第七章社会网络分析

授课教师: 吴翔 wuhsiang@hust.edu.cn

OCT 21 - 24, 2019

🕕 社会网络分析概述 (2 个课时)

- 2 社会网络主要分析角度 (4 个课时)
- 3 案例 (2 个课时)

社会网络分析概述 (2 个课时)

课程存储地址

- 课程存储地址: https://github.com/wuhsiang/Courses
- 资源:课件、案例数据及代码



参考教材

- 斯坦利·沃瑟曼, 凯瑟琳·福斯特. 社会网络分析: 方法与应用. 北京: 中国人民大学出版社. 2012. (注: 对应英文版于 1996 年出版)
- 托马斯. 社会网络与健康:模型、方法与应用. 北京:人民卫生出版社. 2016.
- 埃里克·克拉泽克,加博尔·乔尔迪.网络数据的统计分析:R
 语言实践.西安:西安交通大学出版社.2016.

本节知识点

- 社会网络的基本概念
- 社会网络的符号表示
- 吸烟行为建模: 社会网络视角
- 基本社会网络结构
- 社会网络分析软件

社会网络与健康



图 2: 社会网络与肥胖

- 哪种饮食结构/生活习惯会让人变胖?
- 肥胖会"传染"吗?

社会网络与健康(续)



图 3: 社会网络与抑郁

- 哪种特质的人更容易抑郁?
- 社会支持是否有助于改善抑郁?

社会网络与健康(续)



图 4: 社会网络与卫生服务能力提升

• 医联体/医共体模式是否有助于提升基层卫生服务能力?

社会网络视角

- 行动者之间的关系是主要的, 行动者的属性是次要的
- 行动者和他们的行动被视为相互依赖的,而不是相互独立的 自治体
- 行动者之间的联系是信息和资源的流动通道
- 个体的网络模型将网络结构环境视为个体行动的机遇或限制
- 网络模型将(社会、经济、政治、情感等)结构概念化为行动者之间关系的稳定形式

7.1.1 基本概念

社会网络分析 (social network analysis, SNA) 的关键概念:

- 行动者: 社会网络分析中的社会实体被称为行动者,包括个体、企业、民族国家等
- 关系连接:行动者通过社会关系彼此相连。这些联系包括:评价、资源传输、行为互动等。联系存在于特定的成对行动者之间
- 关系: 群体成员间某种类型的联系的集合
- 社会网络: 行动者 (人、组织等), 及其之间关系的集合

其它关键概念还包括:二元图、三元图、子群、群。

基本特征

SNA 的基本特征:

- 考虑整个网络结构
- 论证网络结构如何影响个体行为
- 运用图表展示
- 运用数学的形式

7.1.2 社会网络数据

社会网络数据包括:

- 行动者集合
- 社会关系
- 行动者属性

社会网络数据的符号表示包括:

- 图论
- 社会计量

图论符号表示法

图 G=(N,L) 由节点的集合 N 和边的集合 L 所定义。

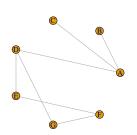
表 1: 社会网络与图论的对应关系

社会网络	图
行动者	点带
社会关系	边

图论符号表示法中,可以采用节点集合和边列表来表示社会网络数据。

图论符号表示法(续)

- 行动者集合 $G = \{A, B, C, D, E, F, G\}$
- 社会关系集合 *L* = {<A, B>, <A, C>, <A, D>, <D, E>, <D, G>, <E, F>, <F, G>}



社会计量符号表示

- 社会计量(sociometric):由人以及被度量的人与人之间的情感关系组成的社会网络数据集合,旨在研究一群人中积极和消极的感情关系
- 社会关系矩阵:邻接矩阵,对应于量化行动者之间的社会关系图

邻接矩阵

	Α	В	C	D	Ε	G	F
Α	0	1	1	1	0	0	0
В	1	0	0	0	0	0	0
C	1	0	0	0	0	0	0
D	1	0	0	0	1	1	0
Ε	0	0	0	1	0	0	1
G	0	0	0	1	0	0	1
F	0	0	0	0	1	1	0

其它情形

- 有值关系
- 有向关系
- 多重关系
- 网络动态性

吸烟行为建模: 社会网络视角



图 5: 要不你也来一支?

案例背景描述

- 吸烟人群, 但同时也认识到吸烟的危害
- 自制力程度有差异,且可以由行动的阈值来刻画
- 行动阈值:周围朋友吸烟的人数达到特定值 (threshold)时, 才会开始吸烟

社会网络符号表示: 图论

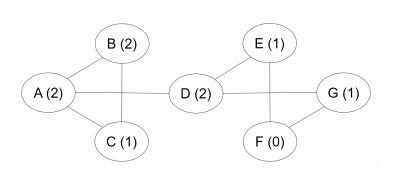


图 6: 图论符号表示

社会网络符号表示: 社会计量

行动者	Α	В	С	D	E	F	G	度	阈值
Α		1	1	1	0	0	0	3	2
В	1		1	1	0	0	0	3	2
С	1	1		0	0	0	0	2	1
D	1	1	0		1	0	1	4	2
E	0	0	0	1		1	0	2	1
F	0	0	0	0	1		1	2	0
G	0	0	0	1	0	1		2	1

图 7: 社会计量符号表示

吸烟行为分析:情境一

slides on smoking behavior

情境二: 网络结构变化

- 假定 A 在某次聚会中认识了 F, 两人成为了好朋友
- 以上社会网络中的吸烟行为规律是否会变化?

社会网络符号表示:图论

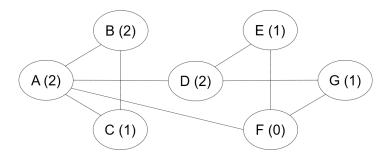


图 8: 图论符号表示

社会网络符号表示: 社会计量

行动者	Α	В	С	D	E	F	G	度	阈值
Α		1	1	1	0	1	0	4	2
В	1		1	1	0	0	0	3	2
С	1	1		0	0	0	0	2	1
D	1	1	0		1	0	1	4	2
E	0	0	0	1		1	0	2	1
F	1	0	0	0	1		1	3	0
G	0	0	0	1	0	1		2	1

图 9: 社会计量符号表示

吸烟行为分析:情境二

slides on smoking behavior

案例总结讨论

- 案例有什么有意思的结论?
- 社会网络分析视角的特点是什么?
- 社会网络分析视角适合哪些健康领域的议题?

节点度

在无向图 G 中, 节点 n_i 的度为

$$\underline{d(n_i)}_{\text{degree}} = \underbrace{\sum_{j} x_{ji}}_{\text{indegree}} = \underbrace{\sum_{j} x_{ij}}_{\text{outdegree}}. \tag{1}$$

对于有向图而言,

$$\underbrace{\sum_{j} x_{ji}}_{\text{indegree}} \neq \underbrace{\sum_{j} x_{ij}}_{\text{outdegree}}.$$

节点度(续)

图 G 中节点度的均值为

$$\bar{d} = \frac{\sum d(n_i)}{g} = \frac{2L}{g}$$

度的方差为

$$S_D^2 = \frac{\sum [d(n_i) - \bar{d}]^2}{q}.$$

 $S_D^2 = 0$ 对应的图称为 d-规则图 (d-regular lattice)。

7.1.4 主要网络模型

参照网络模型:

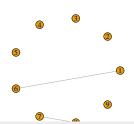
- 随机网络
- 规则网络

现实网络模型:

- 小世界网络
- 无标度网络(优先连接网络)

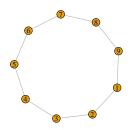
随机网络

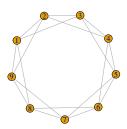
- 基本假定: 节点之间的边是随机构建的
- ullet G(n,p) 模型: 图 G 有 n 个节点, $\binom{n}{2}$ 条边以 p 的概率 随机连接
- 节点的期望度是 (n-1)p,边的期望条数是 $\frac{n(n-1)}{2} \times p$



规则网络

ullet 基本假设:每个节点的度是常数 c





小世界网络

- 基本问题: 社会网络中两个节点之间传递信息将需要几个步骤?
- 基本假设:大多数节点几乎没有联系,但任意两个节点之间 的距离都比预期的短
- 特点:这个世界真小啊!"六度分割"理论

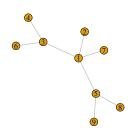


无标度网络

• 基本问题: 加入现有网络时, 行动者对要联系的人有偏好吗?

• 基本假设: 行动者更喜欢连接到网络最中心的位置

特点: "富者愈富"



7.1.5 社会网络分析软件

常用分析工具:

- UCINET
- Pajek
- NetMiner
- STRUCTURE
- MultiNet
- StOCNET

新兴分析工具

- Python-NetworkX
- R-igraph

本课程采用 igraph 包进行演示。

社会网络主要分析角度 (4 个课时)

本节知识点

- 中心性与声望(行动者层级)
- 凝聚子群 (子群层级)
- 评估网络属性 (网络层级)

7.2.1 中心性与声望

- 基本问题:如何识别社会网络中"最重要的"角色?
- 中心性测度的**有效性**
 - 我们是否能够捕捉到实质上所要表示的"重要"?
 - 先有理论基础,再进行量化
- 中心性与声望
 - 中心性: 行动者参与其中, 适用于无向关系和有向关系
 - 声望: 行动者作为接受者, 适用于有向关系
 - 情境(关系本身的性质): 讨厌(接受者,负面)、给出建议 (发送者)

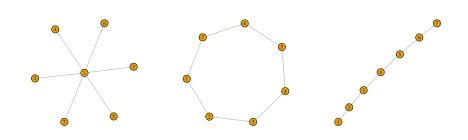
中心性度量

无向关系的社会网络中,主要的几种中心性度量:

- 度中心性 (degree centrality)
- 特征向量中心性 (eigenvector centrality)
- 接近中心性 (closeness centrality)
- 中介中心性 (betweenness centrality)

特殊网络

我们考虑星形网络、环形网络和线形网络。



度中心性

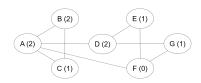
度中心性 (degree centrality) 的测量逻辑:

- 中心的行动者在某种意义上必须是最活跃的
- 节点度可以衡量活跃程度

$$C_D(n_i) = \frac{d(n_i)}{q-1} \tag{2}$$

度中心性(续)

在图
$$G$$
 中,节点个数 $g=7$,度的最大值为 $g-1=6$ $d(A)=4$,故 $C_D(A)=2/3$ $d(D)=d(F)=3$,故 $C_D(D)=1/2$ $d(B)=d(C)=d(E)=d(G)=2$,故 $d(B)=1/3$



度中心性(续)

表 3: Degree centrality for four graphs

star	ring	line	smoking
1	0.33	0.17	0.67
0.17	0.33	0.33	0.33
0.17	0.33	0.33	0.33
0.17	0.33	0.33	0.5
0.17	0.33	0.33	0.33
0.17	0.33	0.33	0.5
0.17	0.33	0.17	0.33

特征向量中心性

特征向量中心性 (eigenvector centrality) 的测量逻辑:

- 如果某个行动者邻居大多是中心行动者,那么他就是中心行动者
- 中心性不仅取决于认识多少人,还取决于认识的人是否重要

图 G 的邻接矩阵为 A,

$$Av = \lambda v$$
,

其中 λ 为特征值, v 为特征向量。

特征向量中心性 $C_e(n_i)$ 定义为最大特征值对应的特征向量。

特征向量中心性(续)

在图 G 中,最大特征值为 2.73。

对应的特征向量为

 $C_e(n_i)=c(0.53,0.31,0.31,0.42,0.31,0.42,0.31)$ 。这一结果可以讲一步归一化。



特征向量中心性(续)

表 4: Eiggenvector centrality for four graphs

star	ring	line	smoking
1	1	0.38	1
0.41	1	0.71	0.58
0.41	1	0.92	0.58
0.41	1	1	0.79
0.41	1	0.92	0.58
0.41	1	0.71	0.79
0.41	1	0.38	0.58

PageRank

Google 搜索引擎

- 采用 PageRank 来度量网页的中心性
- 在检索时,和查询相匹配且 PageRank 值高的网页将最先显示

PageRank 在特征中心性的基础上作了修正:

- 中心节点在传递其中心性时, 考虑其度 (有向图中, 则是出度)
- 每个邻居获取其中心性的一部分 (除以节点度)

接近中心性

接近中心性 (closeness centrality) 的测量逻辑:

- 占据中心地位的行动者在与其他行动者交流信息时更有效率
- 如果行动者能快速地与所有其他行动者产生内在连接,那么 他就是中心行动者
- 最小距离可以用于测量中心性

$$C_C(n_i) = \frac{g-1}{\sum_{j=1}^g d(n_i, n_j)}.$$
 (3)

接近中心性(续)

在图 G' 中,节点个数 g=7,最短距离之和的最小值为 g-1=6。

$$\sum_{n_j \neq A} d(A, n_j) = 1 \times 4 + 2 \times 2 = 8$$
,故 $C_C(A) = 6/8 = 0.75$

$$\sum_{n_j\neq B}d(B,n_j)=1\times 2+2\times 2+2\times 3=12$$
, 故 $C_C(B)=0.5$

$$\sum_{n_j \neq D} d(D, n_j) = 1 \times 3 + 2 \times 3 = 9$$
,故 $C_C(D) = 6/9 = 0.67$



接近中心性(续)

表 5: Closeness centrality for four graphs

star	ring	line	smoking
1	0.5	0.29	0.75
0.55	0.5	0.38	0.5
0.55	0.5	0.46	0.5
0.55	0.5	0.5	0.67
0.55	0.5	0.46	0.5
0.55	0.5	0.38	0.67
0.55	0.5	0.29	0.5

中介中心性

中介中心性 (betweenness centrality) 的测量逻辑:

- 如果某个行动者位于其它行动者的最短路径上,那么他就是 中心行动者
- 最短距离地位具有战略重要性

假定连接 j 和 k 的最短路径共有 g_{ik} 条,而其中包含节点 i 的 有 $g_{jk}(n_i)$ 条

第七章社会网络分析

$$C_B(n_i) = \frac{\sum_{j \le k} g_{jk}(n_i)/g_{jk}}{(g-1)(g-2)/2}.$$
 (4)

中介中心性(续)

在图 G 中,节点个数 g=7,除节点 i 以外,图 G 的路径最大数目为 (g-1)(g-2)/2=15

$$A$$
: $\sum g_{n_j-A} = 4 \times 2 + 1/3 = 8.33$,故 $C_B(A) = 8.33/15 = 0.56$

$$D\colon \sum g_{n_j-D}=2\times 1/2+2\times 1/2\times 2+1/2=3.5$$
 , 故 $C_B(D)=3.5/15=0.23$

$$E$$
: $\sum g_{n_s-E} = 1/3$, 故 $C_B(D) = 1/3/15 = 0.022$



中介中心性(续)

表 6: Betweenness centrality for four graphs

star	ring	line	smoking
1	0.2	0	0.56
0	0.2	0.33	0
0	0.2	0.53	0
0	0.2	0.6	0.23
0	0.2	0.53	0.022
0	0.2	0.33	0.23
0	0.2	0	0.022

中心性测度的比较(续)

表 7: A comparison of centralities for smoking network

	degree	eigen_centra	alit y loseness	betweenness
Α	0.67	1	0.75	0.56
В	0.33	0.58	0.5	0
C	0.33	0.58	0.5	0
D	0.5	0.79	0.67	0.23
E	0.33	0.58	0.5	0.022
F	0.5	0.79	0.67	0.23
G	0.33	0.58	0.5	0.022

有向关系

有向关系的社会网络中, 主要的三种声望测量:

- 度数声望(类似于度中心性)
- 邻近声望(类似于接近中心性)
- 地位或等级声望(类似于特征向量中心性)

吸烟行为的干预策略

开放讨论:

- 在给出的案例中,应当如何有效干预吸烟行为?
- 关系属性与行动者属性是如何协同发挥作用的?

中心性与市场策略

智能穿戴设备可以用作:

- 慢性病管理
- 生活与运动习惯监测

大多数智能手环、智能手表等产品销售状况并不理想。在市场营销中,为了促进新产品的扩散,企业可能采用"product seeding program",亦即:

 选取意见领袖作为"种子客户",免费向其提供新产品,以期 这些"种子客户"能够通过口碑效应,促进新产品的扩散。

智能穿戴设备扩散模型

对于任意行动者 i,购买智能穿戴设备的决策受到两个途径的影响:

• 大众传播: 广告等途径, 全局效果

● 口碑传播: 个体之间传播, 局部效果

智能穿戴设备扩散模型 (续)

如果第 t-1 期末行动者 i 尚未购买智能穿戴设备,那么他在第 i 期购买智能穿戴设备的概率是:

$$\mathsf{prob}_{it} = 1 - (1 - p) \times (1 - q)^{m_{it}},\tag{5}$$

其中:

- p: 创新系数,用以刻画大众传播
- q: 模仿系数, 用以刻画口碑传播
- m_{it} : 在第 t-1 期末,与行动者 i 直接相连的行动者中,已经购买了智能穿戴设备的数量

智能穿戴设备扩散模型(续)

如果行动者 i 在第 t 期购买了智能穿戴设备,则记作 $y_{it}=1$; 否则,记作记作 $y_{it}=0$ 。所以,每期的销售量为:

$$\mathsf{sales}_t = \sum_i y_{it}.\tag{6}$$

进一步,考虑到货币的时间价值,可以通过折现率 r 来计算企业通过销售智能穿戴设备获得的收益净现值:

$$\mathsf{npv} = \sum_{t} \frac{\mathsf{sales}_t}{(1+r)^t}.\tag{7}$$

仿真模型及分析

根据以上仿真模型,模拟智能穿戴设备的扩散过程。

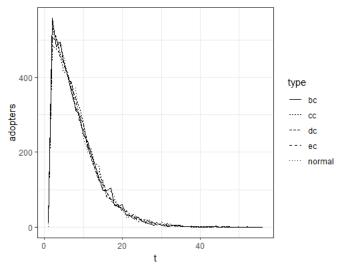
在 5000 个潜在客户中选取 10 个作为"种子客户",并考虑三种策略:

- 选择度中心性最大的 10 位
- 选择特征向量中心性最大的 10 位
- 选择接近中心性最大的 10 位
- 选择中介中心性最大的 10 位

模型的代码及实现请参阅:

cases/case-diffusion Rmd

智能穿戴设备扩散曲线



智能穿戴设备销售净现值

相应地得到销售净现值如下:

```
## normal dc ec cc bc
## 3443 3446 3442 3417 3460
```

考虑到免费向 10 位客户提供了智能穿戴产品,策略四是有效果的。

注: 需要多次重复 (repeat = 30) 求平均值

7.2.2 凝聚子群

社会网络的分析层次

● 行动者:中心性与声望

子群: 凝聚子群

● 网络:评估网络属性

理论背景

社会群体理论

- 结构化凝聚
 - 假设: 两个人存在正向互动时, 存在趋向一致的压力
 - 例子: 党同伐异
- 同质性
 - 社会规范: 凝聚导致同质性
 - 个体选择: 个体选择加入与自己类似的群体

社会群体理念

如何概念化社会群体?

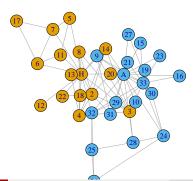
- 联系的交互性
- 子群成员的接近度或可及性
- 成员间联系的频率
- 与非成员相比, 子群成员联系的相对频率

凝聚子群分析方法

- 基于完全交互性
- 基于可及性和直径
- 基于节点度
- 凝聚程度的度量
- 图分割与层次聚类

空手道俱乐部网络

考虑分裂为两个派别的空手道俱乐部网络 karate, 两派领导为 Mr Hi 和 John A。



基于完全交互性

团 (clique)

- 社会学含义:在友谊选择中,由那些彼此相互选择的人们构成的,并且包含了所有与全体子群成员相互选择的人
- 图论定义:节点个数 $g_s \geq 3$ 的最大**完全**子图



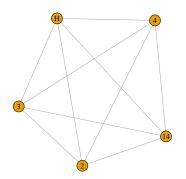
```
# summary of cliques
```

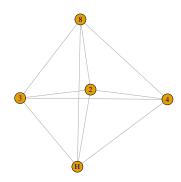
table(sapply(cliques(karate), length)) %>% par

• 5/34 vertices, named, from 4b458a1: [1] Mr Hi

72 / 99

团 (续)





团(续)

缺点:

- 定义过于严格
 - 任意一个联系缺失,则无法成团
 - 现实例子非常少
- 团之间不存在内在的区别
 - 在图论意义上, 都是完全子图
 - 无法探究团的特性带来的影响

改进:

基于可及性和直径

基于可及性,可以定义 n-团

- 基本假定
 - 重要的社会过程可以通过中间人发生
 - 子群成员间的距离是最短的
- 定义
 - 在图 G 中,子图中任意节点距离 $d(n_i, n_i) \leq n$

n-团

右图的 2-团包括:

- 1, 2, 3, 4, 5
- 2, 3, 4, 5, 6

缺陷:

- 节点 4 和 5 的最短路径包含了节点 6
- 节点 6 不在子群中



n-族和 n-社

反思:

- *n*-团作为子图, 其直径可能大于 *n*
- *n*-团可能是非连通的
- n-团未能达到我们希望的凝聚程度

改进:

- n-族: 在子图 G_s 中,任意节点距离 $d(n_i, n_i) \leq n$
- n-社: 直径为 n 的最大子图

n-族和 n-社 (续)

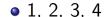
右图的 2-团包括:

- 1, 2, 3, 4, 5
- 2, 3, 4, 5, 6

右图的 2-族包括:

2, 3, 4, 5, 6

右图的 2-社包括:





基于节点度

基本假定的适用性:

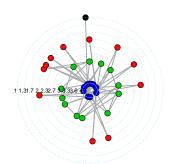
- 可及性: 重要的社会过程可以通过中间人发生(信息与资源 传播)
- 邻接性: 重要的社会过程需要直接接触 (团体内的知识学习)

基于节点度的子群:

- 基本假定: 行动者与子群内相当数量的成员相邻接
- 现实含义:多重冗余的沟通渠道,子图的"脆弱性"问题(星形网络)
- k-丛 (k-plex): 子图 G_s 中 $d_s(n_i) \geq g_s k$

k-核与可视化

核数 (coreness) 为 1 (黑色)、2 (红色)、3 (绿色)、4 (蓝色)



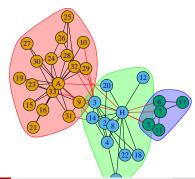
凝聚程度的度量

度量的思路:

- 内外联系的比较
 - 子群内联系集中
 - 子群内外联系的强度或频率之比较大
- 健壮的连诵性
 - 凝聚子群在连通性方面是健壮的(有益的冗余)
 - 移除一定数量的边之后, 子群依然是连通的

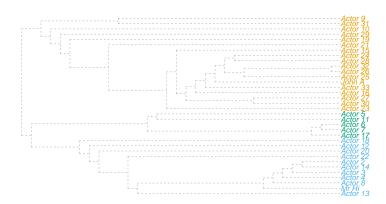
图分割与层次聚类

对空手道俱乐部网络进行层次聚类,发现 3 个社团 (communities),其大小分别为 18、11 和 5。



图分割与层次聚类 (续)

采用树状图展示:



7.2.3 评估网络属性

真实网络的属性:

• 度分布: 幂律

• 聚类系数: 较高

• 平均路径长度: 较短

度分布

真实网络的节点度通常满足**幂律分布**,即度为 k 的节点在网络中的比例为

$$p_k = ak^{-b} (8)$$

或者得到

$$\ln(p_k) = -b\ln(k) + \ln(a). \tag{9}$$

符合幂律分布的网络称之为无标度网络。

聚类系数

聚类系数 (clustering coefficient) 定义为

$$cl_T(G) = \frac{3\tau_{\Delta}(G)}{\tau_3(G)},\tag{10}$$

其中 $au_{\Delta}(G)$ 是图 G 中三角形的个数,而 $au_3(G)$ 为连通的三元组 (即由两条边连接的三个节点,亦即 2-star 网络) 的个数。

聚类系数衡量了"传递三元组"的比例。

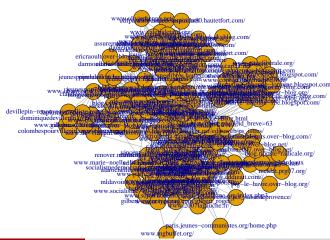
平均路径长度

平均路径长度为

$$\bar{d} = \frac{\sum_{i \neq j} d(n_i, n_j)}{g(g-1)}.$$
 (11)

真实网络的平均路径长度大多在 4-6 之间。

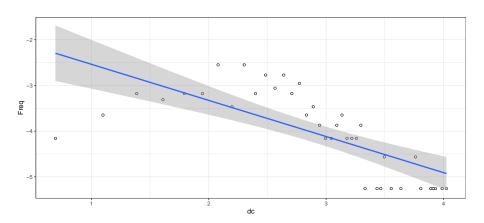
真实网络案例



第七章社会网络分析

度分布

法国的博客网络 fblog, 包含 192 个节点和 1431 条边。



估计幂律指数

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Tue, Oct 22, 2019 - 9:03:11

表 9

	Dependent variable:	
	Freq	
dc	-0.790^{***}	

OCT 21 - 24, 2019

主要参数

我们计算三个主要参数:

- 平均度
- 平均聚类系数
- 平均路径长度

	clustering	
degree	coefficient	distance
14.91	0.3858	2.539

随机网络

设置 n=192, p=15/192=0.078, 创建随机网络。

进而计算三个主要参数。

	clustering	
degree	coefficient	distance
15.14	0.07586	2.198

随机网络:

- 度不是幂律分布
- 聚类系数讨低

小世界网络

设置 n=192, 重链概率 $\beta=0.078\in(0.01,0.1)$, 创建小世界网络。

进而计算三个主要参数。

	clustering	
degree	coefficient	distance
16	0.4434	2.479

小世界网络:

• 度不是幂律分布

优先连接网络

设置 n=192,幂律指数 b=0.79,创建优先连接网络。

进而计算三个主要参数。

	clustering	
degree	coefficient	distance
15.62	0.1567	2.202

优先连接网络:

• 聚类系数过低

典型网络的属性比较

我们最后比较典型网络的主要属性:

	degree	clustering coefficient	distance
fblog	15	0.39	2.5
random	15	0.076	2.2
sw	16	0.44	2.5
pa	16	0.16	2.2

案例 (2 个课时)

本节知识点

- SNA 与文献分析
- SNA 与健康行为分析

7.3.1 医学领域案例: 文献分析

- CiteSpace 中文版指南
- CiteSpace 讲义

7.3.2 医学领域案例: 行为分析

● 孟加拉国 Dhaka 城市贫民社区中青年的精神健康问题研究