社会科学方法前沿（一）

大数据分析

林景 讲师

南京财经大学法学院

定量社会学的危机

- 2004年欧洲社会研究中心的研究方法盛会（ESRC Social Science Festival）
  - Mike Savage
  - 社会网络研究方法（Social Network Methods）
    - 三个大型NGO组织的社会网络
    - 320名志愿者构成的网络关系
  - 英国电信集团公司（British Telecom Group plc.）研究部
    - 数十亿电信通讯交换构成的社会网络
    - 家庭、组织、企业、政府部门……
    - 涵盖社会关系与公共资源分布数据的社会空间地图

大数据从何而来?

- 量化自我（Quantified-Self）
  - Gary Wolf 和Kevin Kelly
  - 借助数字传感器，将日常生活中的状态和表现记录为数据，实现对自己的精控管理。

大数据从何而来?

- 数字足迹（digital footprint）
  - 人们使用互联网产品及其衍生品时产生的一系列数据。

社交网络

访问网页

预定外卖

共享经济

数字支付

在线购物

大数据从何而来?

- 政府部门、科研机构和社会组织等汇总的传统数据

经济统计

人口信息

基础设施

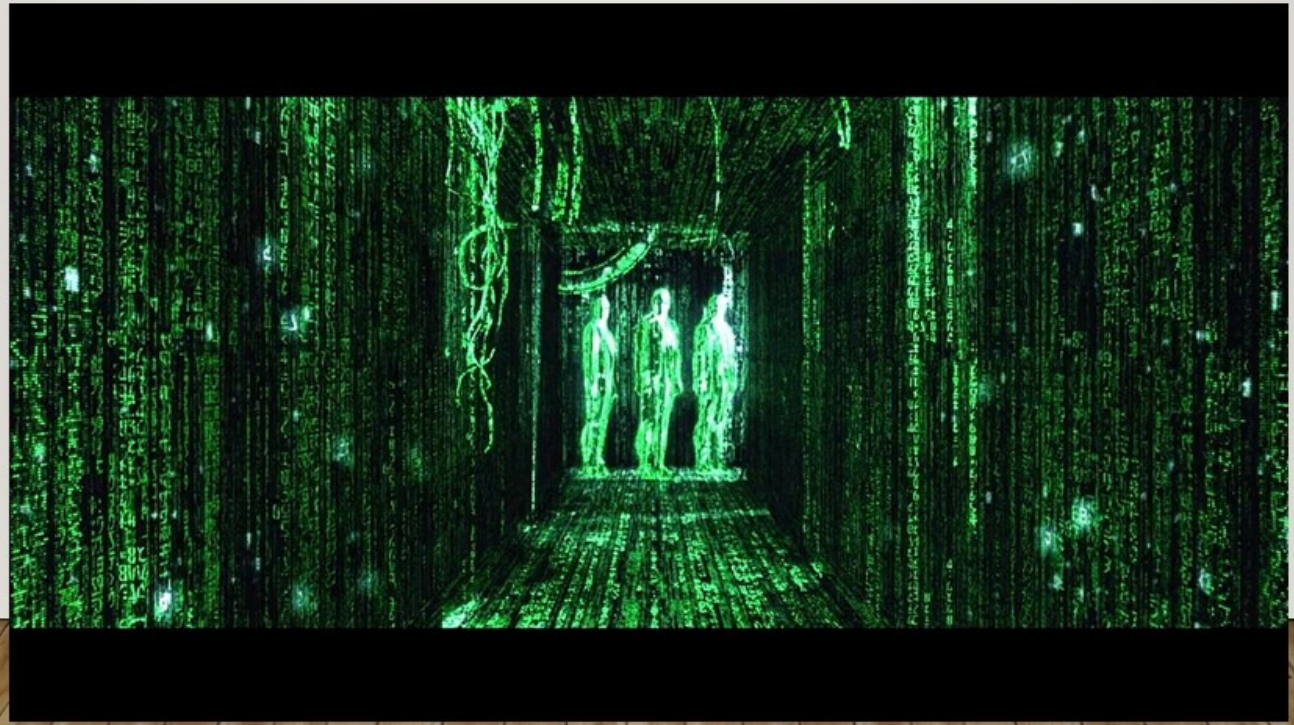
数字监控

印刷出版

历史档案

大数据从何而来?

- 大数据社会科学家眼中的世界



大数据的特征

- 大数据相比传统定量数据或定性资料的特征
  - 体量大（volume）
  - 增速快（velocity）
  - 类型多（variety）
  - 真实性（veracity）
  - 价值高（value）

大数据时代的社会科学

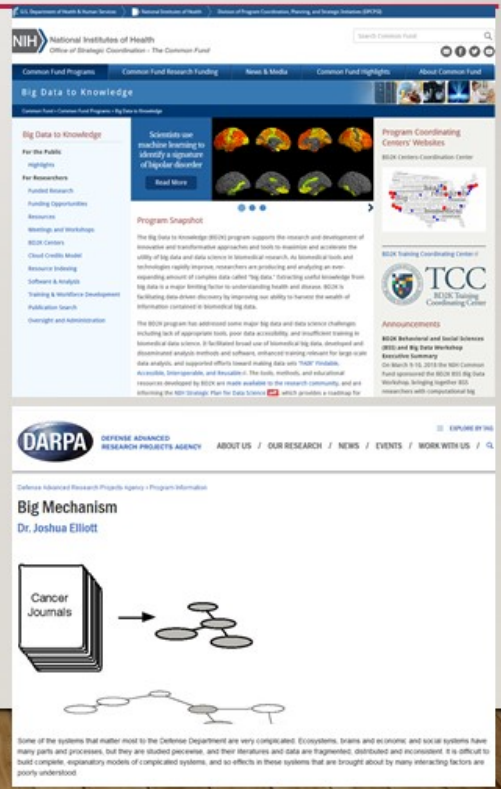
- 大数据为社会科学研究创造的机遇
  - 海量的非结构化数据/信息
  - “全量数据”而非“样本数据”
  - 丰富、高调的方法工具箱
  - 计算机辅助分析与预测技术
  - 更强的时效性
  - 社会科学知识的普及化
  - 社会经济效益

大数据时代的社会科学

- 大数据+社会科学 = ?
  - 全数据驱动的描述研究
  - 诊断或预测的应用研究
  - 机制式研究（微观与宏观）
  - 其他有意义的新关系、新模式或新趋势

大数据时代的社会科学

- 大数据 = 未来社会的石油?
  - 美国国立卫生研究院资助成立大数据中心
    - 使命：将大数据转变为知识（Big Data to Knowledge，BD2K）
  - 美国国防部先进研究项目局（Defense Advanced Research Project Agency）提出利用大数据的大机制计划（Big Mechanism Program）
    - 探寻大数据中的“为什么”；
    - 发展自动化技术，解释驱动复杂系统的因和果；
    - “大机制是科学的未来”。



大数据时代的社会科学

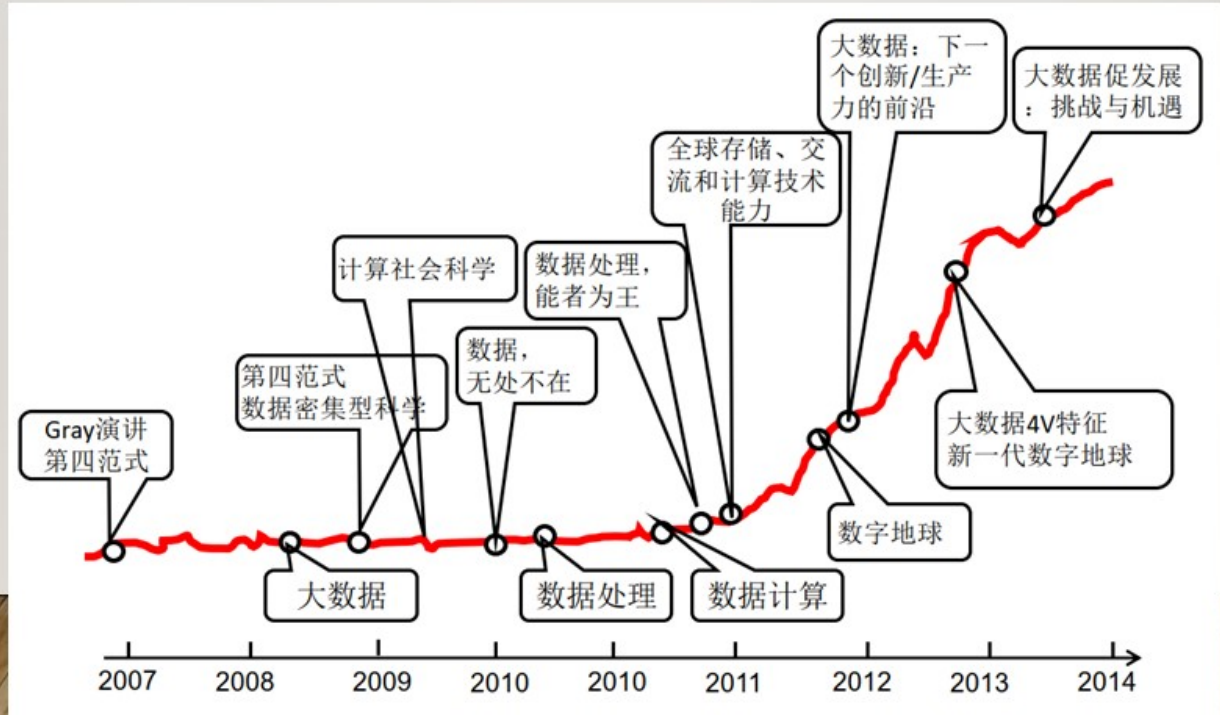
- 美国社会学会的数据马拉松（ASA Datathon）
  - 骇客马拉松（hackathons）的衍生文化；
  - 持续24小时的学术工作坊(24-hour academic workshop)，要求研究者尽最大努力，在尽可能短的时间内将信息转化为知识；
  - 数据马拉松倡导用“研究问题+数据集”来驱动知识进步。

大数据时代的社会科学

- 美国社会学会的数据马拉松（ASA Datathon）
  - 数据马拉松的目标是提交一份接近完整的分析项目；
  - 数据马拉松允许社会科学家检验新的研究方法，并在工作环境中匹配潜在的合作者；
  - 理想情况下，数据马拉松是充满数据和建设性批评的智力对抗赛，它大大压缩了可能需要好几个月才能厘清的分析头绪。

大数据时代的社会科学

- 大数据社会科学的发展



如何采集和整理大数据

- 大数据收集方法：海量收集+压缩降维
  - 网络爬虫
  - 对搜索引擎搜索记录的分析
  - 自动文本分析
  - 视频/图片分析
  - 社会网络分析
  - 地理空间分析

如何分析大数据

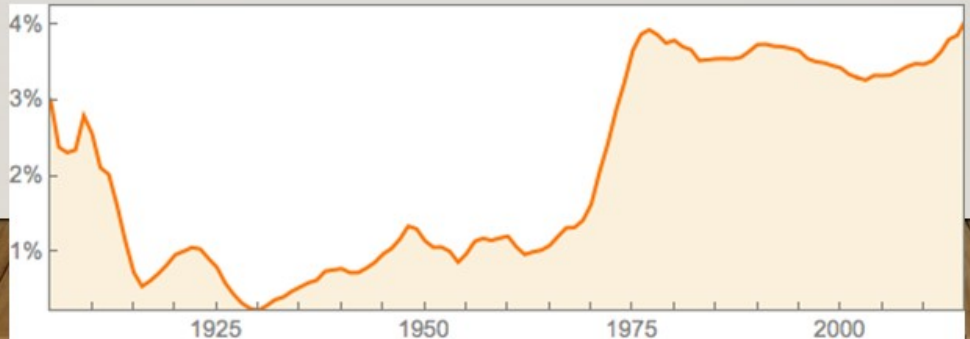
- 大数据分析的方法
  - 分类与聚类
  - 关联分析
  - 回归分析
  - 因果推论
  - 数据可视化

案例（一）：拉马努金与哈代的数学文献研究

- Stephen Wolfram
  - *Who Was Ramanujan?*
  - 哈代（G.H. Hardy），英国数学家
  - 拉马努金（Srinivasa Ramanujan），印度数学家

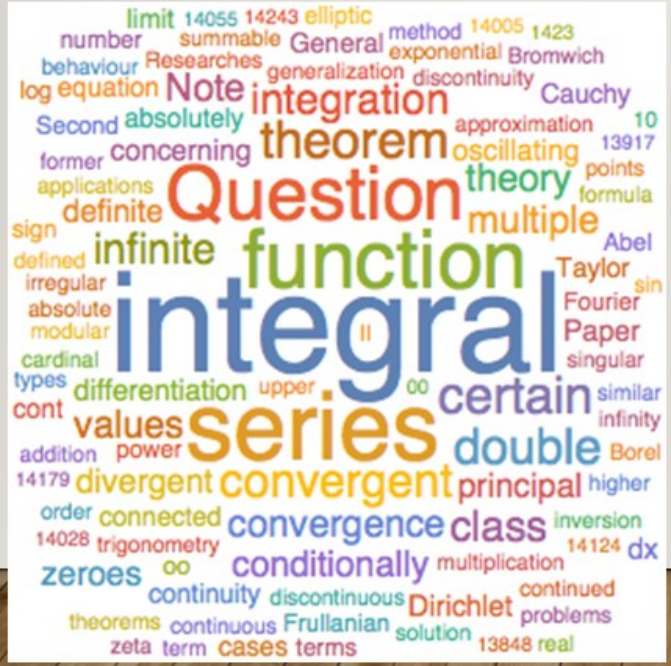
案例（一）：拉马努金与哈代的数学文献研究

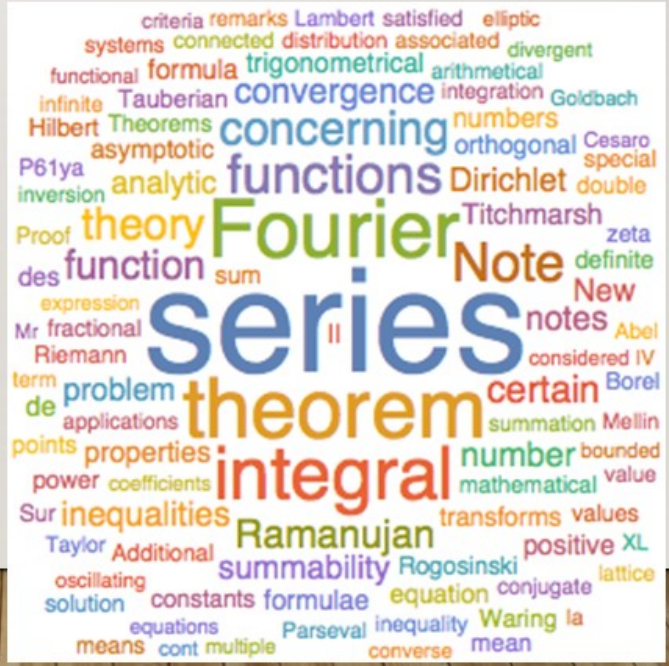
- 哈代7大卷论文集的数字化



案例（一）：拉马努金与哈代的数学文献研究

- 拉马努金去世前后，哈代的数学著作关键词词云（word cloud）



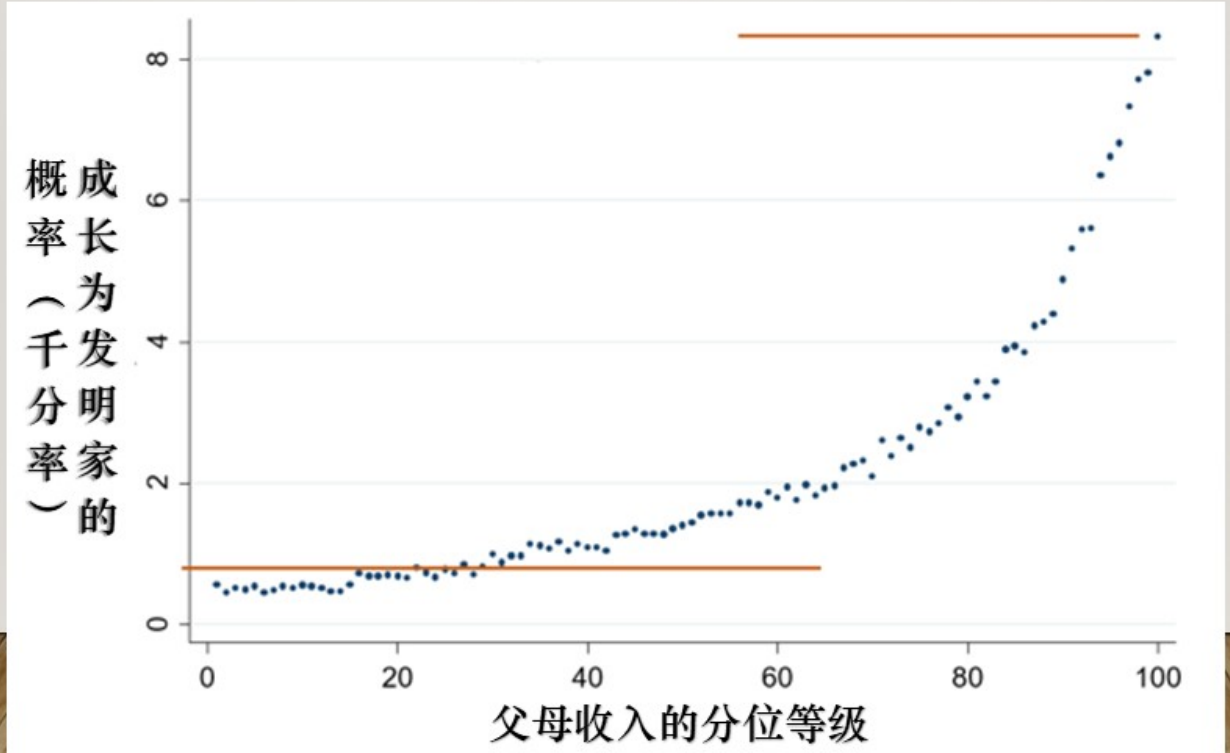


案例（二）：发明家的成长之路研究

- Alex Bell et al.
  - *Lost Einsteins: How Exposure to Innovation Influences Who Becomes an Inventor*
    - 主要数据：美国100多万发明家的成长和生活数据
    - 辅助数据：美国专利税收数据、学校学籍档案数据
  - 研究问题：
    - 社会经济阶层、种族和性别对儿童成长成为发明家的影响
    - 创新接触对成为对儿童成长成为发明家的影响
    - 地区财政激励与税率对儿童成长成为发明家的影响

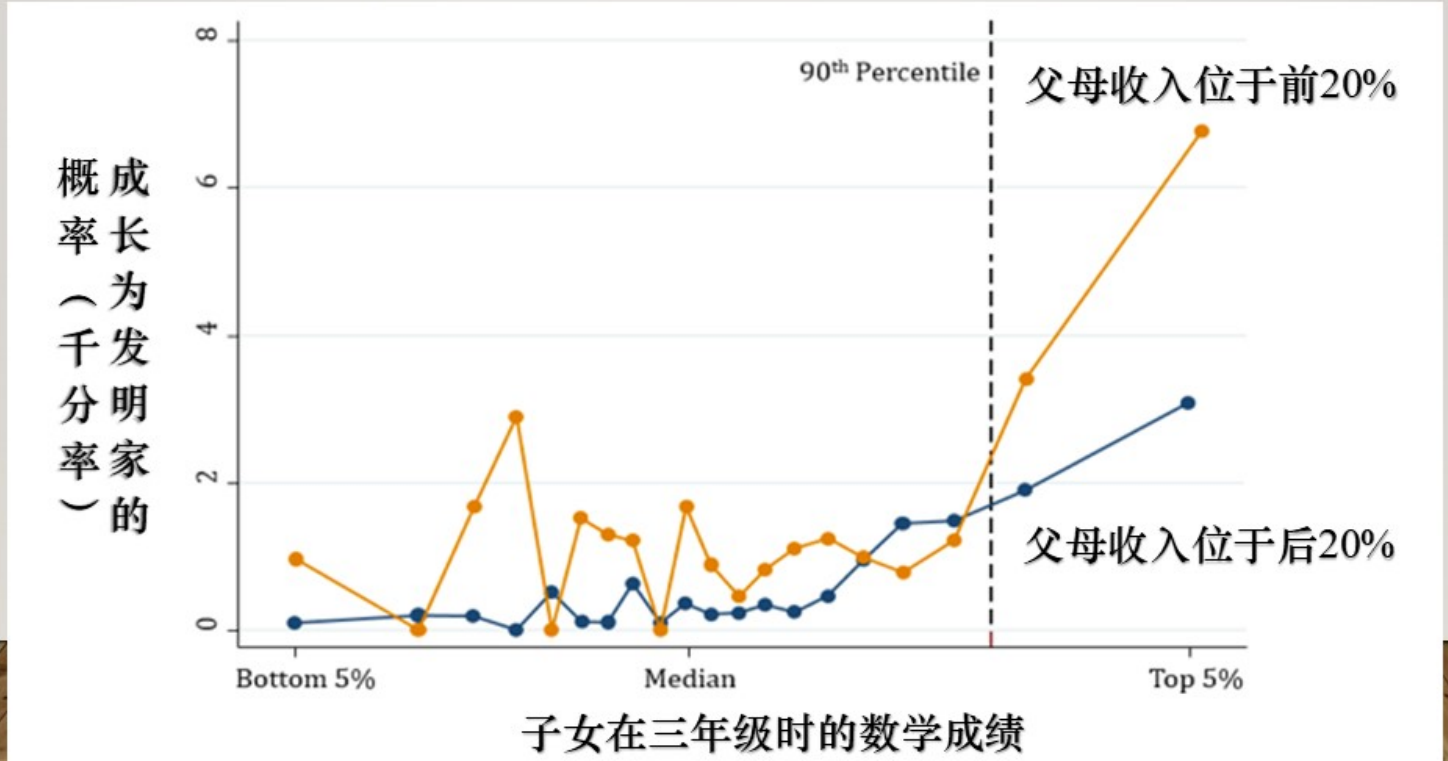
案例（二）：发明家的成长之路研究

- 父母收入对儿童成长成为发明家的影响



案例（二）：发明家的成长之路研究

- 父母收入对儿童成长成为发明家的影响



案例（二）：社会学发展史研究

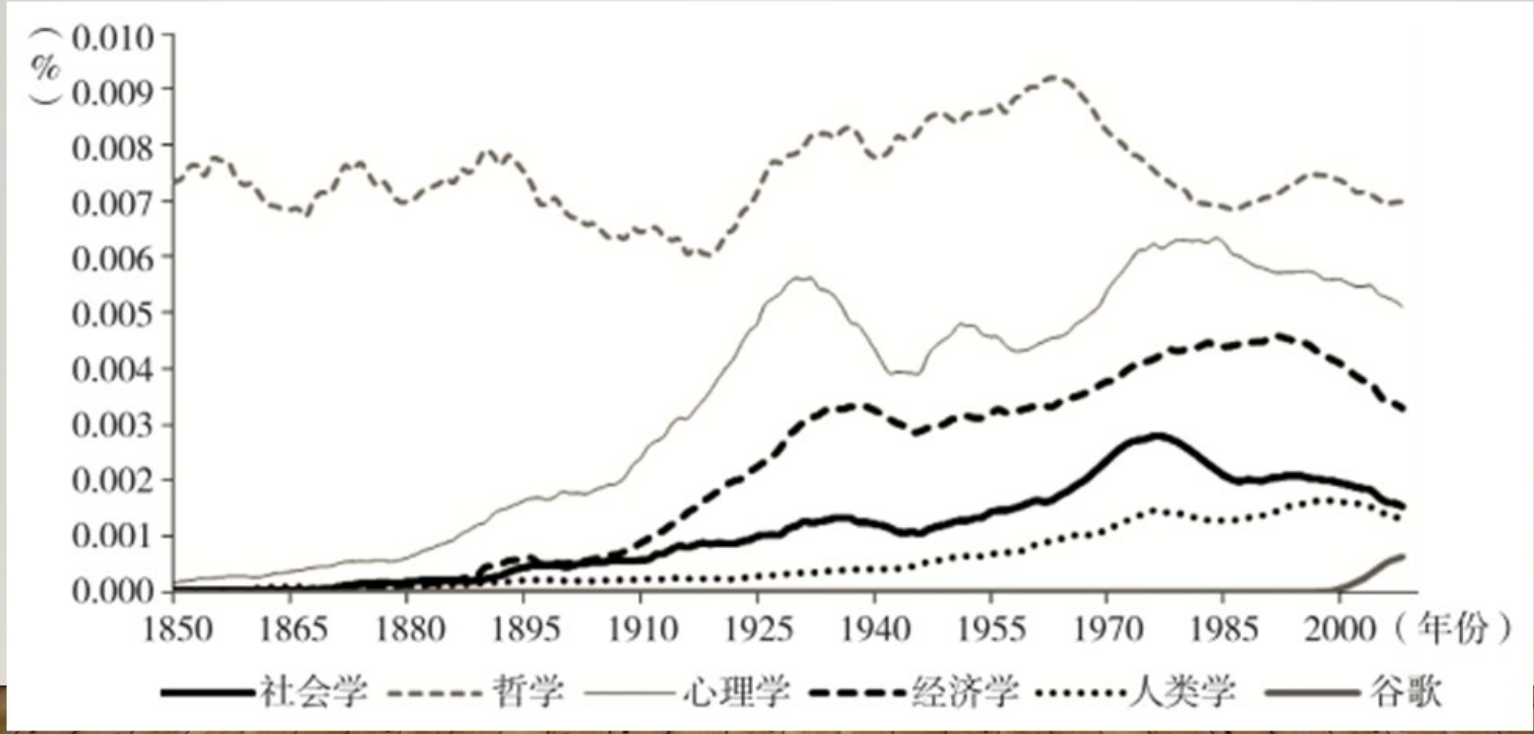
- 陈云松
  - 《大数据中的百年社会学：基于百万书籍的文化影响力研究》
  - 主要数据：谷歌图书语料库，合计811万种图书，8613亿词汇
  - 研究问题：
    - 社会学的学科轨迹、名家大师、理论流派、热点领域、分析方法
    - 中国社会学的百年变迁与文化组学展望

案例（二）：社会学发展史研究

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

案例（二）：社会学发展史研究



案例（二）：社会学发展史研究

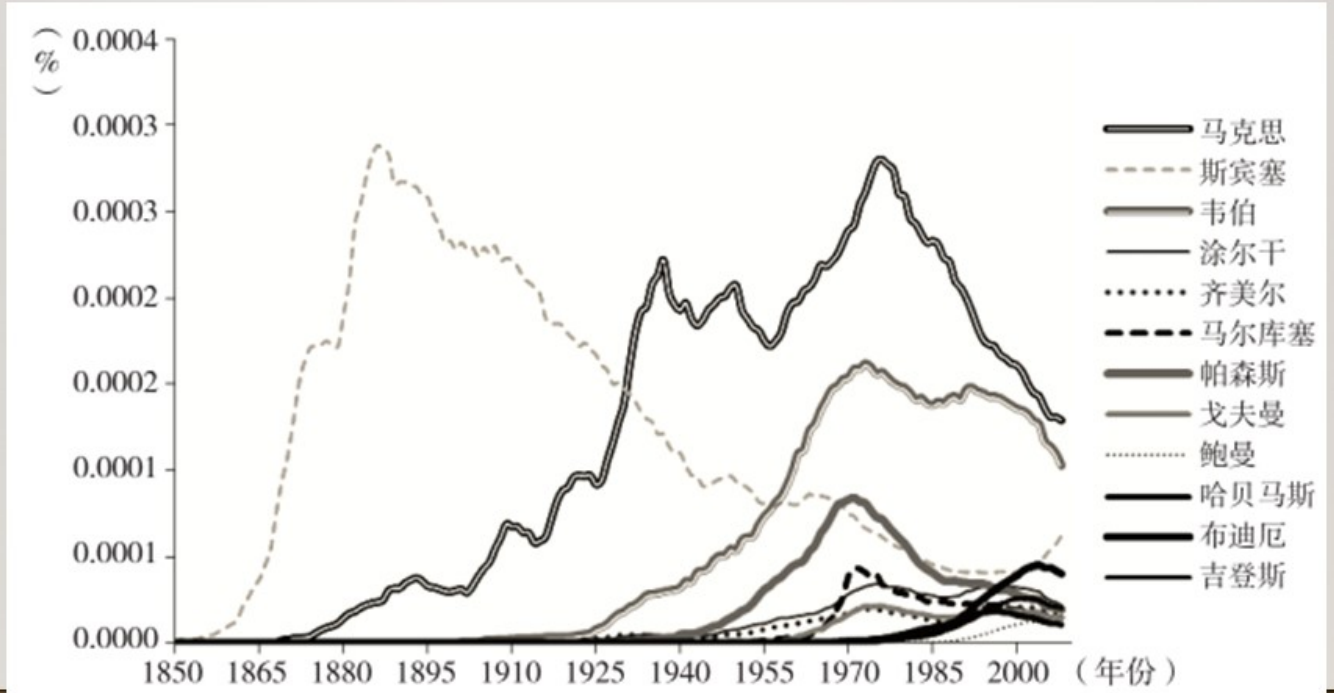
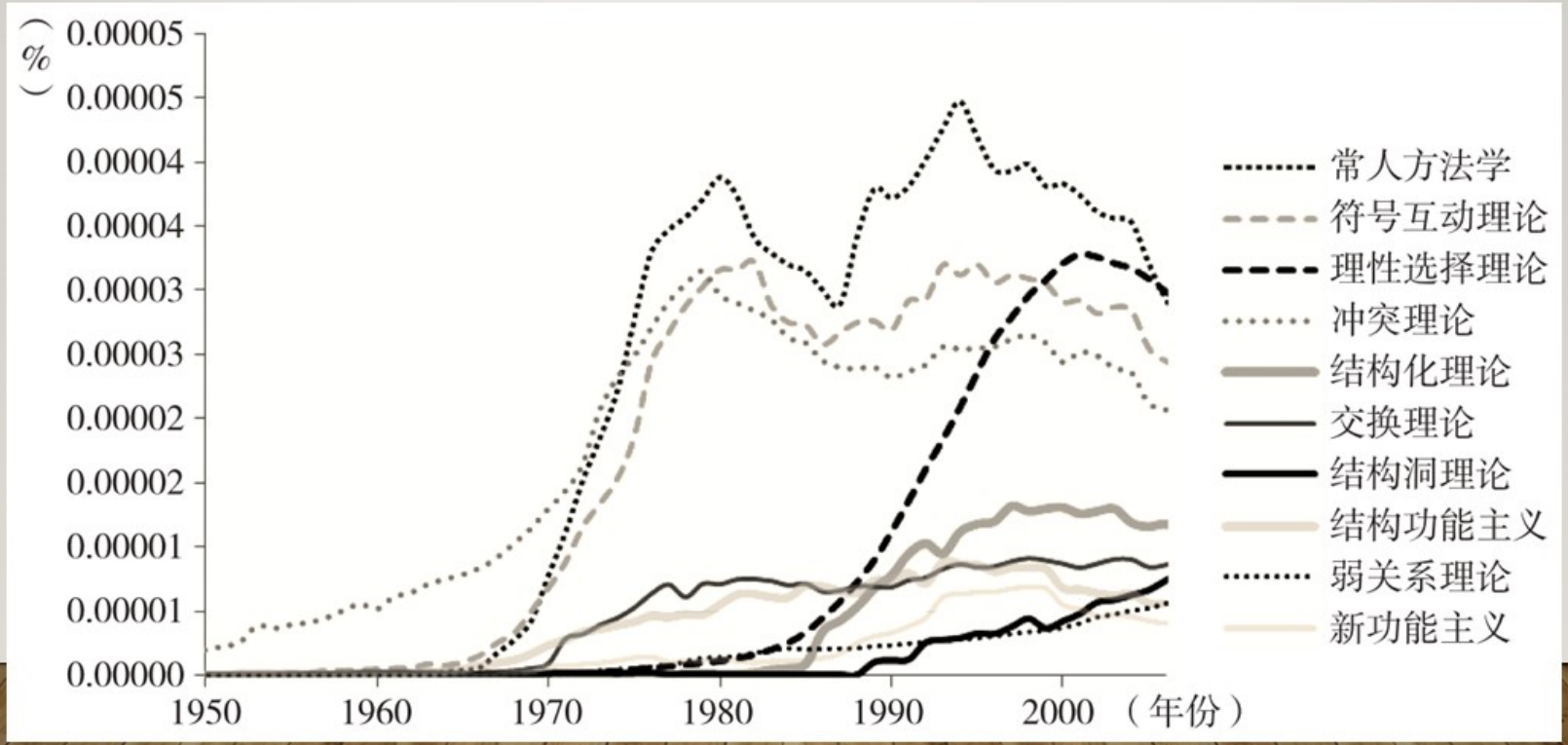
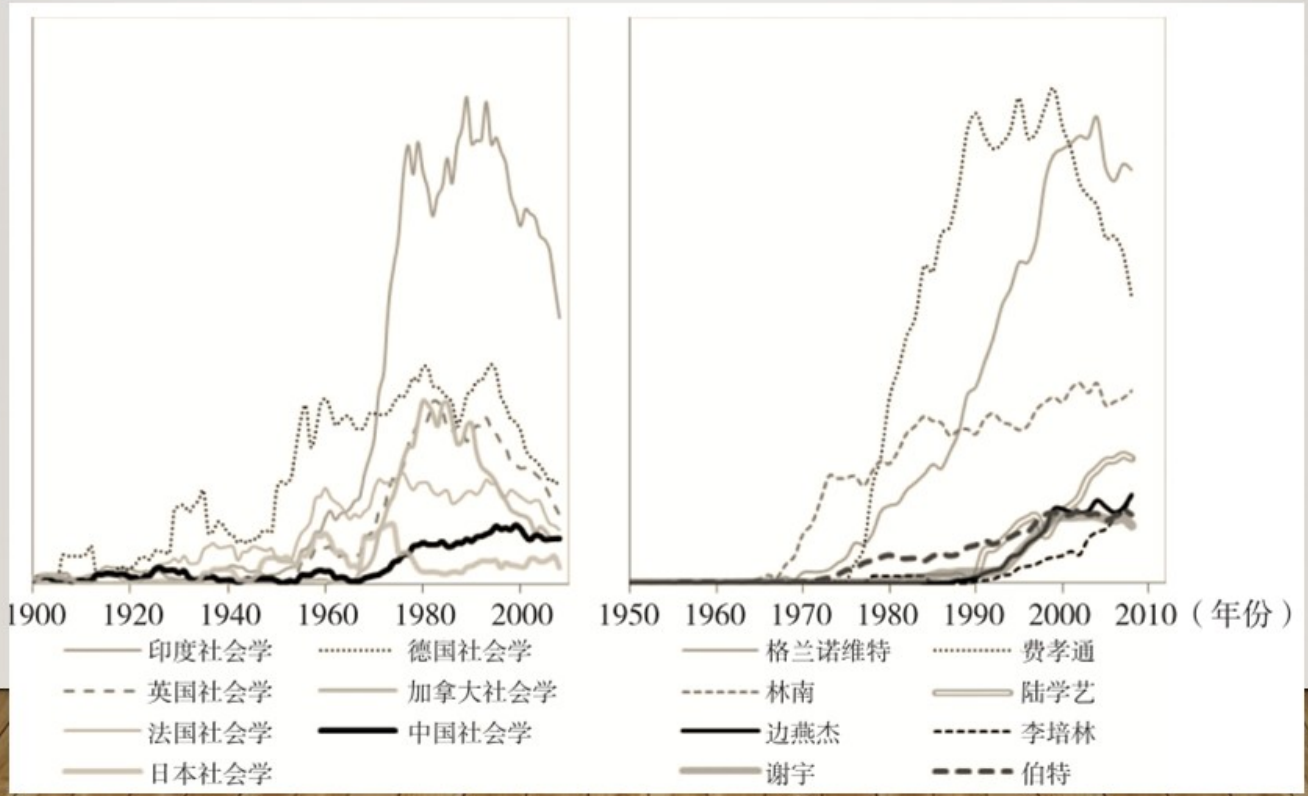


图 3　百年社会学大师的词频比例历史曲线(1850 –2008)

案例（二）：社会学发展史研究



案例（二）：社会学发展史研究



案例（三）：海上丝绸之路的文化影响研究

- 龚为纲
  - 《大数据视野下的19世纪“海上丝绸之路”：以丝绸、瓷器与茶叶的文化影响力为中心》
  - 主要数据：谷歌图书语料库
  - 研究问题：
    - 中国丝绸、瓷器、茶叶在西方发达国家的主要影响力演化；
    - 海上丝绸之路的主要贸易主体及其轨迹；
    - 海上丝绸之路的沿线国家与城市的互动格局。

案例（三）：海上丝绸之路的文化影响研究

图 1　茶叶、丝绸、瓷器在英国文化影响力的变动

案例（三）：海上丝绸之路的文化影响研究

图 2　茶叶、丝绸、瓷器在美国文化影响力的变动

案例（三）：海上丝绸之路的文化影响研究

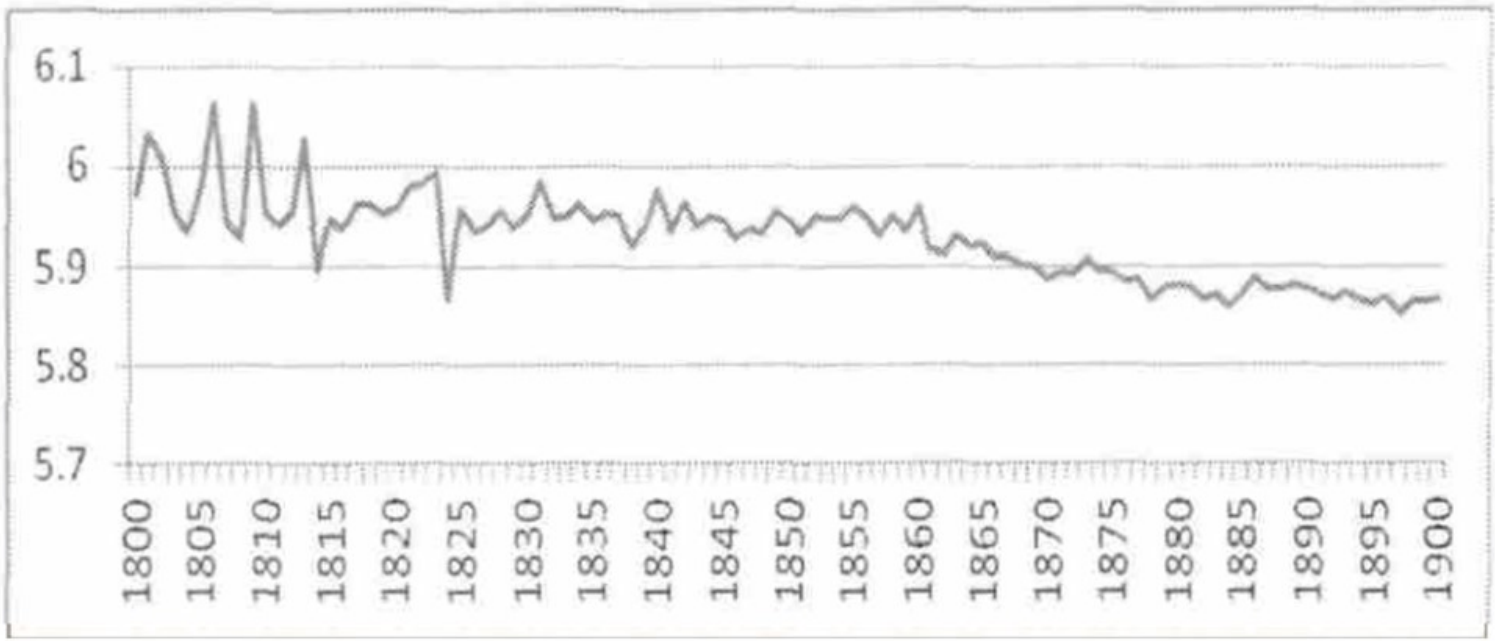


图 8　茶叶在英国情感指数(积极/消极)的变化

案例（三）：海上丝绸之路的文化影响研究

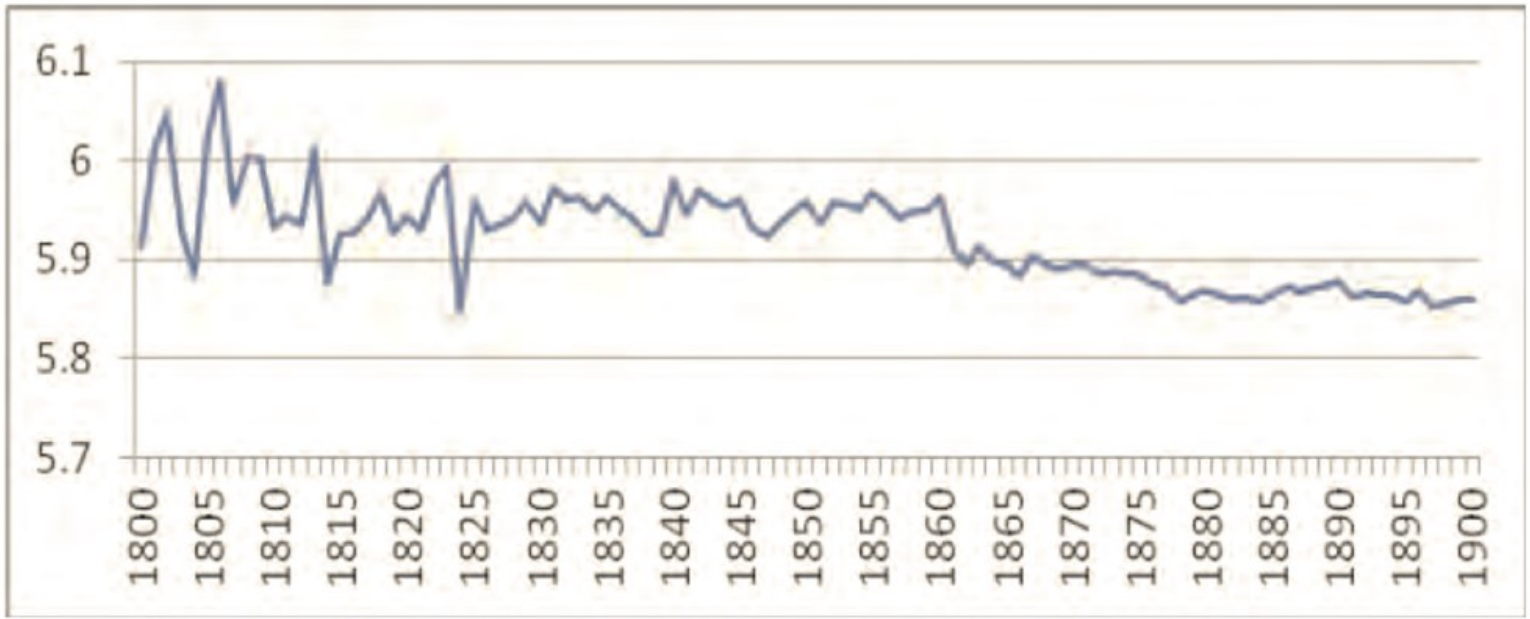


图 9　茶叶在美国情感指数(积极/消极)的变化

案例（三）：海上丝绸之路的文化影响研究

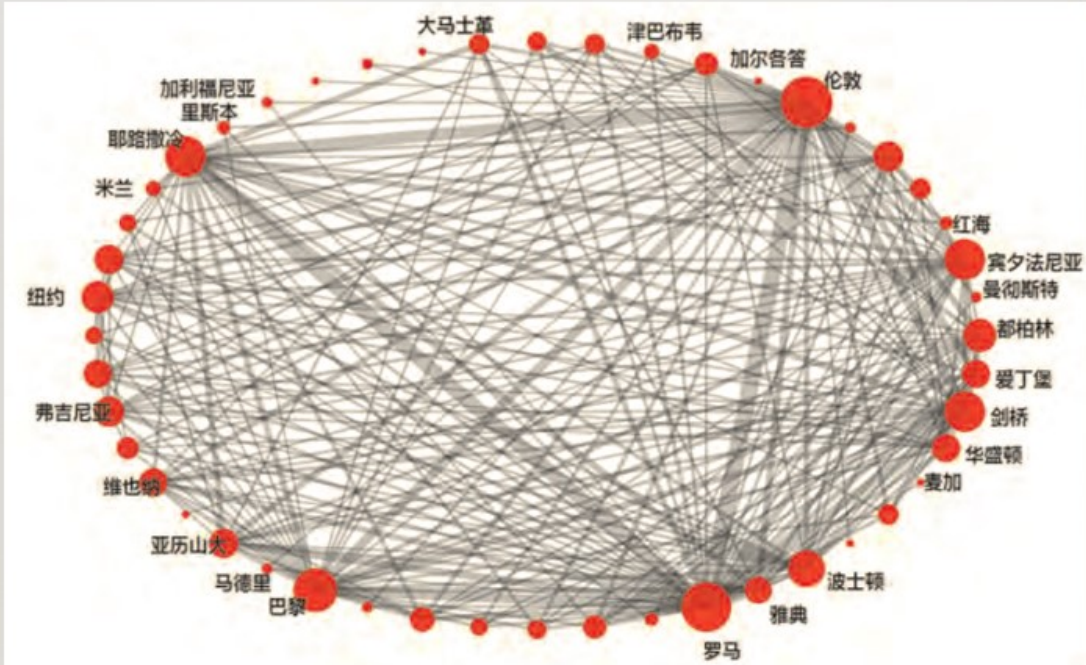


图 14　1800–1810 年间全球茶叶贸易中的互动网络

案例（三）：海上丝绸之路的文化影响研究

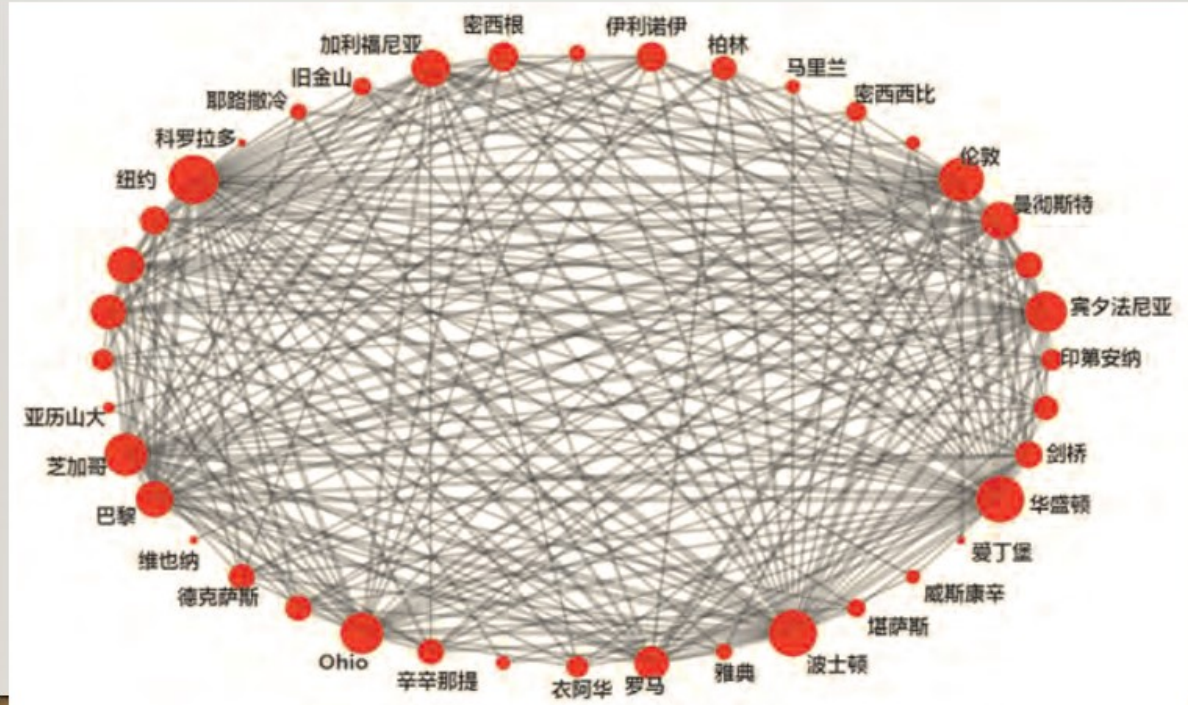


图 15　1890–1900 年间全球茶叶贸易中的互动网络

大数据社会科学的挑战

- 大数据社会科学家的流失
- 从社会学到硅谷
  - Duncan Watts, Columbia U. →
  - Mary Beth Hunzaker, New York U. → Research Scientist of Facebook
  - Zack W. Almquist, UMN → Research Scientist of Facebook
  - Alex Hanna, Toronto → Program Manager of Google
  - Thomas J. Leeper, LSE → Research Scientist of Facebook
  - ……

应对大数据社会科学的挑战

- 建立计算社会科学专业委员会（computational social science section）
  - 建立分支领域的专业学科组织
  - 筹划学科会议
  - 筹办学科期刊
  - 推动学科资讯
- 科际整合（interdisciplinarity）
  - 建立跨学科研究中心
  - 建立跨学科院系

应对大数据社会科学的挑战

- 减少非社会科学期刊/会议的发表障碍
  - 非社会科学的期刊/会议放宽投稿学科限定
  - 非社会科学的会议降低注册会费
- 学术雇佣机构的招聘与升等激励
  - 为跨学科专业学者开放教职
  - 对跨学科期刊的投稿论文有学术认定