第1章　概　论

阳德青

复旦大学大数据学院

Email：yangdeqing@fudan.edu.cn

引　言

社交网络挖掘　vs.　社会网络分析

引　言

- 社会（社交）网络分析的意义何在？

长期以来，社会研究将焦点集中于个体的行为，而忽视了行为的社会方面，即行动者潜入其中的社会关系模式对他们的行动结果有重要的影响。正如社会学家艾伦·巴顿（Allen Barton，1968）所描述的当时社会科学的主流研究情况：

在过去30年里，经验性的社会研究被抽样调查所主导。从一般情况而言，通过对个人的随机抽样，调查变成了一个社会学的绞肉机——将个人从他的社会背景中撕裂出来并确保研究中没有任何人之间会产生互动。这有点像一位生物学家让他的实验动物经过一台碎肉机的处理，然后通过显微镜观察每第100个细胞；在这里，解剖学和生理学用不上了，剩下的只有细胞生物学……如果我们的目的是理解人类的行为而不只是记录它，我们就需要了解他所在的群体、邻里、组织、社交圈、社区以及互动、沟通、角色期望、社会控制。

应从社会的角度去分析和理解个人的行为

引　言

- 社交网络挖掘的意义何在?
  - 社交网络媒体上的虚拟社区很大程度上反映现实人类社会的特质
  - 社交网络媒体的高速发展产生了海量数据，价值有待挖掘
  - 社交网络数据挖掘的巨大商机
    - 精准营销
      - 用户画像、个性化推荐、社会化营销…
    - 舆情管理
      - 危机预警、观点挖掘、情感分析…
    - 精准预测
      - 市场预测、新闻热点预测、用户行为预测…

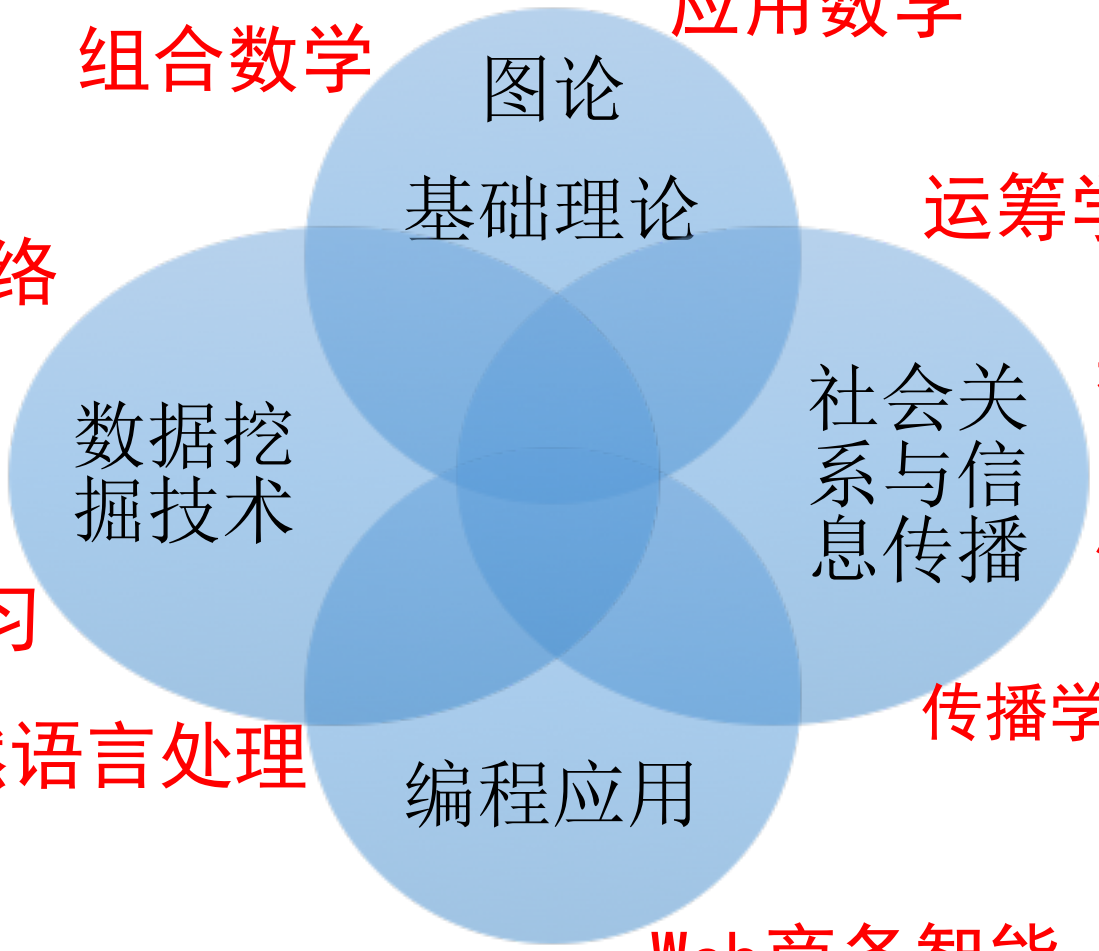
应将数据挖掘等技术手段与传统社会学研究结合

教学大纲

通过课堂教学、上机实践、文献阅读和项目报告等教学方式，让学生学习并掌握社交网络分析与挖掘中所涉及的与计算机科学、统计学、社会学、经济学、传播学等领域有关的原理和方法，了解现今流行的社交网络数据采集和分析技术，并能够运用所学知识和技能开展相关实践项目的应用，让学生掌握社会管理与大数据挖掘等综合技能与素质。

运用计算思维学好交叉学科的知识

教学大纲



应用数学

组合数学

运筹学

复杂网络

社会心理学
→社会计量学
/计算社会学

数据库

机器学习

传播学→计算传播学

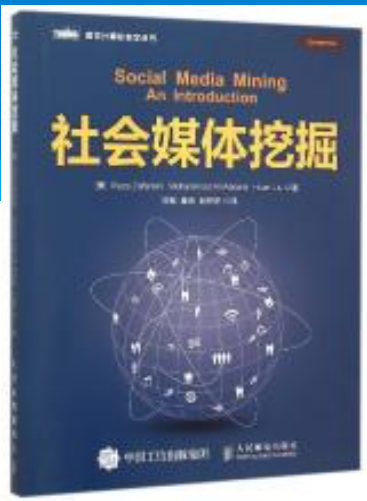
自然语言处理

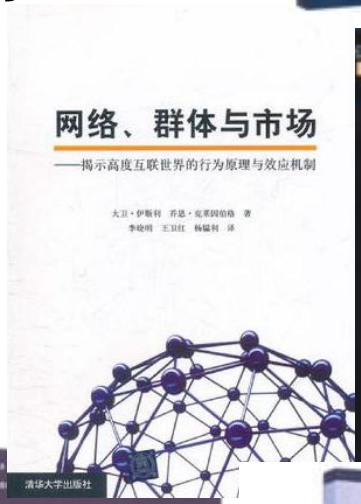
Web商务智能

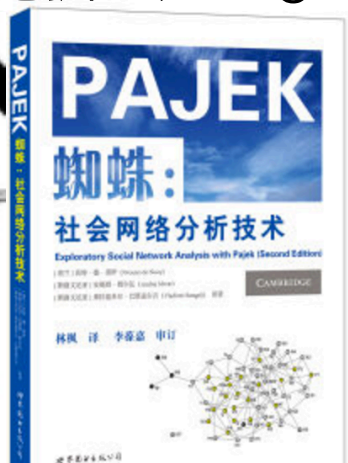
Python/H5/JS/Pajek

参考书目

- 《社会媒体挖掘》，Reza Zafarani等
- 《网络、群体与市场》，大卫·伊斯科，乔恩·克莱因伯格
- 《Pajek: 社会网络分析技术》，沃特·德·诺伊等
- 《社会网络分析：方法与应用》，沃瑟曼等
- 《在线社交网络分析》，方滨兴等
- 《社会网络分析法》，约翰·斯科特
- 《社交网站的数据挖掘与分析》，Russell
- 《Web数据挖掘》，Bing liu







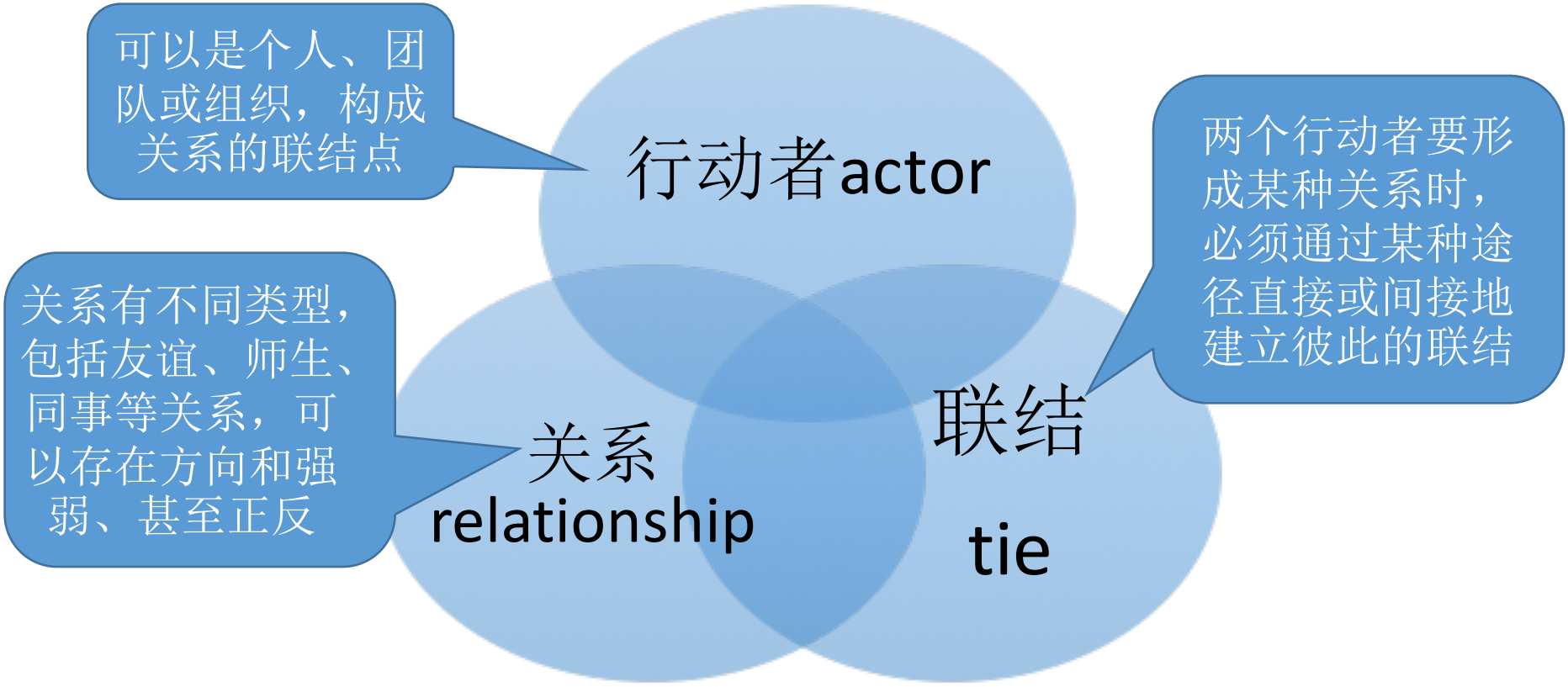
目　录

社交网络概述

- 何为社交网络?
  - SNS （Social Network Services/Systems）
  - 指社会个体成员之间因为互动而形成的相对稳定的关系体系，社会网络关注的是人们之间的互动和联系，社会互动会影响人们的社会行为。——百度百科
  - 由许多节点构成的一种社会结构，节点通常是指个人或组织，网络代表各种社会关系，经由这些社会关系，把从偶然相识的泛泛之交到紧密结合的家庭关系的各种人们或组织串连起来。——维基百科

社交网络概述

- 社交网络的形式化定义
  - 由特定集合的行动者（点）以及行动者之间的关系（线）组成（图），一般包括行动者、关系、联结三个基本概念



社交网络概述

- 行动者的类型

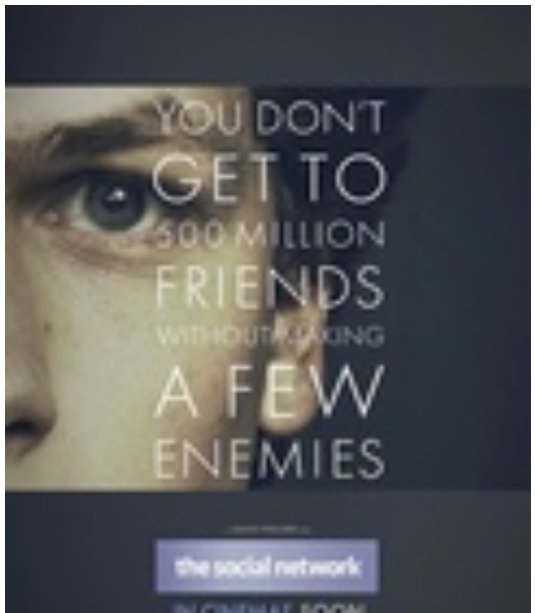
桥 bridge	■ 连接不同沟通群体之间人物
联络人 liaison	■ 不属于任何沟通群体但起着联络作用的人物
孤独者 isolator	■ 很少或不参与任何沟通团体的人物
明星 star	■ 网络中有最多关系联结的人物
结构洞 structural hole	■ 社会网络中的空隙。个体行动者（包括个人、子单位、组织）被认为可通过以下方式提高自己的社会资本：在两个原本不相互联结的小集团之间担当联络员（liason）的角色，或者在他们所隶属的群体和他们所参加的另一个群体之间起架桥（bridging）的作用。

社交网络发展历史

- 人类学家Barnes于1954年首次提出“社会网络”的概念
- 1960年代，产生六度空间/六度分隔/小世界理论

20世纪60年代，哈佛大学的社会心理学家米尔格兰姆(Stanley Milgram)就设计了一个连锁信件实验，将一套连锁信件随机发送给居住在内布拉斯加州的160个人，信中放了一个波士顿股票经纪人的名字，要求每个收信人将这套信寄给自己认为是比较接近那个股票经纪人的朋友，朋友收信后照此办理。最终，大部分信在经过五、六个步骤后都抵达了该股票经纪人。

- 六度分隔理论提出后，引起社会极大关注，到1990年代，同名戏剧和电影先后上演
- 2010年，讲述Facebook创始人故事的电影《社交网络》上映



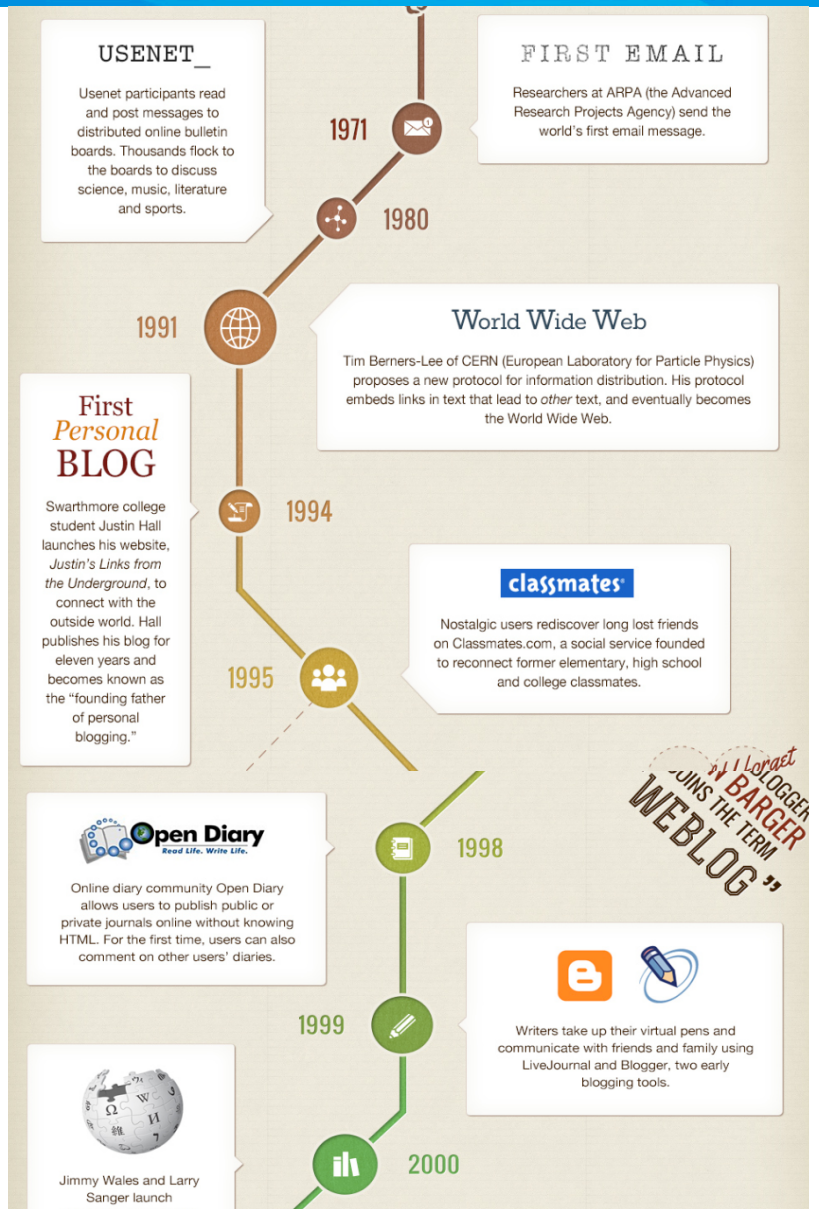
社交网络发展历史

- 150定律
  - 由英国牛津大学的人类学家罗宾·邓巴（Robin Dunbar）于2009年提出：根据猿猴的智力与社交网络推断出：人类智力将允许人类拥有稳定社交网络的人数是148人，四舍五入大约是150人。因此，150又称邓巴数字。
  - 该定律指出：人的大脑新皮层大小有限，提供的认知能力只能使一个人维持与大约150个人的稳定人际关系，这一数字是人们拥有的、与自己有私人关系的朋友数。也就是说，人们可能拥有150名好友，甚至更多社交网站的“好友”，但只维持与现实生活中大约150个人的“内部圈子”。而“内部圈子”好友在此理论中指一年至少联系一次的人。
  - 实例
    - 中国移动动感地带sim卡只能保存150个联系人手机号
    - MSN账号最多联系人数是150

在线社交网络（社交媒体）发展史

- 萌芽与发展期
  - 1971 年，人类第一封电子邮件诞生
  - 1991 年，伯纳斯·李经过多年实践和改进，创办了以“超链接”为特征的万维网（World Wide Web）；国内第一个BBS站“中国长城站”诞生
  - 1997 年，美国在线实时交流工具 AIM 上线，一位名为 Jorn Barger 的先锋博客作者创造了“weblog”一词
  - 1999年，天涯社区诞生
  - 2000 年，Wikipedia成立，这是全球首个开源、在线、协作而成的百科全书
  - 2002年，LinkedIn成立
  - 2003 年，MySpace和WordPress上线

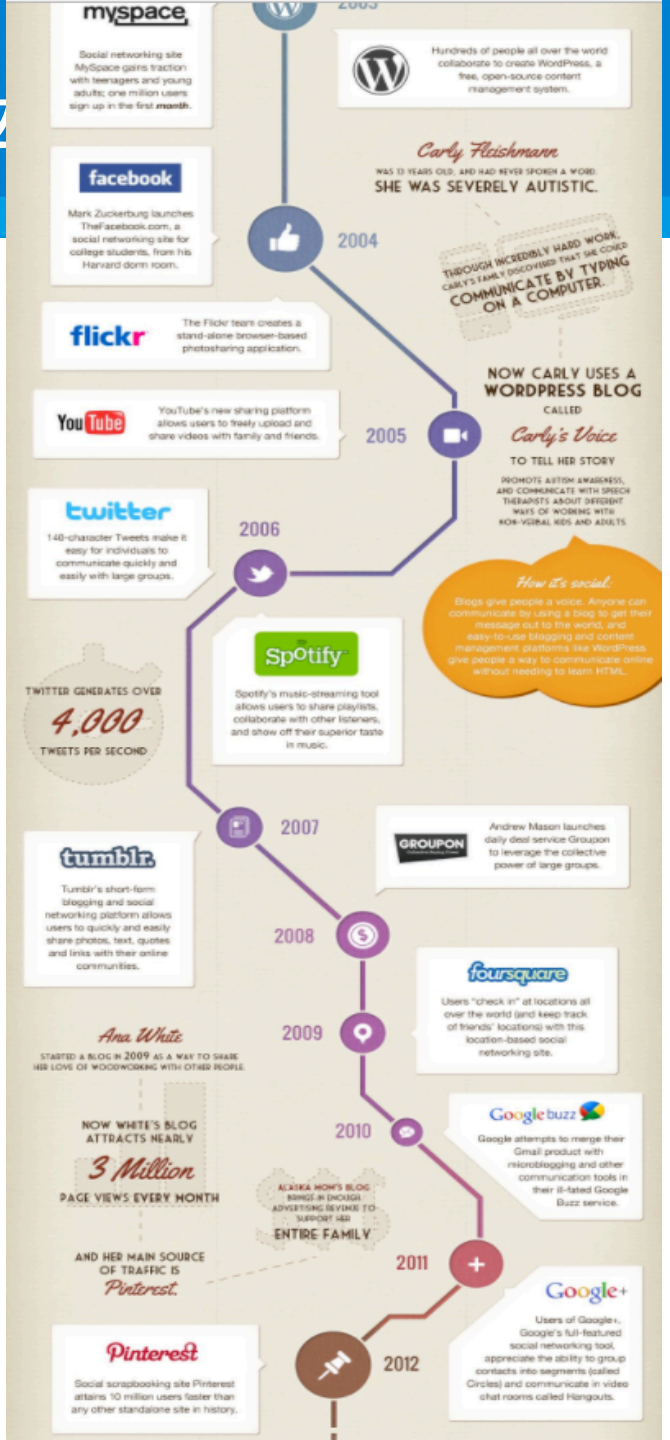
Web1.0时代



在线社交网络（社交媒体

- 鼎盛期
  - 2004 年，Facebook和Filckr成立
  - 2005 年，YouTube成立；校内网、土豆网、豆瓣网成立
  - 2006 年，Twitter成立，千橡集团收购校内网改名人人网
  - 2008年，开心网成立
  - 2009 年，Foursquare（基于check-in）上线，新浪微博推出
  - 2011 年，腾讯推出微信，Google+上线

Web2.0→Web3.0时代

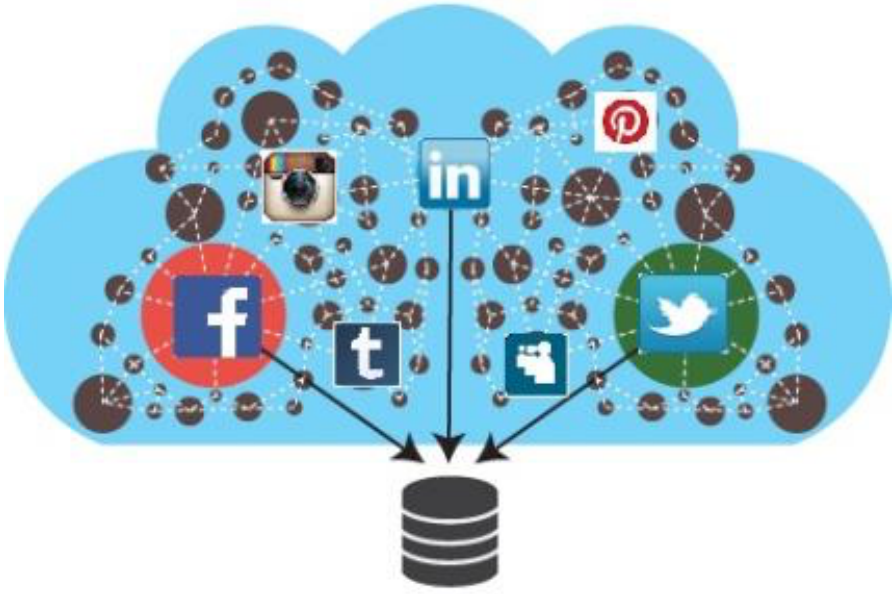


目　录

- 社交网络概述
- 社会网络分析
- 社会计算

社会网络分析的概念

- 何为社会网络分析?
  - 为理解人类各种社交关系的形成、行为特点以及信息传播的规律，采用相关的分析方法，涉及信息学、数学、社会学、管理学、心理学等多学科的融合理论和方法
  - 社会网络分析方法基于一个直觉性观念，即行动者嵌入在其中的社会关系的模式对于他们的行动结果有重要的影响，社会网络分析者则力求揭示不同类别的模式。



社会网络思想的三个来源

物理学力场理论

德国的研究学者库尔特·卢因、弗里茨·海德和雅各布·布雷洛等学者，将网络概念应用于对社会互动的研究中。

数学图论方法

最早的学者是卡特赖特（1956）采用数学图论的方法研究社会互动，推动了社会网络研究从描述性研究转向分析性研究。

社会网络分析法形成的技术基础是图论和社会计量学（Sociometrics）

社会学人类学方法

最典型的代表就是1930年代著名的“霍桑实验”，是首次运用社会网络图（或叫社群图Sociogram）描述个体自由选择的社会互动结构。

“霍桑试验”

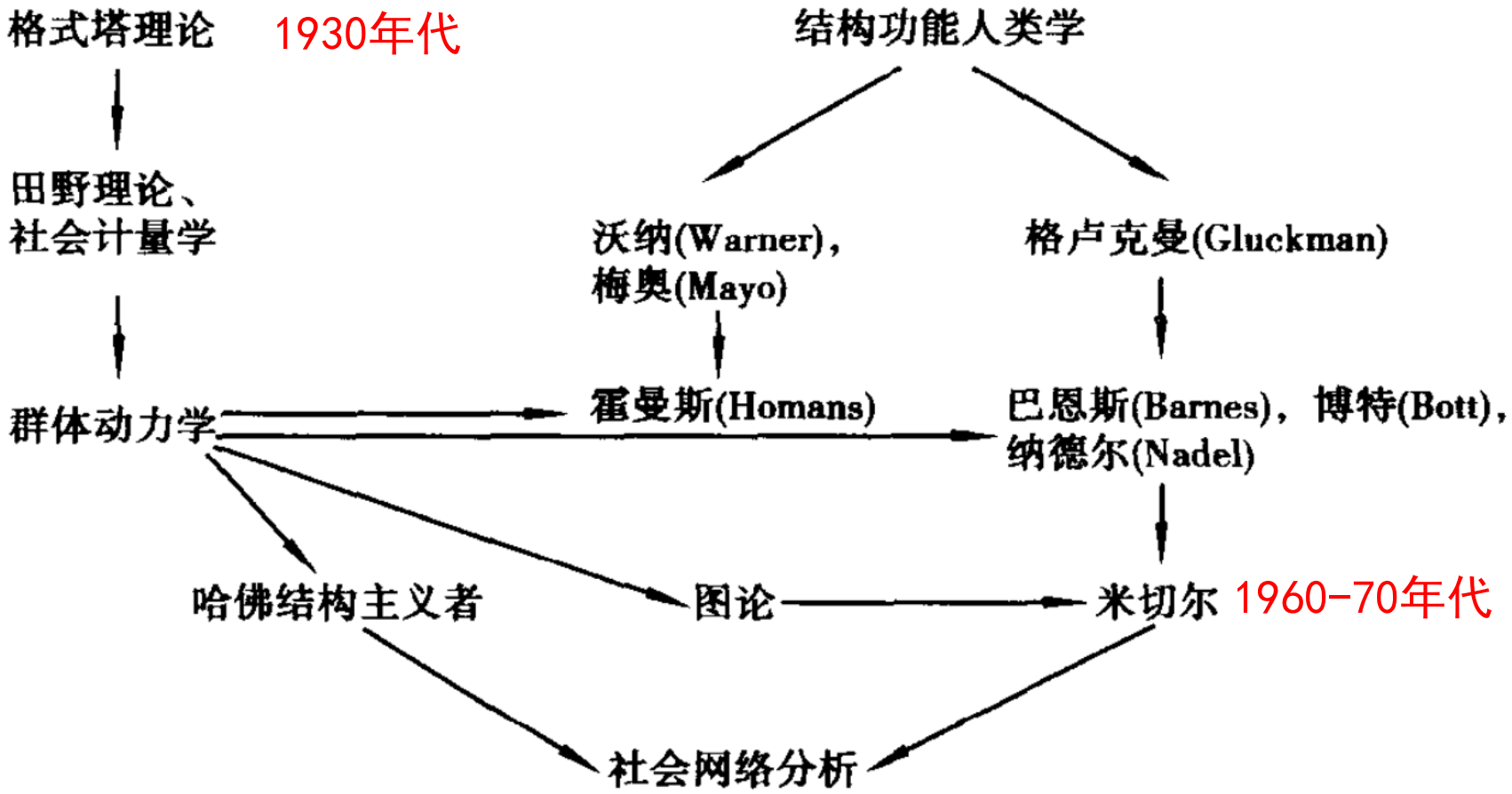
- 在芝加哥霍桑电器厂针对工厂的工人前后进行了五年的实验
  - 第一阶段：对6个女电脑装配工延长工作时间，发现照明、休息等物质条件并不构成影响产量的主要因素，而社会因素和心理因素才是决定工人满意度与生产率的主要因素
  - 第二阶段：用三年时间对两万多人进行面谈，进一步证实了工作环境中人的因素和社会因素的重要性
  - 第三阶段：对14名男性工人组成的小组进行六个月的深入观察研究，研究表明感情、地位和相互间的社会作用是基础的非正式社会组织的作用
- 梅奥等学者通过实验总结出社会组织中人际关系的新观点
  - 不应把工人单纯地看成是“经济人”，而应该是“社会人”
  - 工作效率的高低主要取决于“士气”，而士气来自于人与人之间的关系
  - 不能只重视“正式组织”，还要重视“非正式组织”的作用
  - 领导不仅要善于了解人们合乎逻辑的行为，而且还要善于了解人们不合逻辑的行为

社会网络分析发展溯源

- 从社会学/人类学发展角度，可追溯至1930年代产生的三个流派传统
  - 社会计量学者：结合图论的方法从研究小群体入手
  - 哈佛学者：在人际关系模式时提出了“派系”这个概念
  - 曼彻斯特的人类学家：考察部落和乡村的“社区”关系结构
- 三个流派在1960~70年又汇聚在哈佛大学，产生了当代的社会网络分析

社会网络分析发展溯源

- 格式塔心理学
  - 主张研究直接经验（即意识）和行为，强调经验和行为的整体性，认为整体不等于并且大于部分之和，主张以整体的动力结构观来研究心理现象



社会网络分析的特点

1. 源自于联系社会行动者关系基础之上的结构性思想；
2. 以系统的经验数据为基础；
3. 非常重视关系图形的绘制；
4. 依赖于数学或计算模型的使用。

社会网络分析

- 社交（社会）网络的类型
  - 根据观察的角度不同，社会网络可以分为自我中心社会网（ego-centric networks）和整体社会网（whole-networks）

	自我中心网络	整体社会网络
个体	自我中心网络与个体行为关系研究	整体网络对个体行为的影响
组织	自我中心网络对组织行为的影响	整体社会网络和组织行为的交互影响

社会网络分析

自我中心社会网络

■从研究的个体出发，研究与其直接或间接的联结，找出由中心向外扩展的关系网络。自我中心社会网络通过随机抽样采用问卷方式收集个体社会交往信息，分析这些信息对个体行为产生的影响。

■自我中心社会网络通过提由个体提出在某些领域与自己关系密切的人名及个体与这些人之间关系的类型和特征。其中比较著名的有Ronald Burt（1984）提出的提名生成法自我中心社会网络问卷。

■在个体中心社会网络中关系的类型可以有很多种。台湾心理学家黄光国（1987）研究中国人行为中人情与面子问题，认为中国人有三种关系：**情感型关系**，即情感支持的交换；**工具型关系**，对等的资源交换关系；**混合型关系**，同时包含上面两种关系。

社会网络分析

整体社会网络

■整体社会网络主要研究网络结构问题，圈定范围内所有行动者的相互关系、密度、联系特征和次团体的数量等，范围可以是组织、团体、部门、小组。通过分析整体网络，可以发现网络内不同地位的个体角色，如明星、鼓励者、联络人、桥梁等。

■整体社会网络中的封闭群体中个体间关系是多维的，魁克哈特（1992）认为有三种关系类型：情感关系、情报关系和信任关系，并认为信任关系和情感关系经常会重叠。罗家德（2005）认为一个人在组织的场域中有四维的社会网络：情感网、咨询网、情报网和关系网。

■整体社会网络资料的收集需在一个有边界的团体（企业、部门或者一个团队）中，询问团体中的每个人和其他人之间的关系，问题可以涉及情感网、咨询网、情报网和关系网等方面的内容。

社会学的研究范式

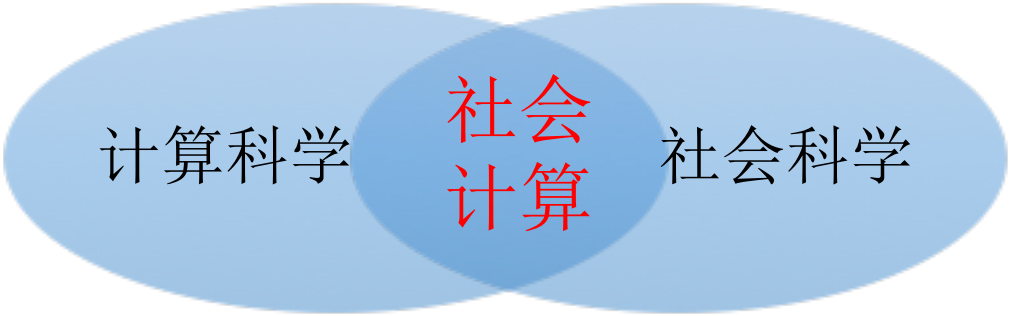
■整体社会网络资料比自我中心社会网络资料更难收集，它要求一整个群体中所有的人都必须愿意填答问卷，而且问卷不能匿名。满足上面的要求，很难进行随机抽样，只能采取便利抽样法。

目　录

- 社交网络概述
- 社会网络分析
- 社会计算

社会计算概念

- 2009年2月，美国哈佛大学大卫·拉泽（David Lazer）等15位美国学者在《Science》上联合发表了一篇具有里程碑意义的文章《Computational Social Science》
- 何为社会计算?
  - 社会计算（Social Computing），又称计算社会学（Computational Social Science）
  - 一般是指社会行为和计算系统交叉融合而成的一个研究领域，研究如何利用计算系统帮助人们进行沟通与协作，如何利用计算技术研究社会运行的规律与发展趋势
  - 涉及学科
    - 计算机科学（复杂网络、数据挖掘、NLP、信息检索等）、社会学、统计学、管理学……



社会计算概念

- 何为社会计算?
  - “利用计算系统帮助人们进行沟通与协作”
    - 复制和重构现实社会中人与人之间的关系
    - 内容包括社会网络服务、群体智慧（维基百科、百度百科）
  - “利用计算技术研究社会运行的规律与发展趋势”
    - 以社交网络/媒体为研究对象
    - 发现社会关系、社会行为的规律
    - 预测政策实施的可行性
    - 内容包括社会网络分析、内容计算等

社会计算的研究目标

1）在深入理解当前社会问题动态性、快速性、开放性、交互性、数据海量性和复杂性的基础上，为解决新兴社会问题建立统一的社会科学基础模型和理论框架；

2）社会科学基础模型和理论“计算化”或建立其到计算技术的映射机制，研究与社会相关应用中的建模与计算方法，自下而上地为解决新兴社会问题提供整套理论和技术支撑；

3）深化学科交叉研究，为网络化社会背景下的社会科学研究提供实验方法；同时，以新兴问题促进相关研究领域内涵和内容的延拓，推动基础理论和方法的创新和突破。

社会计算的研究内容

- 社交网络服务
  - 以SNS为代表，还包括电子邮件、聊天算计、网络论坛等
- 群体智慧（Collective Intelligence）
  - 维基百科、百度百科/百度知道、知乎等
- 社会网络分析
- 内容计算
  - 计算分析的内容以语言文字为主，包括图片、音视频等其它媒体内容
  - 舆情分析、社区发现、人际关系挖掘等

社交网络挖掘

社交网络挖掘的挑战

- 大数据悖论
  - 个体数据是往往是稀疏的，因此才需要汇聚群体/整体的大数据
    - 举例：社会推荐中的冷启动问题
  - 个体行为模式不能简单等同于整体，因此不能是个体的简单相加来代表整体，也不能整体的简单拆分来刻画个体
- 数据采集与获取
  - 爬虫、API等技术需要掌握的重要手段

社交网络挖掘的挑战

- 数据清理与噪音消除
  - 大数据可以允许个体数据的错误，但不等于不需要数据清理
  - 消除噪音个体　trade-off　个体数据价值挖掘
- 性能评价
  - Ground truth（标准答案）如何获取?
  - 没有客观标准答案的如何评价?
  - 数据挖掘出来的模式/结果能否解答“为什么”？