社会网络中的实验与R

Analyzing Social Networks Experiments in R

陈丽云

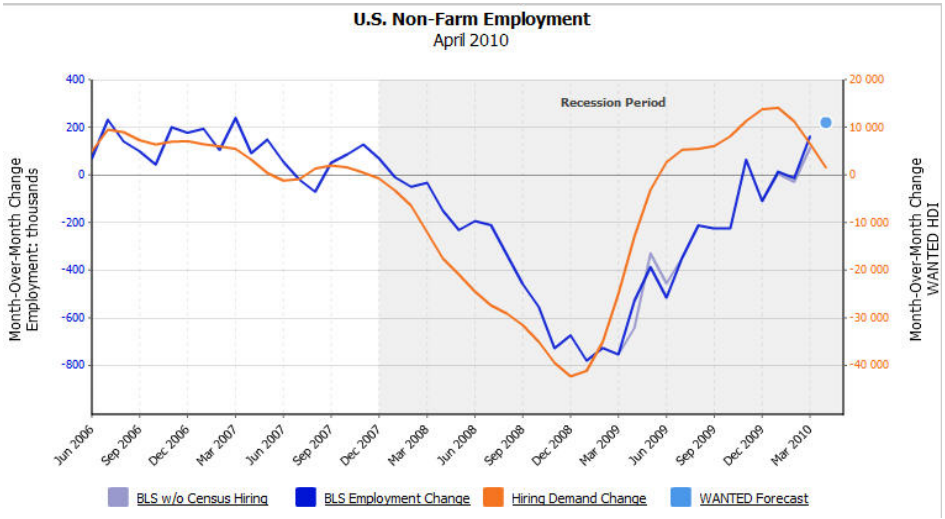
上海河广、统计之都

第四届中国R语言会议（上海会场），2011年11月

The 4th China-R Conference, ECNU, Shanghai

Outline

Correlation v.s. Casuality



- 当有两个变量呈现一致的变化趋势的时候……
  - 到底是X导致了Y，还是Y导致了X?
  - 还是第三个变量从根源上主导了这两个变量的变化?
- 探寻因果关系：更好的估计方法 (Nobel Prize: VAR模型) VS 更高质量的数据(实验数据)

数据→模型，模型→数据

- The object of science is the discovery of relations.., of which the complex may be deduced from the simple.
  John Pringle Nichol, 1840
  ⇒ 高维数据降维：Data Mining, Econometrics, Other Statistical Models.
- 现有的数据不足以完美回答我们关心的问题：内生性、数据的不可直接观测性。
- 与其寻求更好的估计方法，不如寻找更高质量的数据。

——Handbook of Econometrics

数据→模型，模型→数据

- The object of science is the discovery of relations.., of which the complex may be deduced from the simple.
  John Pringle Nichol, 1840
  ⇒ 高维数据降维：Data Mining, Econometrics, Other Statistical Models.
- 现有的数据不足以完美回答我们关心的问题：内生性、数据的不可直接观测性。
- 与其寻求更好的估计方法，不如寻找更高质量的数据。

——Handbook of Econometrics

数据→模型，模型→数据

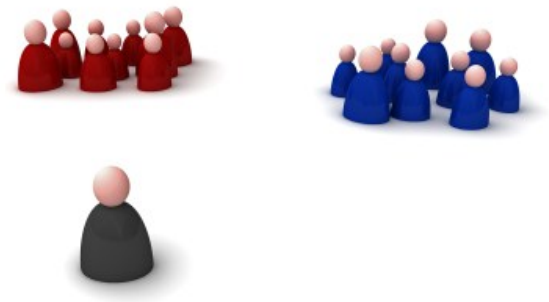
- The object of science is the discovery of relations.., of which the complex may be deduced from the simple.
  John Pringle Nichol, 1840
  ⇒ 高维数据降维：Data Mining, Econometrics, Other Statistical Models.
- 现有的数据不足以完美回答我们关心的问题：内生性、数据的不可直接观测性。
- 与其寻求更好的估计方法，不如寻找更高质量的数据。

——Handbook of Econometrics

现有数据的不完美

- 数据源有限的原因:
  - 历史上的: 几十年前的事情。
  - 本身的性质: 宏观里面常用的gdp等东西。
  - 业界数据: 隐私、API。
  - 成本太高: 全民普查、小规模样本。
  - 非数字记录: 微博、文本挖掘。

社会网络实证研究



- 一个圈子中间各个成员之间的相似性来源于什么?
  - 圈子形成过程的自选择（self-selection）：物以类聚、人以群分?
    →网络构成模型（Network Formation Models）
  - 圈子形成之后的同伴效应（peer effects）：近朱者赤、近墨者黑?
    →社会性学习模型（Social Learning Models Information Spread Models, Social Influencer Models）

豆瓣的例子

- 结果:
  - 用户对于其现有好友评分为1星的书籍，相比于其未来好友评分为1星的书籍，会更倾向于给前者较低的评分；对于现有好友评分较高的如5星的书籍，会相比而言给予更高的评分。
  - 朋友之间影响更强烈的情况：热门书籍；评分较晚的用户；使用时间较短、阅读经验较少的用户；小众圈子用户。

Source: Wang, Alex, Xiaoquan (Michael) Zhang and Il-Horn Hann, 2010, "Social Bias in Online Product Ratings" Workshop on Information Systems and Economics (WISE), December 2010, St. Louis, USA.

Outline

美国高中生性知识教育项目

- 美国青少年艾滋病传播情况：2006年到2009年期间，超过 8,500例新增艾滋病人为13-19岁的青少年。
- 传统性知识教育方式：课堂教育，受诸如宗教信仰、老师讲演水平限制很大。多数青少年不懂的如何利用正确的方式保护自己，不会主动和正确的使用安全套等基本预防手段。
- 欲尝试的新方式：通过青少年网络进行的知识传播——青少年之间性已经是不可避免的话题，相互之间的知识和态度共享非常普遍。
  - 对青少年朋友网络进行调查，选取其中的关键人物和活跃人物。
  - 对活跃人物进行性知识普及教育，引导其正确使用相关措施。
  - 鼓励活跃人物主动向朋友们传播知识。

Outline

费歇尔（Fisher）实验设计三原则

- 费歇尔（Fisher）三原则:
  1. 随机化原则 :
     - 样本偏差
     - 实现方法：随机抽签
  2. 重复原则:
     - “欧洲研究人员发现了难以解释的中微子超光速现象”：“参与实验的瑞士伯尔尼大学的安东尼奥·伊拉蒂塔托说，他和同事被这一结果震惊了，他们随后反复观测到这个现象1.6万次，并仔细考虑了实验中其他各种因素的影响，认为这个观测结果站得住脚，于是决定将其公开”。
  3. 区组化原则:
     - 系统误差
     - 实现方法：分层后随机抽签

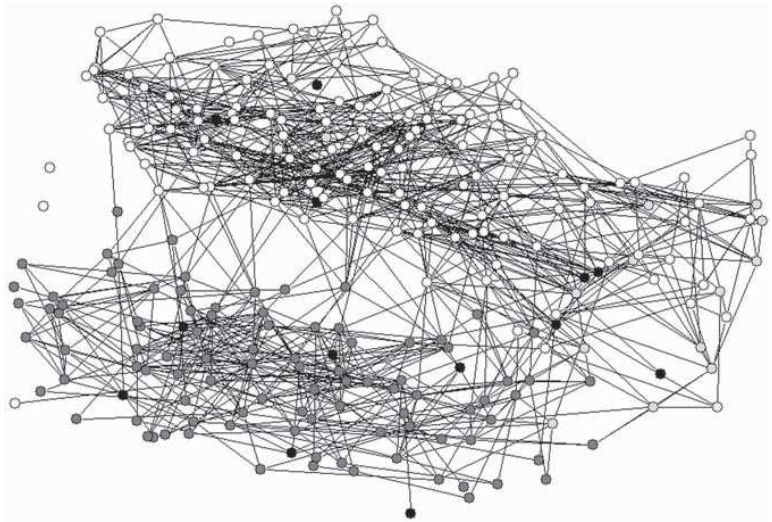
随机化社会实验设计

- 社会实验的特殊性：实验对象为人，关注的是人们的行为决策过程和结果。
- 社会实验的挑战——难以进行独立重复实验：
  - 两群完全相同的人?
  - 人的学习和记忆行为。
- 社会实验的道德约束：
  - 随机抽取还是按需分配?
  - 小额贷款实践
  - 疫苗、书籍等牵涉到一代人命运的分配更是如此。

Outline

美国高中生朋友网络形态

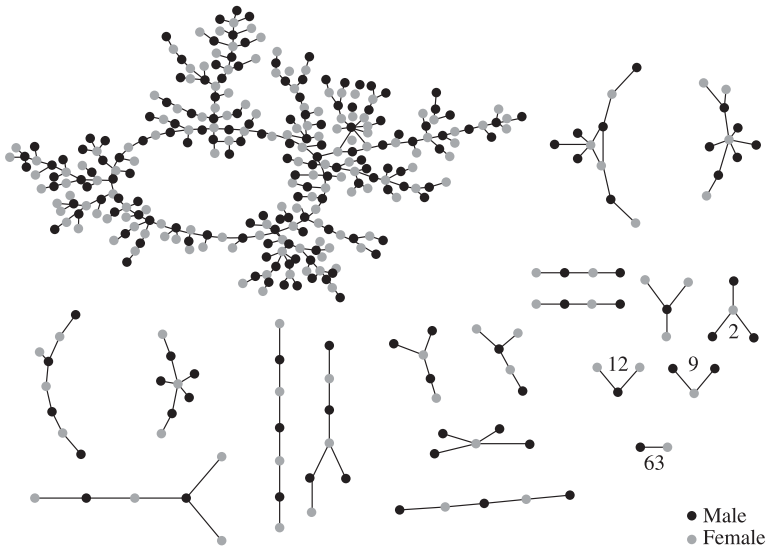
Figure: 美国高中生朋友网络形态



数据来源：ADD Health Database

美国高中生性网络形态

Figure: 美国高中生性关系网络形态



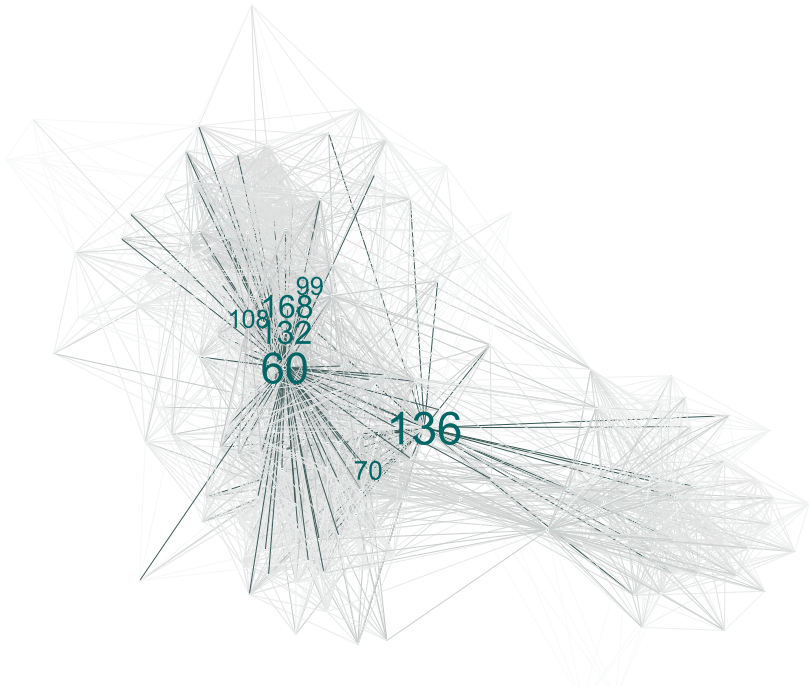
数据来源：ADD Health Database

Outline

美国高中生性知识教育项目：数值模拟

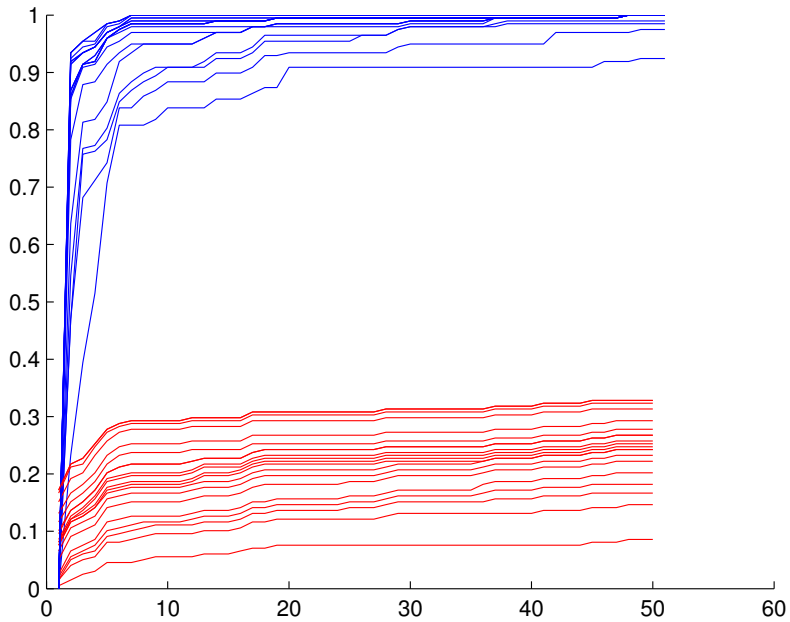
- 数值模拟思路：按照社会网络学习模型，模拟没有外界干涉情况下网络自发的艾滋病传播情况。

Figure: 社会网络学习数值模拟网络



美国高中生性知识教育项目：数值模拟

Figure: 社会网络学习数值模拟结果



Time periods, $t = 50$; *Repetitions* $= 20$; Number of nodes, $N = 198$
Red: proportion of nodes that are being infected
Blue: average risk perception in the network

Outline

美国高中生性知识教育项目：实验设计思路

Figure: 随机分组示意图

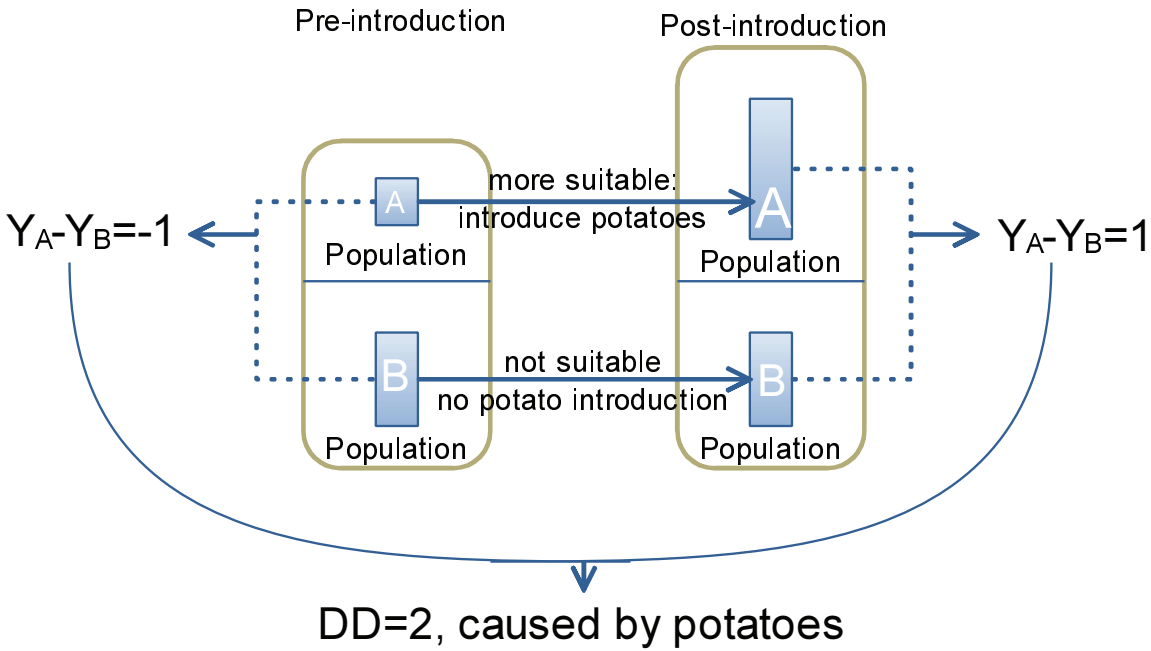
- 评价标准：一段时间之后的问卷调查和知识水平测试；HIV携带率、怀孕率等统计指标

Outline

实验结果分析方法

- 估计方法:
  - 随机分组实验: DID Difference in Difference) with Probit/Logit Models

Figure: DID 示意图



Outline

实验结果分析在R中的实现

优秀的数据源——DID估计是一致的:

$$Infected = \alpha + \beta_1 Network + \beta_2 Traditional + \beta_3 NT + \varepsilon$$

- 数据清理和基本分组统计: `data.table`包
- DID分析: `lm()` `glm()`; `VGAM package - probit() logit()`

Generalized Linear Regression

```
mylogit<- glm(infected~network+traditional+NT,
family=binomial(link="logit"), na.action=na.pass, data=mydata)
summary(mylogit)
library(aod)
wald.test..
#VGAM for multinominal cases
fit.ms <- vgam(infected_level ~ network+traditional+NT,
multinomial(refLevel=1), data = nzmarital)
```

谢谢！