大数据中的百年社会学*

——基于百万书籍的文化影响力研究

陈云松

提要:本文基于谷歌图书的最新语料库(811 万种书籍,8613 亿词汇),通过设计、检索社会学的学科关键词以获得历年词频数据,对 19 世纪中期以来社会学的发展进行追踪,并藉此展示、分析和诠释了社会学的学科轨迹、名家大师、理论流派、领域热点、分析方法,以及中国社会学在文化影响力维度上的百年变迁,同时对建立"社会组学"进行了展望。本研究为利用大数据进行人文社科研究提供了初步经验。

关键词:大数据　社会学史　文化组学　社会组学　文化影响力

一、导　言

"大数据"的应运而生,既给商业、管理和公共行政带来了众多机遇,提出了诸多课题,同时正在全球社会科学界、人文学界引发讨论的热潮。①早在数年前,哈佛的加里·金在在展望政治学的未来 50 年时就预言,随着大数据的出现和使用,整个社会科学研究的实证基础将会出现重大的变化,甚至会加速定性与定量研究的大融合(King,2009)。国内学界也对大数据给社会科学领域带来的机遇充满信心(沈浩、黄晓兰,2013)。不过,"大数据"概念虽热,但社会科学界基于大数据的实证研究却比较薄弱。一方面,大数据往往为政府、大型公司或网络媒体所持有,数据获取渠道和分析方法都与传统的社会科学定量定性分

* 感谢牛津大学维克托·迈尔-舍恩伯格(Viktor Mayer-Schönberger)、杜克大学林南、西安交通大学边燕杰、香港科技大学吴晓刚、中国社会科学院张宛丽、南京大学周晓虹、成伯清、闵学勤、吴愈晓、王浩斌等教授对分析内容和方法的意见建议。感谢匿名审稿人的重要意见。

① "大数据"定义不一而足,总体上是指大小和信息量超过传统规模的海量数据资料,尤其是那些不是通过随机抽样方法所得到的调查数据。具体的定义和对社会科学的价值参见沈浩、黄晓兰,2013。

析存在很大差异;另一方面,现有的大数据往往本身并非为社会科学研究而设立,其在样本代表性、测量可靠性等方面往往受到学界的诟病(Tufekci,2014;Boyd & Crawford 2012)。

2011 年,以让-巴蒂斯特·米歇尔(Jean-Baptiste Michel)为首的哈佛大学、麻省理工学院、大英百科全书的学者与谷歌研究团队,共同署名在《科学》杂志第 331 卷发表了题为《使用百万数字化书籍的文化定量分析》的重要论文(Michel et al., 2011)。该研究借助谷歌图书的海量数字化资料,分析了公元 1500 年到 2000 年间 500 多万本书籍高达 5000 多亿单词的语料库(corpus)。通过对关键词在语料库中的使用频率变化,展示了五百年来人类文化发展史中或鲜为人知或饶有兴味的趋势和现象。这个全新的研究领域,被称为"文化组学"(culturomics)。利用这个文化大数据,国际语言学界和历史学界已经出现了一些跨学科的探索性研究(Bentley et al., 2014;Acerbi et al., 2013;Twenge et al., 2012)。海量的数字化书籍和兄弟学科的最新探索,为社会科学领域的大数据应用研究提供了难得机遇。

社会学自 19 世纪末诞生以来,理论和方法日益丰富,学派和名家不断涌现,其理论和成果对人类经济、政治和社会文化生活的影响也在不断扩大和深入。在学术界内衡量一个学科或者某项研究成果的影响,我们往往依靠学术文献和引用指标(如学术书籍、学术期刊、论文引用影响因子),不过,要在更为宏观的时间、空间维度上观察甚至评估理论的发展、学者的成长乃至整个学科对于人类知识谱系的影响力,也即"文化影响力",则要复杂和困难得多。现在,基于大数据的词频统计技术为这一领域的探索提供了可能。本文将利用谷歌语料库千亿量级的海量数据,通过对社会学关键词的词频分析来初步展示百年社会学发展历程中的现象和规律。本研究也是我国社会学领域的首次大数据分析尝试。

二、数据、概念和策略

让-巴蒂斯特·米歇尔等分析的数据来自谷歌图书(Google Books)。自 2004 底起,谷歌公司陆续对哈佛、牛津等 40 多所顶级大学图书馆藏书及出版社赠书进行了浩大的数字化工程,到 2013 年,谷歌

已对超过三千万种书籍进行了扫描识别，占人类自古登堡印刷术发明以来出版图书的约四分之一，其中数字化质量较好可供全文检索的达八百多万种(8116746)，词汇量8613亿(Lin et al.，2012)。表1分别展示了谷歌图书语料库的主要构成。为实现基于全文检索的词频统计，该语料库采用了词汇连续语音识别中的"n-gram"算法模型①以实现对语料库中海量文本的切分、断句。

书籍是承载人类知识、观念和思维的最主要的载体。只要语料库具有足够的代表性，我们就可以认为一个词汇在书籍中出现的频率，能够近似地反映这个词汇及其相关意蕴的"文化影响力"(涵盖知名度、关注度、影响力等多个维度)，甚至折射出某种社会趋势、风尚或思潮(Twenge et al.，2012)。以"社会流动"一词为例：首先，语言和词汇反映了作者的观点，而书籍作者比一般人拥有更大的文化影响力。作者群体越多地提及"社会流动"，就说明该词的文化影响力越高；其次，书籍出版会考虑读者的需求，因此书籍词汇的总体特征往往能反映大众观念和思维偏好。书籍中"社会流动"出现得越多，就意味着大众对相关的社会现象越为关注。

表1　　谷歌图书语料库的构成(2012年第2版)

	书籍量(万)	词汇量(亿)
英语	454	4685
法语	86	1022
西班牙语	79	840
德语	66	647
汉语(简体)	30	269
俄语	59	670
希伯来语	7	80
意大利语	30	400
合计	811	8613

谷歌语料库为文化研究、语言学研究、观念史研究等提供了难得的

① 在该模型中，n-gram具体表示一个单词或者词组(本文称之为"词汇")。例如，1-gram就是单字词，比如"hello"，而3-grams则是三字词组，比如"how are you"。模型假设第n个单词的出现仅仅与前面n-1个词相关，这样一个完整词汇的出现概率就是各单词出现概率的乘积，而各单词出现的概率可以从语料库中统计出来。

文化大数据。特温格等(Twenge et al., 2012)对美国20世纪书籍中个人化用词的趋势进行了历史解读;阿瑟比等(Acerbi et al., 2013)对人类20世纪书籍中感情用词的演变以及英式美式英语差异进行了分析;宾利等(Bentley et al., 2014)等研究了20世纪美国经济与悲观性词汇使用之间的关联。这批新近的跨学科、跨领域研究,为我们从新的角度观察社会学发展史和探索社会学领域的大数据应用提供了启示。本文将借鉴"文化组学"的研究方法,使用谷歌图书语料库的最新2012版进行社会学词频分析。有关数据特征、概念操作化和分析策略归纳如下。

(一)数据的代表性

谷歌图书语料库2012版拥有1500年以来的811万种印刷图书、8613亿单词。考虑到社会学的诞生是在19世纪末,且英语是百年来全球使用最为广泛的语言之一,我们将检索范围设定为19世纪中晚期到2008年的英语语料库。[①] 由于19世纪以来的图书印刷质量较之早期图书更高、数字化识别率也更好,因此其进入全文检索语料库的比例要较早期图书高出很多。这使得本文检索对象的代表性比谷歌图书语料库跨度五百年的总体代表性要高得多。实际上,本文的检索分析对象几乎囊括英语世界19世纪中晚期以来的绝大部分书籍。最后,尽管书籍内容包罗万象,出于谨慎我们在辅助分析中进一步对非书籍语料库进行了分析:具体而言,我们将利用19世纪中晚期以来的平面媒体(报纸)全文数据库对相关关键词进行检索。如果基于报纸的检索结果和基于书籍的检索结果非常接近,就能进一步证明谷歌图书大数据的代表性。相关结果我们在附录中展示。

(二)数据的针对性[②]

人文社科知识体系的建立、扩张和影响力,以及成果的弥散,比物理、化学等自然科学更借助于文字的形式,也就更多地依托书籍、报纸和杂志等文化载体。不过,读者难免有疑问:为何不直接使用学术期刊

① 谷歌图书语料库的入库书籍目前最晚为2008年。同时,对部分关键词我们也对汉语语料库进行了检索以作为分析的辅证。

② 感谢匿名审稿人对数据库针对性的建议和意见。

来作为社会学关键词的分析对象？实际上，除了谷歌图书语料库更符合大数据的基本特征之外，还有三个方面的原因。第一，书籍内容的覆盖面要比学术期刊广泛得多，而本研究的目的恰恰在于分析百年来社会学的文化影响力变迁而非单纯的学术发展史；第二，作为书面语言的载体，学术期刊的发展、成熟本身要比书籍晚得多，如果用期刊数据库进行分析，早期的社会学相关信息可能会有较大偏误；第三，学术期刊数据库提供的检索功能往往只达到作者、关键词、学科领域级别，[①]有的虽能实现全文检索但又无法提供词频信息。因此，谷歌图书语料库无论在数据规模还是完整性、科学性等方面，都比学术期刊数据库更适合本研究。

（三）概念的操作化

我们正式定义：在某个时间跨度内的具有较好代表性的语料库中，一个社会学关键词的“词频比例”，即其在样本书籍中出现的次数与样本书籍中全体单词总量的比值（考虑到每年书籍总量不一），可以代表该社会学关键词在该时段内的文化影响力。这样，利用谷歌图书语料库对一系列学科关键词进行检索统计，我们就可以获得这些关键词自社会学诞生以来一个多世纪中的历年“词频比例”。在任何一个年份，关键词词频比例越高，就表明其在全社会的使用和提及程度越高，文化影响力越大。

考虑到书籍出版年份越靠后，进入书籍中数字符号等非词汇性内容越多，因此我们用关键词出现频数除以英语单词“the”的出现频数来计算年度词频比例。[②] 具体计算公式为：

$$R_{i\_t} = \frac{C_{i\_t}}{C_{the\_t}}$$

其中，R_{i_t} 为关键词 i 在公元 t 年的词频比例，C_{i_t} 表示关键词 i 在公元 t 年的出现次数，C_{the_t} 为公元 t 年中“the”的出现次数。

（四）检索词的设计

我们的检索分析主要基于英语库。检索方向分为 6 类：学科轨迹、

① 例如，如果一篇论文的关键词和标题未提及马克思，但论文内容却不止一处提及马克思，那么基于期刊数据库的检索就因漏算而低估马克思的影响力。

② “the”在英语书籍中出现频数非常稳定（Bentely et al.，2013）。

名家大师、理论发展、领域热点、分析方法以及中国社会学。关键检索词的设计我们主要参考了斯科特和马歇尔主编的《牛津社会学词典》(Scott & Marshall, 2005)、吉登斯和萨顿的《社会学》(Giddens & Sutton, 2013)、贾春增的《外国社会学史(第三版)》(2008)、谢立中的《西方社会学名著提要》(2007)等辞书和教科书。选取辞书与教科书而非社会学理论专著作为关键词选择依据的原因在于:第一,辞书和教科书本身对学科的总体发展有比较清晰的梳理,其章节、条目为关键词检索提供了良好的备选;第二,社会学辞书、教科书的数量较之社会学著述要少得多,这使得我们可以在前人的总结梳理基础上较为快速和准确地确定关键词。

(五)检索精度的设置

如果关键词在当年书籍中出现少于40次,就被作为0值处理。换句话说,检索得到的词频本身就是"规模性"出现的"热词"词频。40次的门槛设置,除了让数据分析和绘制图形更为简洁之外,对检索精度具有重要的价值:例如,在搜索社会学名家的英文全名之时,通过"热词"筛选就可以排除一些和社会学大师同名同姓的普通人——除非他本身是其他领域的知名人物。此外,我们还根据不同的情况设置了单词字母大小写的严格区分或模糊区分(如人名中区分大小写),对关键词非核心部分进行了有针对性的取舍(如检索"固定效应"而非"固定效应模型"),以确保检索结果的科学性。最后,考虑到图形的视觉效果,我们对词频比例曲线进行了2年平滑处理:以1950年为例,经过平滑后该年份的数值为它与前后两年原始数据一共5年的平均值(1948、1949、1950、1951和1952年的均值)。

三、大数据中的学科轨迹

我们首先分析"社会学"(sociology)这一最重要的学科关键词自1850年以来在英语书籍中的出现频次。为进行对比,我们同时对哲学(philosophy)、经济学(economics)、人类学(anthropology)和心理学(psychology)等四个兄弟学科进行同步检索分析。[①] 图1的横坐标是

① 我们未对政治学(politics或political science)进行搜索,因为相关词汇具有学科之外的含义。

1850－2008 年的时间轴，纵坐标是社会学关键词的词频比例。从图 1 可见，在 150 年来的英语书籍中，“哲学”二字的词频比例总体上保持在 0.008% 上下，也即十万分之八。与其他社会科学门类相比，哲学词频出现更早、占比更高。不过，在 19、20 世纪交替的自由资本主义发展晚期，哲学词频曲线进入了下降通道，直到 20 年代才开始恢复。实际上，哲学史上与此对应的正是 19 世纪中叶德国古典哲学尤其是黑格尔学派的解体。而在哲学词频曲线缓降的世纪之交，其他学科词频则各自崛起。

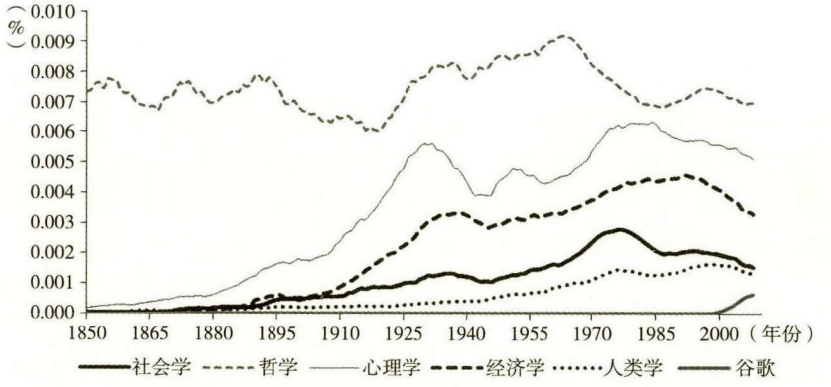

图 1　英文学科名称的词频比例曲线（1850－2008）

社会学、经济学、心理学和人类学的词频自 19 世纪中晚期开始一直到 20 世纪 30 年代初均保持了强劲的上升，而心理学和经济学的势头尤其明显并逐渐拉开与社会学和人类学的距离。不过，在 1870－1880 年间以及 1905 年前后，社会学词频曾经有过高于经济学的辉煌。此外，第一次世界大战期间（1914－1918），社会学、经济学和心理学的词频并未衰减，而二战期间（1939－1945）这三个学科颓势明显，且在 1945 年二战一结束后就迅速上升。这似乎意味着二战对于社会学、经济学和心理学的冲击比一战明显得多。① 同样有趣的是，二战对人类学的词频曲线非但没有负面影响，甚至还微微提升了增幅。这可能是

① 考虑到美式英语书籍多于英式英语书籍，该现象可能反映了美国在两次世界大战中的不同卷入程度。通过对美式英语库的检索，我们验证了这一猜想：美语资料显示，两次世界大战的冲击差异比图 1 更为明显。

因为:与一战相比,二战的交战区域和深度卷入的交战国扩大到了亚洲和大洋洲。空间跨度更大的战争,一方面使得应用人类学得到参战国有目的的资助,另一方面人类学者本身的研究视野也得以从非洲、印第安部落等传统对象里解脱出来,辐射到东欧、东南亚等地区(翁乃群,2000)。①

20 世纪 70 年代末 80 年代初,社会学、经济学、心理学和人类学的词频曲线几乎都达到了整个 20 世纪的高峰。但进入 90 年代之后尤其是世纪交会之际,这四门学科的词频曲线似乎又都开始缓慢下降。不过,考虑到人类书籍的词汇量在快速增加,②在不断扩大的语料库中,词频比例下降可能仅仅代表了一种知识的稀释过程:在不断膨胀的知识海洋中,每个学科或领域的"份额"都可能缩小。另外一个可能就是,因为谷歌语料库仍在对 2000 年之后的书籍进行数字化,所以该时段的样本代表性可能有一定不足。为此我们在检索中专门加入了"谷歌"(Google)字样以进行对比。我们发现,即便在样本代表性可能不足的 2000 - 2008 语料库中,谷歌的词频统计仍然显示出有力的增长。③这间接证明了知识稀释过程作为诠释的有效性。最后,我们还利用汉语语料库进行类似检验,发现所得结果的主要模式和图 1 也非常类似(参见附录 2 图 13)。

利用书籍语料库,我们还进行了更具实质性的社会学发展史研究。例如,在 19 世纪 80 年代到 20 世纪 30 年代的美国社会学草创之初,"社会福音"(Social Gospel)宗教运动的主要倡导者多在大学任社会学教职,并将社会学作为宗教运动的延伸工具。因此,半个世纪以来,社会学史研究学者不断讨论美国社会学的发轫与社会福音运动之间是否存在紧密关联(Morgan, 1969; Maclean & Williams, 2012)。摩根(Morgan,1969)认为:"社会福音和社会学在思想和领军人物上几乎难以区分……而这种紧密联系在欧洲却不明显"。不过,这些研究全是

① 本尼迪克特的《菊与刀》及战后对中国和东欧犹太小镇的研究就是典型案例。

② 在谷歌图书语料库中,19 世纪的词汇年增长量为 6 千万,而 20 世纪的年增长量为 14 亿,21 世纪的每年增长量为 80 亿(Michel et al., 2011)。

③ 为进一步验证以上发现的学科词频比例关系,我们还进行了两种稳健性测试。第一,以"社会学的"(sociological)、"心理学的"(psychological)和"人类学的"(anthropological)为关键词进行检索统计,相关曲线模式和图 1 基本一致。第二,利用简体汉语语料库,对中国改革开放以后社会学、政治学、经济学和人类学汉语词频进行了统计分析,结果也和英语曲线非常接近。

通过案例分析和内容分析等质性方法进行，缺乏数据的经验支撑。

书籍大数据为我们提供了破解这一困局的机会。在本文中，我们以“社会学”、“社会福音”和“霍尔馆”（Hull House，社会福音运动中最为著名的睦邻中心，其创立者简·亚当斯后获得诺贝尔和平奖）为关键词进行检索，同时用“人类学”来作为比照，并分别在美式英语和英式英语数据库中进行对照分析。除了从图 2 中可以看到美国社会福音和社会学较之英国其曲线起伏更为相互呼应之外，我们还计算了这一时段内美国社会学与社会福音、社会学与霍尔馆词频比例的相关性。我们发现，皮尔逊（Pearson）积矩相关系数分别达 0.78 和 0.57，均在 0.001 水平上高度显著。而在英式语料库中的结果则是：社会学与社会福音相关系数为 0.43，仅在 0.02 统计水平上显著，与霍尔馆干脆就无显著相关。在美语库中我们进一步用滞后 10 年的社会福音词频来计算，发现其和社会学词频的皮尔逊积矩相关系数高达 0.85 和 0.75。考虑到计算皮尔逊积矩相关系数的条件是连续数据、正态分布和线性关系，我们进一步放松假设，发现社会学与社会福音及霍尔馆的斯皮尔曼（Spearman）等级相关系数分别为 0.89 和 0.77。这些发现，为验证社会福音运动促进美国社会学发展提供了初步的明确证据。

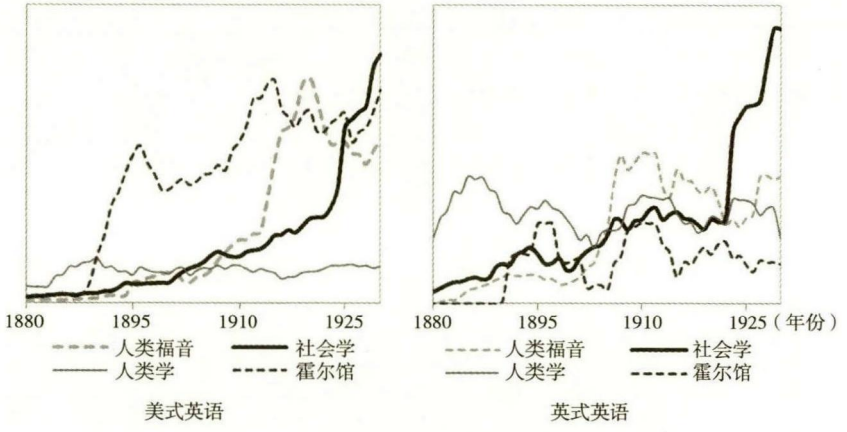

图 2　早期社会学与社会福音运动的关联（1880－1930）

四、大数据中的社会学名家

社会学学者众多,我们对社会学科领域较为知名的30位西方社会学家的英文全名进行了检索。图3中展示的是词频比例曲线总体水平比较高的前12位。[①] 按照出生年月,他们依次是:马克思(Karl Marx)、斯宾塞(Herbert Spencer)、韦伯(Max Weber)、涂尔干(Emile Durkheim)、齐美尔(Georg Simmel)、马尔库塞(Herbert Marcuse)、帕森斯(Talcott Parsons)、戈夫曼(Erving Goffman)、鲍曼(Zygmunt Bauman)、哈贝马斯(Jurgen Habermas)、布迪厄(Pierre Bourdieu)和吉登斯(Anthony Giddens)。[②] 从图3中我们总结如下几点。

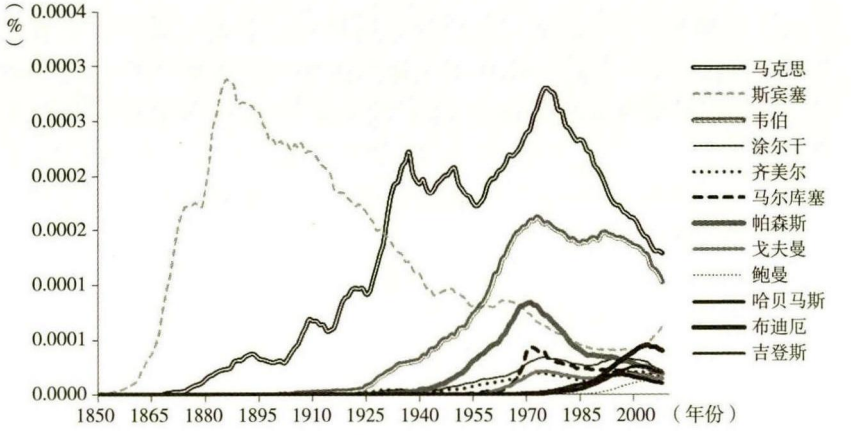

图3 百年社会学大师的词频比例历史曲线(1850-2008)

第一,稀释效应。从马克思到吉登斯,后人似乎再也难以超越前人

① 其他18位社会学家的词频曲线基本都密布在哈贝马斯词频曲线下方:库利(Charles Horton Cooley)、舒茨(Alfred Schutz)、布鲁默(Hebert George Blumer)、米德(George Herbert Mead)、加芬克尔(Harold Garfinkel)、霍克海默(Max Horkheimer)、卢曼(Niklas Luhmann)、卢卡奇(Lukács György)、米尔斯(C. Wright Mills)、默顿(Robert Merton)、达伦多夫(Ralf Gustav Dahrendorf)、伦斯基(Gerhard Lenski)、布劳(Peter Blau)、柯林斯(Randall Collins)、亚历山大(Jeffrey Alexander)、科尔曼(James Coleman)、沃勒斯坦(Immanuel Wallerstein)和埃利亚斯(Norbert Elias)。

② 对姓名中的非英语字母我们也进行了检索,如涂尔干的检索数据系由Emile Durkheim加émile Durkheim计算得来。类似的还有哈贝马斯等。

在文化影响力方面的辉煌。这个发现并非是指社会学家个体的影响无法超越某一位前辈。比如,布迪厄在80年代之后的影响力就超越了前辈的齐美尔与涂尔干,到2003年左右达到0.00005%,在当时仅次于马克思与韦伯。但就社会学家群体进行"代际"比较分析,我们发现70年代帕森斯所达到的巅峰值是0.00008%,而后来者无一能够超越,更不用提达到早期大师斯宾塞和马克思0.0003%左右的水平了。从词频比例曲线的趋势判断,在群体的层次上,后期的大师要超越甚至接近早期大师达到过的巅峰,几乎是不可能的。

我们推测,这种现象可以归因为两个方面:第一,近一百年来人类知识总量和门类的快速增长。进入20世纪和21世纪,尽管社会学本身在不断发展,也在涌现大师级人物,但其在人类总体知识中的相对影响力也即词频比例却比以往下降了。这就好比社会学在人类总体知识水库里被不断稀释。因此,我们称之为"稀释效应"。第二,社会学总体知识也在增长、裂变,所以后来者很难超过前者。实际上,这种现象也可以说是路径依赖或者先发优势。[①] 学界过去常称帕森斯为社会学集大成者,实际上,他或许还是最后一位能够在影响力上和古典大师勉强处在同一个重量级的集大成者——起码在今天,我们仍然看不到布迪厄超越他的可能。

第二,外力效应。和其他社会学家相比,词频比例曲线的上升阶段平均斜率最高的是斯宾塞和马克思。也就是说,他们除创造了有史以来社会学家最高影响力的记录,还是历史上影响力增长最快的大家。不过,他们影响力的迅速崛起,有着截然不同但都异常强大的学术之外的力量支撑:斯宾塞借助了高质量的社会网络并充分发挥了自身的多面手优势,在知识总量相对不多的19世纪末就顺利达到了影响力巅峰;而马克思则依靠其改变20世纪全球政治格局的理论力量,在一个世纪后走向影响力的制高点。

实际上,斯宾塞本人涉猎极广,集哲学家、生物学家、人类学家、社会学家、政治理论家和古典文学家于一身。同时,他一生与社会名流过从甚密,曾由赫胥黎介绍加入著名的"X俱乐部"(X Club),结识了达尔文、胡克在内的一批重要的思想家和权贵。最后,他本人还是维多利亚时代最好辩的思想家——和今天的新媒体时代一样,人脉和舆论焦

① 感谢匿名审稿人对该观点的启发。

点是助力成名的有力工具。相比之下,马克思的个人命运要困窘得多。但马克思主义在现实政治中的实践改变了世界格局。也因此,我们看到马克思的词频比例增长最快的是整个20世纪20-40年代以及70年代,而这两个阶段正是马克思主义在全球快速传播和共产主义运动的高峰期。

第三,加速效应。在20世纪,社会学家的成名越来越早。除了情况特殊的斯宾塞,举凡出生在19世纪的社会学大师,都是"身后成名"。例如,马克思逝世于1883年,而他的词频快速增长在其辞世20年后的20世纪初才出现。韦伯1922年去世,他的名声鹊起,恰恰从其去世后才开始。其他三位出生在19世纪的大师涂尔干、齐美尔和马尔库塞,前两位未能在身前看到自己声名鹊起,马尔库塞也仅仅在去世前10年名声大噪。生于20世纪的晚辈社会学家们则幸运得多。例如,帕森斯在40年代就开始快速成名,其时不过40多岁,而吉登斯也在不惑之年开始成名。哈贝马斯和布迪厄相对属于大器晚成者,但其词频比例在他们50岁之后也即80-90年开始快速增加。而且,他们至今仍然健在。这种个人影响力方面的代际差异,我们称之为加速效应,并归因于20世纪社会学学科体系不断发展和规范化:在19世纪晚期社会学草创之初,学者数量少,学科发展水平较低,传播交流社会学的途径有限,这就使得社会学者发挥影响力所需要的时间大为延长。而随着社会学学科发展加快,随着大学社会学系科的建立和发展,学者拥有越来越好的学术阵地、生活保障以及期刊书籍等媒介来发挥影响力,这使得20世纪的学者能够在健在时就看到成就被社会认可。

五、大数据中的社会学理论

社会学对人类知识的贡献在于一系列具有启发和诠释意义的概念、假说和理论,以及藉此形成的诸多知识流派和体系。因此,我们可以通过对经典社会学理论关键词的词频分析,了解社会学的直接成果对社会的影响和变迁。考虑到19世纪社会学大师多进行的是开创性、奠基性的工作,我们把注意力集中在20世纪中期以来的社会学理论。

在图4中,我们集中展示了常见的八种理论[①]的词频曲线:冲突理论(conflict theory)、交换理论(social exchange theory)、结构功能主义(structural functionalism)、结构化(structuration theory)、符号互动(symbolic interactionism)、理性选择(rational choice theory)、常人方法学(ethnomethodology)、新功能主义(neo functionalism)、弱关系(strength of weak ties)和结构洞(structural holes)。从图4中我们归纳出如下几个现象。

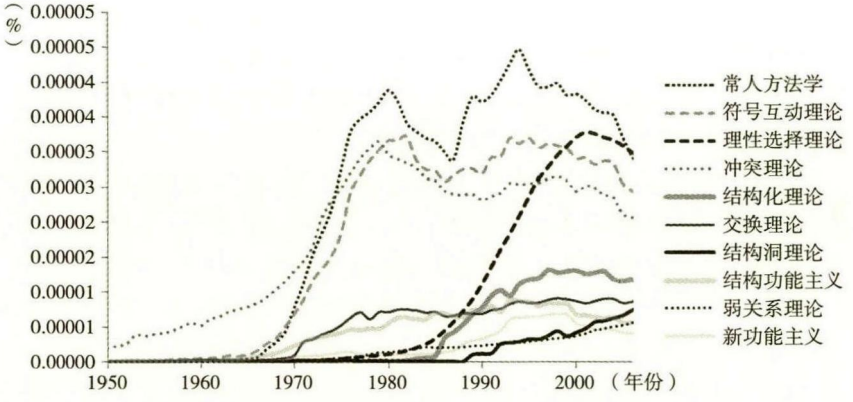

图4　社会学理论的词频比例历史曲线(1940－2008)

第一,理论的生命周期。我们发现,理论从提出到成型、成熟再到式微有一个生命周期。在20世纪中后期,绝大部分理论从提出到达到词频比例的最高点,总体上需要30－40年左右。此后理论的影响力开始缓慢下降。但由于尚未观测到稳定的最低谷,因此我们尚不知理论衰退所需的时间。此外,尽管我们用来分析的理论数量很有限,但该发现和语言学研究的结果契合得较好。彼得森等(Petersen et al., 2012)发现,人类词汇的周期约在30－50年,也即新词汇从出现到消亡或者稳定使用,需要30－50年时间。我们推测,理论的生长

① 在具体设置检索关键词时,我们根据不同理论在文献中的习惯叫法选取最常使用的作为核心检索词。例如,结构化理论我们使用了structuration theory而不是theory of structuration;弱关系理论我们使用了strength of weak ties而不是theory of weak ties。由于冷僻叫法的词频数量级比常用叫法小很多(可达数百倍),因此不纳入冷僻叫法的词频比例并不会影响数据精度。而对于多种叫法并存、数量级接近的,我们则采取了同时检索并相加的方法,例如符号互动同时就汇总了对symbolic interaction的检索。

和衰退周期既和词汇周期有关,同时也取决于社会学理论本身的更新速度。

第二,理论的新陈代谢。例如,结构功能主义、新功能主义词频比例90年代中期就开始下降,而比它们晚出20多年的结构洞理论却已经在词频上超越了前者。此外,70年代兴起的常人方法学、符号互动、冲突理论等也已从90年代开始衰减了约20年,而理性选择约从新世纪开始进入下降通道。90年代以后,新生代理论呈现强劲的增长势头。如果我们把弱关系和结构洞理论相叠加,其词频比例在2008年左右已经可以超过交换理论和结构化理论。也就是说,新兴的社会资本或社会网理论,文化影响力实际已开始超越经典理论。当然,至于它们能不能进一步上扬甚至重现常人方法学、符号互动或理性行动等增长极为迅速的成功理论,尚需时间考验。

第三,理论的解释层次。一般我们会认为,宏观大理论具有更高的概括能力和更宽的辐射使用面,也因此会具备较大的影响力。但是我们发现,起码20世纪中期以来理论世界不再由宏大叙事主导。例如,结构化、结构功能主义、新功能主义均处在词频坐标的中下游,虽然历来是教科书的重点,但和常人方法学、符号互动理论、理性行动理论等基于行动的理论相比存在不小差距。此外,随着时间推移,大理论的空间似乎越来越小,70年代之后兴起的弱关系、结构洞等理论,关注面都非常集中。我们推测,盖因大理论过于野心勃勃而降低了解释力和吸引力,且又越来越缺乏空白的生长点。因此,社会学可能开始进入某种"后大理论"的时代。当然,这一推测是否合理尚待时间检验。

六、大数据中的社会学研究领域

社会学研究领域众多,且非一成不变。一方面,社会学拥有众多的子学科;另一方面,学科的研究热点也随时代进步而不断转移变化。因此,利用大数据我们可以对社会学子学科的结构和变化进行分析,也可对研究热点的变迁进行一些解读。我们首先对教育社会学(educational sociology 和 sociology of education)、农村社会学(rural sociology)、城市社会学(urban sociology)、政治社会学(political

sociology)、经济社会学(economic sociology)、法社会学(sociology of law)、宗教社会学(sociology of religion)和历史社会学(historical sociology)等八大子学科进行检索。

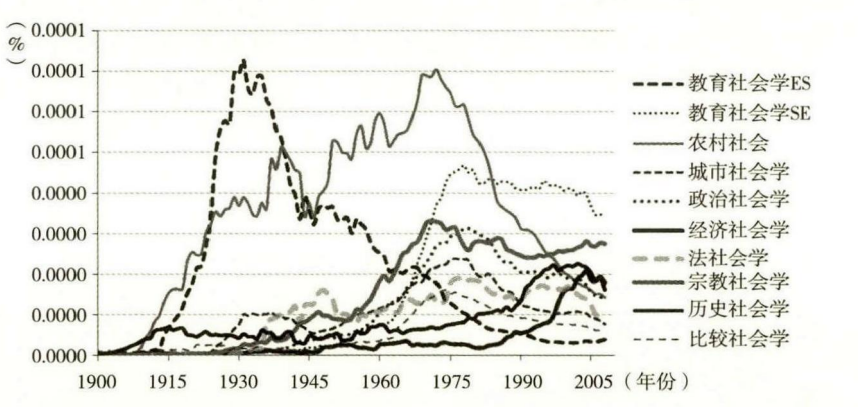

图5　社会学子学科的词频比例历史曲线(1900－2008)

我们从图5发现几个有趣的现象:第一,教育社会学无疑是社会学中最有分量的。不过,从60年代后期开始,教育社会学更多称作sociology of education(SE)而不再是早期的educational sociology(ES)。这主要是因为,早期的educational sociology主要关注的是文化和社会因素,研究如何给公众提供更好的教育,而sociology of education则关心的是国家、政府和个人因素对个体教育结果的影响。第二,我们把90年代后词频不断增长的子学科用实线表示。可见,宗教社会学和历史社会学发展势头比其他领域要强劲,而经济社会学保持平缓发展,其他子学科词频都呈下降趋势。第三,农村社会学60年代词频比例增速极高,进入80年代后期甚至超过教育社会学,完全压倒了其他分支。这个发现为以往研究中的一些观点提供了印证和补充:农村社会学是美国社会学最早也是最大的分支,50－60年代是其最鼎盛时期。实际上,我们发现70年代可能才是它真正的高峰。

除了学科分支,我们还关心社会学研究热点领域的变化。在图6中,我们比较了社会分层和流动、社会资本与网络两大研究领域的8个最具代表性的术语(社会身份social identity,社会地位social status,社会运动social movements,社会流动social mobility,社会分层social stratification,社会资本social capital,社会网络social networks,社会阶层

social class)。这两个领域的研究,集中了社会学近10年来的热点。但它们的词频比例却不尽相同。我们能看到,社会分层和社会流动的词频比例在1975年左右达到高峰,然后开始下降。而社会运动和社会网络则从80年代末90年代初迅速上升,约在世纪之交分别超越了社会地位和社会流动。同样在这段时间附近,社会资本的词频比例也迅速超越社会流动,且增长速度更快,到2003年左右已经超越了社会阶层成为词频最高的领域。为比较中西差异,我们也对汉语语料库进行了同样的检索。结果同样是社会资本、社会网络在2004年前后超越了社会阶层和流动。这反映了社会学研究的跨国传播现象和中国社会学与世界社会学的接轨程度(参见附录2图14)。

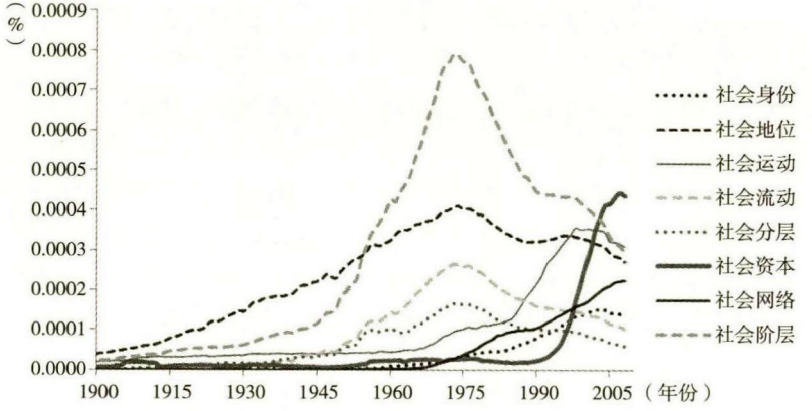

图6 社会学研究热点领域的词频比例历史曲线(1880-2008)

七、大数据中的社会学研究方法

本节我们探索社会学研究方法在书籍中的出现频次。对于定量方法,我们检索了最小二乘法(OLS),[①]对数线性模型(logit)、概率比模型(probit)、主成分分析(principal component analysis)、结构方程(structural

① 我们采取无大小写区别检索。OLS检索结果中,全大写占超过98%的绝大多数。我们没有使用因素分析(factor analysis),因为考虑到这一词汇检索可能带来较多的非定量研究方法内容。

equation)、社会网分析(social network analysis)、固定效应模型(fixed effects)、随机效应模型(random effects)、工具变量(instrumental variable)、事件史分析(event history)、倾向性匹配(propensity score)。[①] 对于定性方法,我们检索了深度访谈(depth interview)、焦点组访谈(focus group interview)、会话分析(conversation analysis)、内容分析(content analysis)、叙述分析(narrative analysis)和扎根理论(grounded theory)。在图7中,左侧为定量方法,右侧为定性方法。为便于两图比较,我们在定性方法坐标中也加入OLS词频(灰色粗线条)。

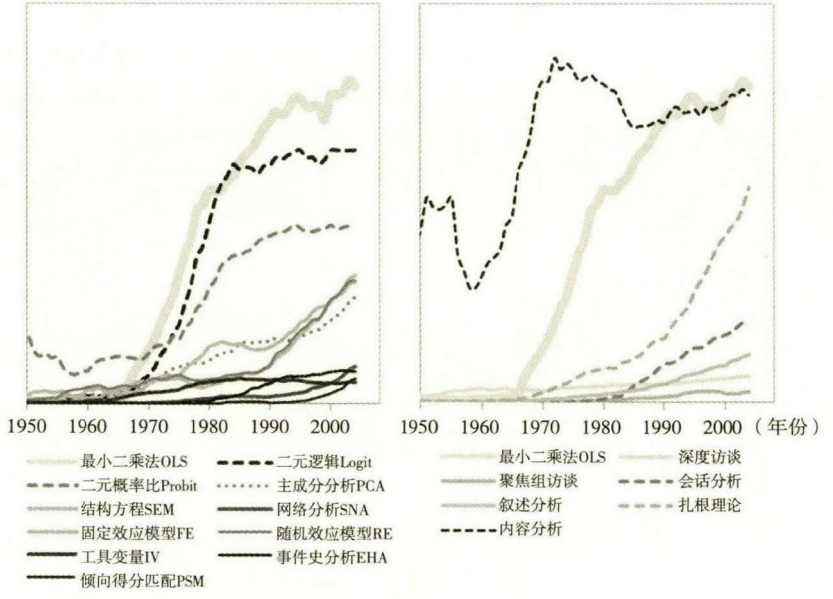

图7　定量与定性研究方法词频历史曲线(1950－2008)

从图7中我们发现了几个特点:第一,无论是定性还是定量,词频曲线几乎全部在持续增长。第二,定量方法和定性方法在语料库的词频比例存在差别。用OLS的词频为基准进行比较,我们就会发现除了

① 注意,英文检索关键词中我们尽量避免使用模型(model)、方法(method)这样的词汇,而把搜索关键词压缩到最核心部分。这样可以尽可能精确地检索到和核心词有关的方法应用。比如,涉及到随机效应方法的语句里很可能是random effects model而不是method。

历史悠久的内容分析方法，总体上其他定性方法词频都不高。[①] 第三，定性方法中扎根理论的扩张速度十分可观，超过了其他定性甚至不少定量方法。其他定性方法发展势头平缓或走向式微。第四，定量方法中，影响力最大的就是使用率最高的 OLS，Logit 和 Probit 模型。其余依次是固定和随机效应模型、结构方程和主成分分析法。其他方法的影响力则和一般的定性方法相差无几。第五，社会网分析和倾向性匹配在 2000 年左右异军突起。

值得注意的是，由于数据限制，社会学定量研究中使用固定或随机效应模型的还比较少，而主成分分析法、结构方程和社会网只能适用于特定研究主题。我们也测试了多层模型（multi level model）、潜类分析（latent class analysis）和赫克曼方法（Heckman selection）等其他关键词，但词频比例都比较小。因此，我们认为倾向性匹配、工具变量、多层模型和赫克曼方法是社会学定量分析中最富有潜力的方法群。

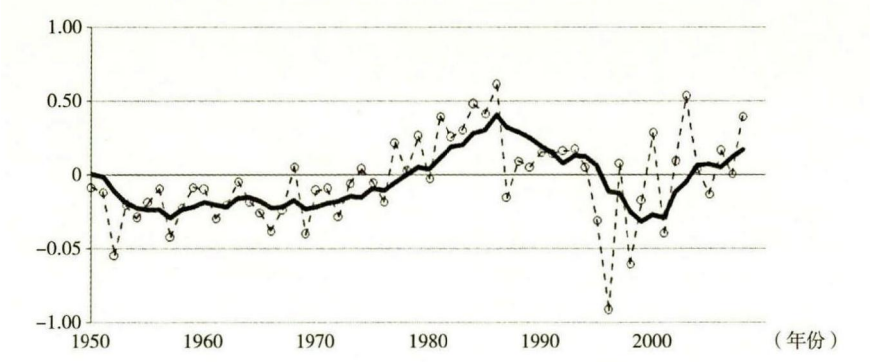

图 8　定量分析指数曲线（1950－2008）

半个世纪以来，定量方法和定性方法在总体上究竟哪个占优？趋势是否还有变化？为回答这个问题，我们采取去量纲方法来计算“定量分析指数”。首先，对于定量方法组，在每个时间点上，我们对 11 种具体方法的词频求组均值；对于定性组，我们对前述 6 种具体方法也进行同样操作。然后，我们分别对两组均值进行 1950 到 2008 年的标准化，获得各自 Z 值。在每个时间点上，我们用定量组的 Z 值减去定性

① 当然，这并不说明社会学领域的定量研究比定性研究普及。这是因为，我们的检索并非局限在社会学书籍。

组 Z 值,获得差值就是"定量分析指数"。我们将该指数的历年散点以及平滑趋势曲线绘制在图 8 中。总体上,两类方法呈交替主导的状态。从 50 年代到 80 年代,定性方法占据优势。但 80 年代到 90 年代定量方法的使用超过了定性。不过很快在 95 年左右被定性超越。到 2000 年左右,定量方法再次超越定性。不过,值得注意的是,由于从事定性研究的学者有可能更偏重于著书。因此图 8 中的定量指数仍然是比较保守的。

八、大数据中的中国社会学

一般我们认为中国社会学的诞生标志是严复翻译《群学肄言》或更早的社会学著作,这一时间点在 1894 - 1897 年左右(李培林,2000)。而我们的检索结果表明,英语世界里第一次规模性提及"中国社会学"(Chinese sociology)早在 1854 年;第一次规模性提及"中国社会学家"(Chinese sociologist)是在 1927 年;第一次规模性提及"中国的社会学家们"(Chinese sociologists)是在 1928 年。尽管由于语料库和谷歌图书网页并未同步,因此我们尚不清楚检索到的具体书籍以及相关短语基于上下文的准确含义,但该发现有一定可能使中国社会学发轫时间的历史锚定提前。

我们接下来观察一下 20 世纪中国社会学在全球社会学舞台中的位置,主要比较对象是北美的加拿大,[①]欧洲的英、法、德和亚洲的印度与日本。从图 9 中我们看到,欧洲的总体座次依次为德国、英国和法国。加拿大和英国比较接近。但出乎意料的是,"印度社会学"的词频统计在 70 年代后甚至超过了欧洲诸国。这可能要归因于印度庞大的人口和英语母语。70 年代末起,中国社会学的词频开始快速增长并超越日本,目前已和法国、加拿大与英国持平且仍在强劲攀升。

我们还在英语语料库中检索了费孝通、林南、边燕杰、谢宇、陆学艺、李培林等学者的名字。考虑到中英文姓名顺序差异,我们对每名学者均组合搜索"名 + 姓"和"姓 + 名"。同时,为与西方社会学家进行比较,我们同时也检索了提出"弱关系"与"结构洞"理论的两位美国当代著

① 美国社会学曲线在绝大多数时段远远超过其他国家,故图中未加绘制以方便阅读。

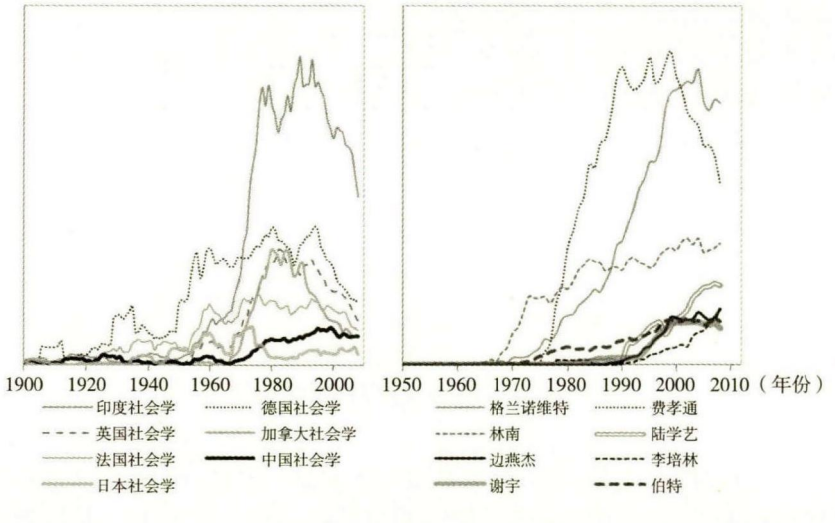

图9 中国社会学与社会学家词频曲线(1900－2008)

名社会学家:马克·格兰诺维特(Mark Granovetter)和罗纳德·伯特(Ronald Burt)。从图9中我们大致能推测出如下几个关联和特点。

第一,国际知名度变化和国家政治经济因素有关。能和格兰诺维特的词频统计相当的是费孝通。但我们随机查阅了费孝通名字出现的资料,发现有约四分之一的内容是因为费老担任的国家领导人职务。此外,费老的词频高峰出现得非常晚。相比之下林南早在70年代词频统计就开始增长。[①] 这表明,改革开放之后,随着国力的增强和社会学的重建,大陆社会学家才得以享有国际学术界的知名度。[②]

第二,国际知名度不完全等同于西方学术评价标准。例如,陆学艺的词频统计大大超过了谢宇、边燕杰等曾多次在英文权威期刊发表重要论文和出版英文专著的学者。这个排序恰恰表明:基于书籍大数据的词频数据比单纯学术评价指标更能反映文化影响力、知名度,陆学艺提出的十大阶层,虽未辅以复杂的数据和模型,但深切现实的观点、敏锐的洞察力和理论构建的勇气,就已奠定了他作为中国当代最重要的

① "林南"为两字拼音更易混淆,为此我们专门对照了相关检索结果的数字化书籍,发现主要混淆对象是他人名字(如Lin Nan LEE),但总体数量较少。为此,我们将林南70年代之后的词频减去其此前的峰值(1962年)词频以确保检索结果不会被放大。

② 我们对费孝通在上世纪早期使用的英文名字(Siaotung Fei)也进行了检索,但词频不高。

社会学家之一的历史地位和国际影响力基础。

第三，华人学者的国际影响力不断上升。在70年代，只有台湾学者林南的词频统计比较高。70年代末期，随着改革开放，老一代学者费孝通迅速取得了较高的词频比例，而新一代学者里李培林在80年代就已取得了一席之地。90年代之后，以李培林等人为代表的本土学者和以谢宇、边燕杰等人为代表的海外华人学者的词频比例大幅度上升。实际上，他们的词频均已接近或超过美国著名学者、结构洞理论提出者伯特。①

九、展望“社会组学”

尽管词频统计本身在文本分析领域并非新鲜事物，但基于空前规模的大数据进行时间空间大跨度的观察与分析，无论对习惯于抽样调查、回归分析的定量研究者，还是对习惯于纯理论推演、有限文本分析和深度访谈的定性研究者，都是一种新颖而有力的工具。在本文中，我们使用了全球规模最大也最为成熟的书籍语料库，对19世纪中叶到2008年的百万书籍进行了一系列社会学关键词的检索分析，从学科、名家、理论、领域等几个方面，初步梳理出了这百年来社会学发展的吉光片羽以及中国社会学不断崛起的良好势头。为让研究过程和诠释更具有科学性和说服力，我们还同时利用百年来的数字化报纸语料库进行对比分析。本文虽仅是大数据方法在学科发展史领域的一次探索性尝试，但无论是数据还是检索方法，都可以在更实质性的人文社会科学分析领域中使用。

不过，我们的研究仍然存在诸多不足：第一，文化影响力本身是一个综合性指标，用词频比例来代表文化影响力虽是一种可行的操作化方法，但不一定是最准确的测度；第二，限于篇幅和时间，我们仅对英语和汉语语料库进行了检索，而对德语、法语等子库则没有兼顾；第三，由于谷歌图书语料库提供的检索功能相对有限，尽管我们采取诸多技术性手段（如筛选核心词、设定频数阈值、进行年代段随机抽检、对照谷歌搜索引擎数据）来对检索过程进行科学控制，但检索精度仍待提高；第四，限于时间我们只能对社会学领域较具代表性的名家、理论和领域

① 限于篇幅我们没有全部展示检索的学者名称，但能够出现在图表中，也即被“规模性”提及的华人学者名字已经不少，这里未能一一列举。

进行梳理,因此分析对象、检索条目可能挂一漏万;第五,对社会学关键词的文化影响力变迁的解读和诠释,往往是基于时间曲线所启发的直觉,未能有更多的理论或实证证据来进一步阐发和验证观点。

严格意义上说,本文是基于大数据的内容分析(Lewis et al.,2013),采用了语义学(semantics)中的词频分析方法。① 本文的研究目的并非为社会学理论、名家、领域和方法进行影响力、知名度的排名,也不是要用一篇短文概括社会学的百年发展:这本身就是不可能的任务。从方法的角度,我们的研究试图表明,在现阶段利用现有的大数据,通过词频统计来对社会科学发展史进行分析,是一种相对可行、又能带来新发现的研究路径。随着大数据的进一步丰富、相关分析工具的进一步优化,实现更复杂更有价值的大数据研究将会摆上议事日程。比如,除了学科发展史,我们还可以进行社会思想史、政治思想史、文化观念史以及社会学理论发展史、政治学思想发展史等更偏重文化、学术语言学等方面的研究。再如,把词频的时间序列与反映经济社会发展指标的时间序列进行相关分析、格兰杰因果检验等,有助于我们发现文化现象和经济社会现象之间的关联。甚至,我们还可以把"原分析"(meta-analysis)的深度与广度提升到空前的程度:利用未来的学术期刊大数据,我们可以对海量的定量分析进行关键统计量提取,实现超大规模、超长跨度的"超级原分析"。

回到最初让-巴蒂斯特·米歇尔提出的文化组学(culturomics)。这个词,实际是"文化"(culture)和"基因组学"(genomics)二词的合并。其意义在于,单个的词汇 n-gram 就好比人类的基因,通过它们的排列组合,决定功能异常复杂的人类机体。如果我们把文化组学理解为一个最新的泛研究门类,那么,社会科学领域的"基因组学"也应该呼之欲出了。对于这个新的子学科,我们不妨称之为"社会组学"(societalimics)。它之所以有建立和研究的价值,是因为社会科学工作者以阅读文献的方式只能接触社会科学知识总体中非常有限的一部分。作为人的内在的学习能力瓶颈,这种不可避免的以管窥豹,会阻碍我们对宏观层面社会科学思想发展趋势的理解,不利于我们发现大尺度、大结构上的社会科学、社会思想发展规律。而通过词汇的"基因"序列分析,基于越来越完善、开放和准确的大数据,我们有可能获得过

① 感谢匿名审稿人对于本文研究性质和文献等方面资料的建议。

去完全不可能获得的理论启发和学科知识。因此，我们呼唤学界重视并早日建成“社会组学”。

附录1：基于数字化报纸的检索结果

尽管本文使用的谷歌图书语料库规模极大，为确保大数据分析结果能够稳健地反映社会学关键词的文化影响力，我们在辅助分析中用《纽约时报》(*New York Times*，简称 NYT)的语料库来进行对照分析。利用该报的全文数据库(自 1851 年创刊至今)，我们可以得到社会学关键词出现在某一年份《纽约时报》文章中的次数。考虑到每年文章数量的波动，我们将其除以当年《纽约时报》刊发文章总数，从而获得基于该报纸的相关词频比例。其计算公式为：

$$R'_{it} = \frac{C'_{it}}{C'_{t}}$$

其中，R'_{it} 为关键词 i 在公元 t 年《纽约时报》中出现的词频比例，也即知名度，C'_{it} 表示在 t 年的《纽约时报》中提及 i 的文章数量，C'_{t} 为整个 t 年中该报纸刊发的全部文章总数。

由于媒体更侧重新闻和评论，因此涉及社会学方法、理论等的关键词的出现次数会比较少，较难与基于谷歌图书的检索结果进行对比。但是，我们对社会科学学科(图 10)、社会学大师(图 11)、社会学研究热门领域(图 12)等最为核心的关键词的分析表明，尽管基于媒体的曲线局部坡度、形状等等与基于书籍的并不完全相同，但关键词之间的总体高低层次、时段变化和趋势等都呈现出高度的统一。

更重要的是，即便有一些差异，也往往可以通过书籍和媒体在内容导向方面的区别得到解释。例如，“经济学”在 1980 年以来的《纽约时报》中出现频率超过了“哲学”，这显然和 20 世纪晚期新闻媒体的内容取向有关。而 60 年代末和 70 年代马尔库塞在《纽约时报》具有极高的词频，原因主要在于当时《纽约时报》的报导倾向和美国媒体对他个人的大量宣传(郑春生，2009，2012)。总体上，基于《纽约时报》的辅助分析结果表明基于谷歌图书语料库的词频可以客观、可靠地反映社会学关键词的文化影响力。

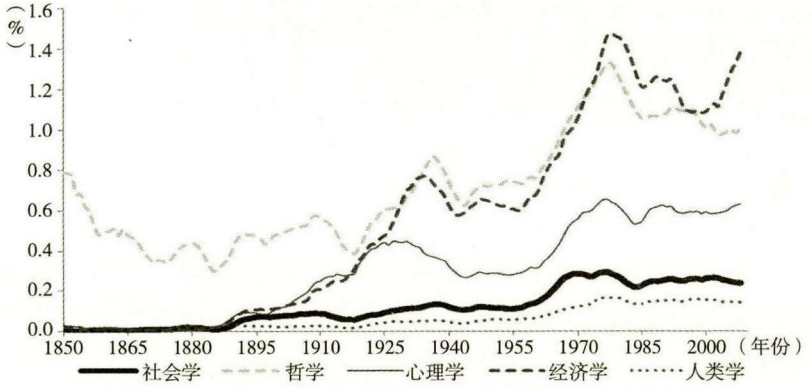

图 10 英文学科名称的词频比例曲线(NYT 1851 –2008)

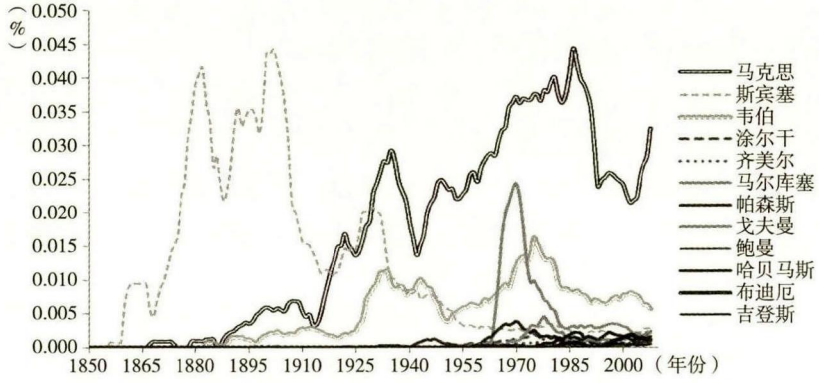

图 11 百年社会学大师的词频比例历史曲线(NYT 1851 –2008)

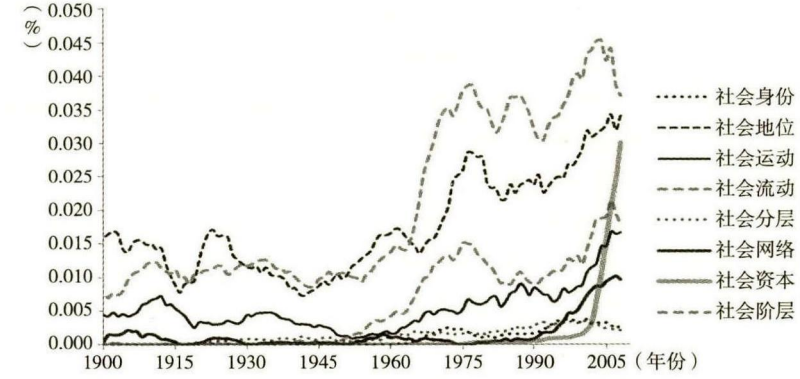

图 12 社会学研究热点领域的词频比例历史曲线(NYT 1900 –2008)

附录2　基于中文书籍的相关搜索

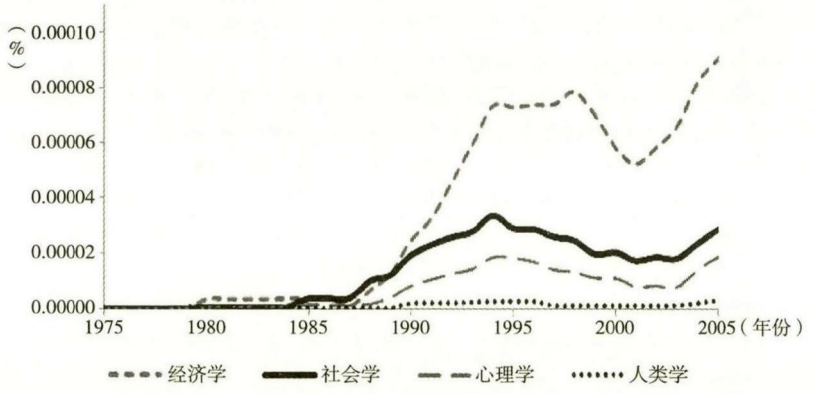

图 13　中文学科名称的词频比例历史曲线(1975－2008)

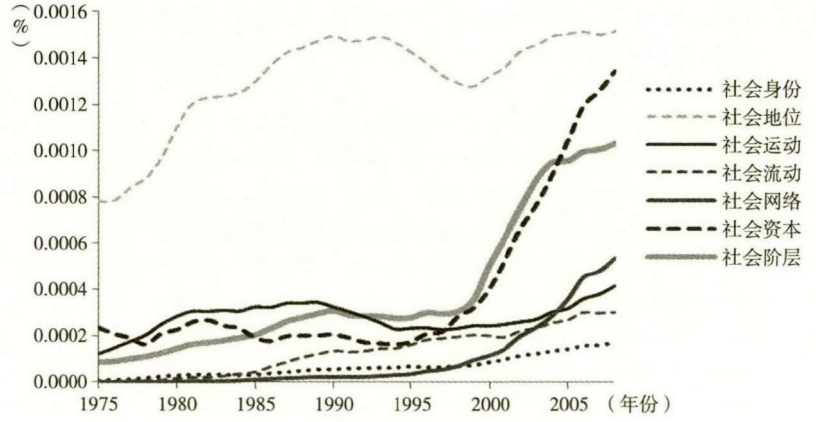

图 14　中文书籍社会学研究热点领域的词频比例历史曲线(1975－2008)

参考文献：

贾春增,2008,《外国社会学史(第三版)》,北京:中国人民大学出版社。

李培林,2000,《中国早期现代化:社会学思想与方法的导入》,《社会学研究》第 1 期。

沈浩、黄晓兰,2013,《大数据助力社会科学研究:挑战与创新》,《现代传播》第 8 期。

翁乃群,2000,《美、英社会文化人类学研究的时空变迁》,《民族研究》第 1 期。

谢立中,2007,《西方社会学名著提要》),南昌:江西人民出版社。

郑春生,2009,《试论 20 世纪 60 年代美国媒体对马尔库塞的形象塑造》,《世界历史》第 5 期。

——,2012,《试论马尔库塞思想的传播》,《宁夏社会科学》第1期。

Acerbi, A., V. Lampos, P. Garnett & R. A. Bentley 2013, "The Expression of Emotions in 20th Century Books." *PLoS ONE* 8(3).

Bentley, Alexander, Alberto Acerbi, Paul Ormerod & Vasileios Lampos 2014, "Books Average Previous Decade of Economic Misery." *PLoS ONE* 9(1).

Boyd, Danah & Kate Crawford 2012, "Critical Questions For Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication and Society* 15(5).

Giddens, Anthony & Philip Sutton 2013, *Sociology*. Cambridge: Polity Press.

King, Gary 2009, "The Changing Evidence Base of Social Science Research." In G. King, K. Lehman Schlozman & N. Nie (eds.), *The Future of Political Science: 100 Perspectives*. New York: Routledge.

Lewis, Seth C., Rodrigo Zamith & Alfred Hermida 2013, "Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods." *Journal of Broadcasting and Electronic Media* 57 (1).

Lin, Yuri, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, William Brockman & Slav Petrov 2012, "Syntactic Annotations for the Google Books Ngram Corpus." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics:2 (ACL 12)

Maclean, Vicky M. & Joyce E. Williams 2012, "Ghosts of Sociologies Past: Settlement Sociology in the Progressive Era at the Chicago School of Civics and Philanthropy." *The American Sociologist* 43(3).

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak & Erez Lieberman Aiden 2011, "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331.

Morgan, J. Graham 1969, "The Development of Sociology and the Social Gospel in America." *Socilolgical Analysis* 30(1).

Petersen A. M., Tenenbaum J., Havlin S., Stanley H. E. & M. Perc 2012, "Languages Cool as They Expand: Allometric scaling and the decreasing need for new words." *Scientific Report* 2.

Scott, John & Gordon Marshall 2005, *Oxford Dictionary of Sociology*. Oxford: Oxford University Press.

Tufekci, Zeynep 2014, "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls." Proceedings of the 8th International AAAI Conference on Weblogs and Social Media.

Twenge, J., K. W. Campbell & B. Gentile 2012, "Increases in Individualistic Words and Phrases in American Books, 1960 – 2008." *PLoS ONE* 7(7).

作者单位:南京大学社会学系

责任编辑:闻 翔

SOCIOLOGICAL STUDIES 2015

(Bimonthly) 1

Vol. 30 January, 2015

PAPER

Abstract: This paper integrates CCP history, political history, educational history and social history of Republican China, and studies the organizational form of Jiangxi Communist Party and Communist Youth League before the Northern Expedition. The formation and development of the organization of early CCP was on the one hand embedded in the situation of Republican China's politics and education, on the other hand embedded in traditional social relations. First, modern schools controlled by KMT personages from aristocratic family provided the base of legitimacy for CCP's organizational development. Different types of secondary schools and normal schools had generated different organizational structures of CCP. Second, in the development of CCP's organization, the combination of local relations and scholastic relations depended on several factors, such as recruiting students in certain regions, and the leader of the CCP organization having plenty of social connections in the local area. There is close relevance between the coastal areas which had convenient transportation and the organizational network of early CCP. Finally, the formation of early CCP's organizational network drew support from the resources of the descendant of aristocratic family or rich family. The study of the origin of the organizational form of early CCP will contribute to the understanding of CCP's organizational transition in the Soviet Revolution period.

Abstract: Using the updated Google Books Corpus containing around 8-million books, this paper surveys trends in the usage of sociology-related keywords in English language books from the middle 19th century. It tracks the semantics of cultural influence of sociology over centuries, focusing on discipline terrains, important figures, major