科学学数据、问题和研究方法的新框架

吴金闪

北京师范大学系统科学学院

2018年9月20日

$$\frac{e^{\pm\beta H}}{Z}$$

等着你来提出新问题发展和实现新视角和新工具

1. 报告的目的
2. 网络科学的三层网络框架
3. 运用三层网络框架来表示数据、表述问题和发展算法
4. 几个研究工作的例子
    1. 国家（城市、学科、国家×学科）之间的知识创造依赖关系
    2. 论文的影响力，从论文引用本身来看（类PageRank）
    3. 论文的影响力，从专利（应用）引用来看
    4. 论文的影响力，从书籍（文明）引用来看
    5. 论文和专利统一起来看，部分数据的问题
    6. 概念层：概念本身的重要性，从论文获取概念频次
    7. 概念层：论文影响力，论文首次提出概念，从书籍中获取概念频次
    8. 作者层用于更好的作者识别、作者学校国家贡献计数和影响力分析
    9. 作者-论文（专利、书籍）-概念多层网络整体来看影响力，经济
    10. 直接间接影响视角来提问和计算：职业生涯大事、人才培养和流动
5. 带回家的消息：大概一致的问题数据研究方法的框架推动学科发展

报告的目的

- 介绍网络科学最核心的思想是什么
- 介绍科学学三层网络框架
- 通过例子来展示三层网络框架能够为科学学做什么
- 寻找合作者，也顺便推动网络科学和科学学两个领域的发展

直接联系

- 一个网络就是通过一条条的边相连的一群顶点
- 边表示联系、相互作用、影响、物质思想或者货币的传递
- 顶点表示主体
- 例如论文之间的引用关系
- 例如概念之间的逻辑关系
- 通常用矩阵（A）来表示，元素（A_{ij}）非零就表示（i, j）有联系
- 用网络来描述系统是对实际相互作用的一种简化

间接联系

- 如果你构建了一个网络，就是为了算度，那么，你没有真的用网络
- 平均最短距离、社团聚类、PageRank等就开始用到了间接联系
- 综合直接和间接联系：PageRank、投入产出分析等的核心思想，$(1-B)^{-1} = 1 + B + B^2 + \cdots$
- 看问题的视角，提问题，举例职业生涯大事、人才培养和流动

作者-文章-概念三层网络

- 三类顶点：作者、文章、概念
- 层内联系：作者学术传承、文章引用、概念逻辑
- 层间关系：作者写了文章、文章工作在概念上

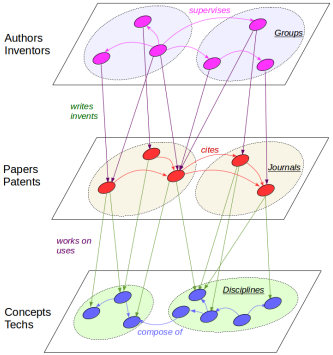

Figure: 科学计量学三层网络模型

作者-文章-概念三层网络，续

- 合作关系不是直接关系，而是作者→文章→作者
- 共施引和共被引，以及很多种共现，也不是直接关系
- **将来所有的这些非直接关系，由网络上的算法来处理**
- 类似的，发明人-专利-技术三层网络，作者-书籍-概念三层网络
- 将来可以考虑把这些网络联合起来
- 甚至加上从论文专利到产品、产品生产、从产品到利润再到基金
- 开放和封闭系统：某些研究中某些顶点被当做外界，忽略反作用
- 更多信息见“Big Physics”网站，以及《基于网络的科学学》
- 这个框架需要**概念层数据、作者层数据**

有了这个数据、问题和方法的框架会怎样?

- 把最基础的关系找出来，甚至搞得更加准确
- 把问题表示成这个关系数据上的问题
- 发展这个表示下的方法来解决这个框架表示下的问题
- 发展方法的时候运用好网络的核心精神：直接和间接联系的综合
- 甚至提出问题促进网络科学分析方法的发展

职业生涯大事的影响、人才培养和流动

- 问题：科学家得奖、当领导、被人才、被撤稿之后的影响
- 对其他科学家以及对整个科学界,
- 问题：国家地区的人才培养和流动的影响
- 对其国家以及对整个科学家劳动力市场
- 直接对象和直接数数
- 间接对象和间接影响
- 正在和杨立英（前者）、李江（后者）开展合作研究

几个传统上就关心的例子

- 论文分类问题
  1. 概念层最好有，也可以用粗糙的自然语言处理得到的词矢量表示来代替
  2. 论文引用层可以考虑网络的矢量表示
- 作者学校国家贡献计数和影响力分析
  1. 论文本身的影响力也考了间接联系
  2. 过间接传播把论文影响力分配作者头上，考虑作者的主概念主领域
  3. 统计作者的国籍算到国家
- 概念层用于更好的作者识别、发现重要边的算法来找到真引用?

领域之间的知识创造依赖关系

- 问题：从领域之间的论文引用关系来看一个领域的影响力
- 对某个其他领域和整个学科的
- 数据描述：网络；计算方法：直接数数、PageRank、投入产出

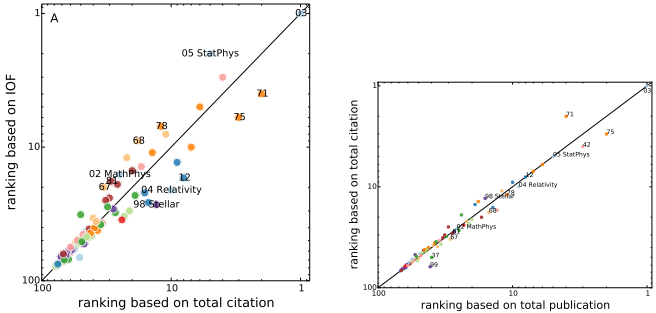

Figure: 关注05 (StatPhys) 、02 (MathPhys)、04 (Relativity) 以及98 (Stellar)这些突出的领域。还可以看期刊、论文、城市、国家、学校等。

论文的影响力，从专利（应用）引用来看

- 论文内部的影响已经可以综合考虑直接和间接，在不同的主体层面
- 问题：现在考虑论文对技术发展的作用，专利当做技术的计量
- 数据描述：论文内部引用网络，加上，专利对论文的直接引用
- 计算方法：开放系统PageRank和开放系统投入产出
- 正在和沈哲思和李梦辉开展合作研究
- 也讨论不同层次的主体，学科、学校、国家等
- 还可以反过来把论文当外界，专利引用当内部网络

论文的影响力，从书籍（文明）引用来看

- 问题：从论文对文明的发展的作用来讨论论文的影响力
- 把书籍当做文明的计量
- 数据描述：论文内部引用网络，加上书籍对论文的引用
- 计算方法：开放系统PageRank和开放系统投入产出

论文和专利统一起来看，部分数据的问题

- 忽略了系统外部的元素之间的相互作用以及系统对外部的反作用
- 问题1：把系统外部放进来形成更大的系统专利-论文多层网络
- 问题2：任何研究实际上都是部分数据下的研究，需要研究如何用好部分数据
- 例如，对比专利直接、当做外界、完全不纳入系统三种方式的结果
- 最好还能有基于Ground Truth的检验

概念本身的重要性，从论文获取概念频次

- 问题：概念之间逻辑上是相联系的，哪一个最值得学，最应该先学
- 数据：概念之间的逻辑关系构成内部网络，学术论文中出现的概念的次数是外界
- 计算方法：直接数数、PageRank、投入产出
- 考虑在科学学上先做一下？需要概念集合、概念关系
- 最好还能有检验

论文论文首次提出概念带来的影响力

- 现象：某篇文章首次提出某些后来得到了学界的使用的概念
- 问题：考虑了首次提出之后，论文的影响力计量
- 数据描述：论文-概念网络，相当于把论文-概念连边分成是否首次提出
- 计算方法：直接数数、PageRank、投入产出
- 可以在论文-概念网络之内研究，还可以加上专利和书籍
- 数据工作量比较大

从多层网络整体来看各个主体的影响力

- 问题：一起考虑了作者、论文（专利、书籍）、概念之后，各个主体的影响力
- 理念：重要作者的文章更重要，工作在重要概念上的文章更重要，被重要工作研究的概念更重要，发表重要文章的作者更重要
- 数据描述：多层网络
- 计算方法：多层网络上的PageRank和投入产出，还要发展
- 科学学的研究问题反过来促进网络科学的发展
- 将来还可以考虑把产品、基金数据放进去
- 例如讨论美国基金对中国科学家和整个科学界的影响

带回家的消息：一起来推动大概一致的科学学学科的框架

- 内容数据需要进入科学学
- 网络科学的精神就是直接联系用网络表示、分析方法综合考虑直接和间接联系
- 科学学典型研究对象和对象之间的直接关系是多层网络
- **多层网络框架可以帮助内容进入科学学，可以在科学学上用好网络科学的精神**
- 到了形成大概一致的问题数据研究方法的框架，来推动学科发展的时候了
- 这个框架需要概念层数据、作者层数据

感谢您的时间与注意力

- Xiaoyong yan, Ying Fan, Zengru Di, Shlomo Havlin, Jinshan Wu, Efficient learning strategy of chinese characters based on network approach, PloS ONE, 8, e69745 (2013)
- Zhesi Shen et al. , Interrelations among scientific fields and their relative influences revealed by an input–output analysis, Journal of Informetrics 10, 82-97(2016)

- 感谢合作者：闫小勇、沈哲思、李梦辉、杨立英、Ronald Rousseau、曾安、狄增如、裴建锁、韦添、鲍建樟、郭金忠、刘晓玲等

- 要了解我更多，请访问“系统科学人之吴金闪”，吴金闪在北师大系统科学学院的主页
- 研究工作的整理在这里“Big Physics（大数据大物理研究小组）”