究的“第四范式”（the fourth paradigm），认为这一范式可以在大数据的支撑下从理论驱动（theory-driven）走向数据驱动（data-driven），归纳推理的逻辑由此大显身手（详见 Kitchin 的论述[42]）。事实上，本文回溯的大部分高质量计算传播研究并未脱离理论的指导，这些研究往往倾注着研究者明晰的问题意识和理论关怀，且其智识根基源自传播学、社会学、政治学、心理学等多个领域。如同 Gould[43]所言：“数据永远不会为自己代言，我们在调查、分析、阐释数据时总是有意或无意地跟随着概念框架的指导，无论这些框架是结构严密、正式提出的，还是直觉驱动、尚处襁褓的。”因此，传播学研究者必须对“新经验主义”“范式变革”等诱人且精巧的话术保持警惕，无论是传统量化研究抑或是计算传播研究，其本质都是对复杂的社会事实和认识过程进行简化。现阶段的计算传播研究仍应在理论或学说的指导下构建缜密的分析框架，进而让数据为问题服务，逐渐开辟发展理论的契机。相比于断言某种逻辑或范式取代了另一种逻辑或范式，笔者更愿意相信当下的研究场域中孕育着一类混合模式，溯因（abductive）、归纳（inductive）、演绎（deductive）路径并驾齐驱，并给予彼此相互支持、相互验证的动力[42]。对应于国内的计算传播人才培养，授课者应该鼓励学生广泛涉猎多学科文献，将理论支点与社会情境对照，进而培育出一系列有研究价值的命题，再借助计算传播研究提供的一系列新颖方法来提升研究的创新水准以及信效度。计算传播研究秉持开放的视野，相应地，专业教育亦要积极打破学科与研究视角的藩篱，孕育交叉的、反思的、语境敏感的研究思维，杜绝纯粹追求精致方法的数据探索，追求研究价值、社会意义与方法创新的高度统一。

参考文献

[1] Domahidi, E., Yang, J., Niemann-lenz, J., & Reinecke, L. (2019). Outlining the way ahead in computational communication science: An introduction to the IJoC special section on “computational methods for communication science: Toward a strategic roadmap”. *International Journal of Communication*, *13*, 3876-3884.

[2] Waldherr, A., Geise, S., & Katzenbach, C. (2019). Because technology matters: Theorizing interdependencies in computational communication science with actor-network theory. *International Journal of Communication*, *13*, 3955-3975.

[3] van Atteveldt, W., & Peng, T.Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2-3), 81-92.

[4] Hilbert, M., Barnett, G., Blumenstock, J., Contractor, N., Diesner, J., Frey, S., et al. (2019). Computational communication science: A methodological catalyzer for a maturing discipline. *International Journal of Communication*, *13*, 3912-3934.

[5] Shen, C., Ratan, R., Cai, Y. D., & Leavitt, A. (2016). Do men advance faster than women? Debunking the gender performance gap in two massively multiplayer online games. *Journal of Computer-Mediated Communication*, *21*(4), 312-329.

[6] Kahn, A. S., & Williams, D. (2016). We’re all in this (game) together: Transactive memory systems, social presence, and team structure in multiplayer online battle arenas. *Communication Research*, *43*(4), 487-517.

[7] Wu, A. X., & Taneja, H. (2019). How did the data extraction business model come to dominate? Changes in the web use ecosystem before mobiles surpassed personal computers. *The Information Society*, *35*(5), 272-285.

[8] Majó-Vázquez, S., Nielsen, R. K., & González-Bailón, S. (2019). The backbone structure of audience networks: A new approach to comparing online news consumption across countries. *Political Communication*, *36*(2), 227-240.

[9] Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. *ICONIP 2013*. Retrieved from https://scholar.princeton.edu/files/bstewart/files/stmnips2013.pdf

[10] Li, M., & Luo, Z. (2020). The ‘bad women drivers’ myth: The overrepresentation of female drivers and gender bias in China’s media. *Information, Communication & Society*, *23*(5), 776-793.

[11] Peng, Y. (2018). Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, *68*(5), 920-941.

[12] van Atteveldt, W., Margolin, D., Shen, C., Trilling, D., & Weber, R. (2019). A roadmap for computational communication research. *Computational Communication Research*, *1*(1), 1-11.

[13] Zhu, Y., & Fu, K. (2019). The relationship between interdisciplinary and journal impact factor in the field of communication during 1997-2016. *Journal of Communication*, *69*(3), 273-297.

[14] Song, H., Eberl, J. M., & Eisele, O. (2020). Less fragmented than we thought? Toward clarification of a subdisciplinary linkage in communication science, 2010-2019. *Journal of Communication*, *70*(3), 310-334.

[15] Blei, D. M., & Lafferty, J. D. (2005). Correlated topic models. *Advances in Neural Information Processing Systems 18 (NIPS 2005)*. Retrieved from http://papers.neurips.cc/paper/2906-correlated-topic-models.pdf

[16] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

[17] Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1), 17-35.

[18] Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., et al. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2-3), 93-118.

[19] Scikit-learn. sklearn.decomposition.LatentDirichletAllocation. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html#re25e5648fc37-1

[20] Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

[21] Ogan, C., & Varol, O. (2017). What is gained and what is left to be done when content analysis is added to network analysis in the study of a social movement: Twitter use during Gazi Park. *Information, Communication & Society*, *20*(8), 1220-1238.

[22] Ji, Q., & Raney, A. A. (2015). Morally judging entertainment: A case study of live tweeting during Downtown Abbey. *Media Psychology*, *18*(2), 221-242.

[23] Bao, T., Chang, T. S., Kim, A. J., & Moon, S. H. (2019). The characteristics and business impact of children's electronic word of mouth in marketing communications. *International Journal of Advertising*, *38*(5), 731-759.
[24] Keller, T. R., & Klinger, U. (2019). Social bots in election campaigns: Theoretical, empirical, and methodological implications. *Political Communication*, *36*(1), 171-189.
[25] Jackson, S. J., & Welles, B. F. (2016). #Ferguson is everywhere: Initiators in emerging counterpublic networks. *Information, Communication & Society*, *19*(3), 397-418.
[26] van der Meer, T. G. L. A. (2018). Public frame building: The role of source usage in times of crisis. *Communication Research*, *45*(6), 956-981.
[27] Chan, C., & Fu, K. (2017). The relationship between cyberbalkanization and opinion polarization: Time-series analysis on Facebook pages and opinion polls during the Hong Kong occupy movement and the associated debate on political reform. *Journal of Computer-Mediated Communication*, *22*(5), 266-283.
[28] Liu, J., Siegel, L., Gibson, L. A., Kim, Y., Binns, S., Emery, S., et al. (2019). Toward an aggregate, implicit, and dynamic model of norm formation: Capturing large-scale media representations of dynamic descriptive norms through automated and crowdsourced content analysis. *Journal of Communication*, *69*(6), 563-588.
[29] Allen, W. L., & Blinder, S. (2018). Media independence through routine press-state relations: Immigration and government statistics in the British press. *The International Journal of Press/Politics*, *23*(2), 202-226.
[30] Baden, C., & Tenenboim-Weinblatt, K. (2017). Convergent news? A longitudinal study of similarity and dissimilarity in the domestic and global coverage of the Israeli-Palestinian conflict. *Journal of Communication*, *67*(1), 1-25.
[31] Ørmen, J. (2019). From consumer demand to user engagement: Comparing the popularity and virality of election coverage on the Internet. *The International Journal of Press/Politics*, *24*(1), 49-68.
[32] Buhl, F., Günther, E., & Quandt, T. (2019). Bad news travels fastest: A computational approach to predicators of immediacy in digital journalism ecosystems. *Digital Journalism*, *7*(7), 910-931.
[33] Choi, S., Yang, J. S., & Chen, W. (2018). Longitudinal change of an online political discussion forum: Antecedents of discussion network size and evolution. *Journal of Computer-Mediated Communication*, *23*(5), 260-277.
[34] Gibson, A. (2019). Free speech and safe spaces: How moderation policies shape online discussion spaces. *Social Media + Society*, *5*(1). Retrieved from https://doi.org/10.1177/2056305119832588
[35] Ng, Y. M. M., & Taneja, H. (2019). Mapping user-centric Internet geographies: How similar are countries in their web use patterns? *Journal of Communication*, *69*(5), 467-489.
[36] Wu, A. X., & Taneja, H. (2016). Reimagining Internet geographies: A user-centric ethnological mapping of the World Wide Web. *Journal of Computer-Mediated Communication*, *21*(3), 230-246.
[37] Mukerjee, S., Majó-Vázquez, S., & González-Bailón, S. (2018). Networks of audience overlap in the consumption of digital news. *Journal of Communication*, *68*(1), 26-50.
[38] Possler, D., Bruns, S., & Niemann-Lenz, J. (2019). Data is the new oil- But how do we drill it? Pathways to access and acquire large data sets in communication science. *International Journal of Communication*, *13*, 3894-3911.
[39] King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, *107*(2), 326-343.
[40] Pak, C. (2019). News organizations' selective link sharing as gatekeeping: A structural topic model approach. *Computational Communication Research*, *1*(1), 45-78.
[41] Röchert, D., Neubaum, G., Ross, B., Brachten, F., & Stieglitz, S. (2020). Opinion-based homogeneity on YouTube: Combining sentiment and social network analysis. *Computational Communication Research*, *2*(1), 81-108.
[42] Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1). Retrieved from https://doi.org/10.1177/2053951714528481.
[43] Gould, P. (1981). Letting the data speak for themselves. *Annals of the Association of American Geographers*, *71*(2), 166-176.