

社会计算

童超

计算机学院

联系方式

- 实验室 : 北航计算机学院分布与移动计算实验室
- 地点 : 新主楼G1029房间
- 电话 : 010 - 82316575 (O)
- Email : tongchao@buaa.edu.cn
- 课程网址 : <http://course.buaa.edu.cn/portal/site/865ce42f-5fbb-4ec0-b930-4f47db063243>
- 百度网盘 : <http://pan.baidu.com>

博客 : <http://blog.sina.com.cn/buaasocialcomputing>

百度网盘 : 下载课件和参考资料 , 提交作业

博客 : 提交博文

用户名 : buaasc2013@sina.com

密码 : bUaasocial2013

考核方式

- 考试（闭卷笔试） 50%
 - 课后作业 30%
 - 1篇课程报告 20%
-
- 命名规则：班号-学号-姓名.pdf docx

一本优秀的跨学科教材

Over the past decade there has been a growing public fascination with the complex connectedness of modern society. This connectedness is found in many incarnations: in the rapid growth of the Internet, in the ease with which global communication takes place, and in the ability of news and information as well as epidemics and financial crises to spread with surprising speed and intensity. These are phenomena that involve networks, incentives, and the aggregate behavior of groups of people; they are based on the links that connect us and the ways in which our decisions can have subtle consequences for others.

The introductory undergraduate textbook takes an interdisciplinary look at economics, sociology, computing and information science, and applied mathematics to understand networks and behavior. It describes the emerging field of study that is growing at the interface of these areas, addressing fundamental questions about how the social, economic, and technological worlds are connected.

"The field of information networks is an emerging discipline of immense importance that combines graph theory, probability and statistics, microeconomics, and facets of the social sciences. Easley and Kleinberg present a panoramic view of this field, from basic graph theory all the way to the state of the art in research."

— Prabhakar Raghavan, Head of Yahoo! Labs

"David Easley and Jon Kleinberg have given us a totally new kind of basic economics text, where students learn how to analyze social networks and crowds as well as games and markets. This book covers a remarkable range of topics and offers a broad new vision of what economics can be about."

— Roger Myerson, Glen A. Lloyd Distinguished Service Professor of Economics at the University of Chicago and winner of the 2007 Nobel Prize in Economics

"In my three decades plus of teaching, I cannot recall an urge to teach a new undergraduate course like the one I felt upon leafing through this wonderful introduction to everything that is new and important and intellectually challenging in our world."

— Christos Papadimitriou, C. Lester Hogan Professor of EECS, University of California, Berkeley

"In this remarkable book, David Easley and Jon Kleinberg bring all the tools of computer science, economics, and sociology to bear on one of the great scientific challenges of our time: understanding the structure, function, and dynamics of networks in society. Clearly written and covering an impressive range of topics, Networks, Crowds, and Markets is the ideal starting point for any student aspiring to learn the fundamentals of the emerging field of network science."

— Duncan Watts, Principal Research Scientist, Yahoo! Research, and author of *Six Degrees: The Science of a Connected Age*

"Networks are everywhere, in our social lives, in our economic relations, and in nature; they are now finally arriving to our classrooms. Easley and Kleinberg have written a masterful introduction to networks. This book successfully combines the game-theoretic and algorithmic approaches to the study of social, economic, and communication networks. It is a lively, interesting, readable, and accessible. It is a pleasure to teach using this book and never a dull moment for the students."

— Daron Acemoglu, Charles P. Kindleberger Professor of Applied Economics, Massachusetts Institute of Technology

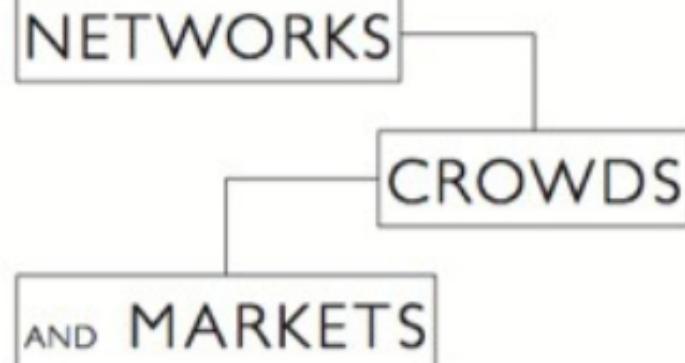
CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-0-521-79535-1

EASLEY
KLEINBERG

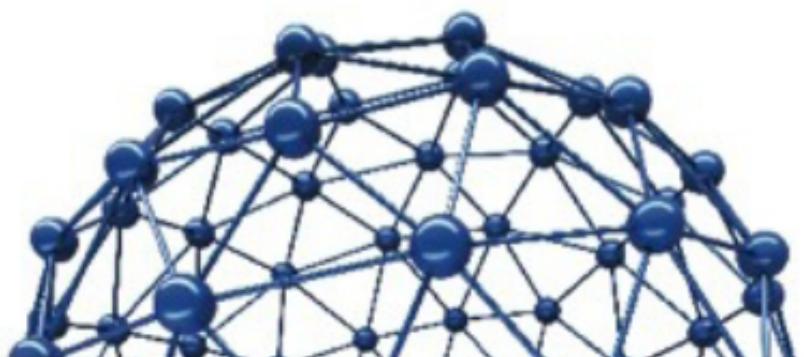
NETWORKS, CROWDS, AND MARKETS

CAMBRIDGE



Reasoning about a Highly Connected World

DAVID EASLEY
and
JON KLEINBERG





信息级联
(information cascades)

(基于第16章)

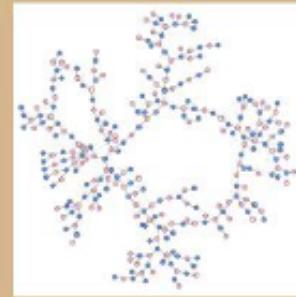


文件(F) 编辑(E) 查看(V) 收藏夹(A) 工具(T) 帮助(H)

★ 收藏夹 | Internet Explorer... 建议网站 免费 Hotmail 获取更多加载项

② Networks (Econ/Soc/CS/Info , D...

书签(F) 图片(I) 打印(P) 页面(E) 安全(S) 工具(Q) ?



High School Dating
(Bearman, Moody, and Stovel, 2004)
(Image by Mark Newman)



Corporate E-Mail Communication
(Adamic and Adar, 2005)

Networks

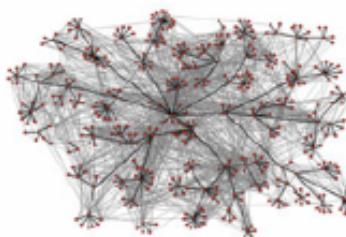
Economics 2040 / Sociology 2090 / Computer Science 2850 / Information Science 2040
Cornell University, Fall 2010
Mon-Wed-Fri 11:15-12:05

David Easley (Economics) and Jon Kleinberg (Computer Science)

The class meets in Statler Auditorium (Mon/Wed) and Ives 305 (Fri)

A course on how the social, technological, and natural worlds are connected, and how the study of networks sheds light on these connections. Topics include: how opinions, fads, and political movements spread through society; the robustness and fragility of food webs and financial markets; and the technology, economics, and politics of Web information and on-line communities.

The course is designed at the introductory undergraduate level with no formal prerequisites; it satisfies the Arts & Sciences Social and Behavioral Analysis (SBA) distribution and the Engineering Liberal Studies (SBA group) distribution. (See also the poster announcing the course.)



HONORS 493 Introduction to networks

Resources

syllabus

[home](#)

[cTools site](#)

instructor:

Lada Adamic

ladamic@umich.edu

Networks are everywhere: from social networks to the Web, from food webs to corporate ownership structures. This course will introduce you to the basic concepts of network analysis and applies them, using software tools, to a variety of real-world examples. We will learn how to represent networks, how to measure properties of networks, how to compare different networks, and how to model how networks grow and evolve and how processes such as information diffusion and disease spread through them, depending on the structure of a social network.

learning objectives:

- empirically analyze real-world networks: Which nodes are central? Is there local or global connectivity?
- understand how networks form
- model processes on networks

Schedule

Fall 2009:

Tues&Thurs 10:00-11:30

in WH 317A

6 week course: Sept. 8 -

Oct. 15th

grading scheme:

40% (4 homework assignments)

40% final

10% in class/online participation

10% student demo

reading:

PDFs of relevant book chapters and articles will be available on cTools.

By Jure Leskovec

STANFORD
UNIVERSITY

CS224W: Social and Information Network Analysis Autumn 2011

- Home
- Handouts
- Course info
- FAQ
- Resources

World Wide Web, blogging platforms, instant messaging and Facebook can be characterized by the interplay between rich information content, the millions of individuals and organizations who create and use it, and the technology that supports it.

The course will cover recent research on the structure and analysis of such *large social and information networks* and on models and algorithms that abstract their basic properties. Class will explore how to practically analyze large scale network data and how to reason about it through models for network structure and evolution.

Topics include methods for link analysis and network community detection, diffusion and information propagation on the web, virus outbreak detection in networks, and connections with work in the social sciences and economics.

• Announcements:

- 9/29: Review of probability (Gates B01, 4~6pm) [[Slides](#)]
- 9/30: Review of basic linear algebra (Gates B03, 4~6pm) [[Slides](#)]
- 10/6: Introduction to SNAP (Gates B01, 4~6pm) [[Slides](#)] [[Website](#)]
- 10/6: First homework comes out [[Homework 1](#)] [[Cover sheet](#)]
- 10/7: Introduction to NetworkX (Gates B03, 4:15~6:15pm) [[Slides](#)]
- 10/13: Homework 1 Due (9:30am, at class)
- 10/20: Reaction Paper Due

UNIVERSITY OF
CAMBRIDGE

Search



Contact us | A-Z | Email & phone search

Computer Laboratory

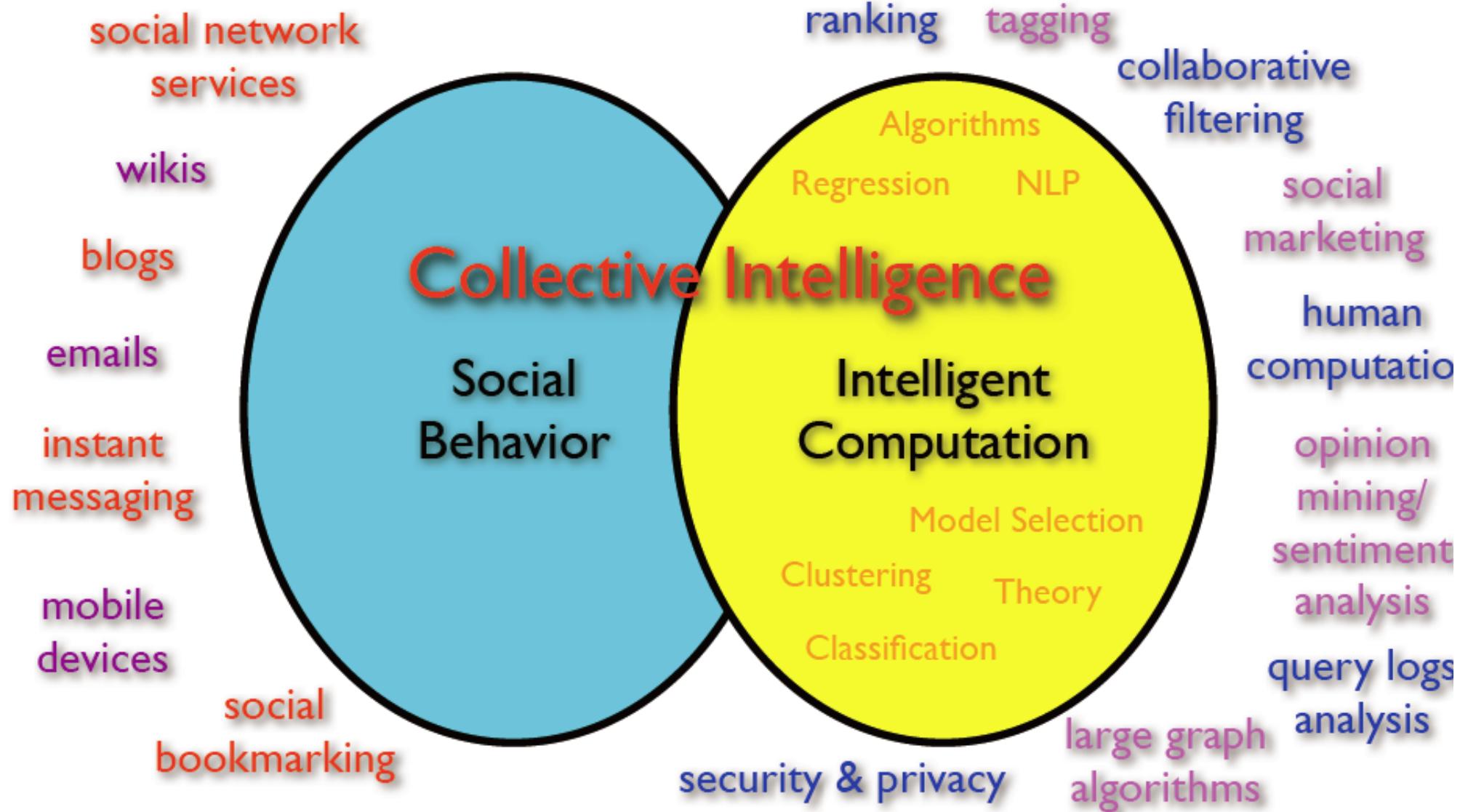
University of Cambridge > Computer Laboratory > Research > Networks and Operating Systems Group

[Home](#)[Profile](#)[Publications](#)[Group](#)[Professional Activities](#)[Projects](#)[Teaching](#)

Social and Technological Network Analysis (Lent Term 2011-2012)

[Lecture 1: Networks and Random Graphs](#)[Lecture 2: Small World and Weak Ties](#)[Lecture 3: Network Centrality and Community Detection](#)[Lecture 4: Modularity and Overlapping Communities](#)[Lecture 5: Structure of the Web and Power Laws](#)[Lecture 6: The Internet and Robustness](#)[Lecture 7: Information Cascades on Networks](#)[Lecture 8: Epidemic Dissemination on Networks](#)[Lecture 9: Cascades and Epidemics: Applications](#)[Lecture 10: Temporal Social Networks](#)[Lecture 11 and 12: Practical NetworkX Analysis of Datasets](#)

Social Computing



Social Computing (SC)

- Social computing is a general term for an area of computer science that is concerned with the intersection of *social behavior* (社会行为) and *computational systems*. Wikipedia
- A *social structure in which technology puts power in communities, not institutions.* Forrester
- *Forms of web services where the value is created by the collective contributions* (集体贡献) *of a user population.*

Issues

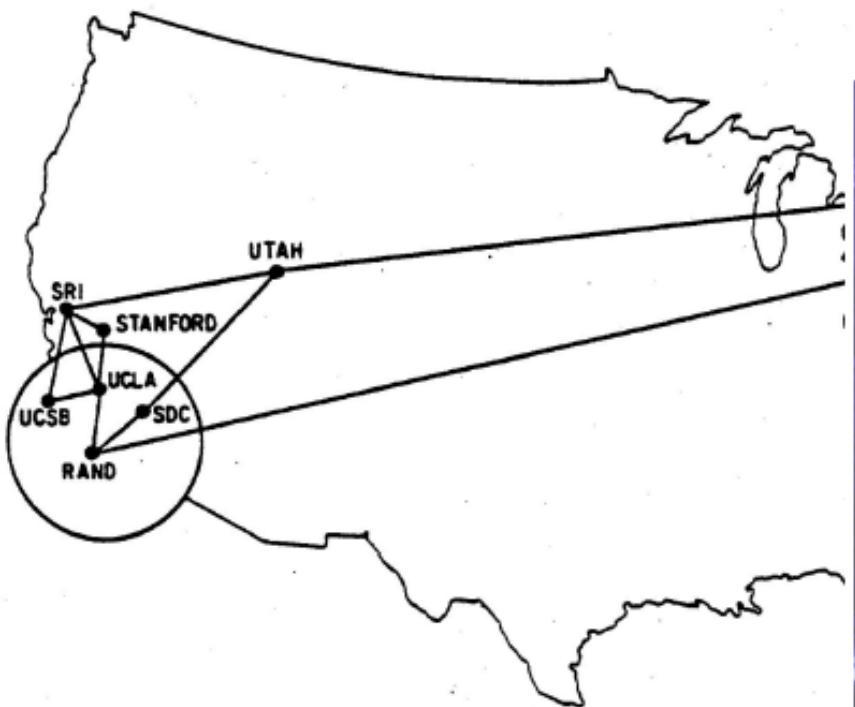
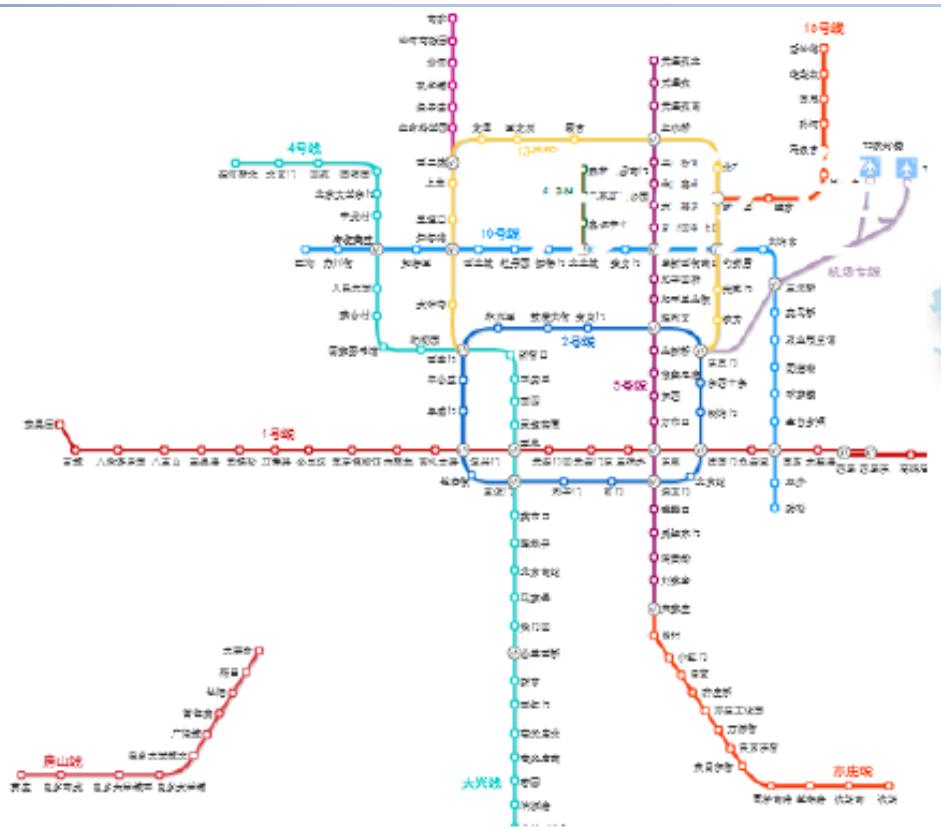
- Theory and models (理论与模型)
- Mining (数据挖掘) of existing information, e.g., spatial (relations) and temporal (time) domains
 - Dealing with partial and incomplete (局部和不完整) information, e.g., collaborative filtering, ranking, tagging, etc.
- Scalability and algorithmic (可扩展性和算法) issues
- Security and privacy (安全和隐私) issues
- Monetization (货币) of social interactions

概述 图论

课程概要

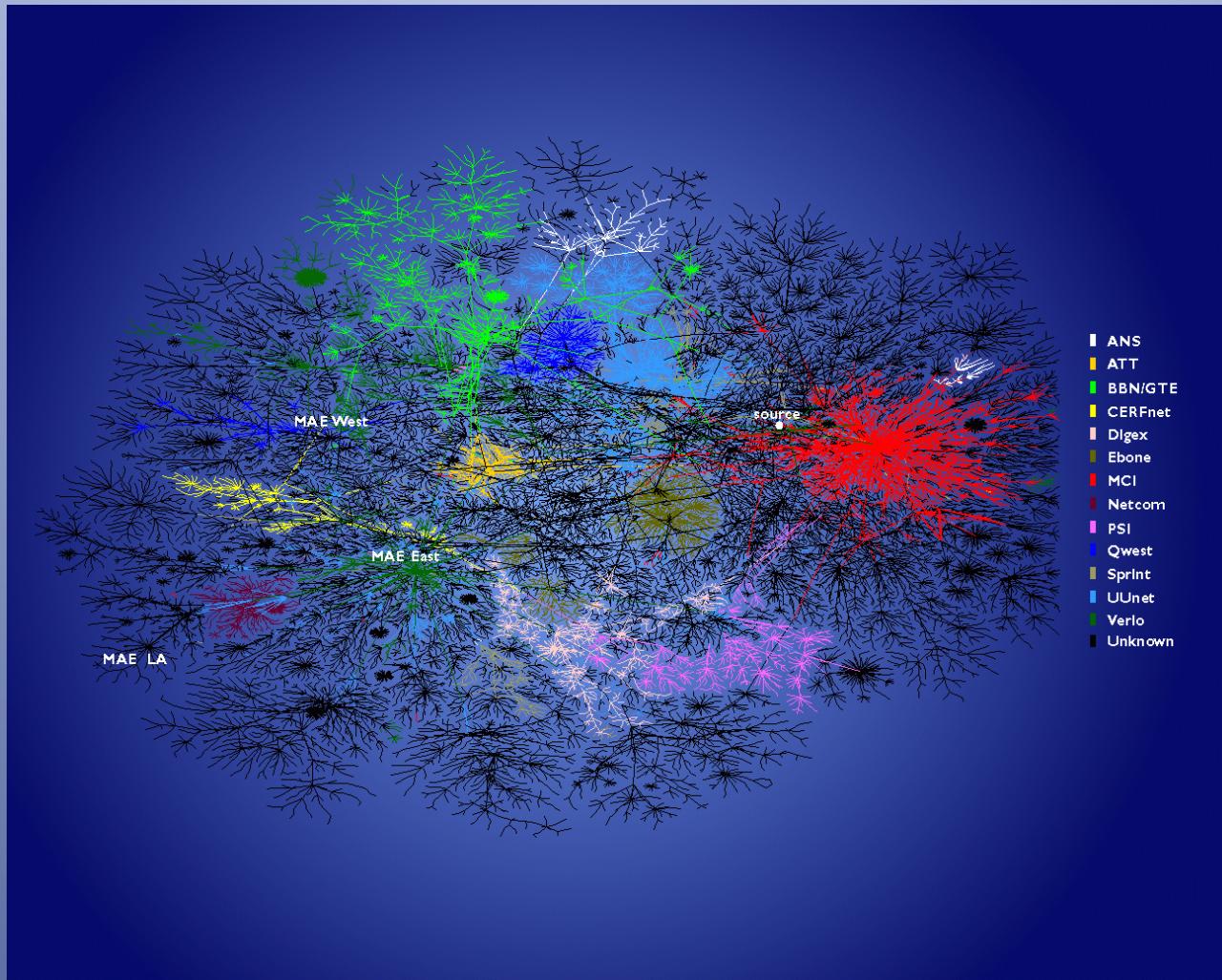
- 出发点
 - 网络，一个已经渗透到人类生活方方面面的概念，理解其作用原理不仅必要且已有可能
- 主要内容
 - 讨论网络的语言（基础）
 - 网络结构分析
 - 网络行为推理
- 学习预期
 - 能学到什么？被要求什么？

针对若干经济学与社会学经典问题



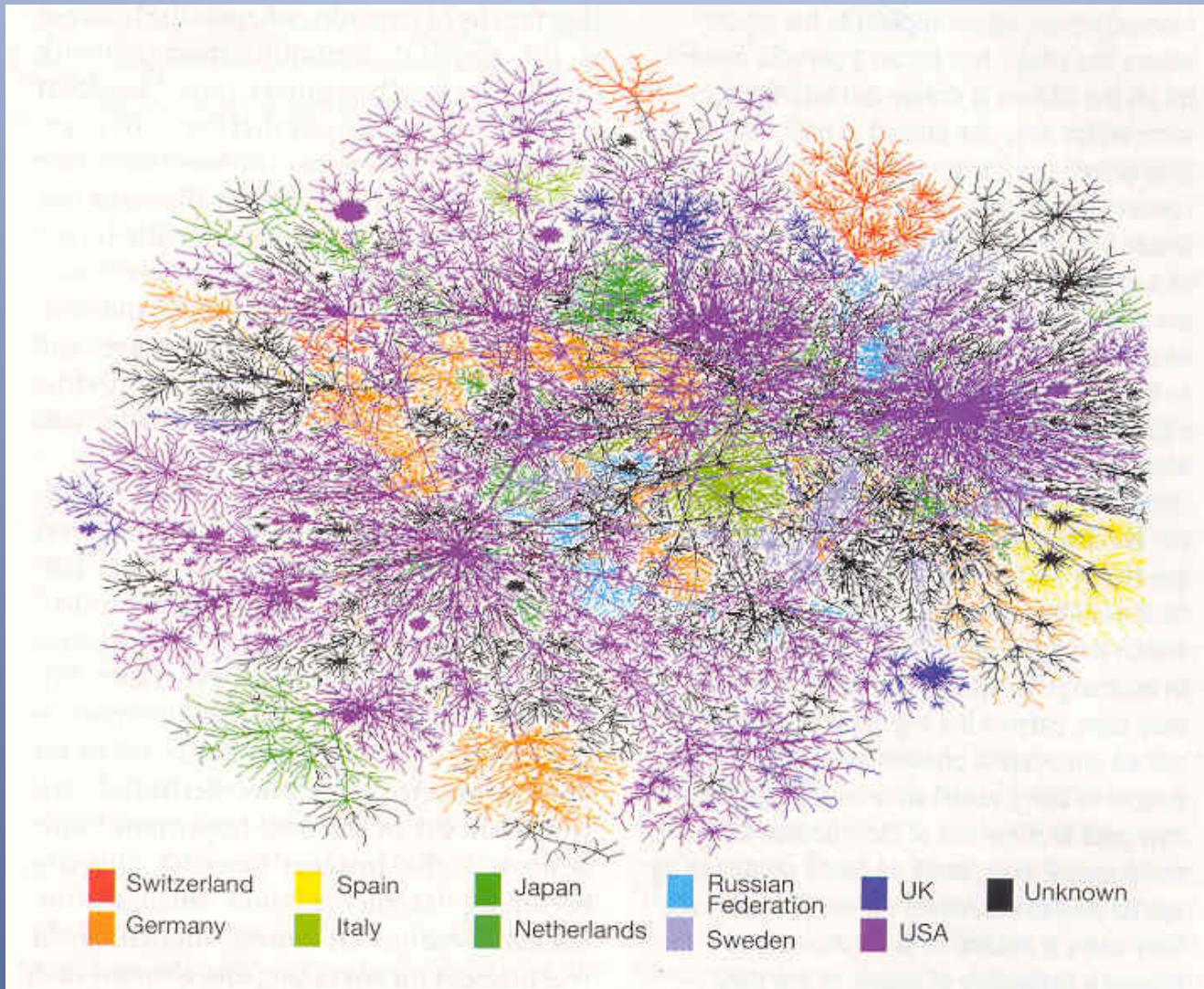
Internet

(K. C. Claffy)

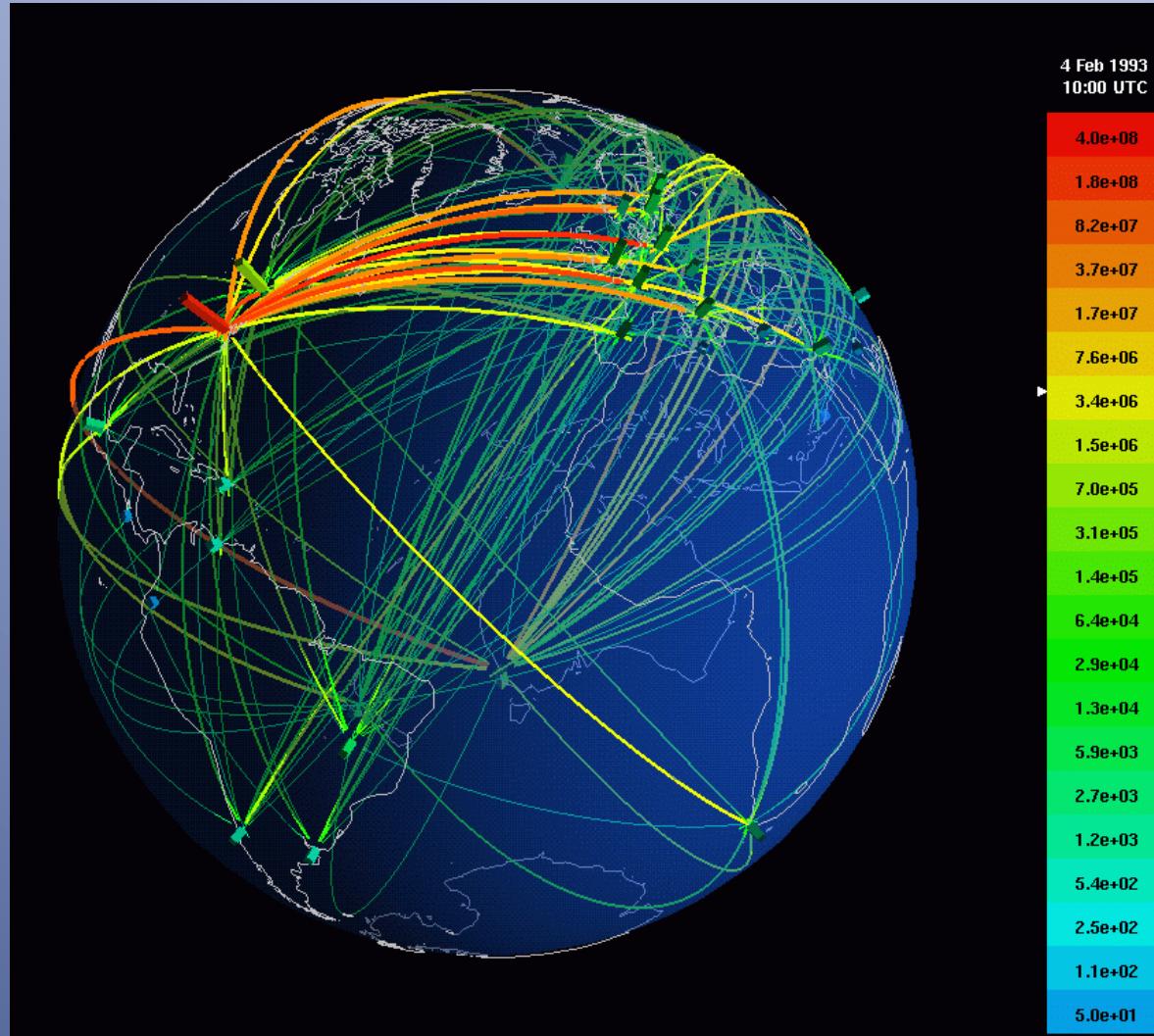


万维网World Wide Web(www)

(William R. Cheswick)

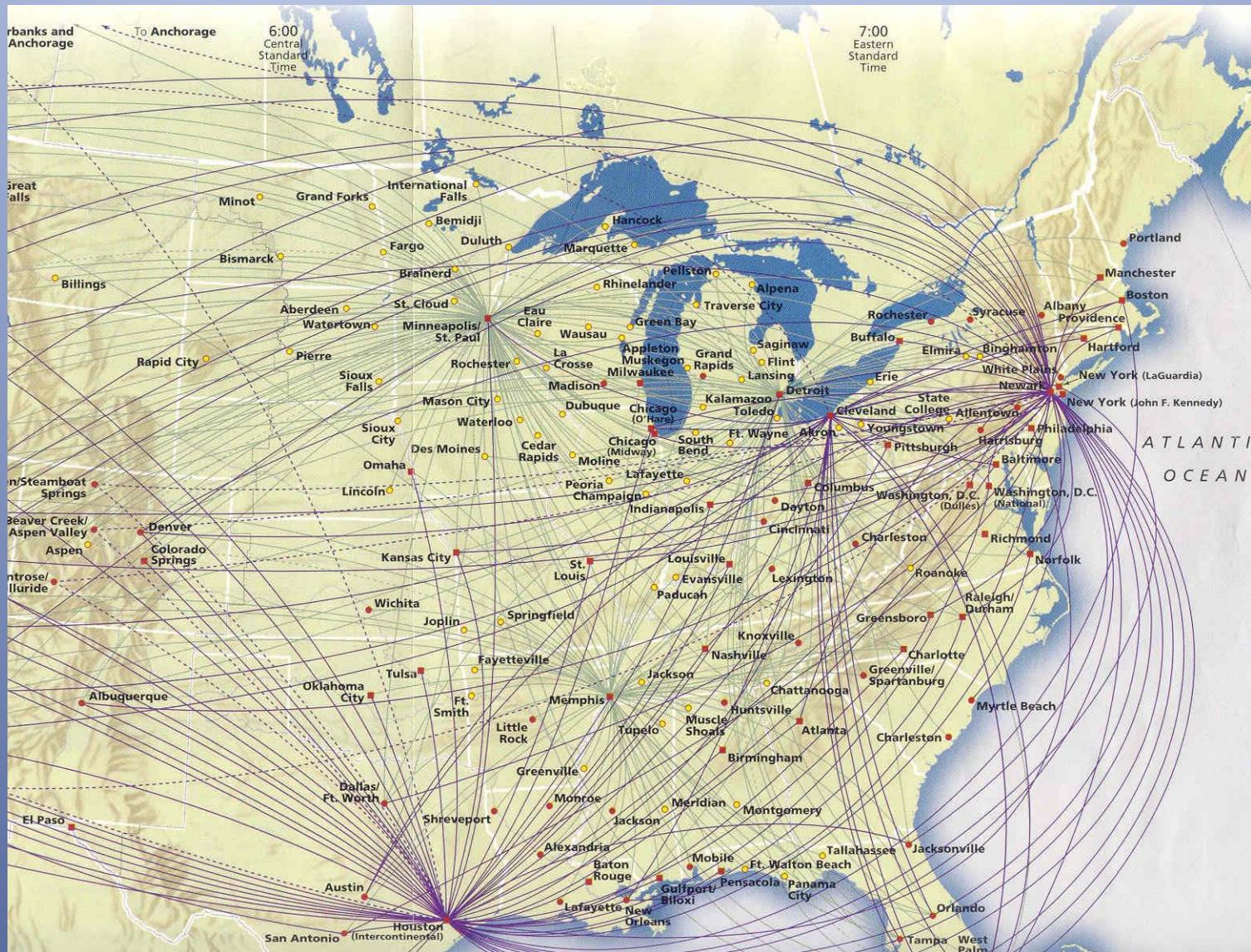


电信网络

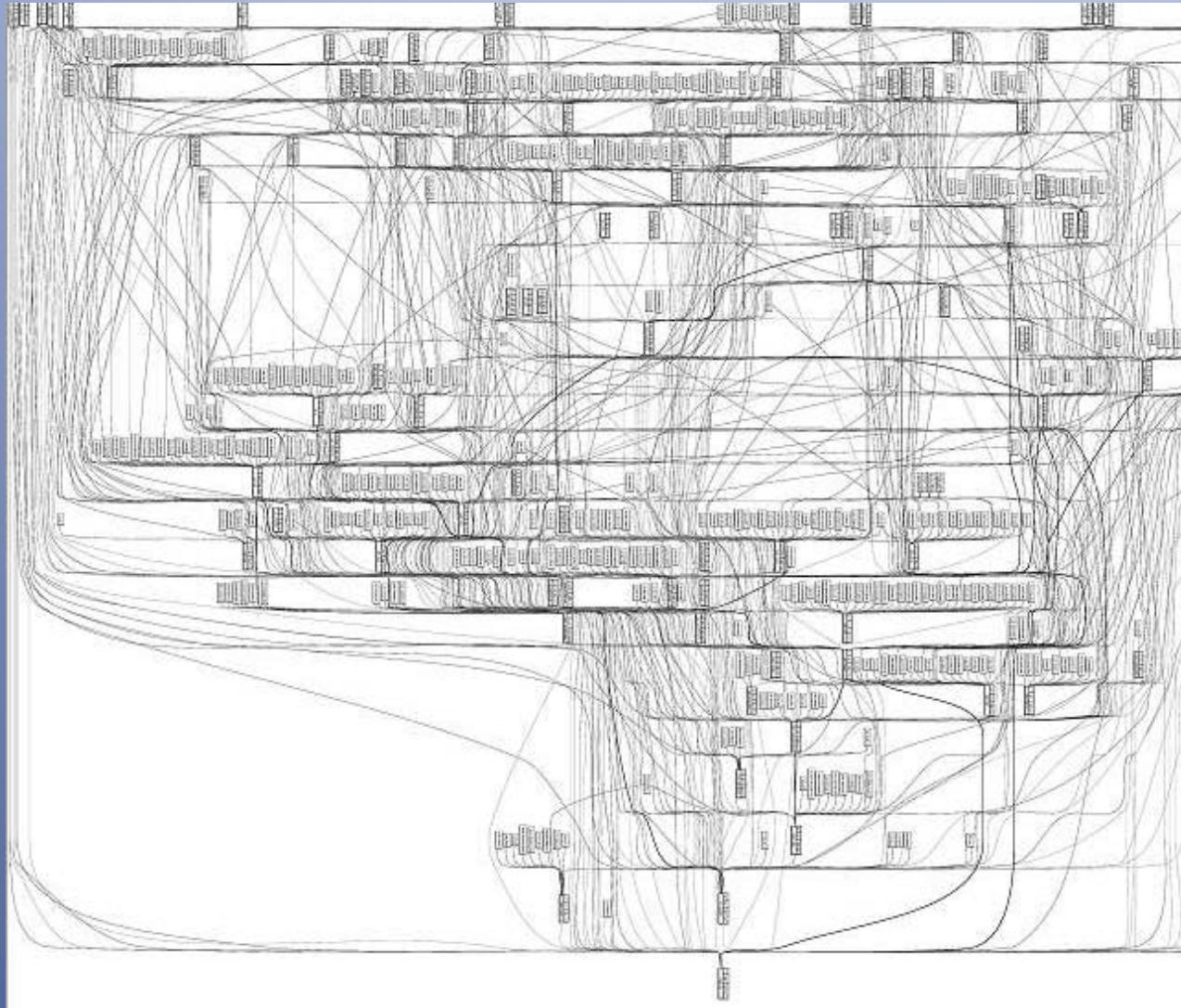


(Stephen G. Eick)

美国航空网

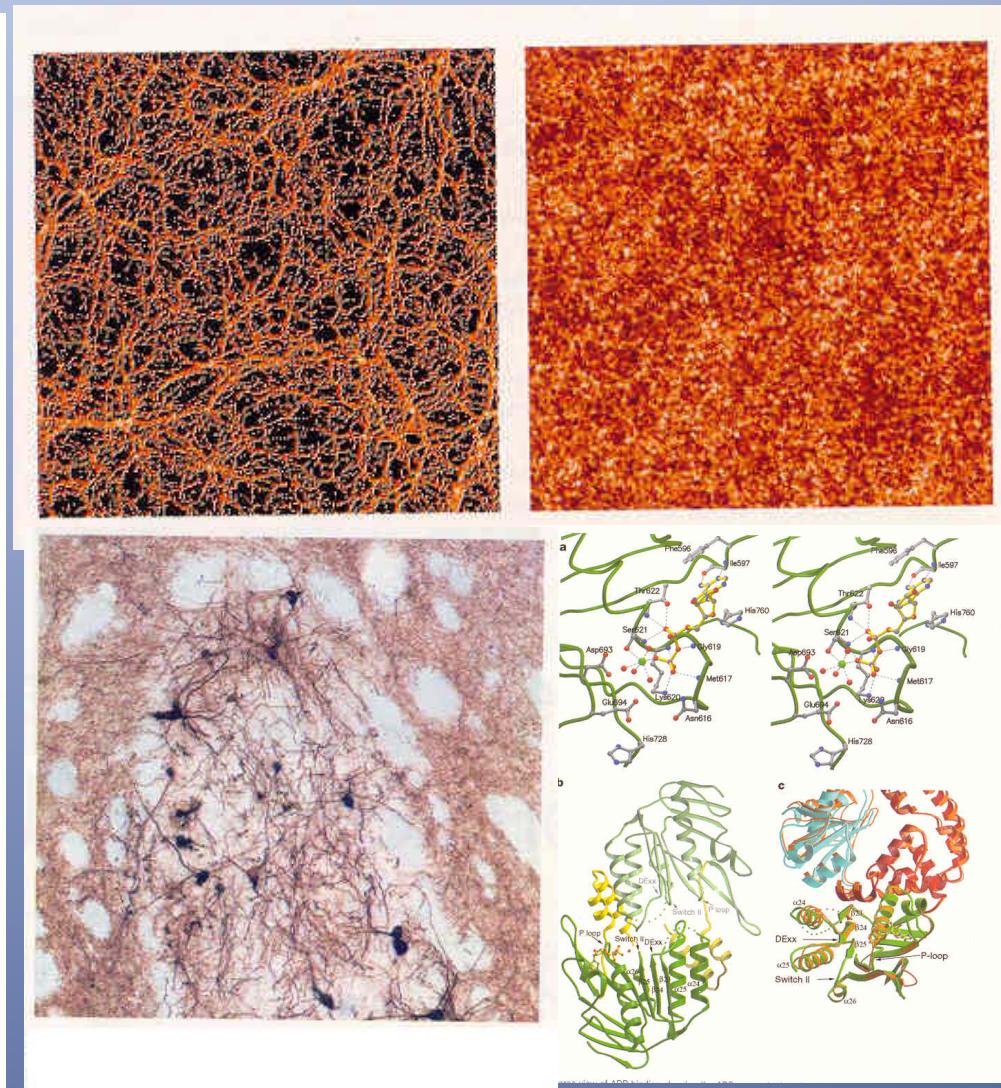
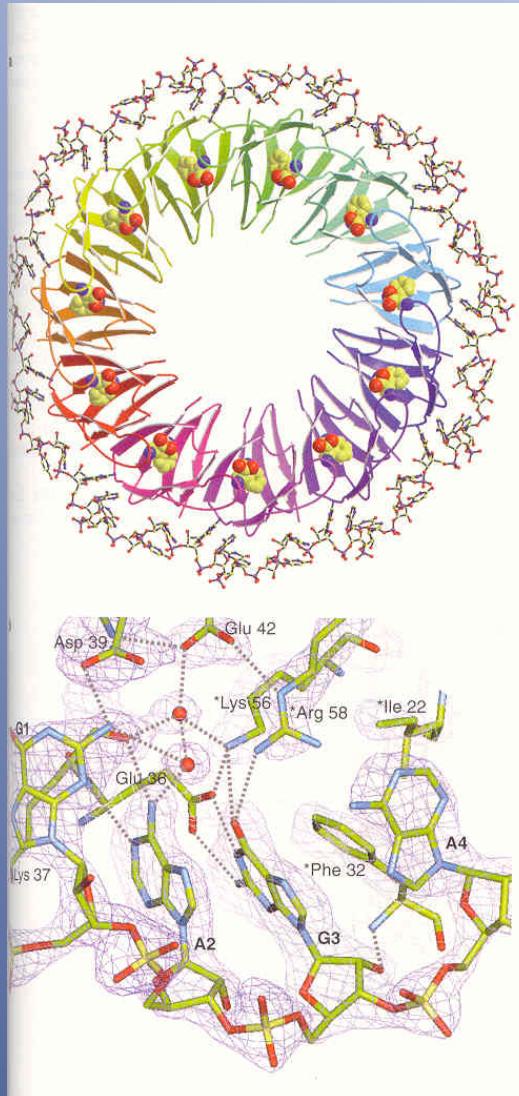


世界性的新闻组网络



(Naveen Jamal)

生物网络



人际关系网络



形形色色的网络

- 交通运输网，邮政网，电话通信网，计算机网，互联网，万维网
- 社会关系网，产品供销网，金融借贷网
- 智能电网，无线网，传感网，物联网
- 神经网，生物代谢网，食物链（网）
- 攻守同盟网，恐怖主义网络
- ...，
- 人人网，新浪微博网，QQ，...，团购网

不同类型网络的依存

- 互联网（Internet）—物理的、技术的
- 万维网（World Wide Web）—基于互联网的信息网络
- 人人网—基于万维网的在线社会（社交）网络（Online Social Network, OSN）
- 社会关系网（尤其在古代），常常与地域相关，地域上接近，见面的机会多，建立关系的可能性就大，随着交通工具的发达，社会关系的地域范围扩大

技术发展的两重作用

(尤其是计算机与通信技术 (ICT) 的发展)

- 催化了各种网络的发展
 - 规模变大，范围变广
 - 新型网络的涌现
- 使分析和理解大规模网络的行为成为可能
 - 行为数据与网络运行伴生
 - 海量数据分析的能力（计算设施，算法工具）

当我们想到“网络”这个词语...

“事物、情况”

+

“联系”

- 节点 (vertex,point)
- 边 (连接,链接,关系,联系; edge,link,tie)

网络的一些基本区分

- 节点的性质（类型）是否一样?
 - 例如，有些代表人，有些代表机构
- 边是否有“方向”？
 - 例如，道路的单行线，粉丝->明星
- 节点是否有区别（即便相同性质）？
 - 例如，节点各有特定名称，名称乱了含义不同
- 边是否有“强弱”（即便没有方向）？
 - 例如，关系的亲疏程度，友好或者敌对
- ...

图论：讨论网络的语言之一

- 节点，边
 - 完全图，散点图；（群体）
 - 同构，异构
- 无向图（graph），有向图（directed graph）
- 连通，连通分量（分枝），强连通分量
- 路（path）、桥（bridge）、捷径（local bridge）、距离
- 二部图（bipartite graph）
- 图上的广度优先搜索（breadth-first search）

博弈论：讨论网络的语言之二

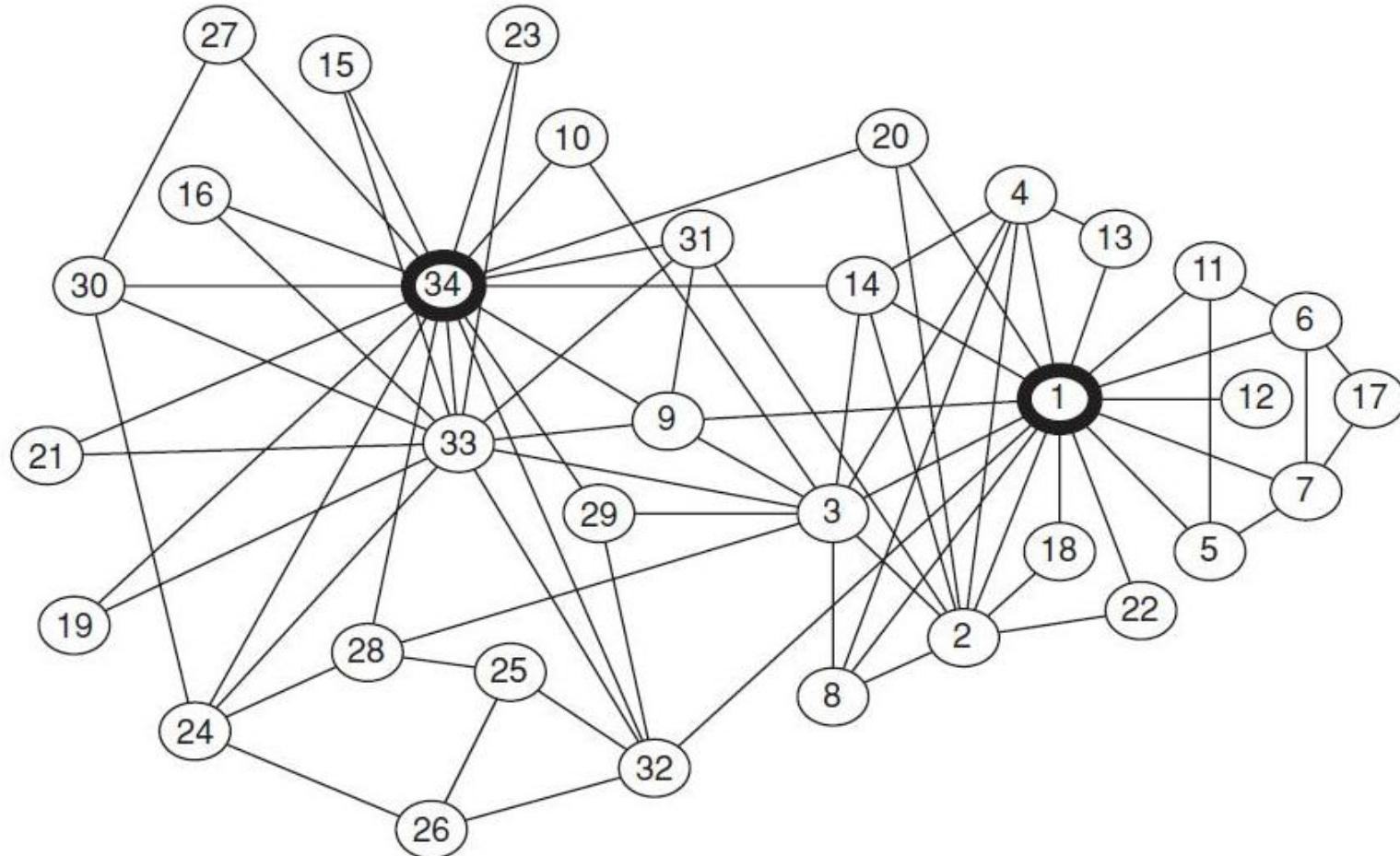
- 参与者、策略、回报（收益）
- 应对、最佳应对；占优策略，非优策略
- 均衡、纳什均衡
 - 预期的自我实现
- 纯策略、混合策略
- 帕累托最优、社会最优
- 进化博弈，进化稳定，进化稳定混合策略
 - 与经典博弈的关系

网络结构分析

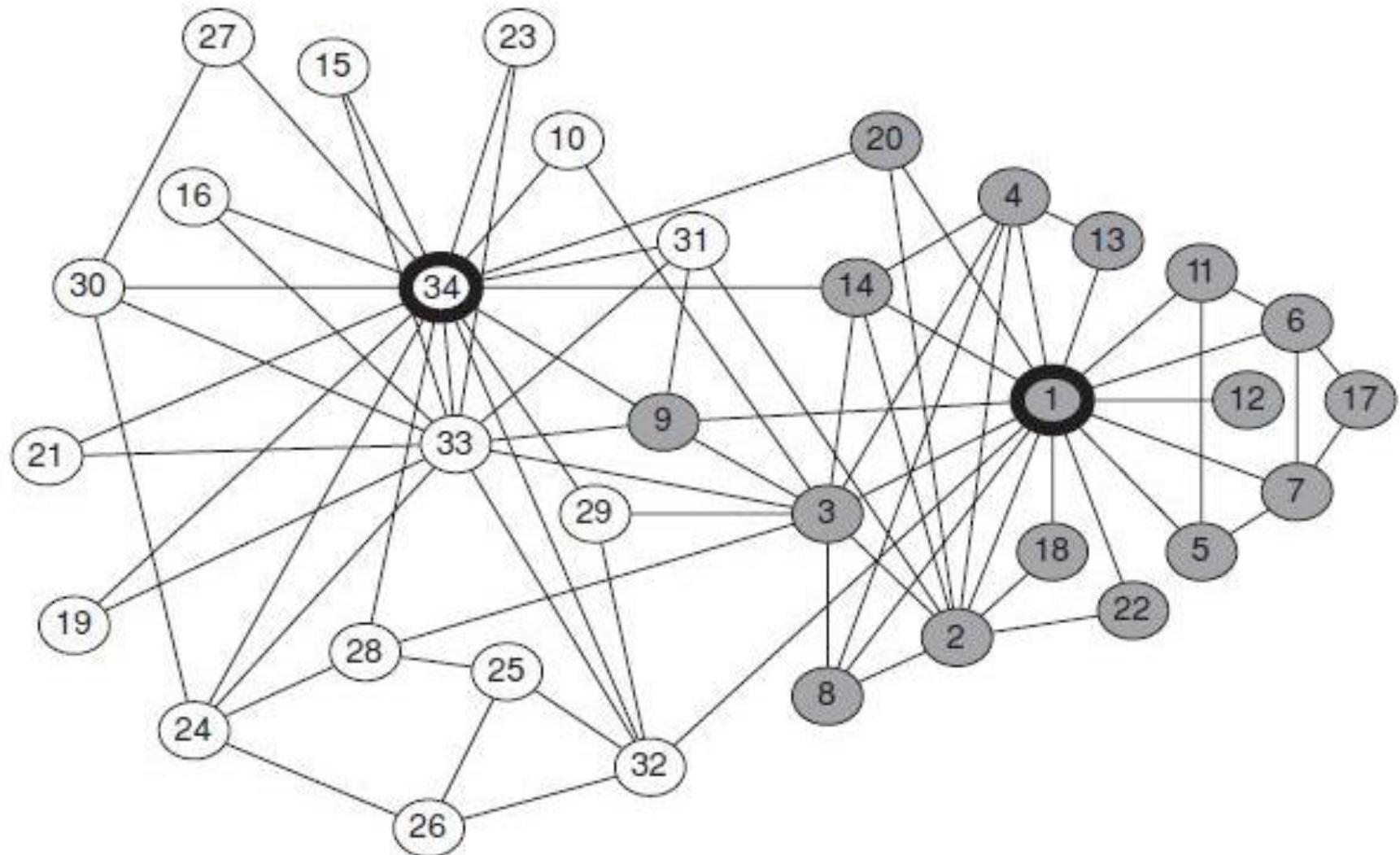
(社会学和经济学背景下)

- 边的成因 (edge prediction)
 - 哪些当前没有联系的两个节点将来形成联系的可能性比较大? (三元闭包, 同质性)
- 边的意义
 - 网络中不同位置的边的不同作用 (关系强度)
- 节点的重要性
 - 由结构特征带来的节点的权力, 或“权重”
- 结构的划分, 结构的平衡

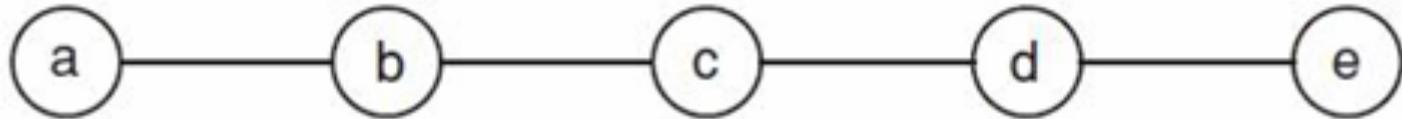
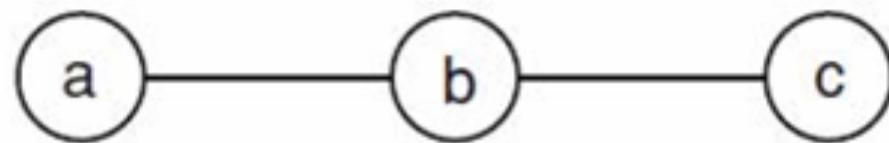
一个空手道俱乐部成员之间的社会 网络



网络结构中的社会性含义示例



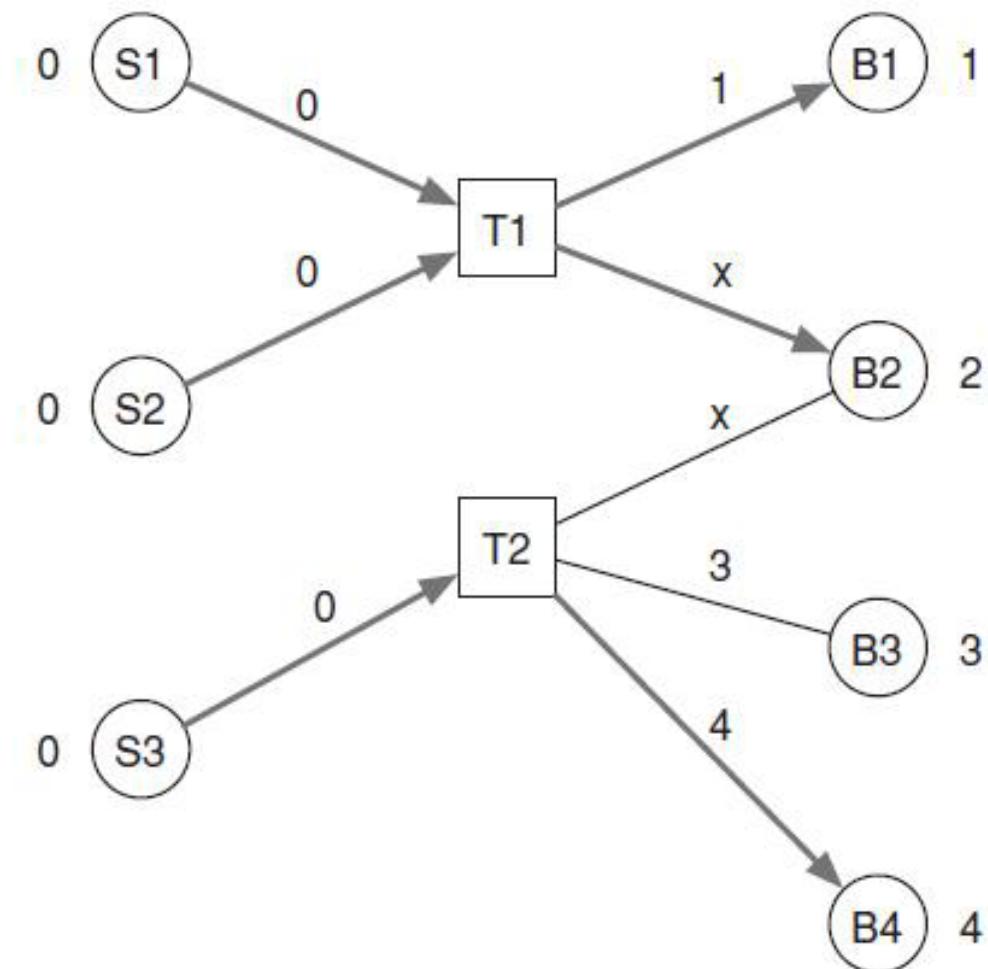
社会网络中的权力分析示例



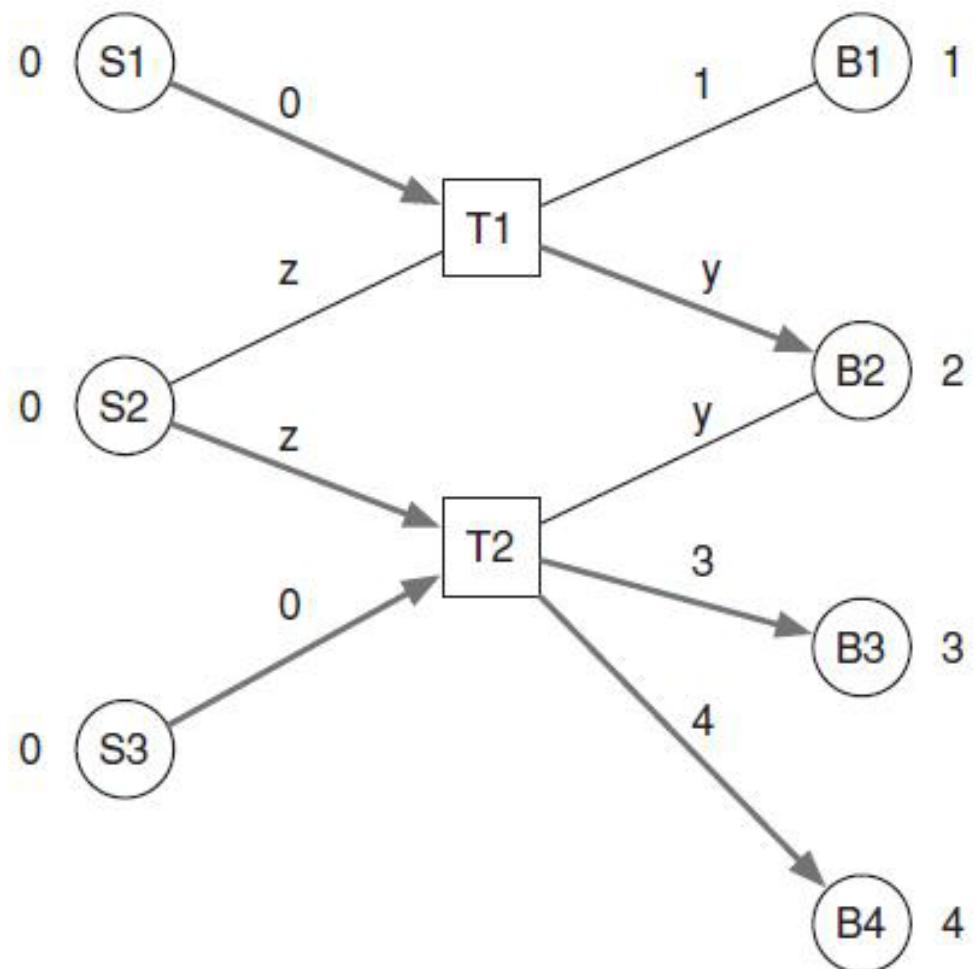
网络行为推理

- 群体效应（总体效应）
 - 交通拥堵问题
- 结构效应
 - 级联，波及，效用
- 在一定制度下的聚合行为
 - 市场，表决，产权
- 行为反过来对结构的影响

商业网络中的波及效应示例

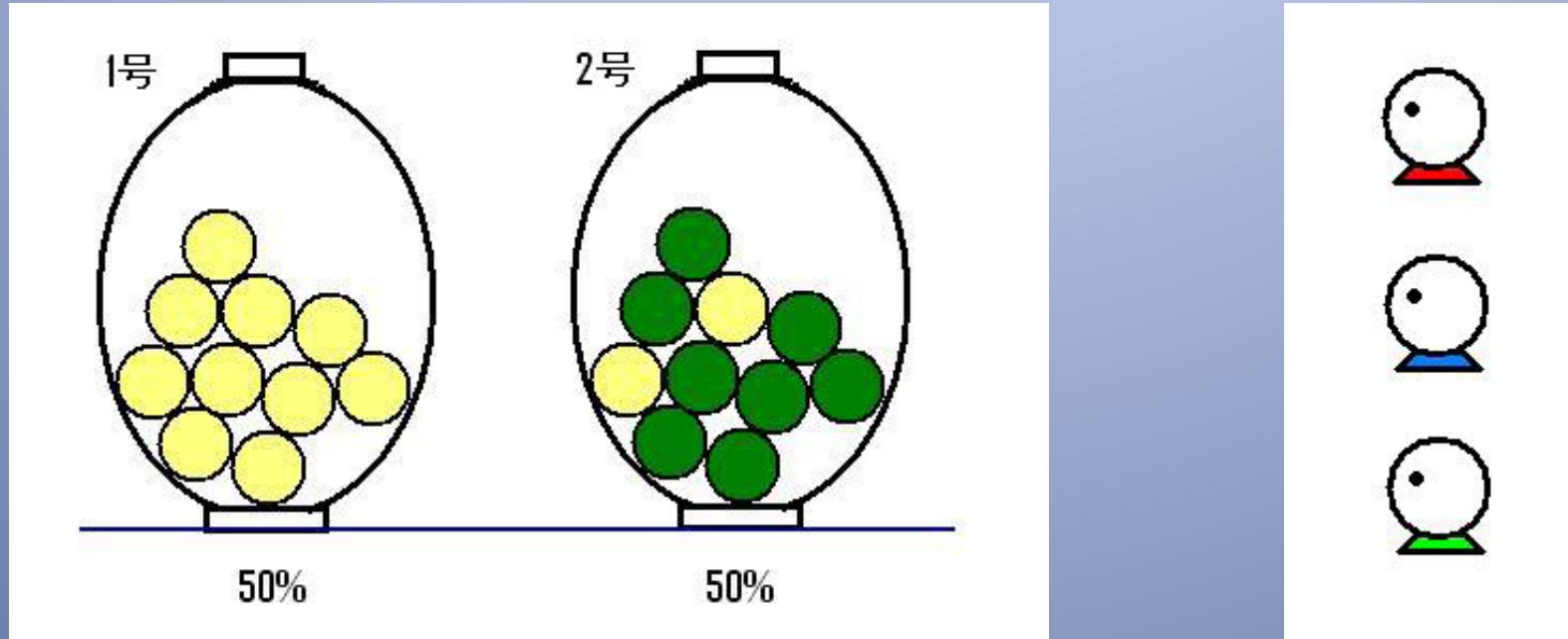


(a)



(b)

制度示例：关于诚实表决问题



- 以50%概率拿出其中一个坛子供三人表决用
- 三人依次，随机取一个看看，放回；不交换意见
- 每人给出关于坛子是1号还是2号的判断
- 若多数对了，3人都得奖；否则，3人都受惩罚

推理 (reasoning)

- <Reasoning about a highly connected world>
- Reason –the capacity human beings have to make sense of things, to establish and verify facts, and to change or justify practices, institutions, and beliefs.
- “推理”不同于“直觉”。推理总是要试图讲出道理
- ... is a way by which thinking comes from one idea to a related idea ...

能学到什么？

- 知识内容
 - 宽、广、交叉；对常见问题的重新认识
- 理工方法及其工具在社会科学问题上的应用
 - 图论、博弈；抽象，自动化
- 以学术研究的精神面对问题的态度
 - $A \Rightarrow B$
 - “study” vs “research”
- ...

课程学习后能分析的问题（5）

- 假设人类学家正在研究三个互为邻里的小村庄组成的集合。每个村庄都有30人，包括2-3个大家庭。每个村庄的人们都互相了解自己村庄的人。
- 当人类学家在这三个村庄建立一个社会网络，会发现人们都会和自己村庄的人成为朋友，和其他两个村庄的人敌对。这就给出了90人形成的网络（因为每个村庄30人），该网络中的边也会带有正关系和负关系的标识。根据所学的概念，这个90人形成的网络是平衡（均衡）的吗？请做一个简明的分析。

课程学习后能分析的问题（15）

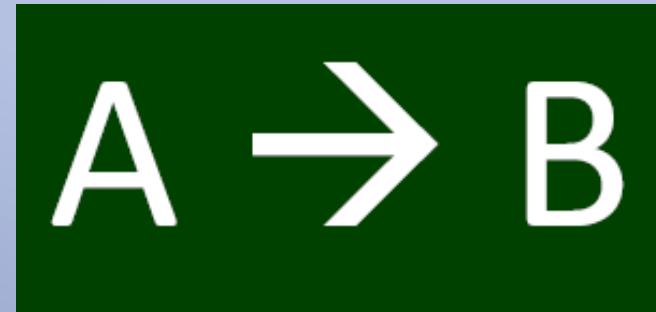
- 假如一个搜索引擎有两个广告位可以出售。广告位a的点击率为10，b的点击率为5。有三个广告商有兴趣购买这两个广告位。广告商x对每点击的估价为3，广告商y对每点击的估价为2，广告商z的每点击估价为1。分别计算这两个广告位的社会最优分配价格和VCG价格，对你的答案给出简要的解释。

课程学习后能分析的问题 (22)

- 考虑一个博彩市场，有两匹赛马A和B；有两个赌客，1和2。设每个赌客的财富量都是 w 。赌客1相信A赢的概率是 $1/2$ ，因此B也是 $1/2$ 。赌客2相信A赢的概率是 $1/4$ ，因此B是 $3/4$ 。两个赌客的财富效用都是对数函数，都根据自己的信念下注，要使财富效用的期望最大化。
 - (a) 赌客1, 2分别应该在赛马A, B上投多少钱？
 - (b) 求A和B的均衡赔付率倒数。
 - (c) 如果A赢了，赌客1会有多少钱？若B赢了呢？

课程学习后能分析的问题（24）

- 考虑一个机场，要将它的无线网络的排它运营权卖出去。取决于有多少人用这网络，它有可能变得拥塞，从而导致低质量的用户体验。为分析简单起见，假设任何时候机场都有 N 个旅客，如果其中 x 部分同时用这个网络，那么对每个人的回报就是 $1/2-x$ 。（我们可以将这回报看成是他们愿意支付的服务价格）
 - (a) 当机场将这个运营权卖给第三方运营商，该运营商会要通过对机场中上网的旅客收费来挣回它所支付的运营权费用。机场可以指望这运营权卖多少钱？运营商该向旅客收费多少？所有旅客的回报总和如何？请解释。
 - (b) 假设机场让人们免费使用这服务。此时所有旅客的回报总和如何？请解释。



- 学术研究的一种范式，不仅理工科
- 常见的错误（经验不足）
 - 仅看重B，不重视->，特别不重视A
 - 有了结果找所针对的问题（根据打中的地方画靶）
- 道理 =道 +理

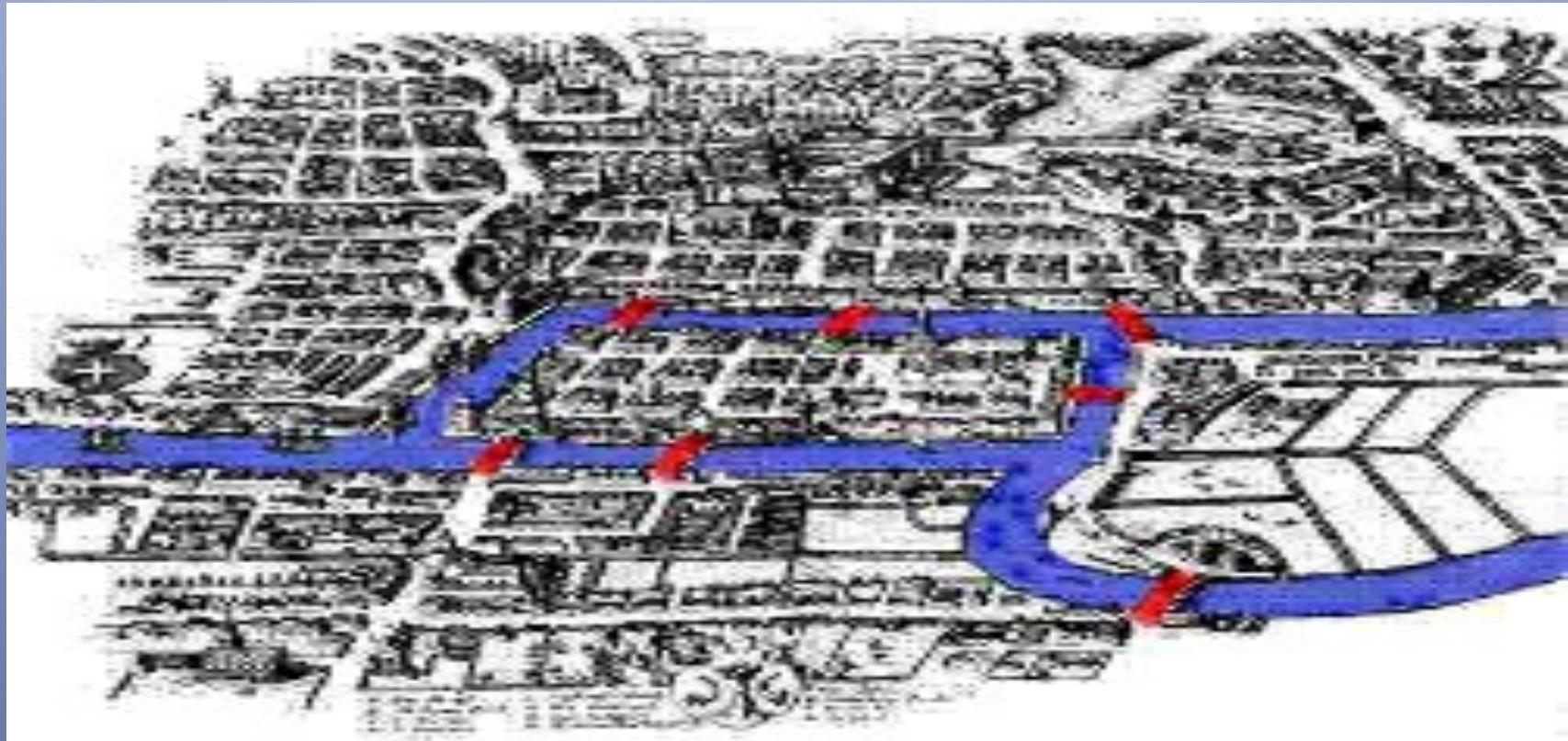
图论基础

与网络分析有关的概念

研究简史

- 1736 Euler 七桥问题
- 1959 Erdos, Renyi 随机图理论
- 1967 Milgram 小世界实验
- 1973 Granovetter 弱连接的强度
- 1998 Watts, Strogatz 小世界模型
- 1999 Barabasi, Albert 无标度网络

Konigsberg 七桥问题

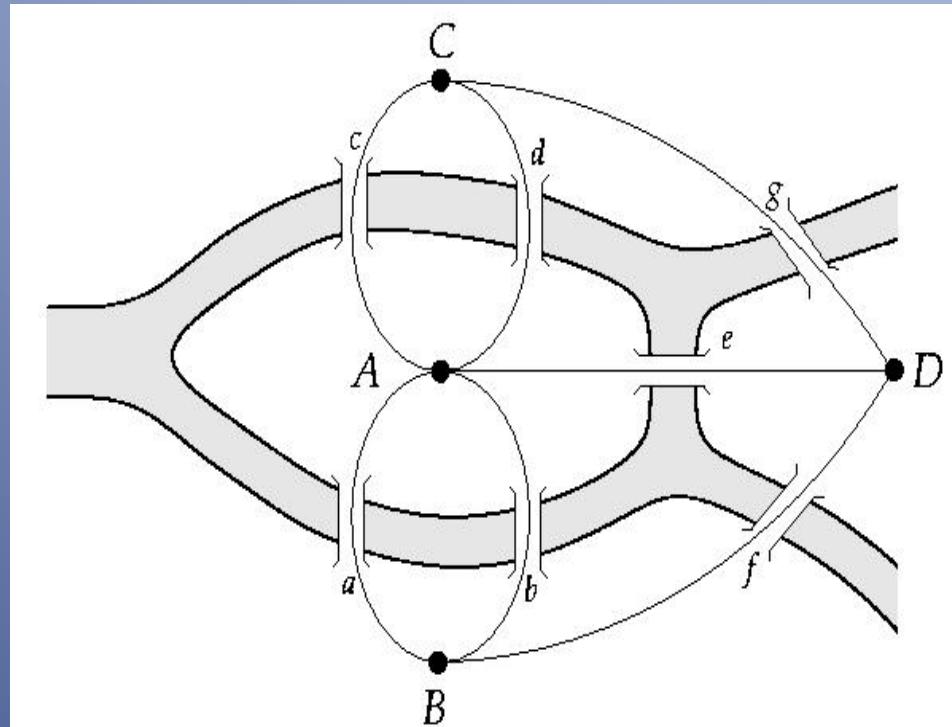


Can one walk across the seven bridges and never across the same one twice?

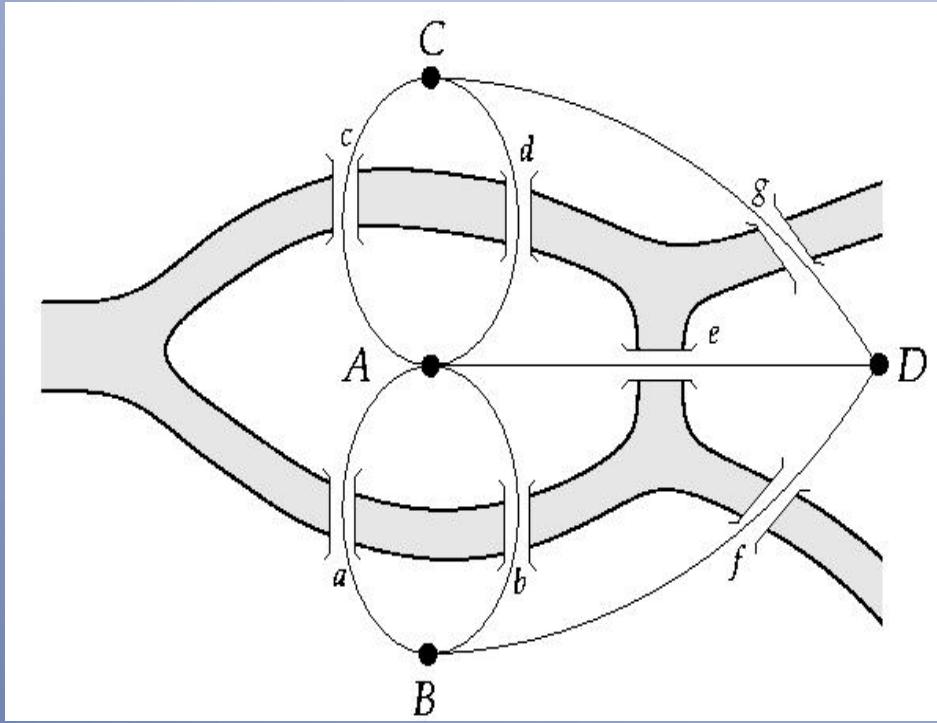
--- No! (Proved by Euler in 1736)

Leonhard Euler (1707-1783)

“图论之父”



看作4个节点，7条边的
图

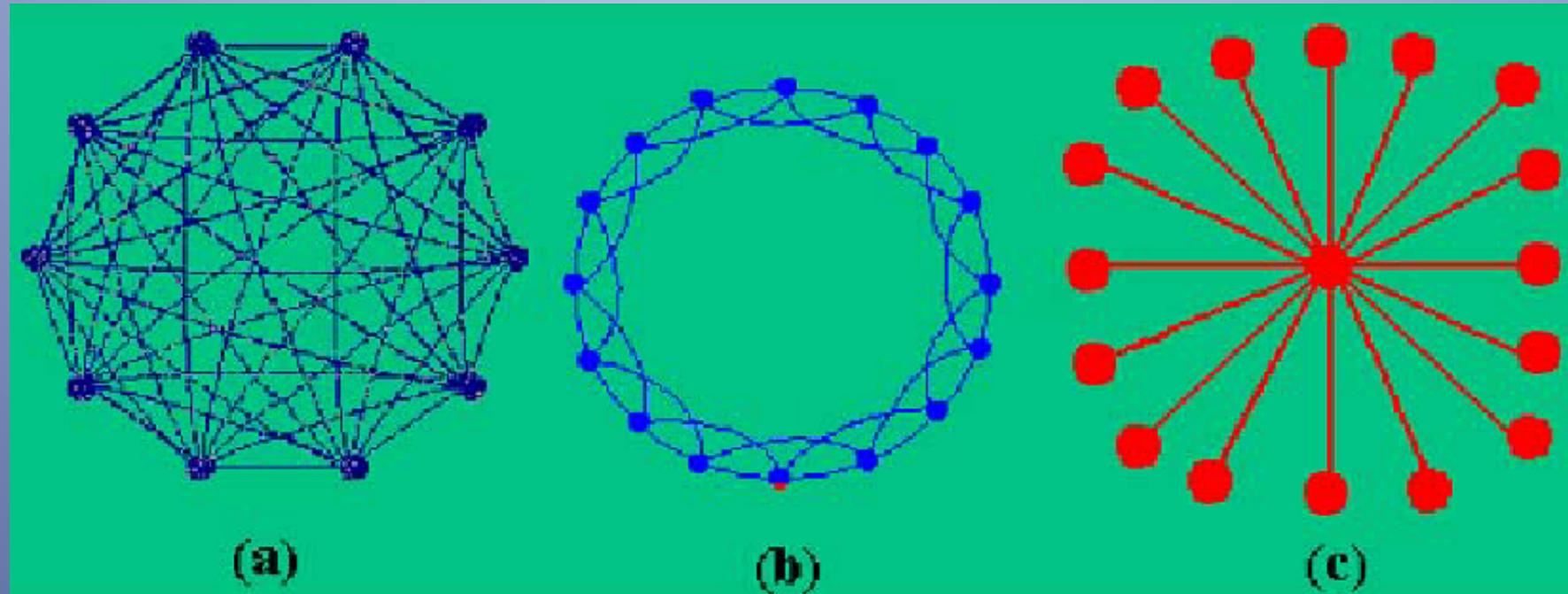


- ✓ 路必须有起点和终点。
- ✓ 一次走完所有的桥，不重复，除起点与终点外，其余点必须有偶数条边，所以七桥问题无解。
- ✓ 1875年，B 与 C 之间新建了一条桥解决了该问题！ 😊

Euler 对网络的贡献

- ✓ Euler 开启了数学图论，抽象为顶点与边的集合
- ✓ 图论是网络研究的基础
- ✓ 网络结构是理解复杂世界的关键

规则网络

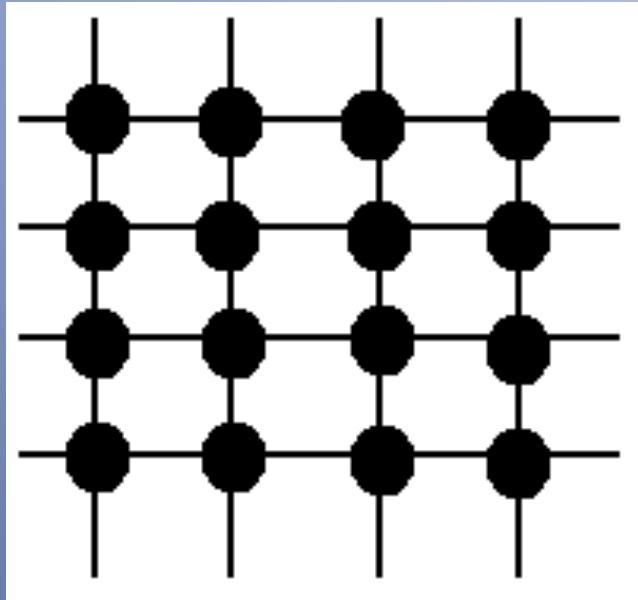


(a) 完全连接;

(b) 最近邻居连接;

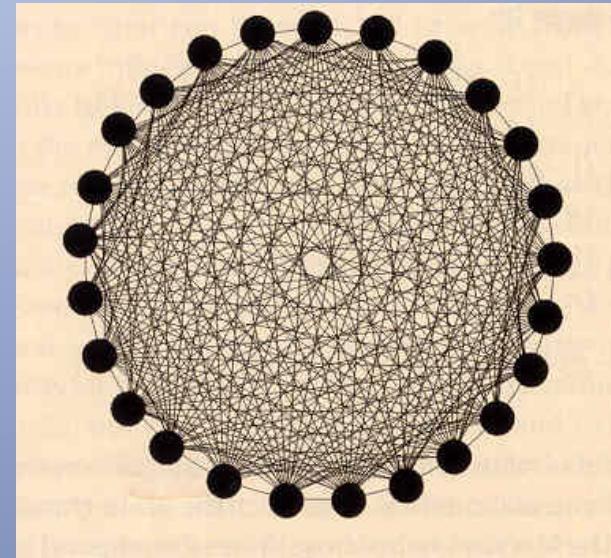
(c) 星形连接

规则网络



(d) Lattice

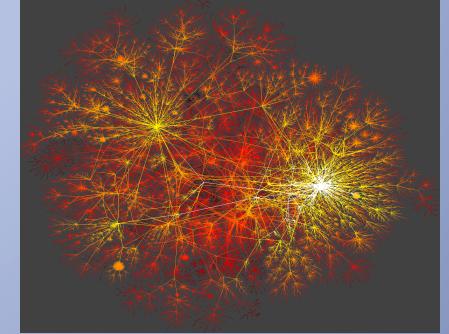
▪ ▪ ▪



(z) Layers

■ ■ ■

随机图理论



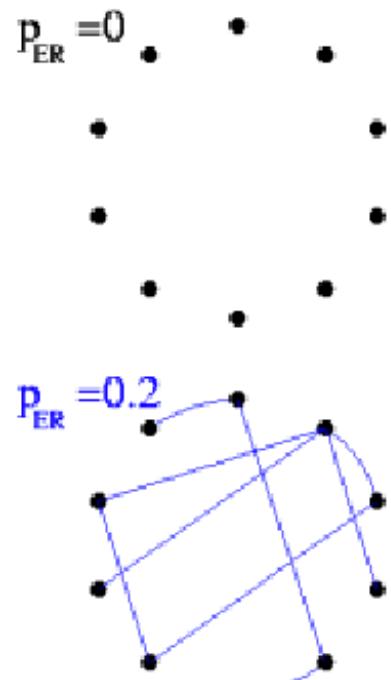
- 随机图论 - Erdős and Rényi (1960)
- ER 随机图模型统治四十余年..... 直到今天
- 当今大量可获取的数据 + 高级计算工具，促使人们重新考虑随机图模型及其方法

ER 随机图模型

Erdős-Rényi

(Publ. Math. Inst. Hung. Acad. Sci. 5, 17
(1960))

N nodes, each pair of node is connected with probability p



特征:

» 连通性:
Poisson 分布

» 齐次特征:
每个节点大约有相同的连接数

» 节点数不增加

小世界实验

- **Milgram小世界实验**

- 上世纪60年代哈佛大学社会心理学家Stanley Milgram 社会调查后推断出：世界上任意两个人平均距离是6(6度分离).
 - Milgram实验：信件传递.

- **Kevin Bacon游戏**

- **Kevin Bacon**美国著名演员
 - 任意一个演员与Bacon一起演过电影则其Bacon数为1.
 - 平均Bacon数为2.944.
 - 周星驰的Bacon数是3.

- **Erdös数**

Granovetter弱连接的强度

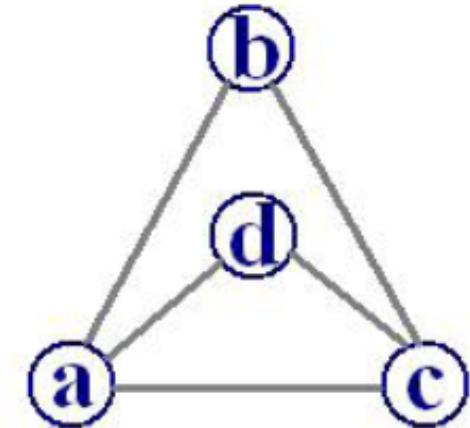
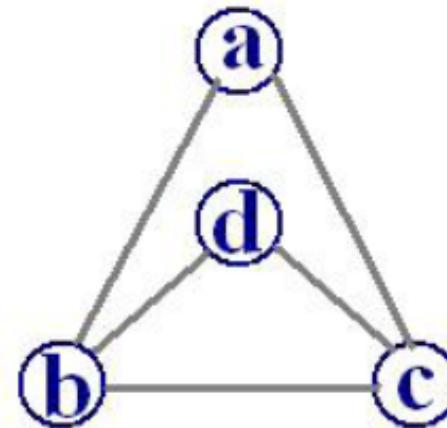
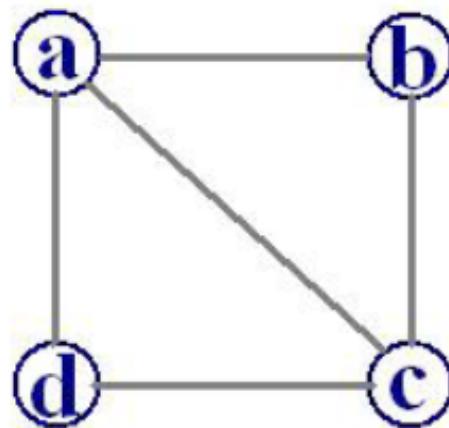
- 人们在**寻找工作**时，关系紧密的朋友(强连接)反倒没有那些关系一般甚至只是偶尔见面的朋友(弱连接)更能发挥作用
- 《Strength of Weak Ties》发表在《美国社会学》，有史以来最有影响的社会学论文之一

近年来的重大发现

- 小世界效应 (Watts and Strogatz, Nature, 1998) 《Collective Dynamics of ‘Small-World’ Networks》
- Scale-Free 特征 (Barabási and Albert, Science, 1999) 《Emergence of Scaling in Random Networks》

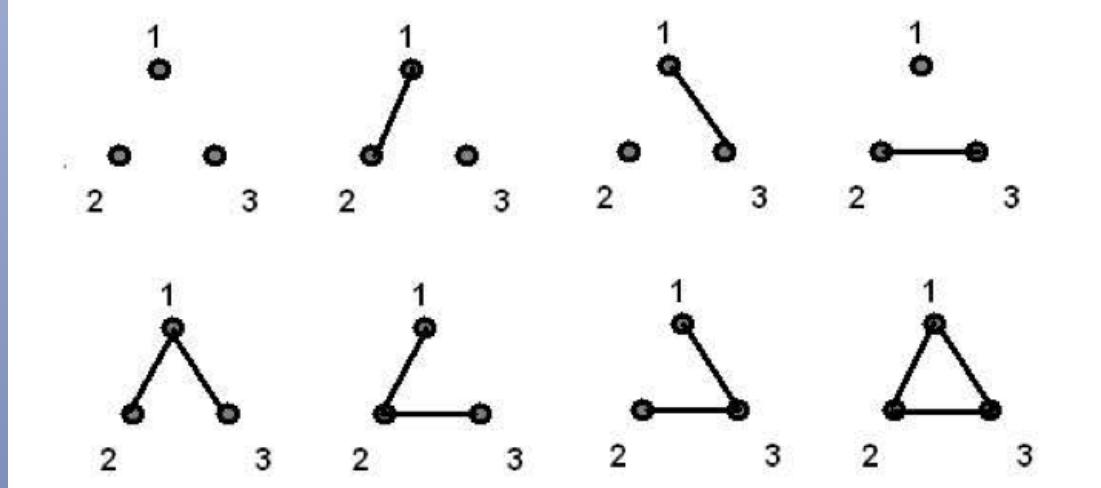
作为一个数学概念的“图” (graph)

$$G(V, E), \quad V = \{a, b, \dots\}, \quad E \subseteq \{(x, y) \mid x, y \in V, x \neq y\}$$

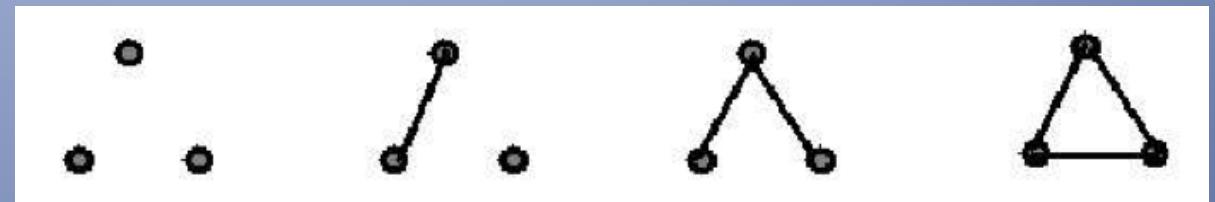


- 节点，边（圆括号表示 (x,y) 中的元素次序无关），有些场合也允许（需要）在一个节点上的“自环”，即 $x=y$
- 标号图 (labeled graph)，无标号图 (graph)
- 同构，异构

(不一样的) 图的个数 (枚举)



$$2^{\binom{n}{2}}$$



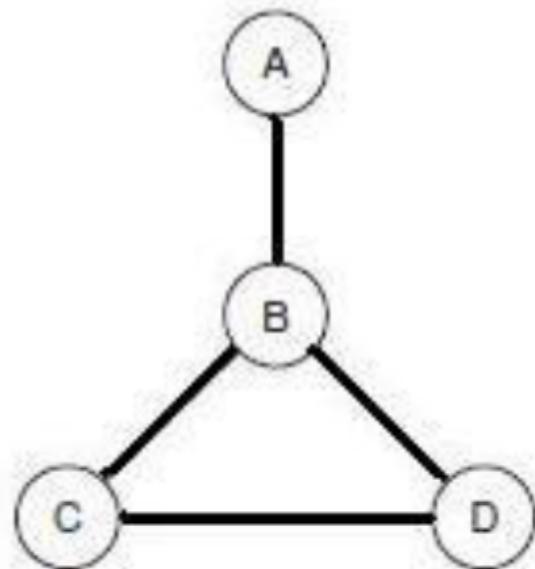
- 给定节点数 (n)
 - 标号图? 无标号图?
- Polya定理, 告诉我们如何计算无标号图的个数
- 如何判断两个图是否“同构” 依然是图论的最基本挑战之一

非同构图的个数

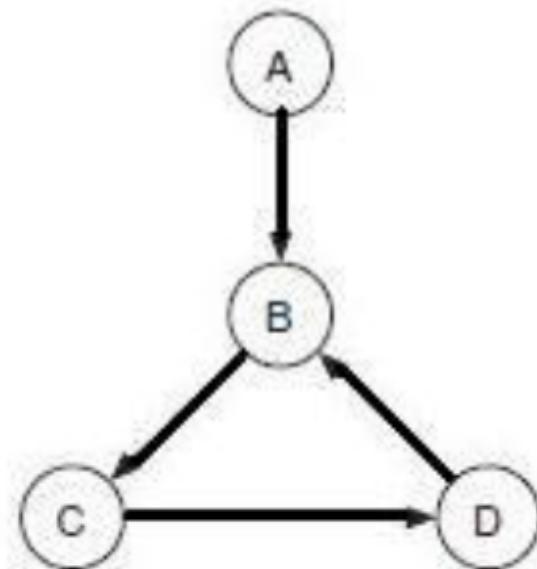
#vertices	Connected graphs	All graphs
1	1	1
2	1	2
3	2	4
4	6	11
5	21	34
6	112	156
7	853	1044
8	11117	12346
9	261080	274668
10	11716571	12005168
11	1006700565	1018997864
12	164059830476	165091172592
13	50335907869219	50502031367952
14	29003487462848061	29054155657235488
15	31397381142761241960	31426485969804308768
16	63969560113225176176277	64001015704527557894928

无向图，有向图 (directed graph)

$G(V, E), V = \{a, b, \dots\}, E \subseteq \{<x, y> | x, y \in V, x \neq y\}$



(a)



(b)

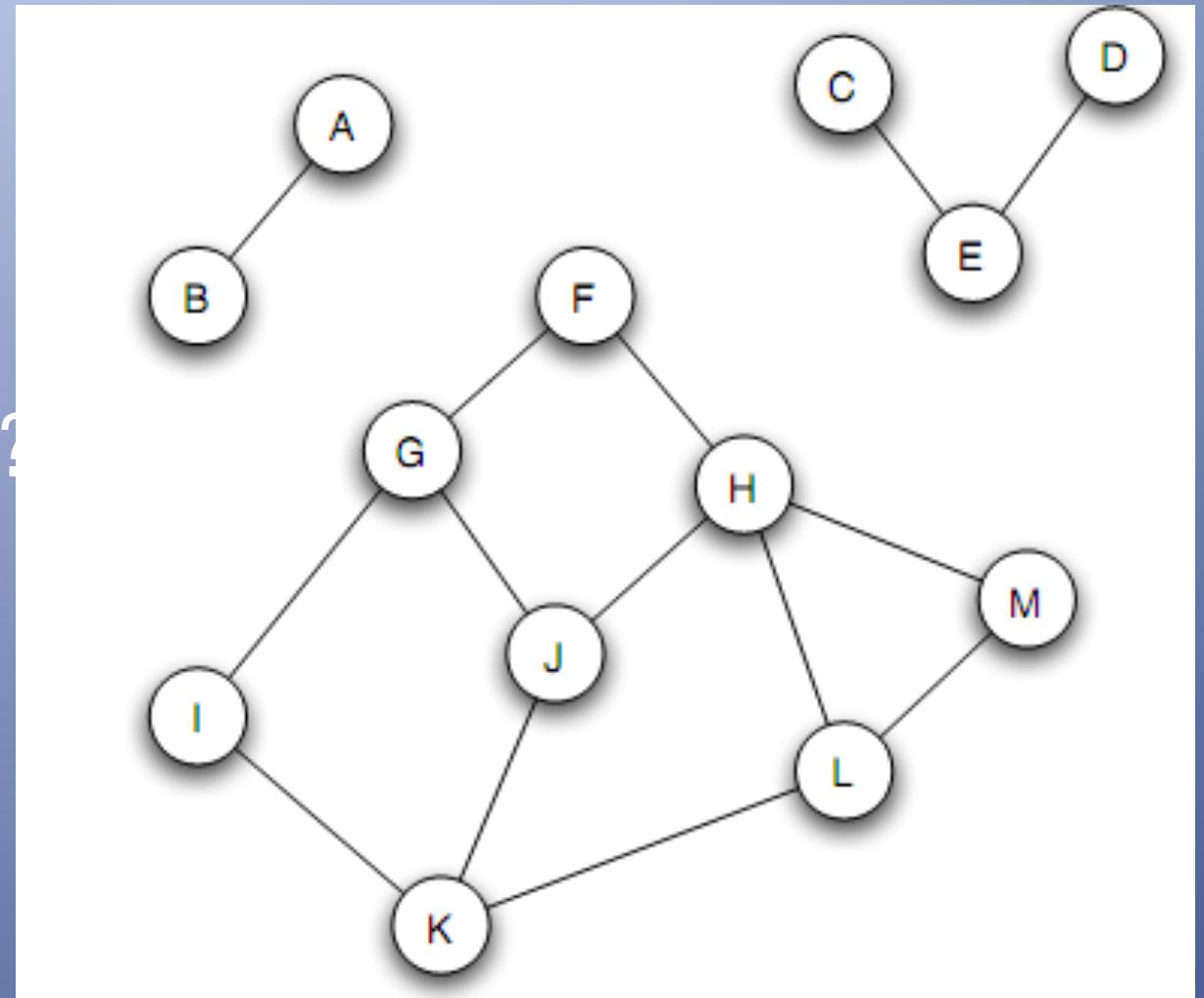
- 也可以是标号或者无标号的
- $<x, y>$ 和 $<y, x>$ 有可能同时存在

路， 距离， 连通， 连通分量

- 路 (path, 路径, 通路) : 节点序列, 相邻两个节点之间存在一条边
 - 长度: 节点数减1, 所涉及边的条数
 - 简单路径, 回路 (仅相同的路径)
- 距离: 两个节点之间最短路径的长度
- 连通图: 任何两个节点之间都存在一条路
- 连通分量
 - 连通子图
 - 不被真包含在任何其他连通子图中

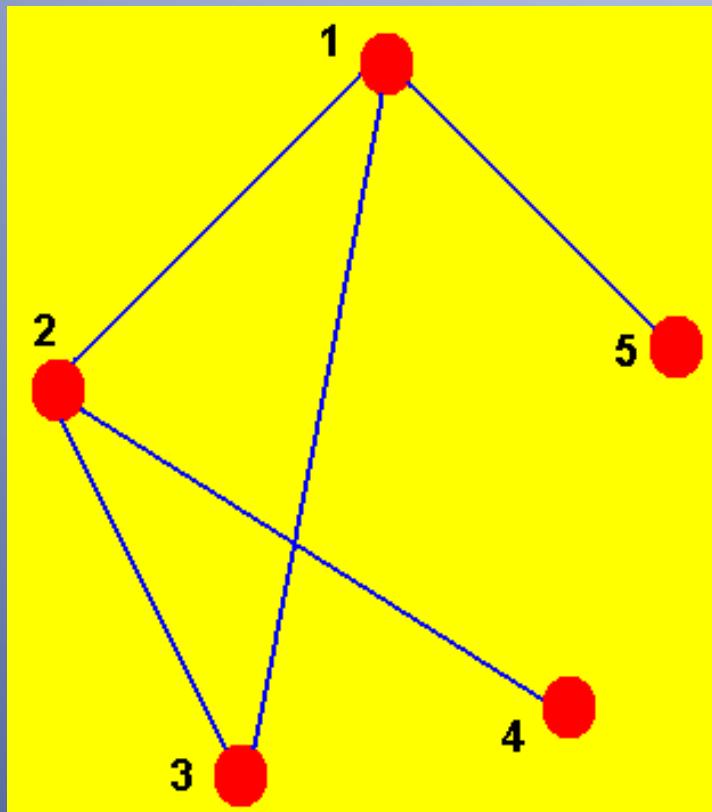
例子：路，距离，连通分量

- 节点I和M之间有多少不同的路?
 - 有多少不同的简单路?
- 它们之间的距离?
- ($\{A, B\}, \{(A, B)\}$)是不是连通分量?
- ($\{H, L, M\}, \{(H, L), (L, M), (H, M)\}$)是不是连通分量?



例子：

一个具5个节点5个边的网络：



$$L = \frac{1}{\frac{1}{2}N(N-1)} \sum_{i>j} d_{ij}$$

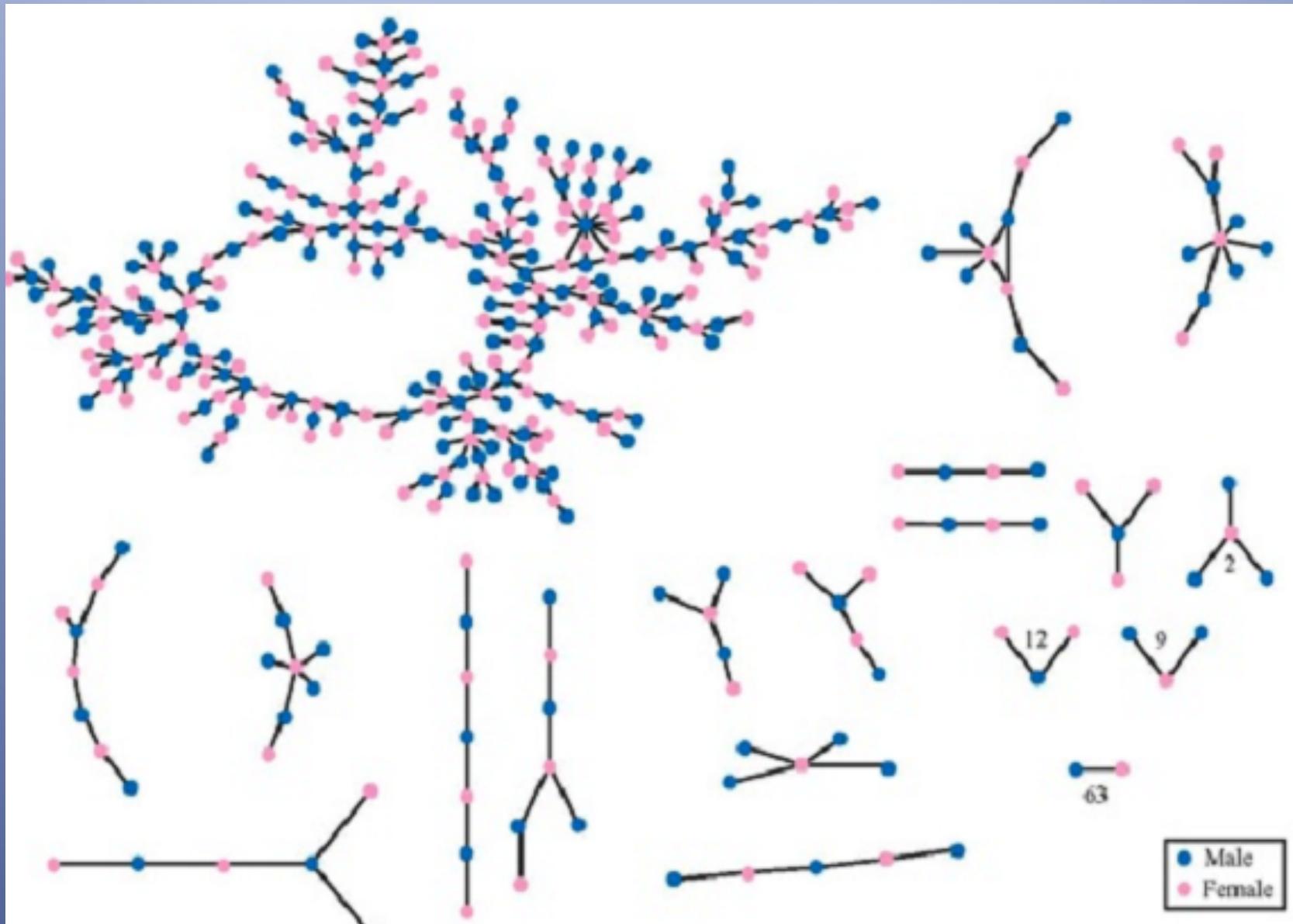
$$\begin{array}{llll} d_{12} = 1 & d_{13} = 1 & d_{14} = 2 & d_{15} = 1 \\ d_{23} = 1 & d_{24} = 1 & d_{25} = 2 & \\ d_{34} = 2 & d_{35} = 2 & & \\ d_{45} = 3 & & & \end{array}$$

Total = 16

Average: $L = 16 / 10 = 1.6$

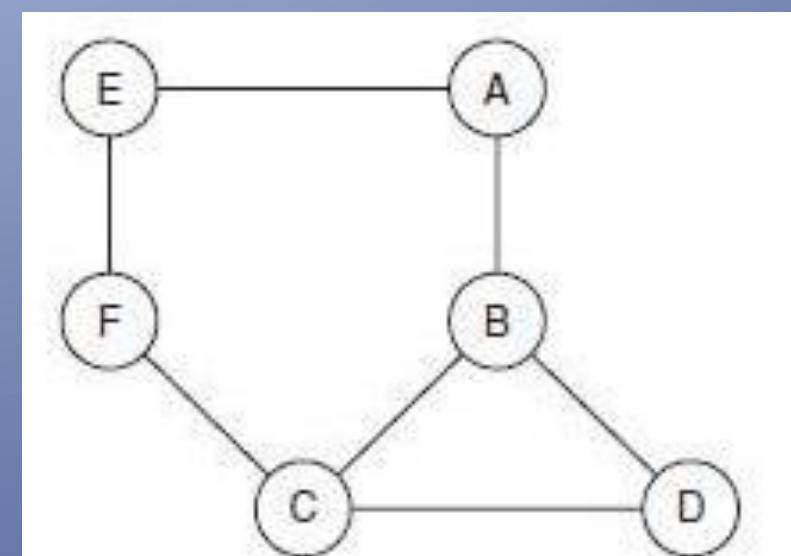
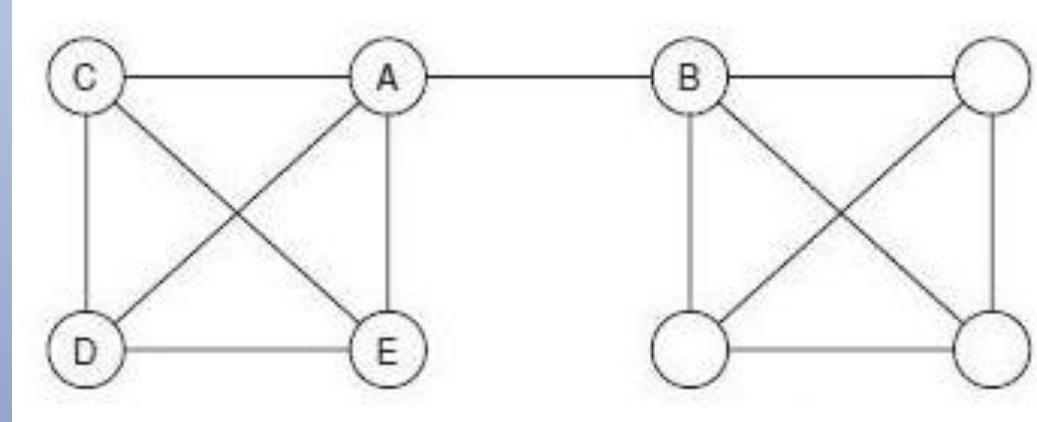
L 较小 → 小世界网络

超大（巨大）连通分量示例



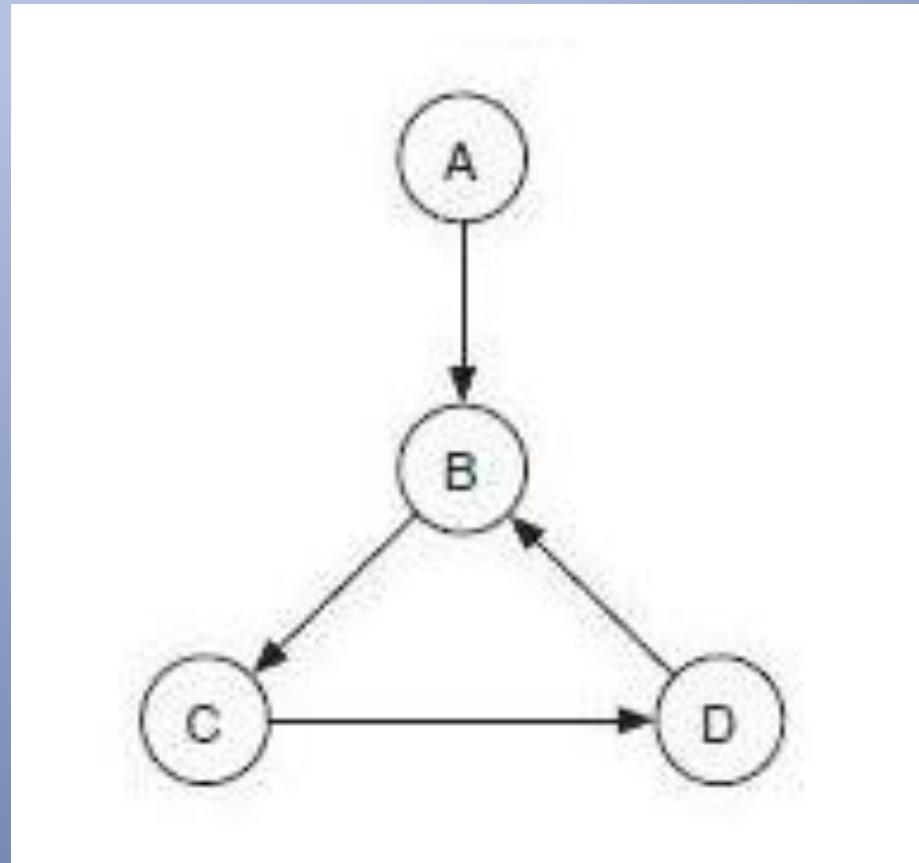
桥，捷径 (local bridge)

- 桥：具有特别性质的边，删除它，其两个端点之间就不再有路
 - 删除它，增加图的连通分量的个数
- 捷径：也是一种边，删除它，其两个端点之间的距离至少为3。
 - 桥可以看成是捷径的一个特例



对于有向图：有向路径，强连通分量

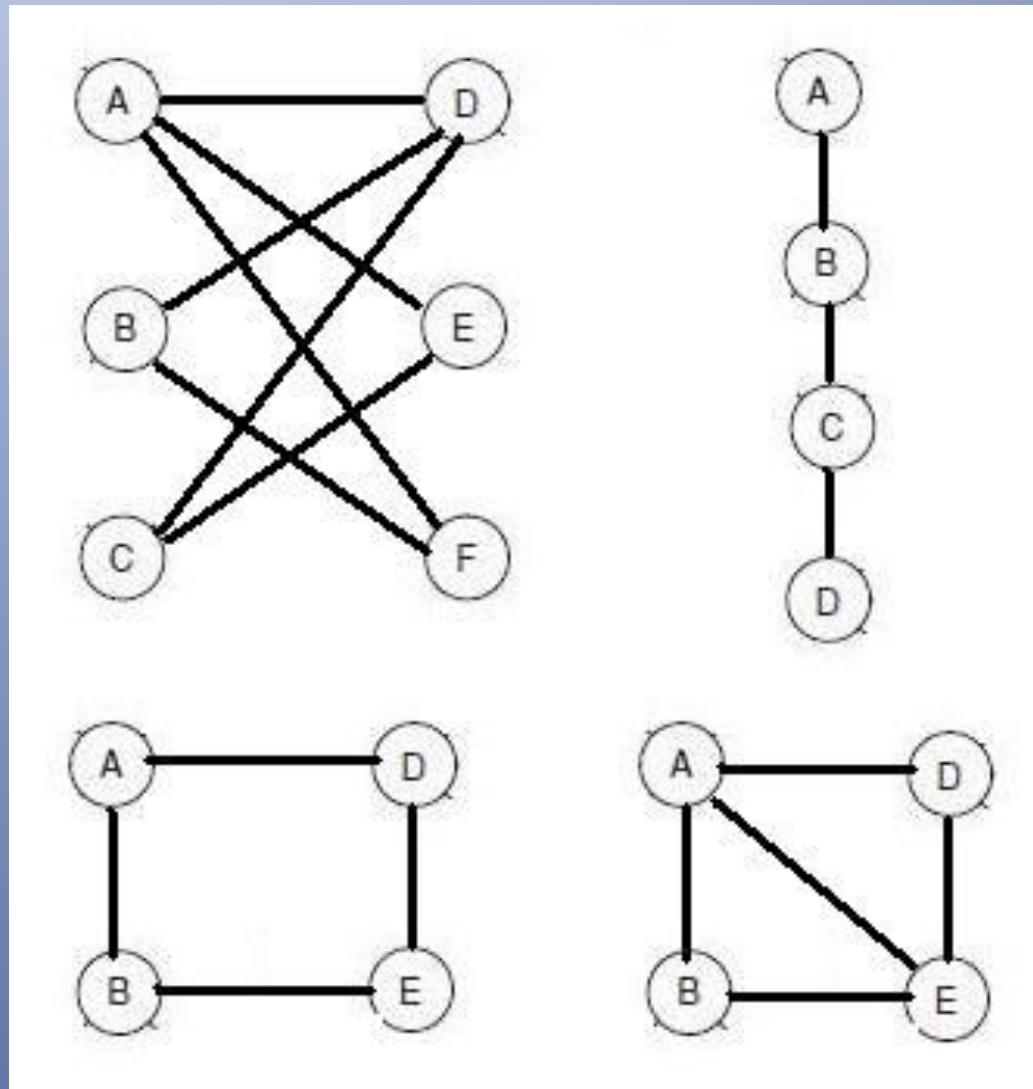
- 有向路径：节点序列，相邻节点之间有从前往后的有向边
- 强连通分量
 - (1) 任意两个节点之间存在有向路径的有向子图
 - (2) 不被真包含在任何其他满足性质(1)的子图中

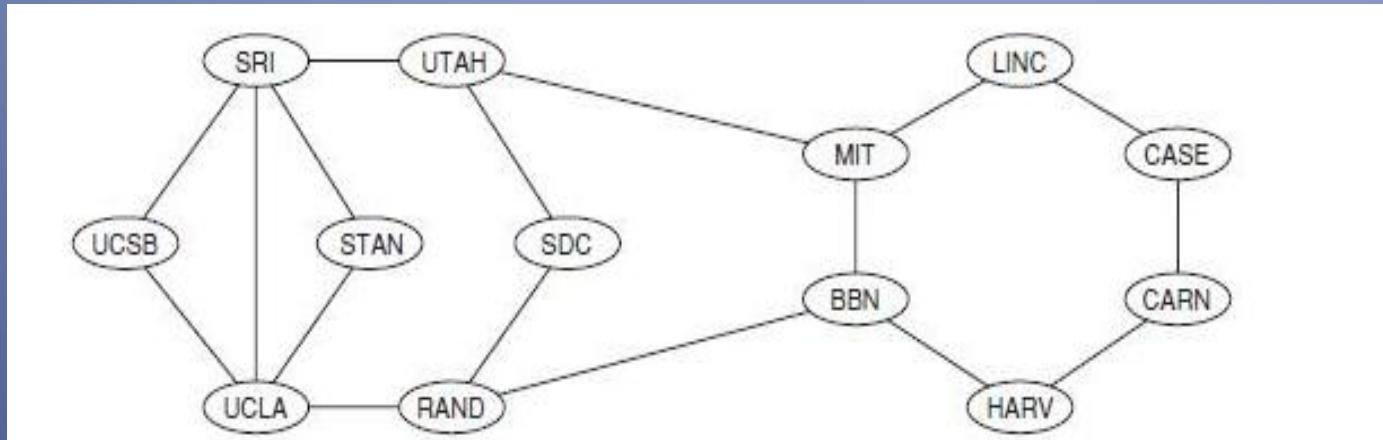
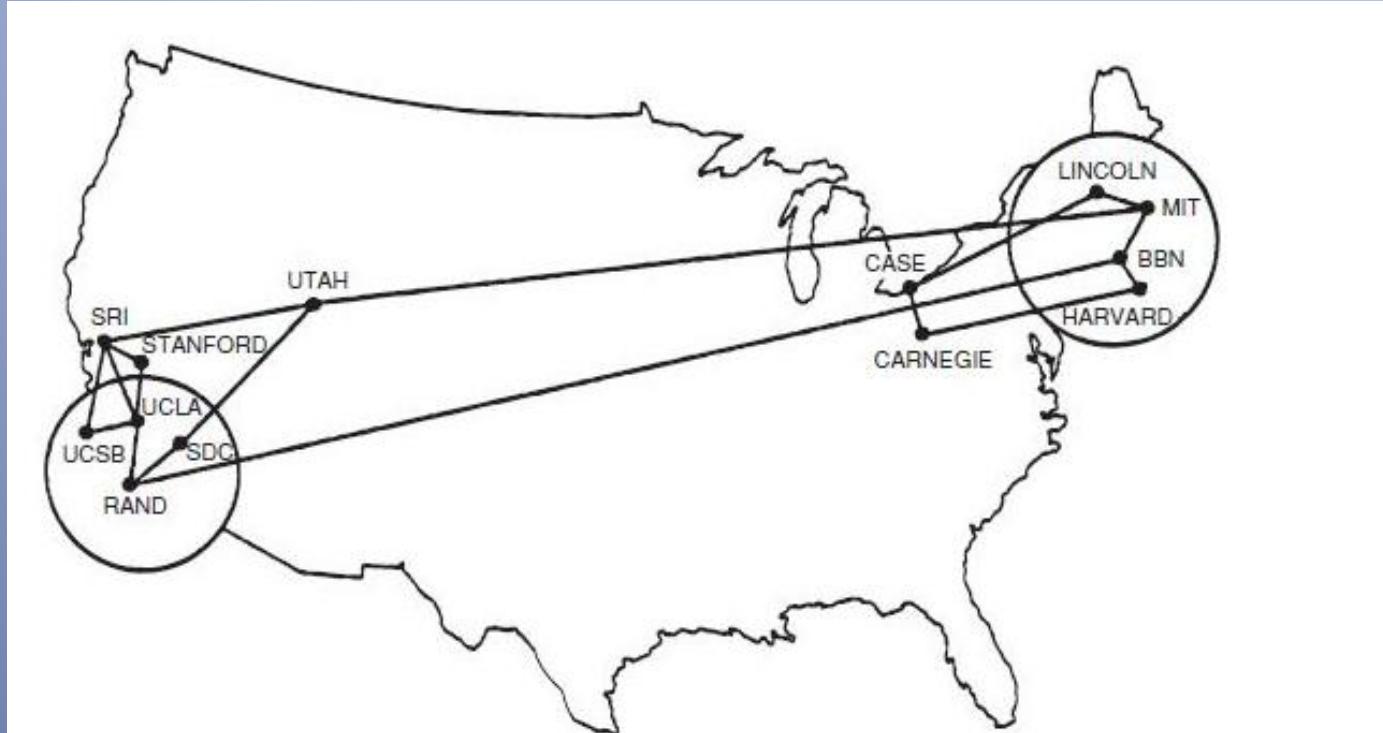


$\langle A, B, C, D \rangle?$ $\langle A, B, D \rangle?$
 $(\{B, C, D\}, \{\langle B, C \rangle, \langle C, D \rangle, \langle D, B \rangle\})$

二部图，图上的广度优先搜索

- 二部图 (bipartite graph)：节点可以分成两组，组内没有边
- 图上的广度优先搜索 (breadth-first search)
 - 从某一点开始，对图的节点的一种“遍历”方式





- 从LINC开始
广度优先搜索
- {LINC}
 - {MIT,CASE}
 - {CARN,BBN,UTAH}
 -
 - {HARV,SDC,RAND,S
RI}
 - {UCSB,UCLA,STAN}

BFS从概念上
对图中的节点
进行了一个
“分层”，所
涉及到的边
“自然形成了”
一个二部图

典型现实网络（图）的数据表示

- 合作图
 - 例如，一群学者之间co-authorship
 - 节点：人；边：当且仅当两个人有合著的文章
- 交流网
 - 例如，一所大学师生之间的电子邮件关系网
 - 节点：人；边：两人之间发过一定量的往返邮件
- 信息链接网
 - 万维网上的网页之间的链接关系
 - 论文之间的引用关系
- ...

网络数据的计算机表示

- 邻接矩阵 (adjacency matrix)

- 相邻节点列表

- 关联矩阵 (incidence matrix)

- 边序列

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

边序列

1,2

1,3

1,4

2,4

3,4

$$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

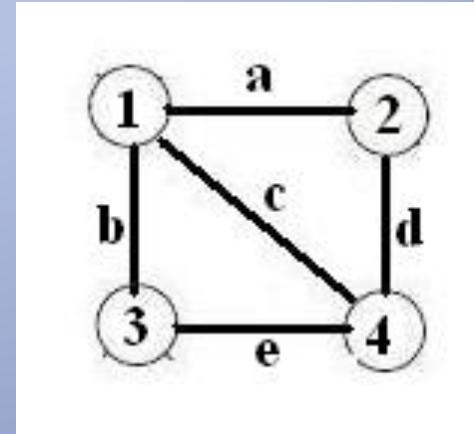
相邻节点列表

1: 2,3,4

2: 1,4

3: 1,4

4: 1,2,3



小结

- 网络无处不在，行行色色
- 两个重要方面：结构、行为；不同但相互影响
- 图论：讨论网络结构的语言

作业

- 第2章 1,2,3

社会网络的基本概念 及其在OSN上的体现

三元闭包，关系的强度及其与网络结构的关系，
同质性及其影响，正负关系及其平衡，幂率，
小世界，节点的地位与关系的均衡，...

社会网络，不仅是人类社会的一个属性

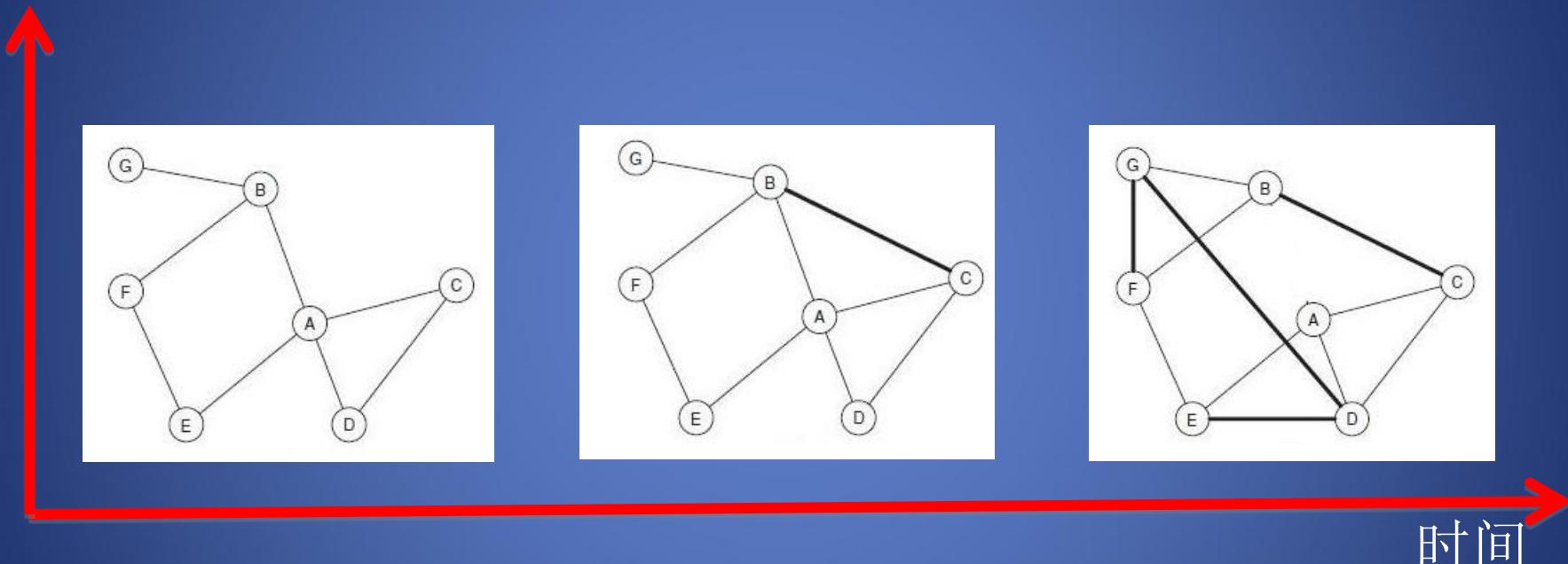


- 但在人类社会体现得最丰富，最多姿多彩
- 人类在社会网络中的行为，能否在基因中找到原因？



讨论社会网络的空间

现象
原理



- 不仅考虑一个时刻（“快照”）上的性质
- 更要研究随时间发生的变化

提要

- 三元闭包（triadic closure），社会网络的基本成因
- “关系”的强弱与网络结构的关系
 - 启示，假说，假说的论证（抽象形式化的，数据支持的）
 - OSN中的关系强度
- 在社会网络中跨越“结构洞”的节点的性质分析（“责权利”）

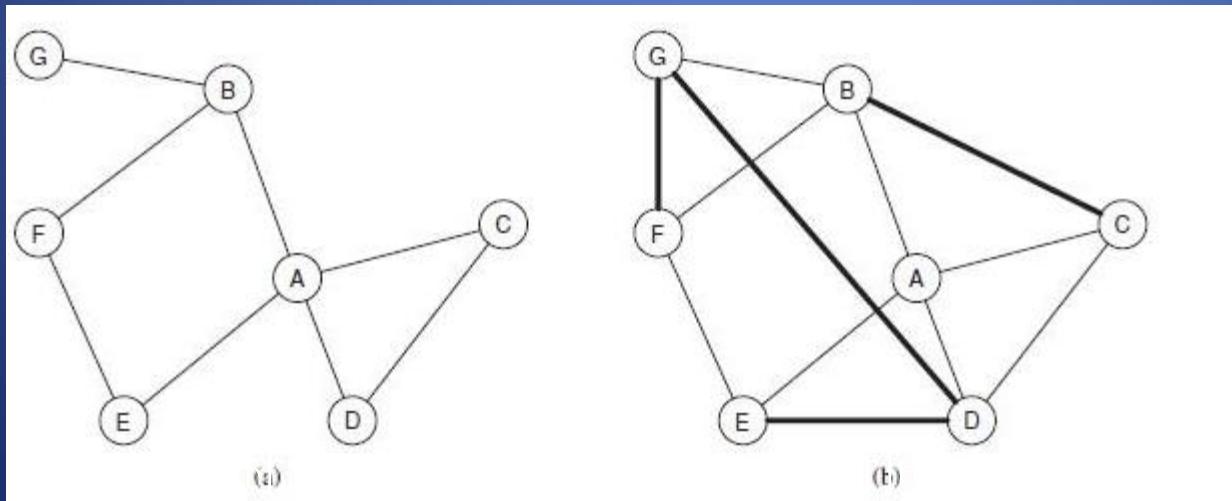


三元闭包（闭合）



- 社会网络的最基本成因
 - Anatole Rapoport (阿纳托尔·拉波波特, 1953)

如果两人在社会网络中有一个共同的朋友，则他们俩将来成为朋友的可能性提高。



机会?
opportunity
信任?
trust
动机?
incentive



三元闭包原理的拓展

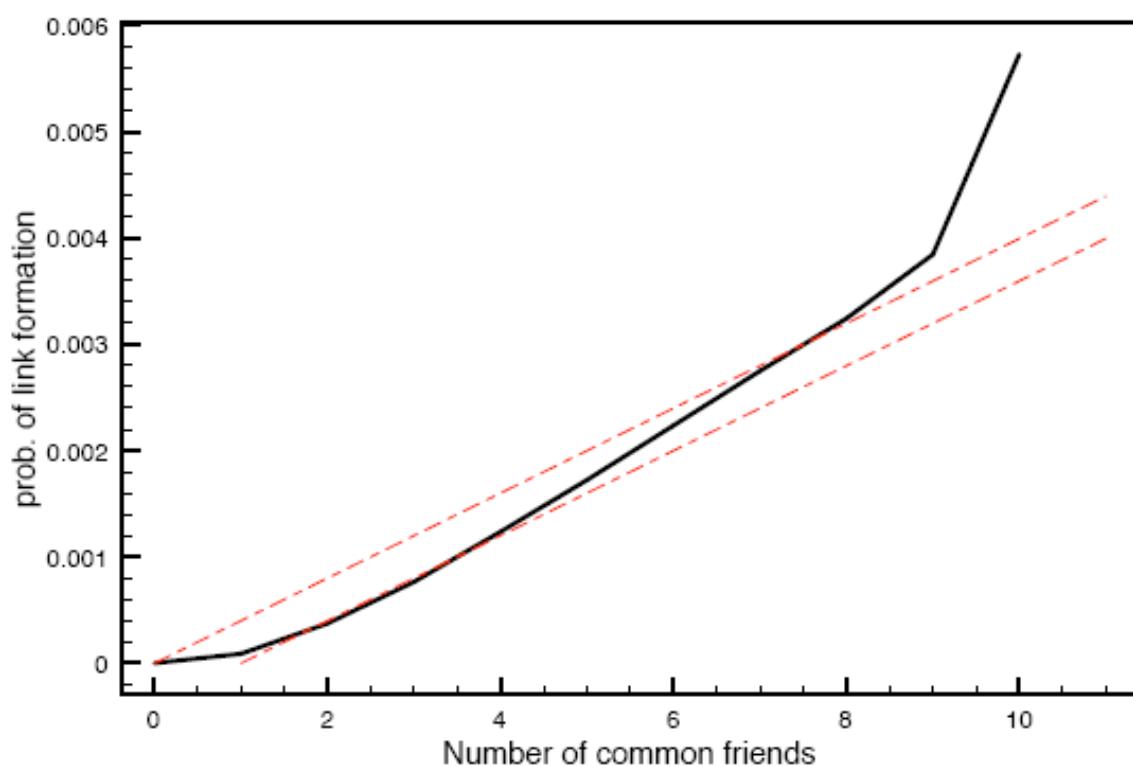
- 两个人的共同朋友越多，则他们成为朋友的可能性越高
 - 这是从“量”方面的拓展
- 两个人与共同朋友的关系越密切，则他们成为朋友的可能性越高
 - 这是从“质”方面的拓展
- 三个原因（机会、信任、动机）的作用在这些拓展的意义上保持一致

如何验证？

一个利用在线数据研究三元闭包的例子

- 电子邮件网络≈社会网络
 - 节点：一定范围的邮件地址（例如一个大学）
 - 边：一段时间（例如一个月）里有邮件通信
 - 单向 vs 双向？多少才算？...
- 网络的演化
 - 什么叫两个相继的网络快照？
 - 两个相继的快照就能说明问题？（回避偶然性事实，大量快照对的平均）
- 如何定义考察三元闭包现象的测度？
 - 简单统计闭包个数 vs 共同朋友数的影响

结果及其含义



- * 在电子邮件网络上三元闭包迹象明显——共同朋友有助于关系的建立；
- * 突出体现在1—2个共同朋友情形；
- * 为什么8—9—10也突出？

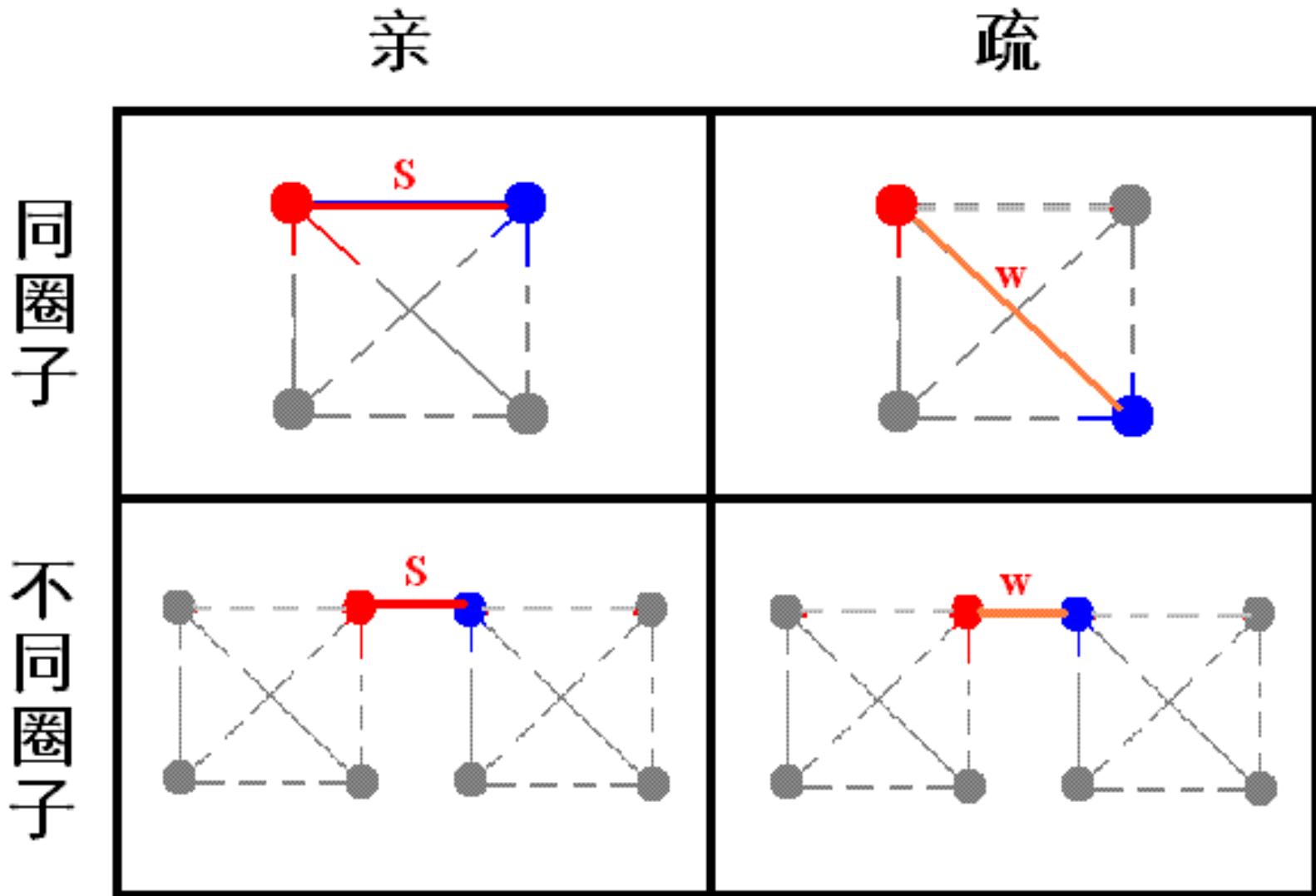
- 特定电子邮件网，其他网络如何？
- 定量分析 vs 定性结论



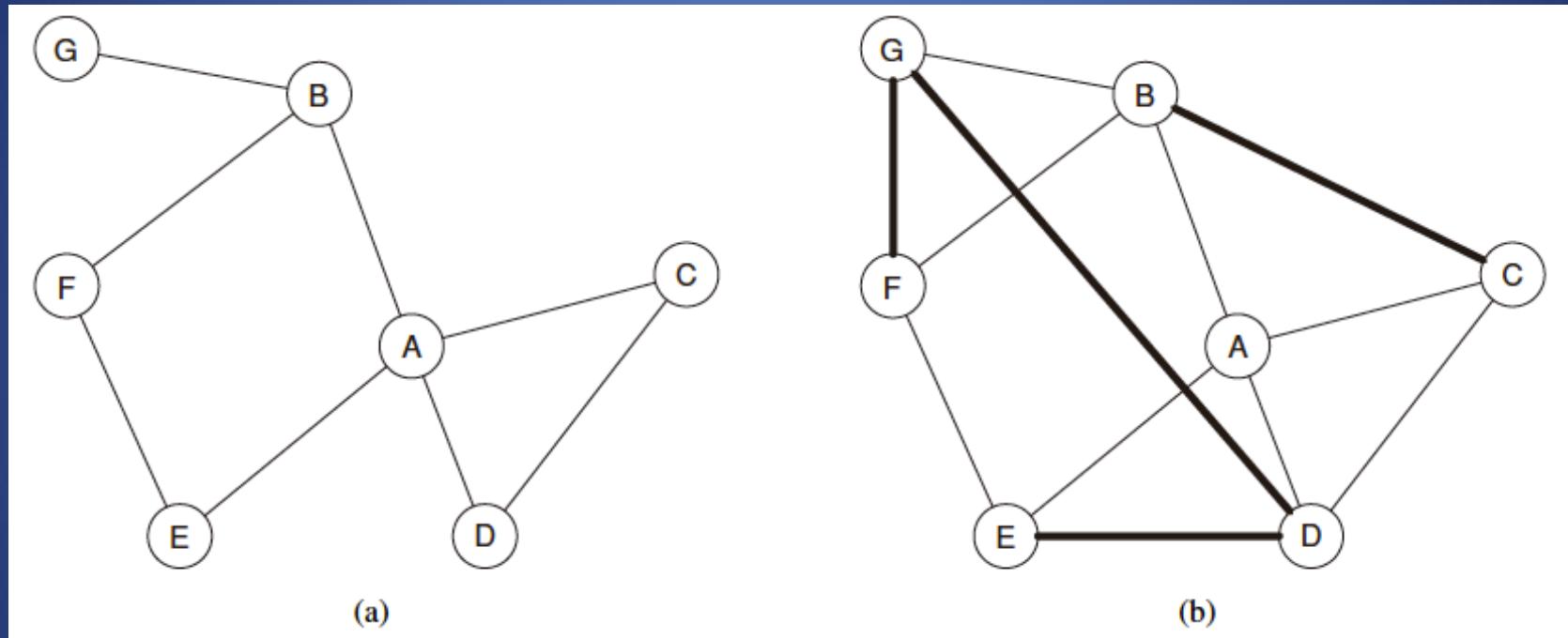
格兰诺维特的诧异

- Mark Granovetter, “The Strength of weak ties”
American Journal of Sociology, 1973.
- Mark Granovetter, *Getting a Job: A study of Contacts and careers*. University of Chicago Press, 1974.
- 为什么对找工作提供有效帮助的人更多只是一般熟人，而不是亲密朋友？
 - 两个层面的认识，导致对社会关系（网络）两个维度的视角

社会关系的两个视角



节点的聚集系数： 邻居间三元闭包体现的强度



节点A的聚集系数 = A的任意两个朋友之间也是朋友的概率（即邻居间朋友对的个数除以总对数）

聚集系数

- 一个网络的聚集系数 C 满足:

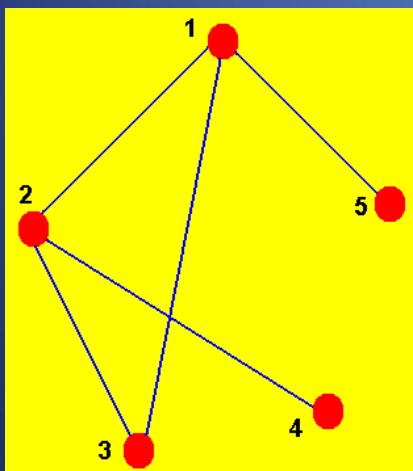
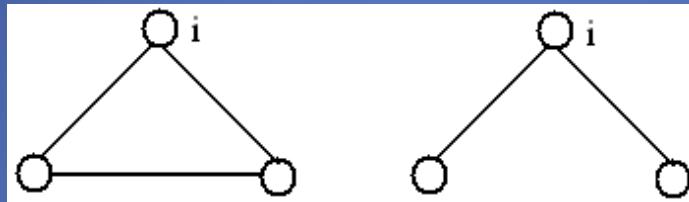
$$0 < C < 1$$

- $C = 1$ if 任意两个节点有连接
- $C = 0$ if 无三角形连接

- 大部分复杂网络有较大的 $C \rightarrow$ 小世界特征
- 富者越富, 马太效应

聚集系数的定义：

$$C(i) = \frac{\text{number of complete triangles with corner } i}{\text{number of all triangular graphs with corner } i}$$



Node-1 has 1 complete triangle and 3 triangular graphs, so $C(1) = 1/3$

Node-2 has 1 complete triangle and 3 triangular graphs, so $C(2) = 1/3$

Node-3 has 1 complete triangle and 1 triangular graph, so $C(3) = 1$

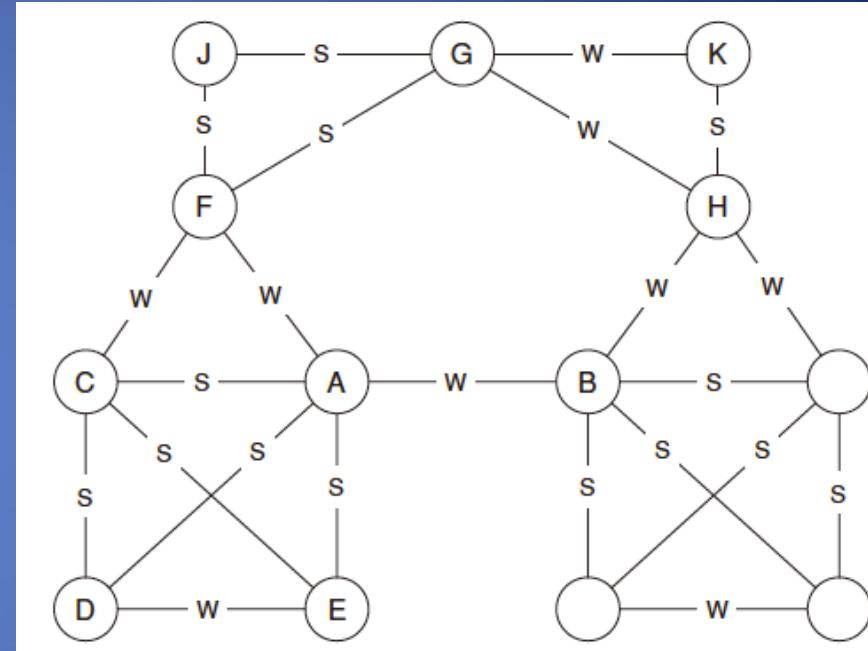
Node-4 has 0 complete triangles, so $C(4) = 0$

Node-5 has 0 complete triangles, so $C(5) = 0$

Average $C = (1/3+1/3+1+0+0) / 5 = 1/3$

关系的强度

- 含义：亲密程度 vs 这一阶段联系频度
 - 尽管“亲密”与“联系的频度”并不是独立的
- 程度：一定范围的数值 vs “强”和“弱”
 - 这里只用强弱，以突出核心思想；后面，会看到也可能用可以测量的某种数值来表达

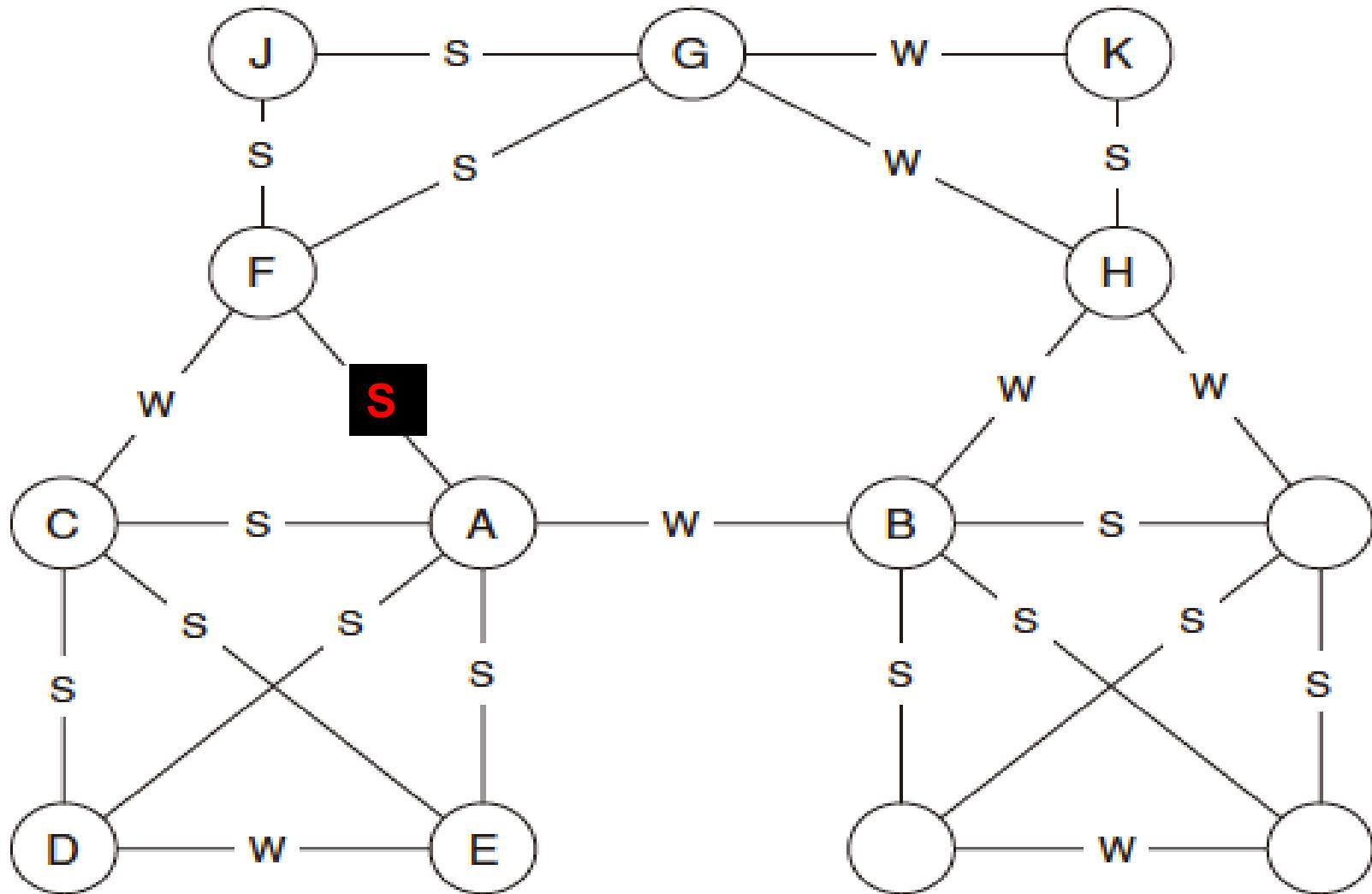


强三元闭包：在标注了关系强弱网络中的节点的一个属性

- 强三元闭包原理（假设）
 - 如果A-B和A-C之间的关系为强关系；则B-C之间形成边的可能性应该很高；
- 若A有两个强关系邻居B和C，但B-C之间没有任何关系（s或w），则称节点A违背了强三元闭包原理；
- 如果节点A没有违背强三元闭包原理，则称节点A符合强三元闭包原理。

注意：如同聚集系数，一个节点是否符合强三元闭包也是严格定义的，即每个节点要么“符合”，要么“违背”。

哪些节点符合 / 违背强三元闭包？



捷径 = 弱关系？

- 断言：若节点A符合强三元闭包，且至少有两个强关系邻居，则与A相连的任何捷径必定意味着是弱关系。
- （证明虽然很简单，但结论的意义重要，以及得到这个结论的思路漂亮）
- 纯数学的证明，得到了一个具有社会学意义的结论
- 这个结论将一个局部概念（关系）和一个全局概念（捷径）连接了起来

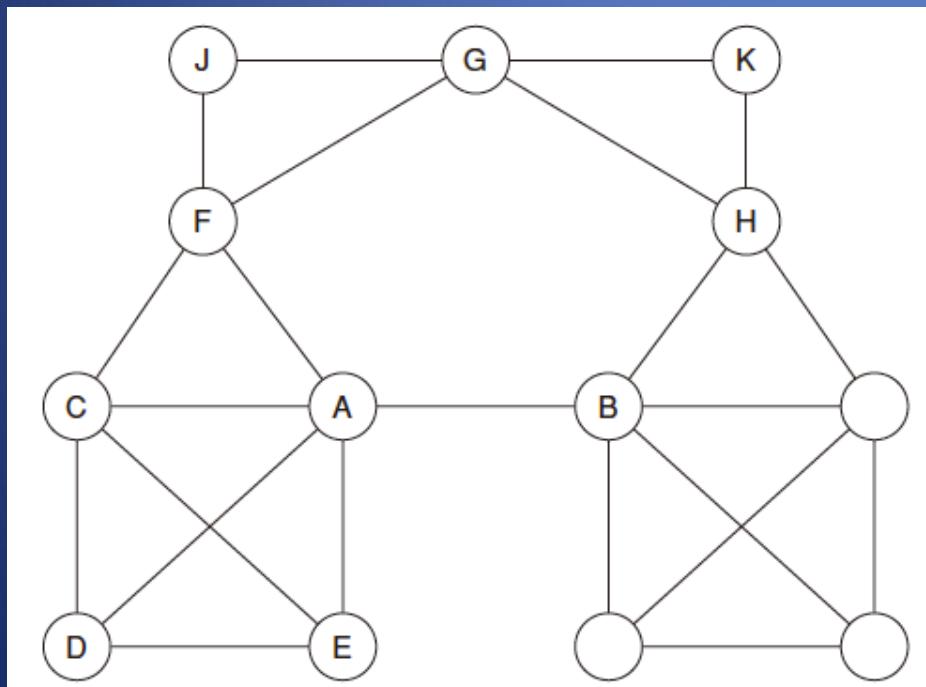
有没有数据来支持这结论？

- 上述结论的精神：两人关系的强度与是否有共同朋友直接相关
 - 捷径意味着没有共同朋友，强度为“弱”。
- 推论：共同朋友数越多，关系的强度越高
 - 精细一些，可以说共同朋友数在总朋友数中的占比（邻里重叠度）
- 我们来找一个能验证这个推论的场景，从而也就间接验证上述结论

用什么社交网络？如何定义关系的强度？

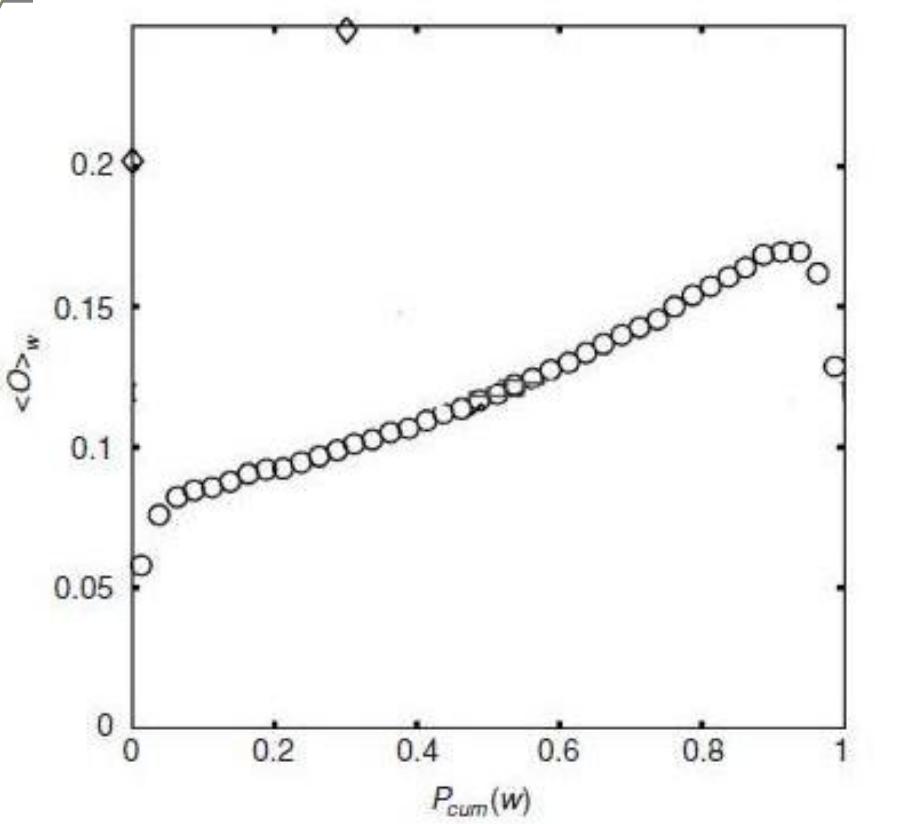
边 (A, B) 的邻里重叠度

- 与A和B相邻的节点数/与A或B相邻的节点数（不算A和B本身）



A-F边的邻里重叠度：
C既与F相邻，也与A相邻
与A或F相邻的则有B, C, D, E, G, J
则邻里重叠度为1/6

捷径 = 邻里重叠度为0的边



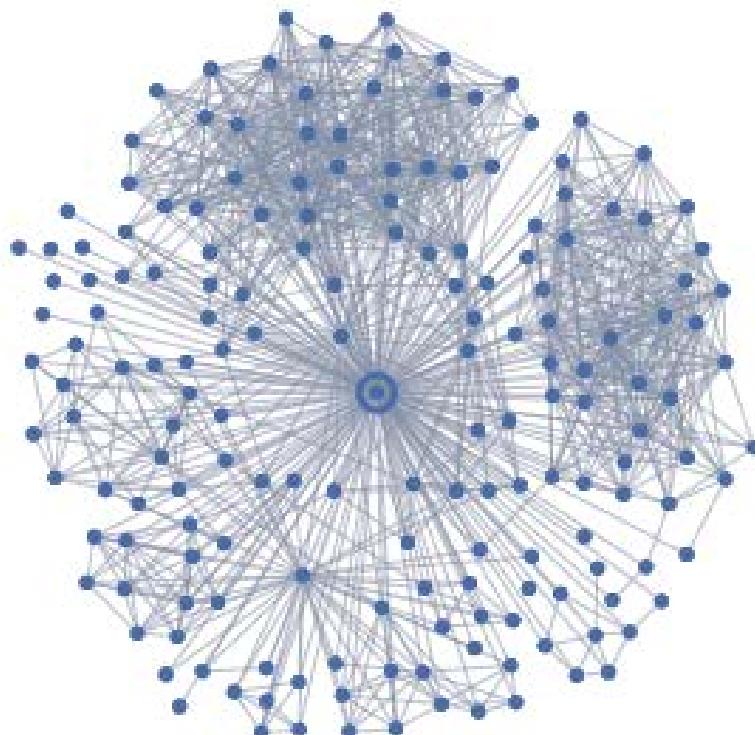
在手机通信网上的数据结果

- 美国全国人口的20%，
18周的通信数据
- 节点：手机号
- 边：通话关系
- 关系强度：通话时长

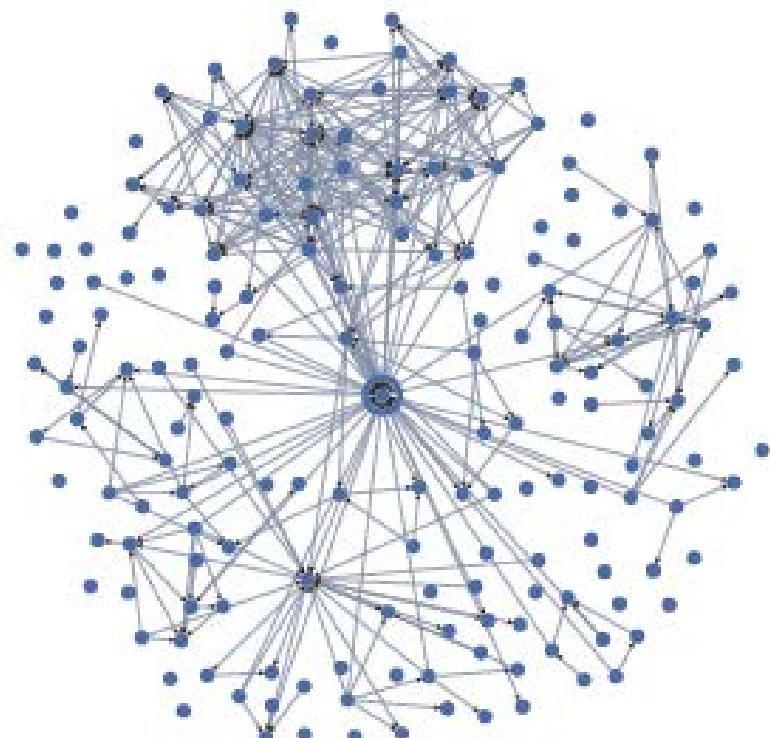
- 横轴表示边的关系强度（由低到高， %）
- 纵轴表示邻里重叠度
- 曲线表明这两个量正相关

OSN上关系强度的不同体现形式

All Friends



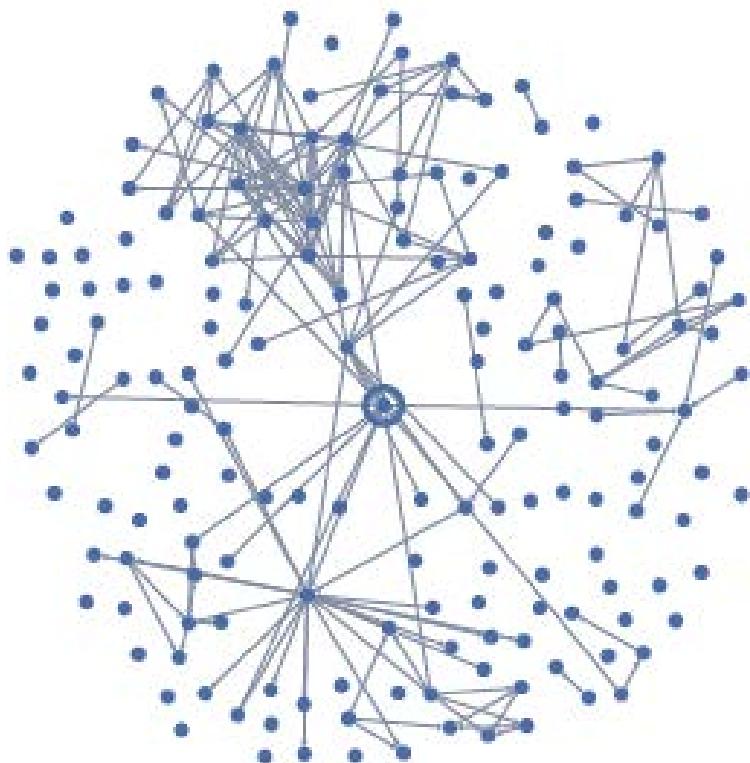
Maintained Relationships



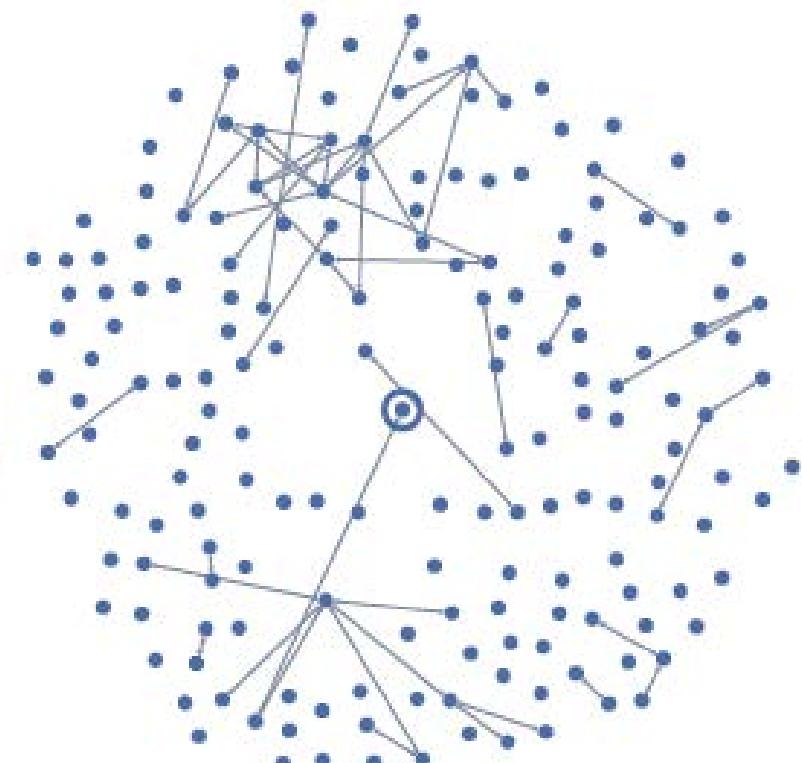
以Facebook为例，图中给出的是一个用户及其“朋友”之间的关系情况。
左图表表示有关用户自己给出的“好友”情况，其中许多实际没发生任何通信联系

Facebook的关系强度体现形式

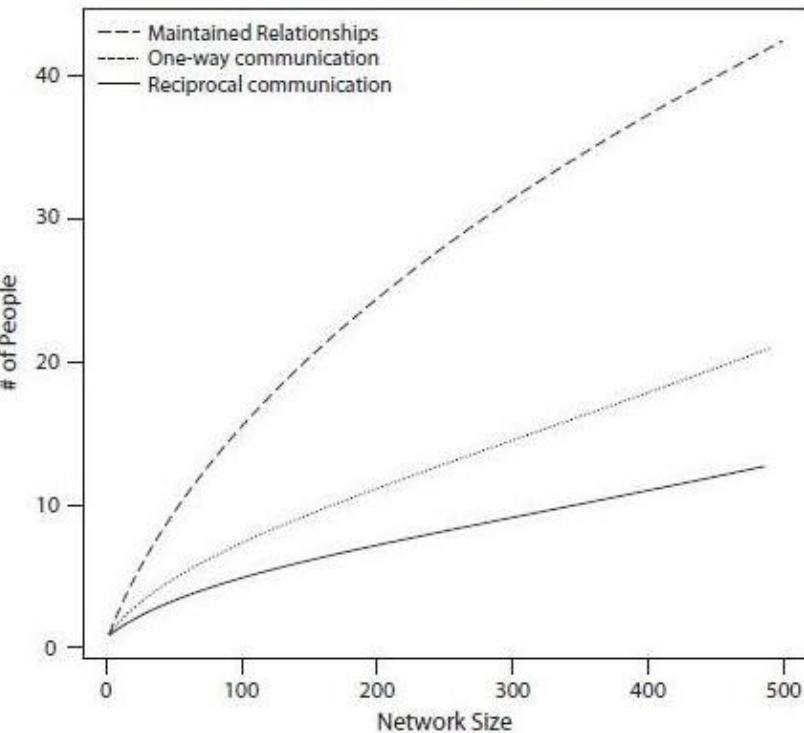
One-way Communication



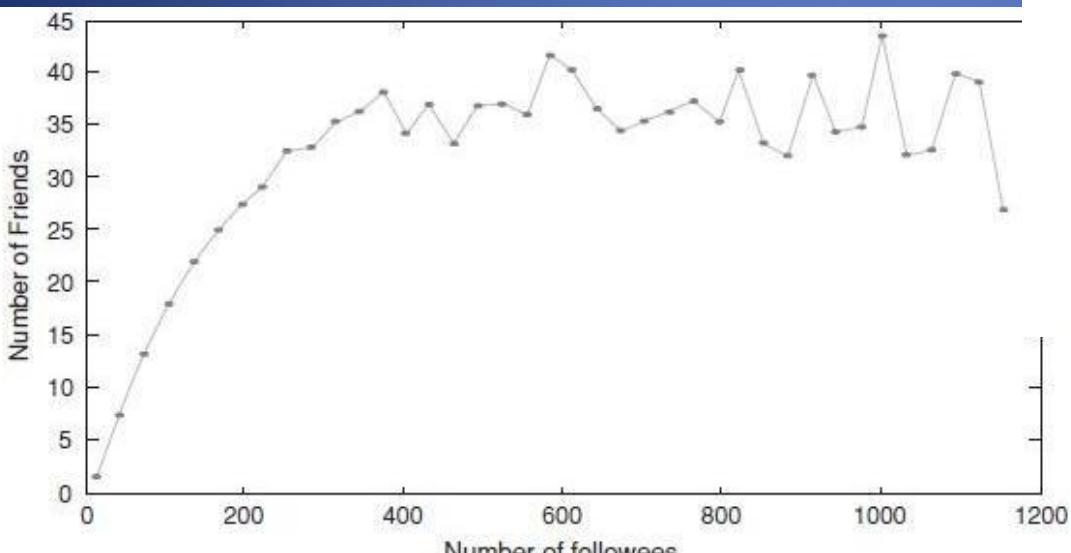
Mutual Communication



按照该项工作研究人员的定义，one-way包含mutual，因此我们看到左图包含右图所有的边。



Twitter上一个用户追随对象的个数与他实际联系的人数之间的关系

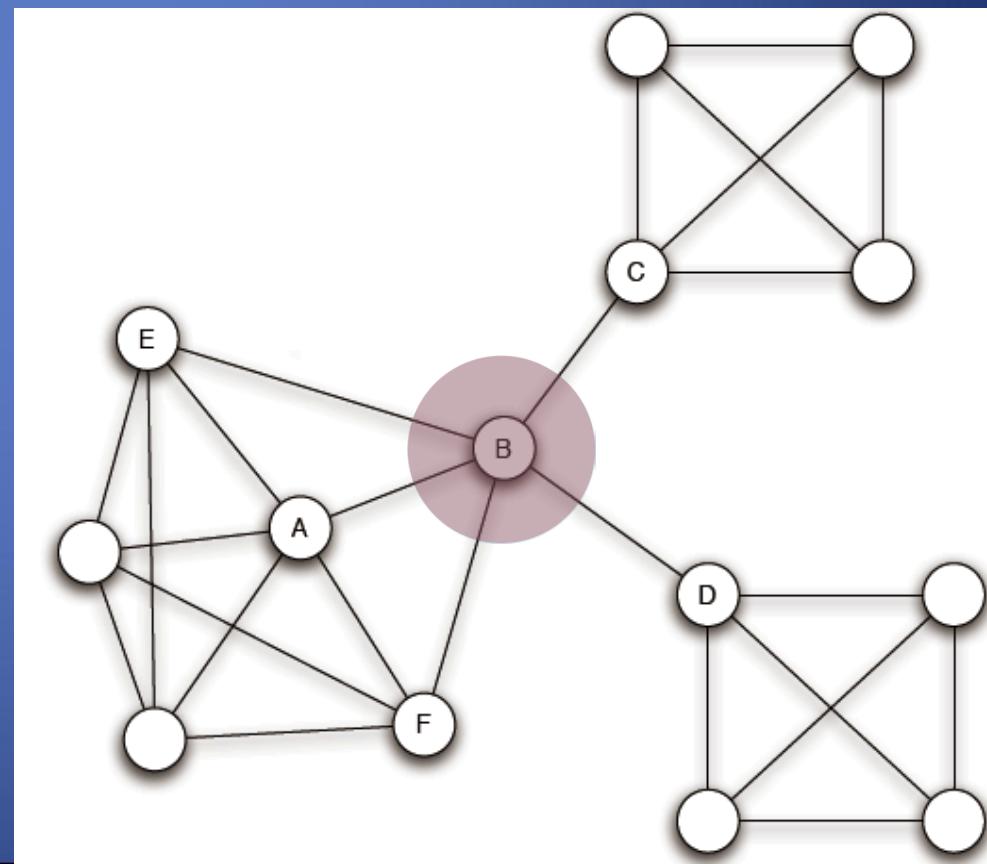


Facebook上一个典型用户好友之间联系的强度情况

从上述可以得到的一个定性结论是：在OSN上，尽管一个用户可以声明他关注大量（几百）其他用户，但实际关注的大约在**50**以下，而真正有联系的则更少，在**20**以下。

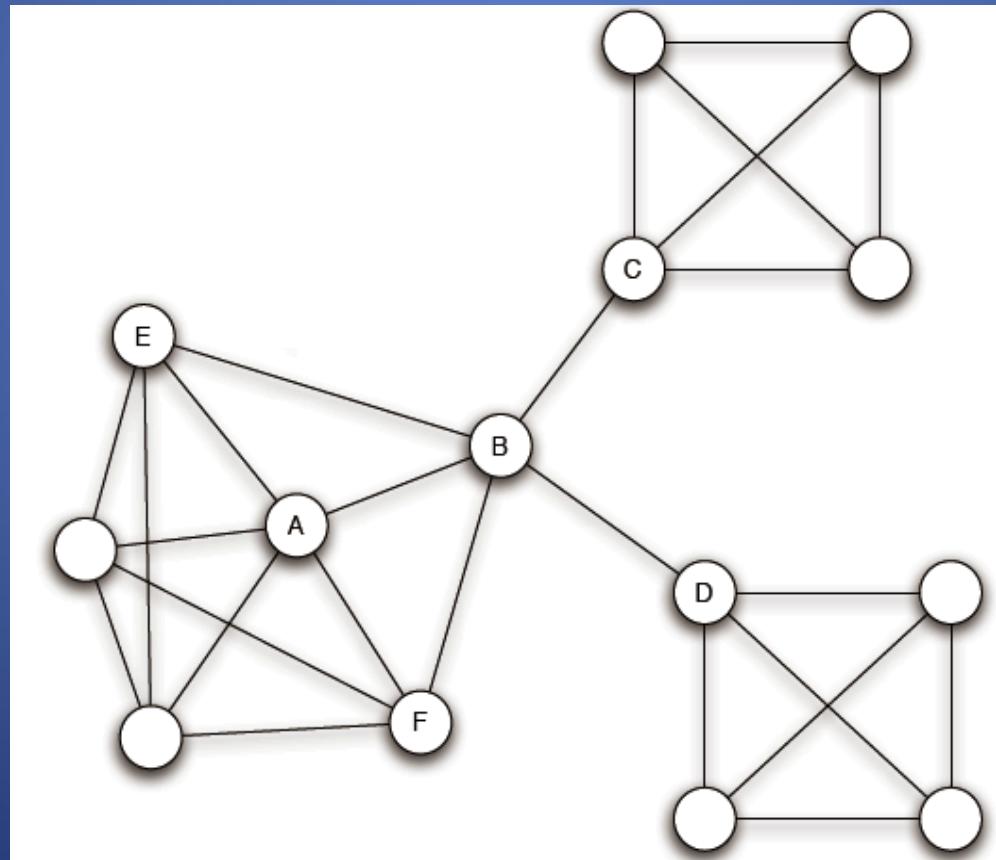
由上述，可得社会网络结构的一个基本意象

- 用桥（或者捷径，或者邻里重叠度很低的边，弱关系）连接起来的相对比较密集互连的节点群
 - 边的**嵌入性**：两个端点共同邻居的数量
- 其中，那些是多个桥的端点的节点（B）值得特别讨论
 - 聚集系数较低
 - 她与群组内部的节点（A）相比，有什么利弊？
 - 怎样与她打交道？



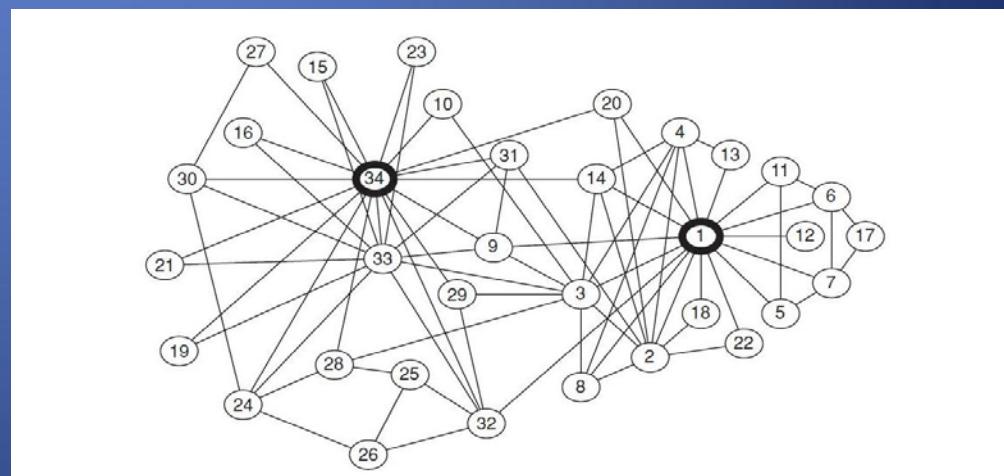
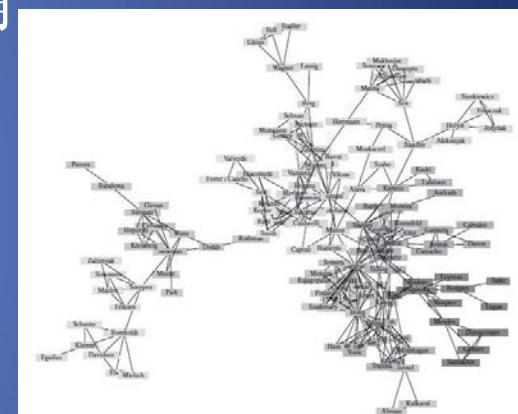
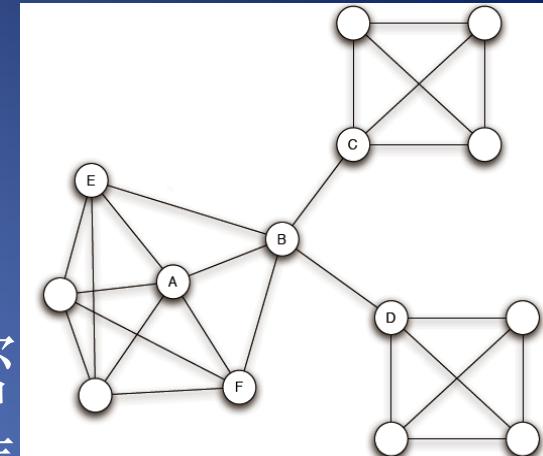
结构洞

- 结构洞：存在网络中两个或多个没有紧密联系的节点集合之间的“空地”

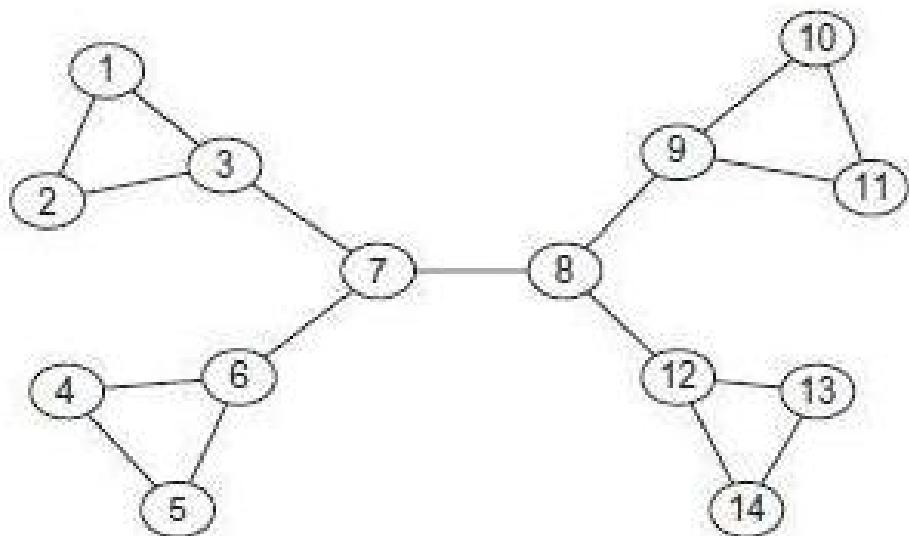


图划分算法

- 如何刻画社会网络中“相互紧密连接的节点群”？能否有一种精确的方法将它们找出来？
 - 分割法
 - 逐步去掉“跨接边”
 - 聚集法
 - “滚雪球”
 - 近似
 - 准确与效率的平衡

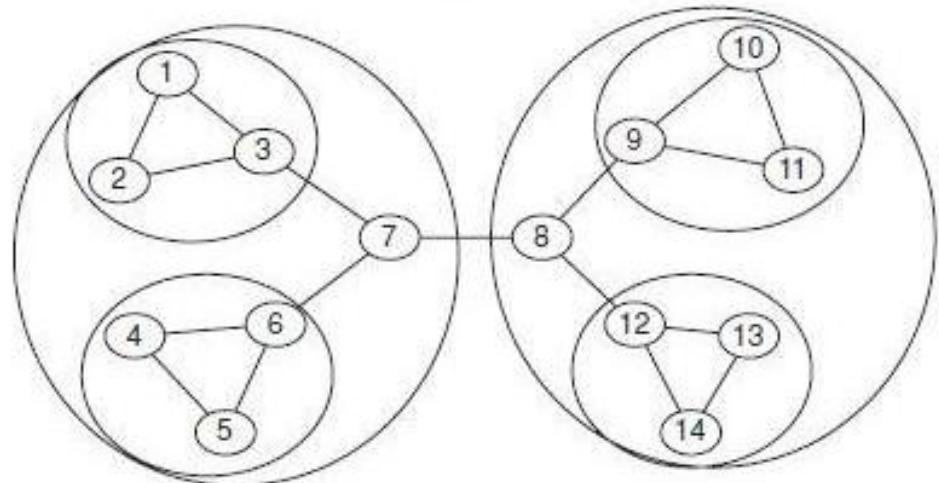


Girvan-Newman方法



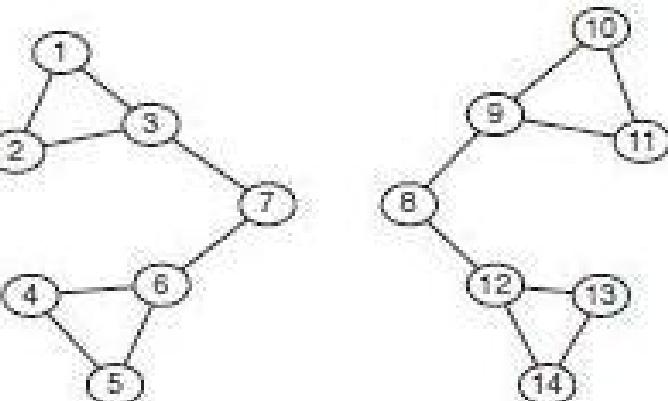
- 最先应该删除哪条边？

- 可以“一层层”进行

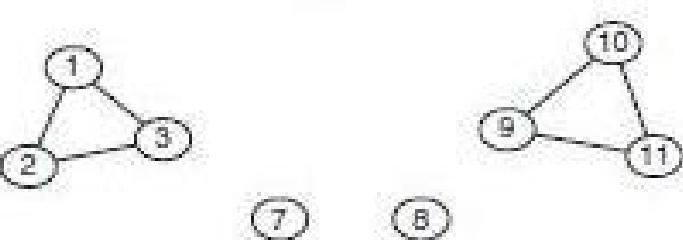


如何发现那些最“弱”的边？

- 或者“最关键”的边：许多节点之间的最短路径都要经过它
- 介数——一条边承载的一种“流量”
 - 两个节点A和B，设想1个单位的流量从A到B，均分到它们之间所有的最短路径上
 - k 条路径，则每条路径上分得 $1/k$,
 - 若一条边被 m 条路径共用，则在它上面流过 m/k
 - 所有节点对都考虑后，一条边上的累计流量就是它的介数（betweenness）



(a)

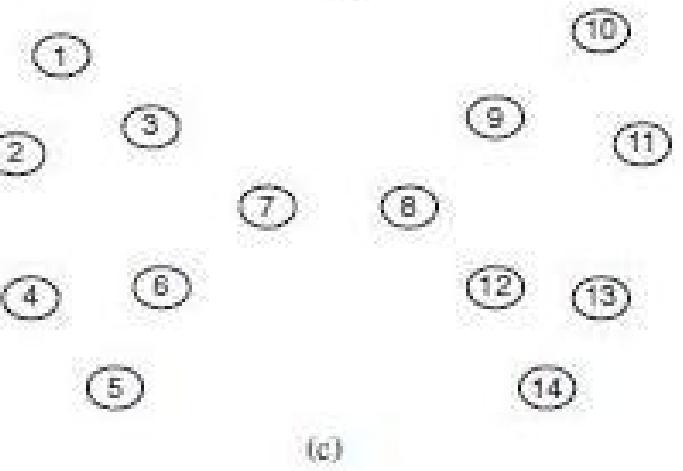


⑦

⑧



(b)



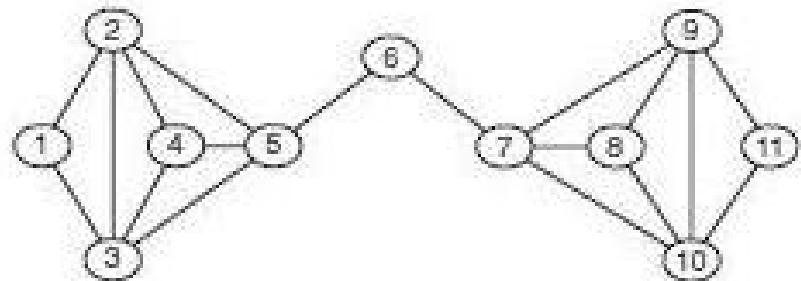
(c)

逐步删除高介数边：例

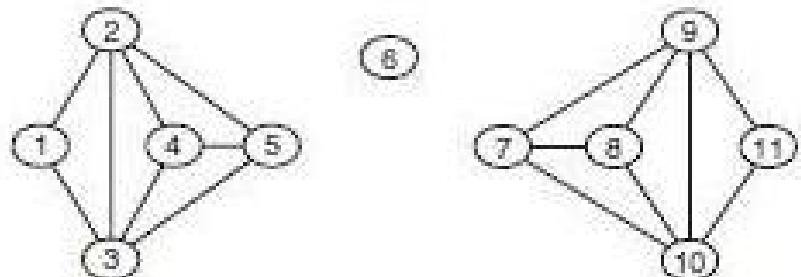
- $b(7,8) = 49$
 - 两边各7个节点，都要经过它， $7 * 7$; 7个节点内部则不经过
- $b(3,7)=b(6,7)=b(8,9)=b(8,12) = 33$
 - $3 * 7 + 3 * 4$
- $b(1,3)=\dots = 12$
 - 涉及1和3-14等12个节点
- $b(1,2)=\dots b(13,14) = 1$
 - 仅涉及1和2两个节点

去掉最高介数边后，重新计算剩下的...

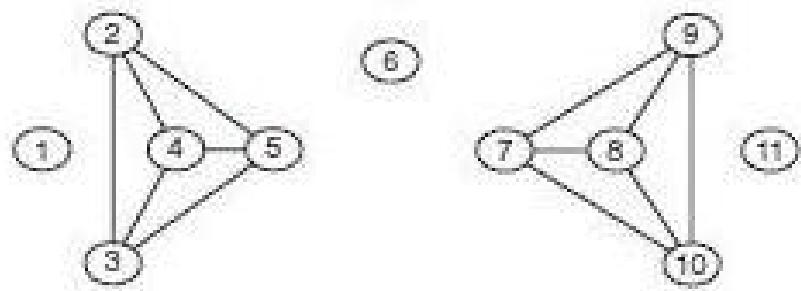
课堂练习



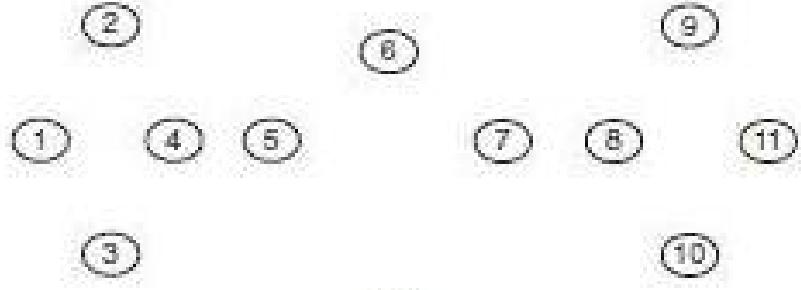
(a)



(b)



(c)



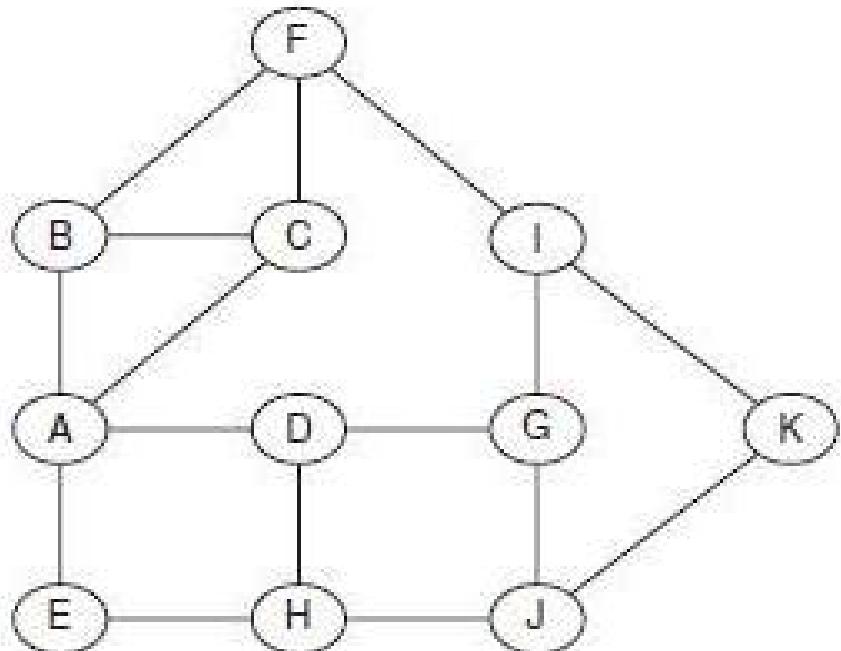
(d)

介数计算的一种算法

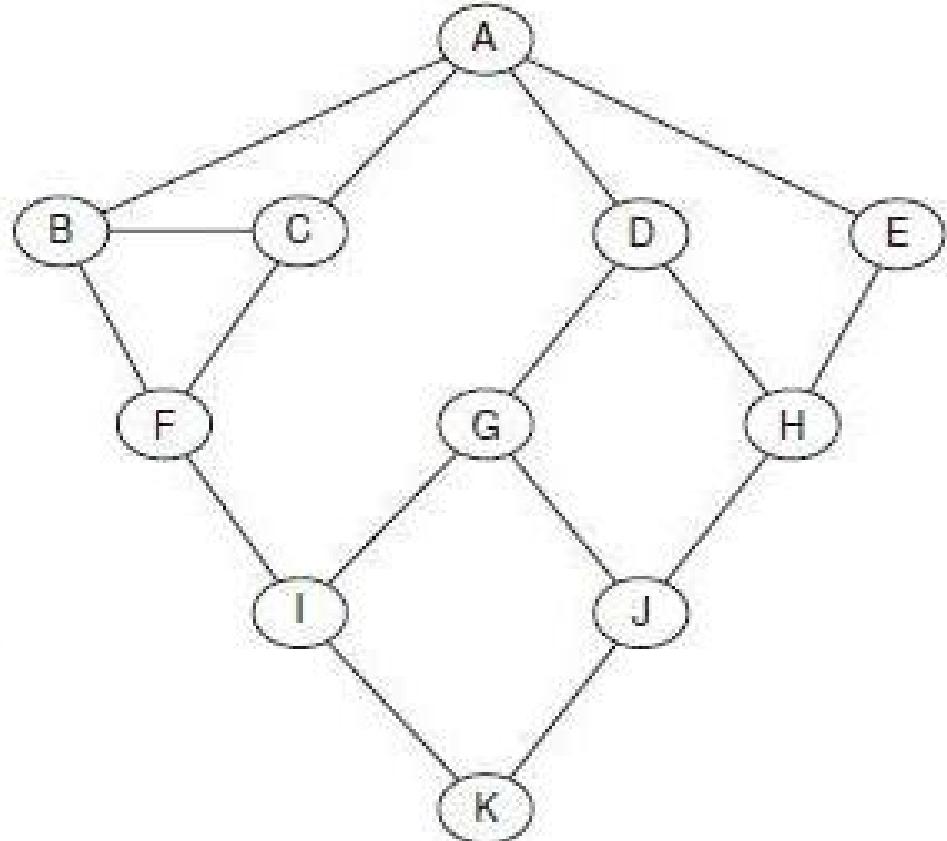
- 从一个节点（A）开始，做宽度优先搜索，将节点分层（以便于下面的步骤）
- 确定从A到其他每个节点的最短路径的条数
- 确定当从节点A沿最短路径向其他所有节点发送1个单位流量时，经过每条边的流量。

对每一个节点，重复上述过程，累计，除以2，即得每条边的介数。

例子：从A开始先宽搜索结果



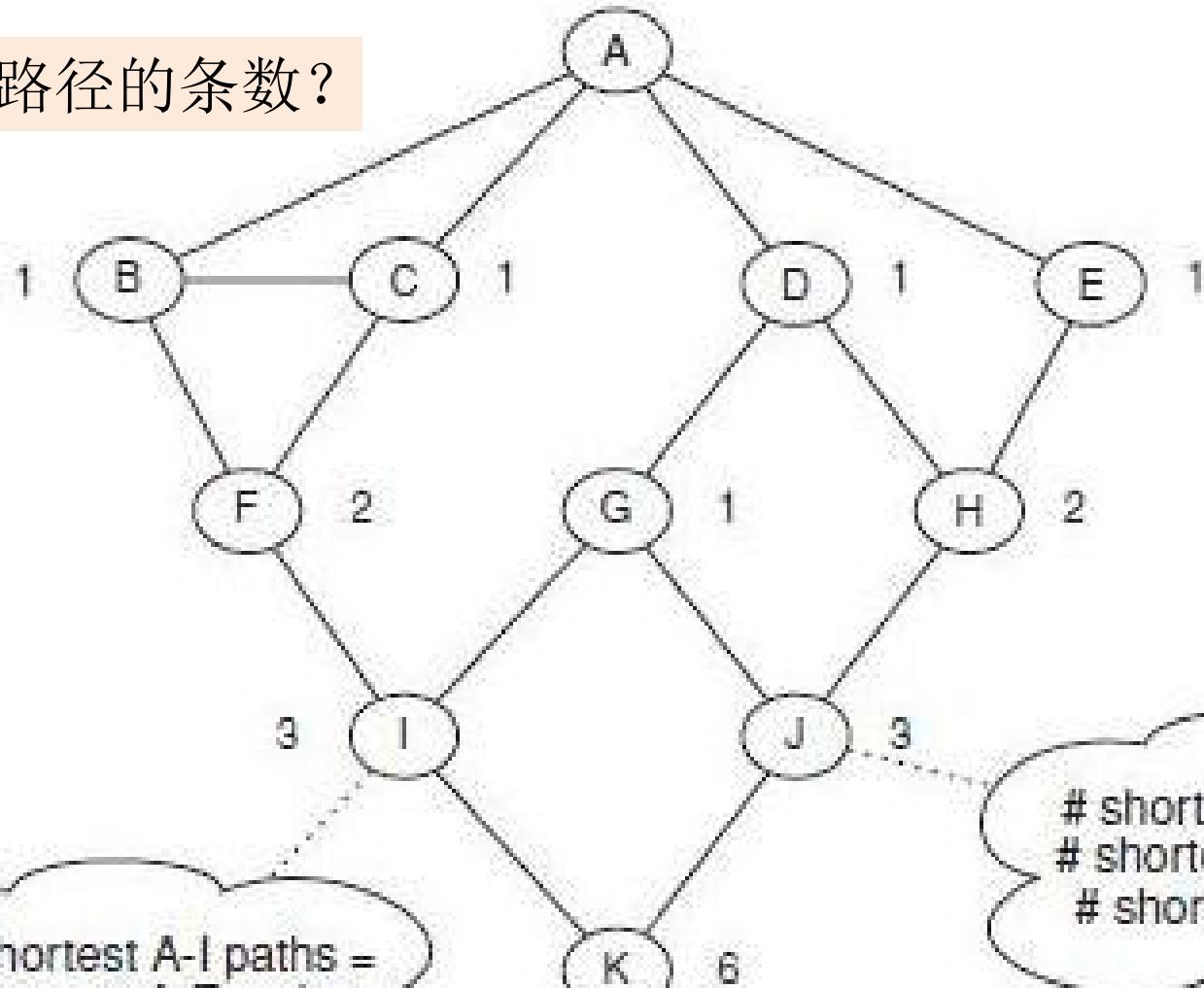
(a)



(b)

- 从A到K有多少条最短路径？（系统化方法）
- 层次就是最短路径的长度（距离）

最短路径的条数？



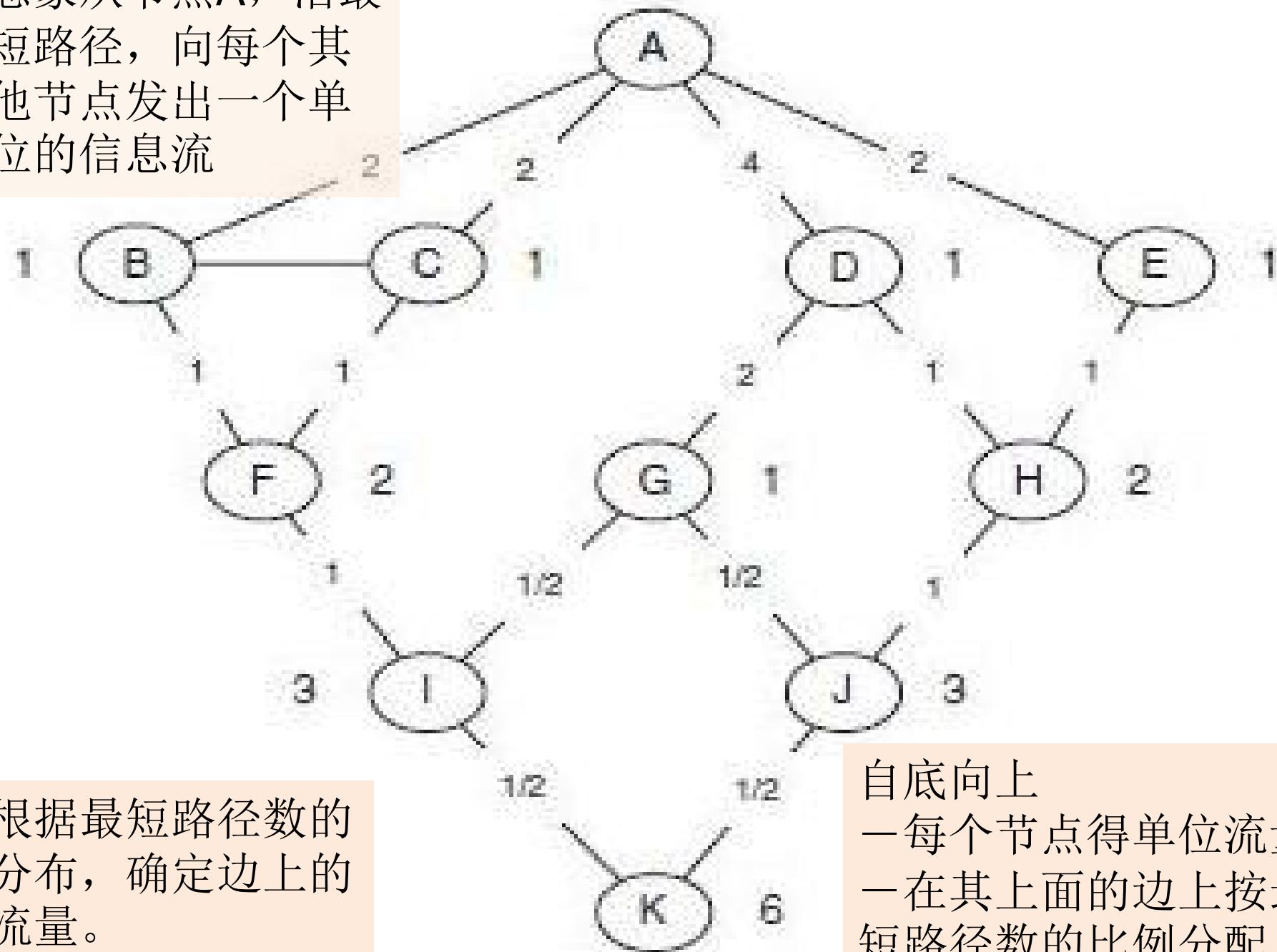
shortest A-I paths =
shortest A-F paths +
shortest A-G paths

shortest A-J paths =
shortest A-G paths +
shortest A-H paths

shortest A-K paths
= # shortest A-I paths
+ # shortest A-J paths

自上而下：
每个节点到A
的路径数，等
于它上面节点
路径数之和。

想象从节点A，沿最短路径，向每个其他节点发出一个单位的信息流



根据最短路径数的分布，确定边上的流量。

自底向上
— 每个节点得单位流量
— 在其上面的边上按最短路径数的比例分配

作业

- 第3章 1,3,5

同质性：影响社会网络结 构的重要外部因素

前面讨论的“三元闭包”影响网络结构，属于
网络自身的因素（内部因素）

提要

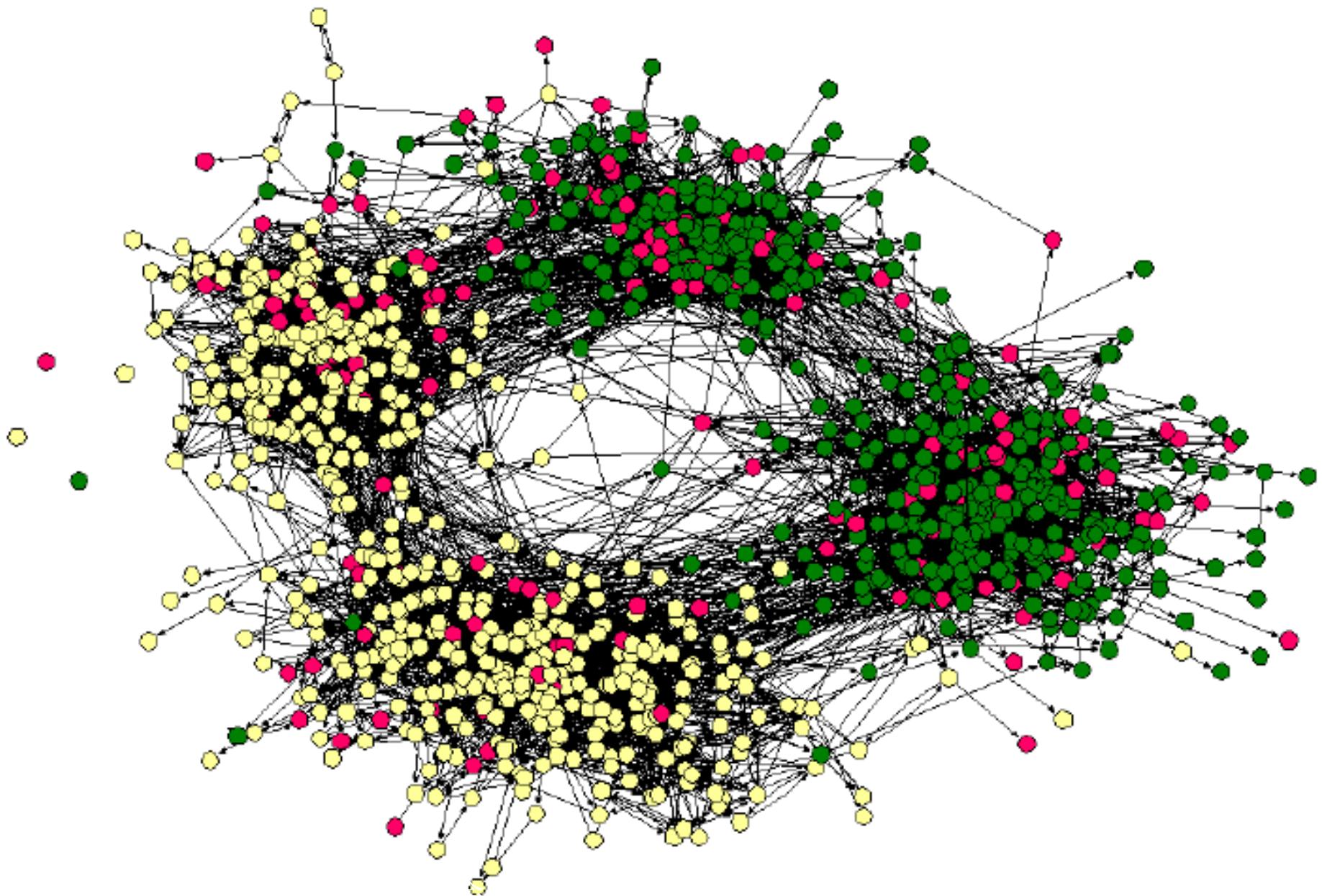
- 同质性的基本含义
- 同质性在网络中的一种评估方式
- 归属网、社会归属网对同质性的影响
- OSN数据分析例子
- 隔离——同质性的一种现象，谢林模型

同质性 (homophily)

物以类聚，人以群分
近朱者赤，近墨者黑

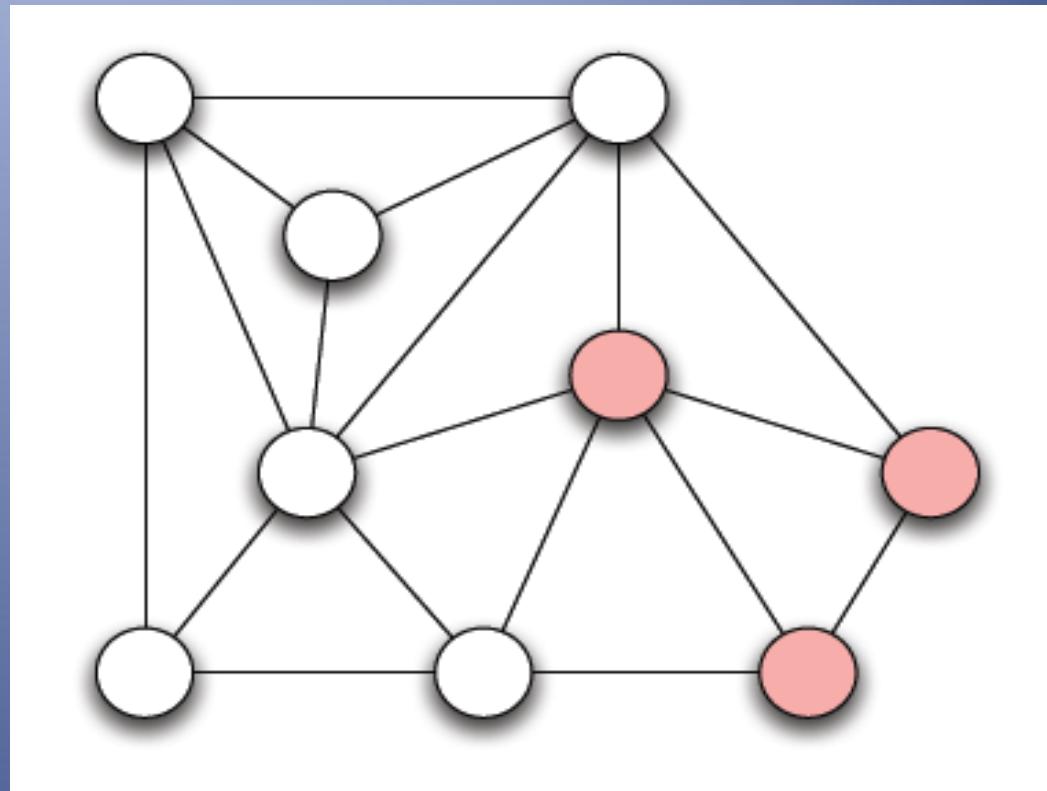
- 每个人的特质可分为两种：
 - 固有特质：性别、种族、母语等
 - 可变特质：居住区、爱好、专长、思想等
- 基本问题：因为相似，才成为朋友（selection）；还是因为成为朋友后变得相似（social influence）？
 - 给定相似性，区别“选择”与“社会影响”的作用程度是社会学研究中一个经典问题

一个反映较强同质性的社会网络

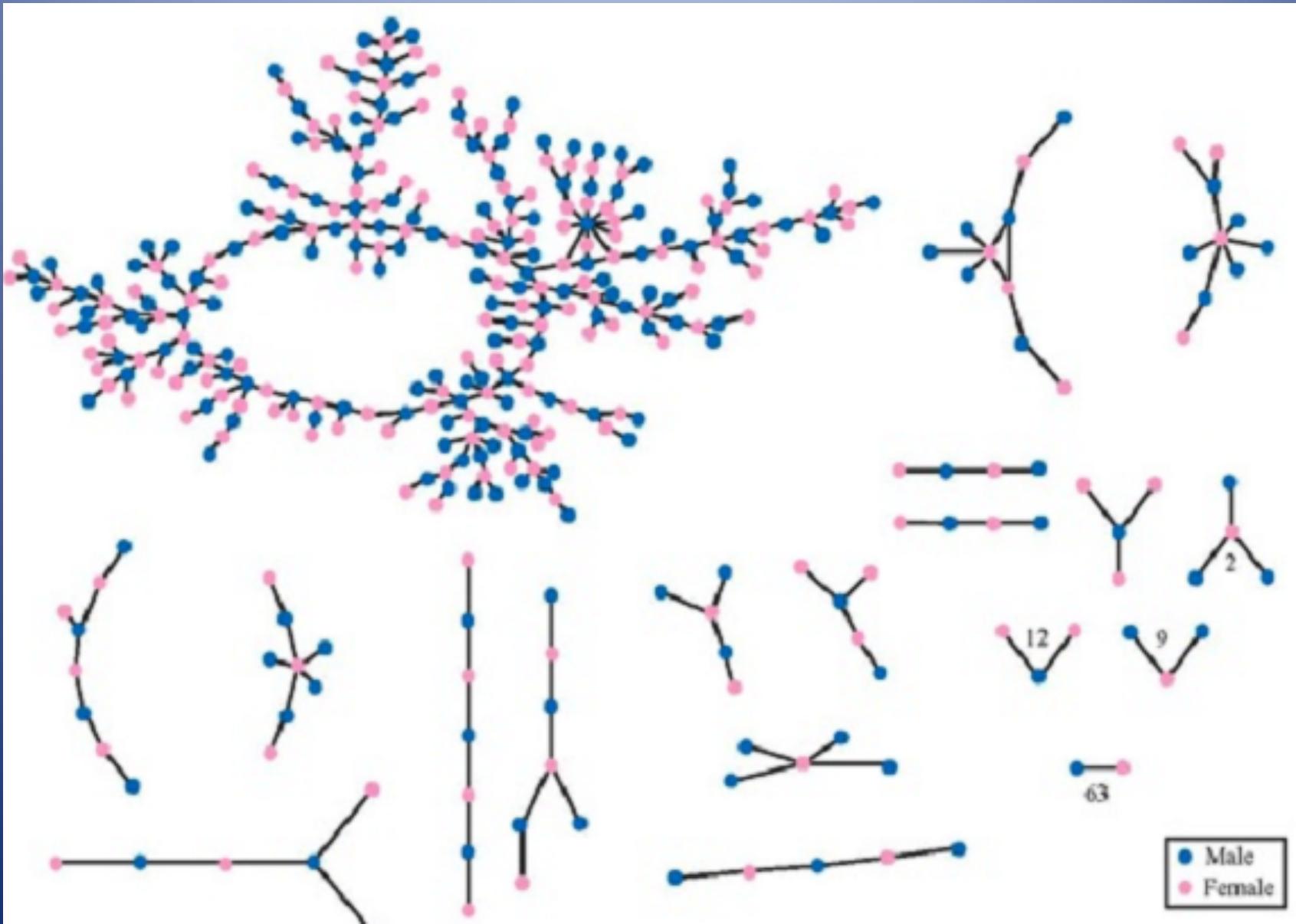


社会网络中同质性判别的一个测度

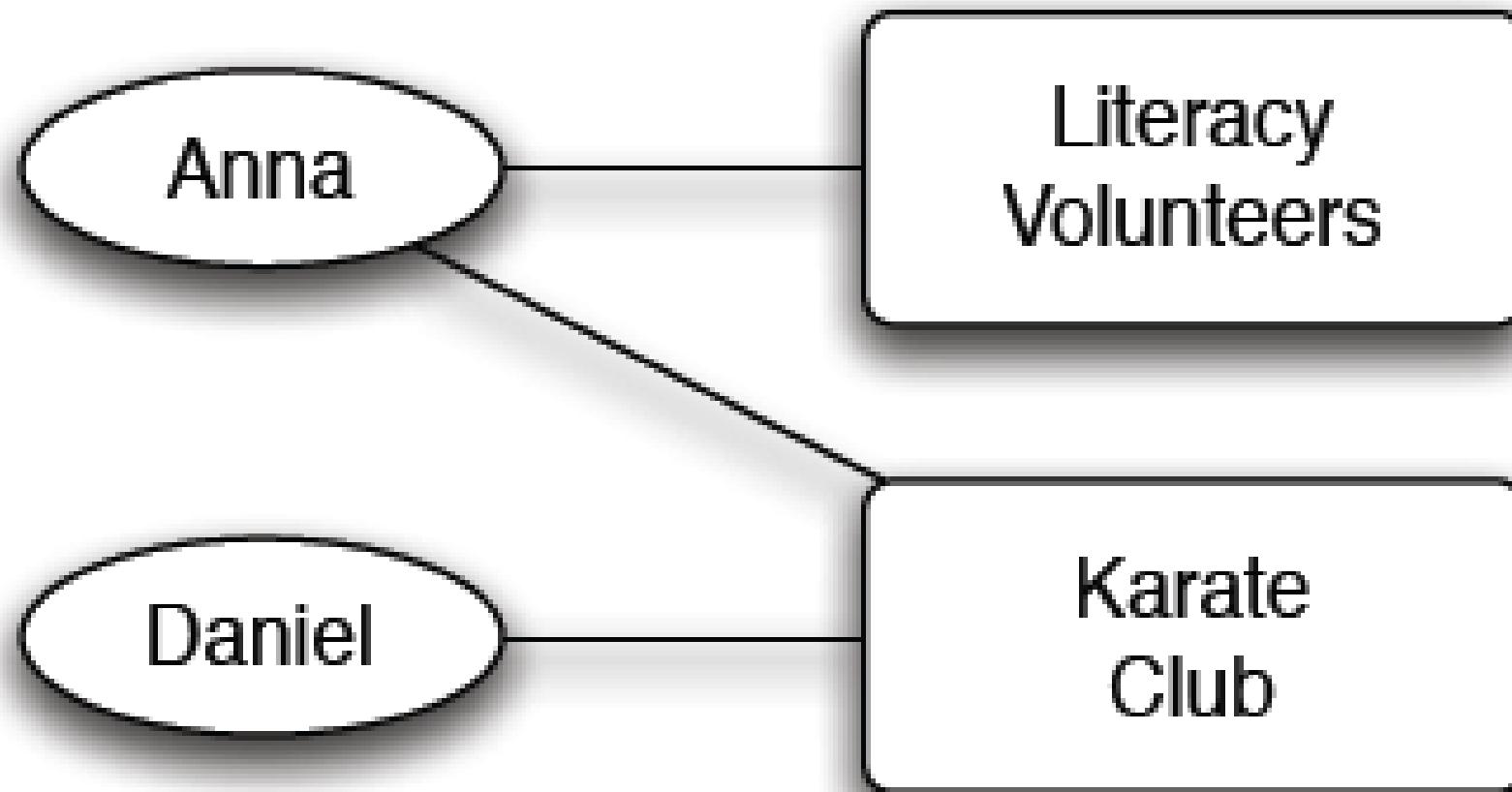
- 假设人们按照两种属性区分（例如性别），我们来看如何判断社会网络中同质性体现的程度
 - 直观上：相同颜色的节点的聚集程度
 - 如何定量把握？
- 端点颜色相同的边如果太多，则同质性迹象明显
- 多少算“太多”？（什么是基准？—“混合得比较均匀”）
 - 考虑不同颜色节点的占比，随机给节点着色的情况

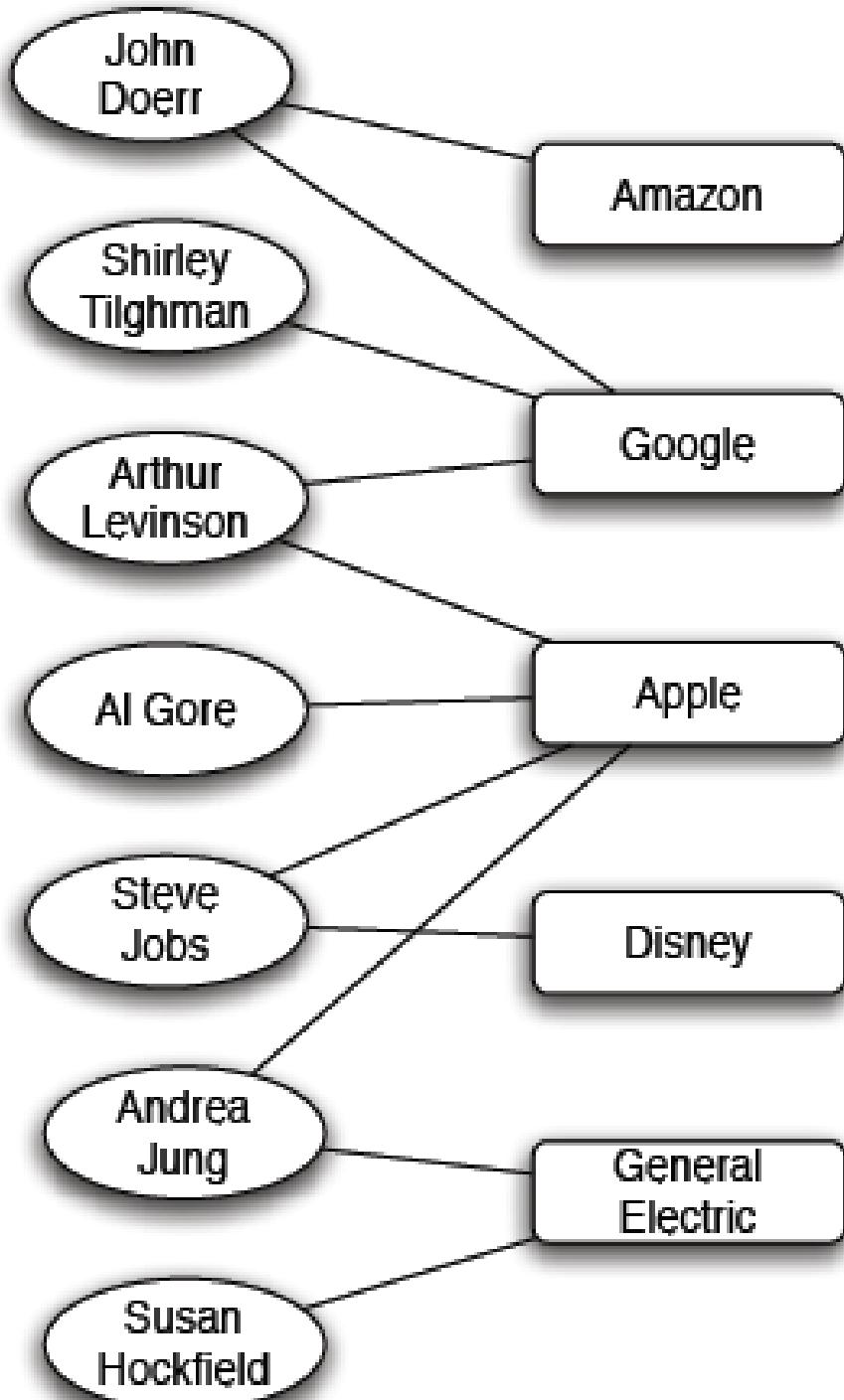


逆同质性社会网络的一个例子



归属网络：描述人与相遇机会的关系

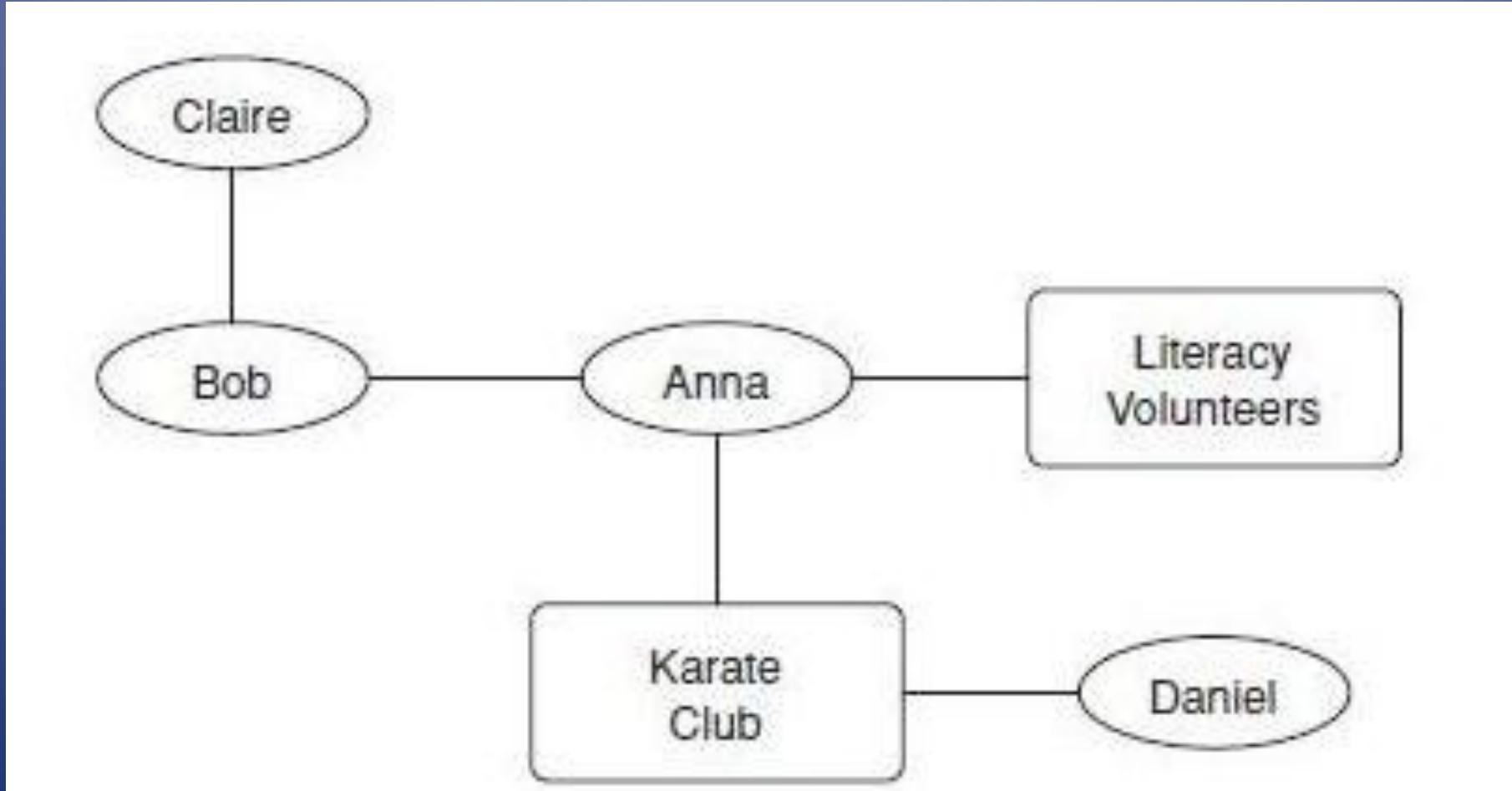




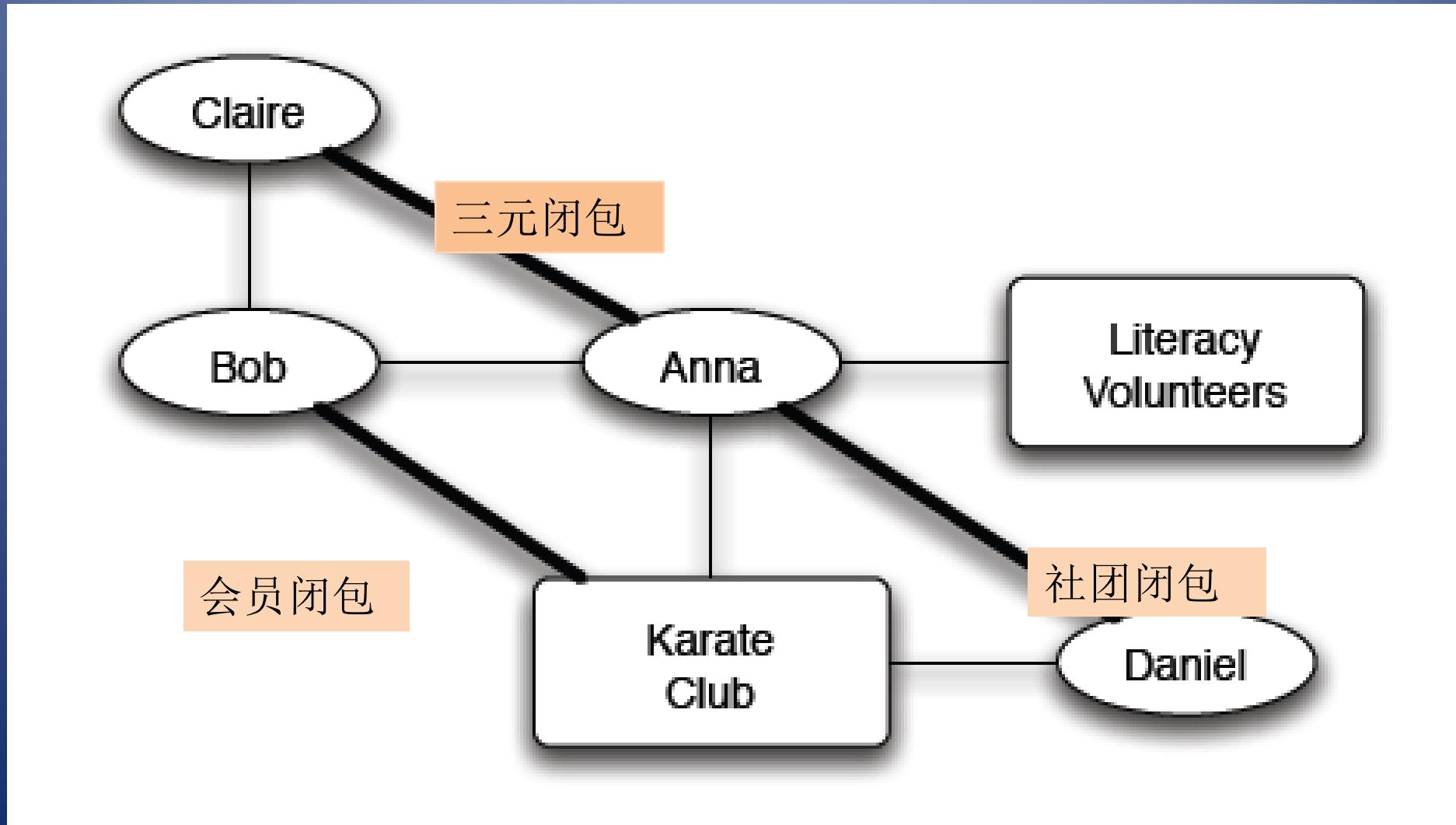
归属网络的一个实际例子

- 名人可能在多家公司的董事会中兼职，分析这种结构，以及他们之间的个人关系，对理解公司的行为有帮助
- 这种现象也可能造成一些丑闻

社会归属网：考察同质性的一个工具

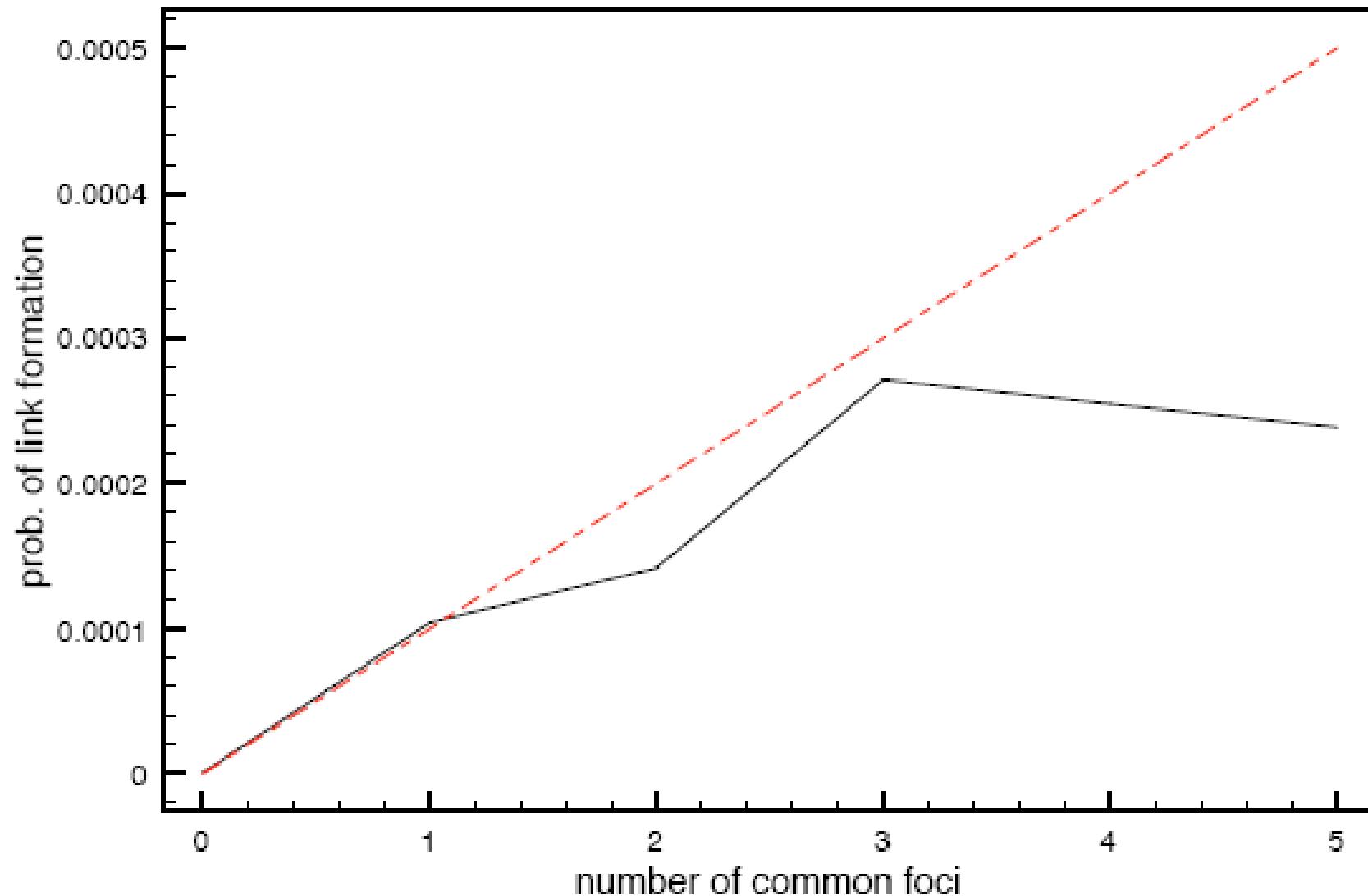


社会归属网： 三元闭包、社团闭包、会员闭包可同时显现



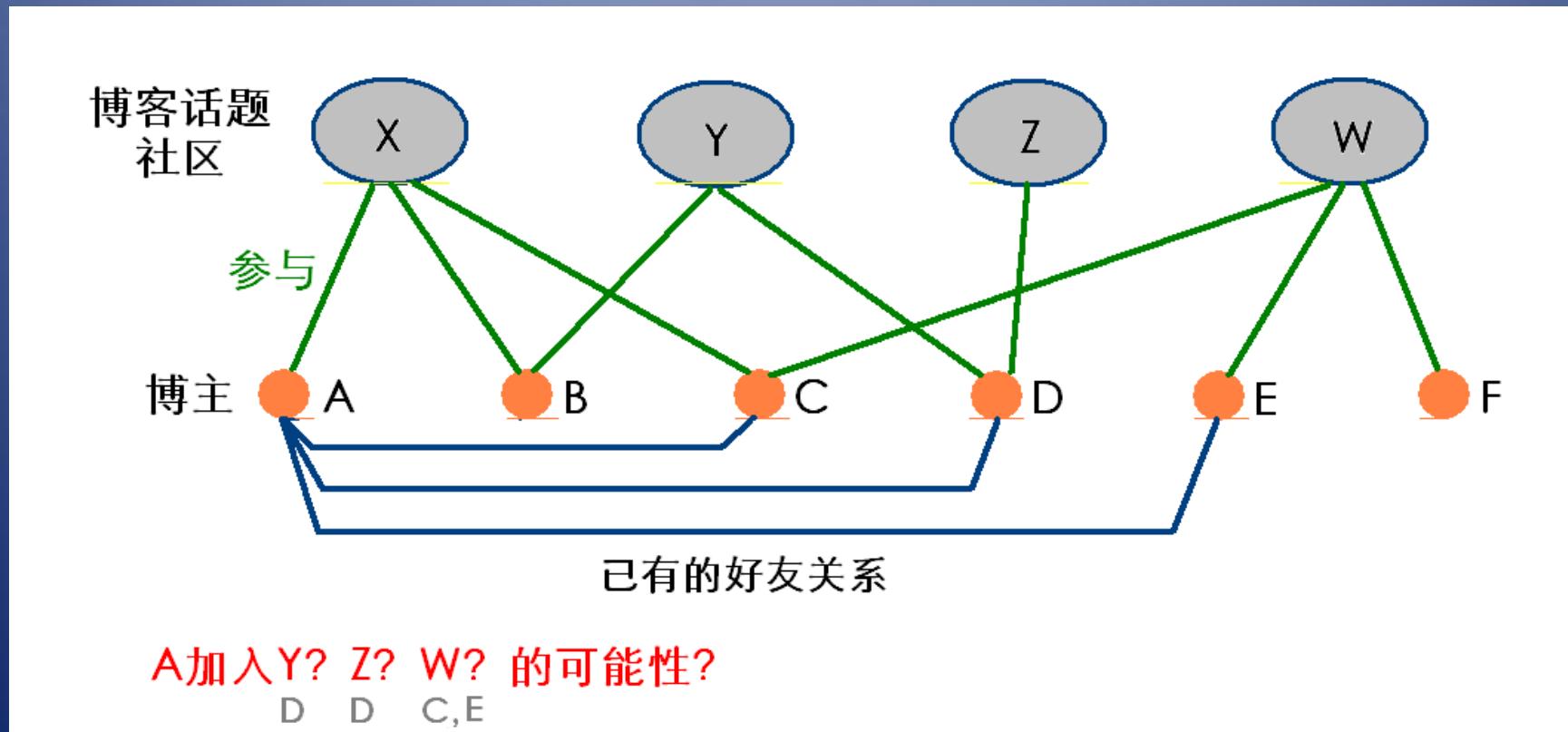
- 社团闭包→选择； 会员闭包→社会影响

社团闭包验证：基于共同兴趣建立联系的概率

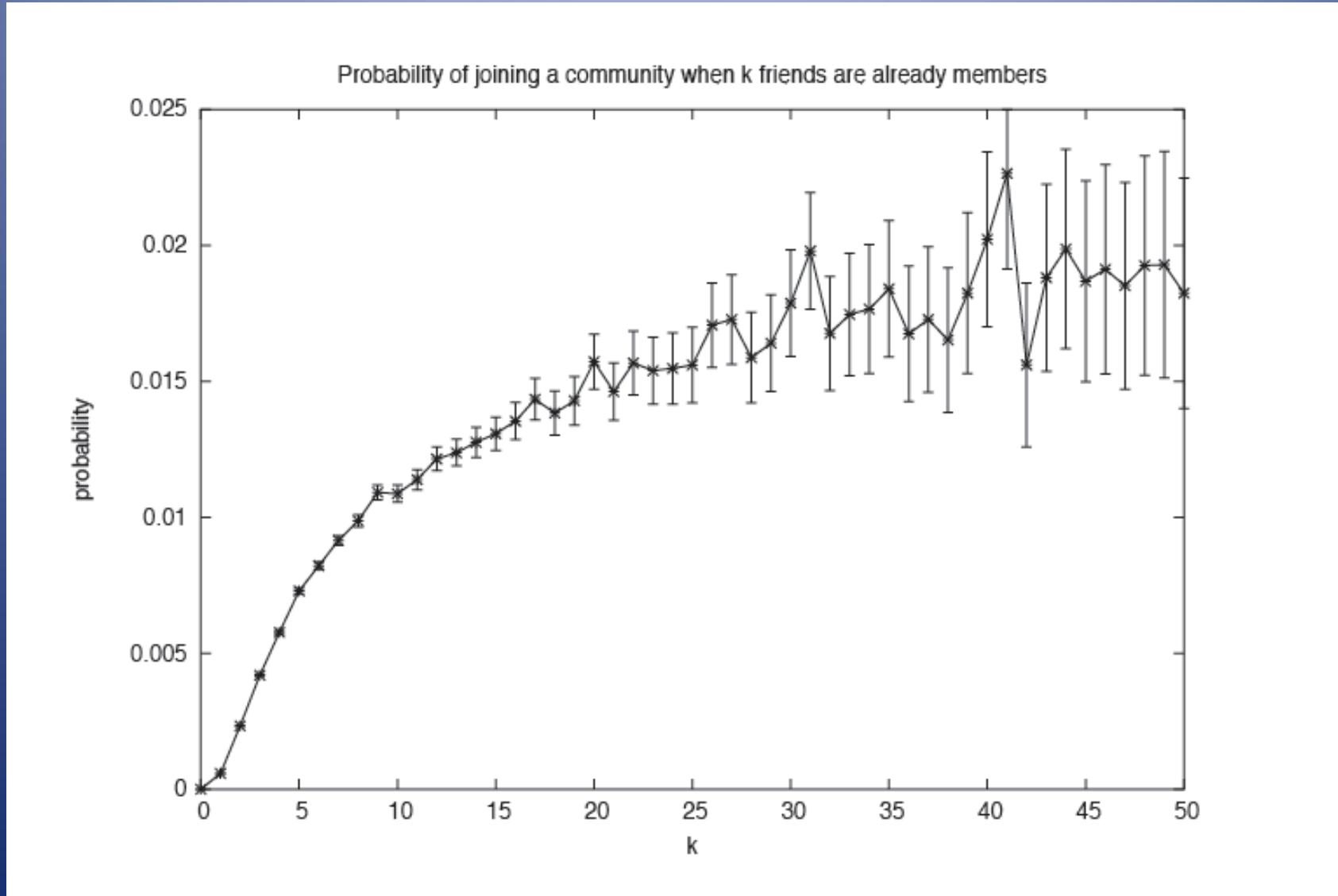


利用在线数据支持同质性研究

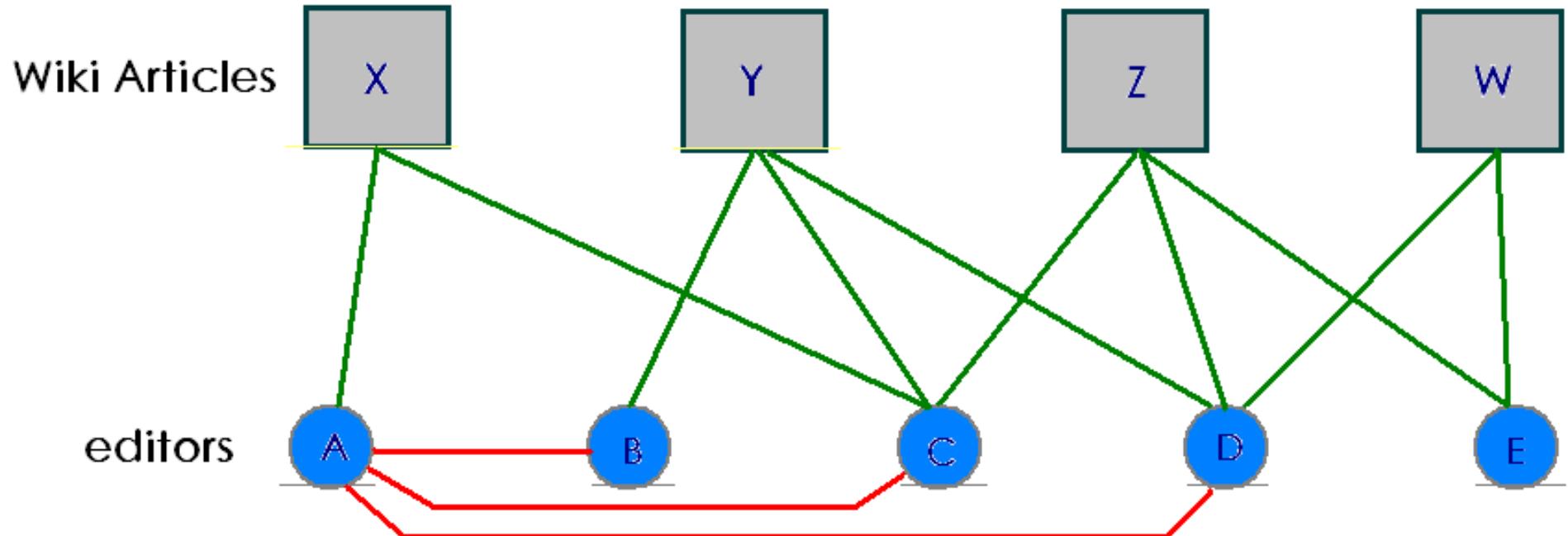
- 定义（社会归属网）：人、社团；人和人的关系，人和社团的关系
 - 社团：某些人共同参与的活动



会员闭包（社会影响）： 基于相关朋友数，参与一个博客话题社区的概率



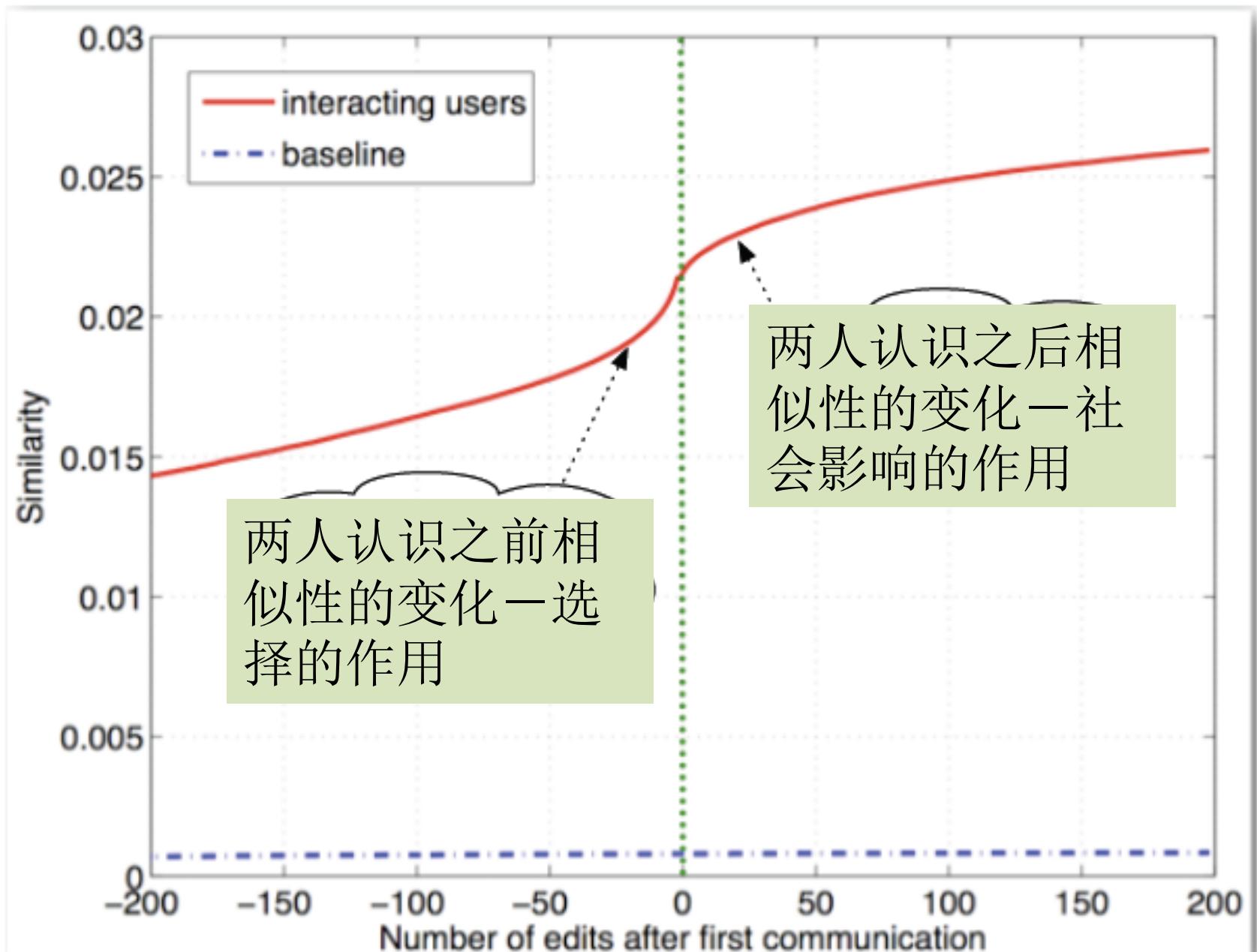
利用在线数据研究选择与社会影响的互动



每个编辑有一个“user talk page”，其他编辑可以在上面留言，从而构成通信（社会网络）关系。

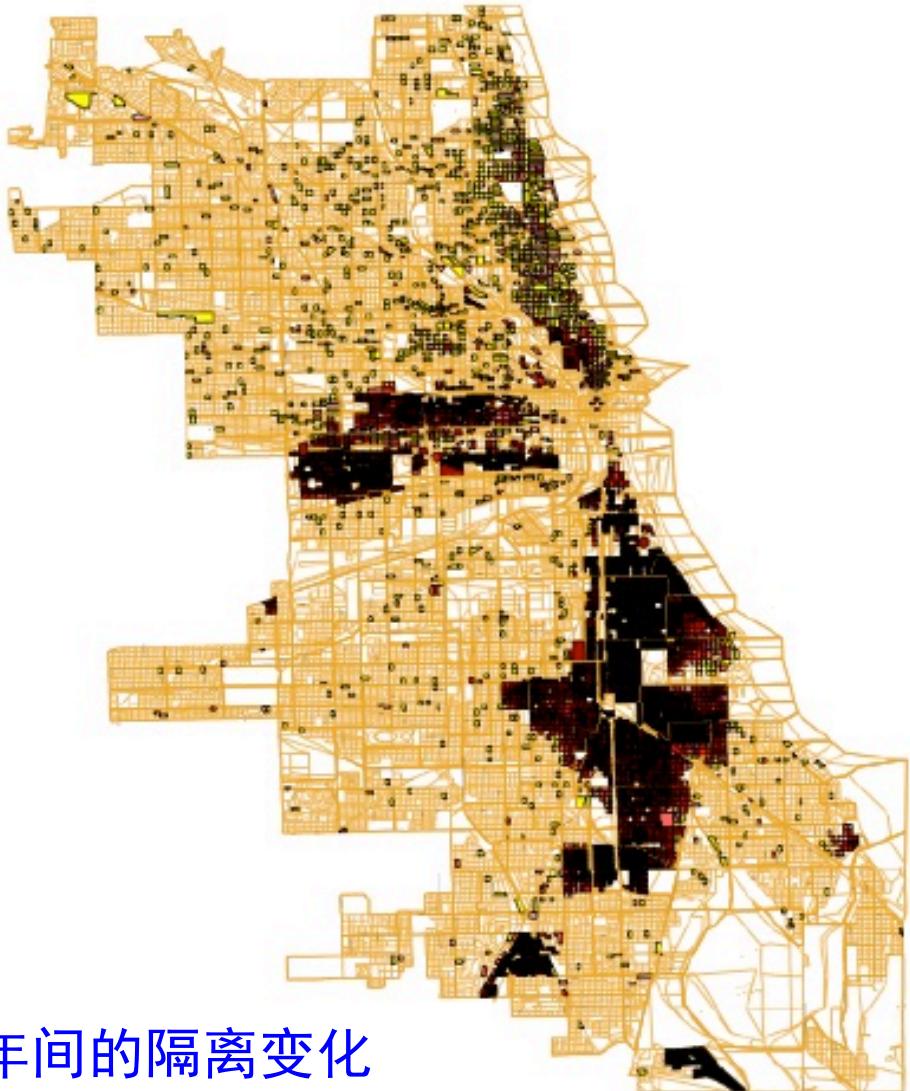
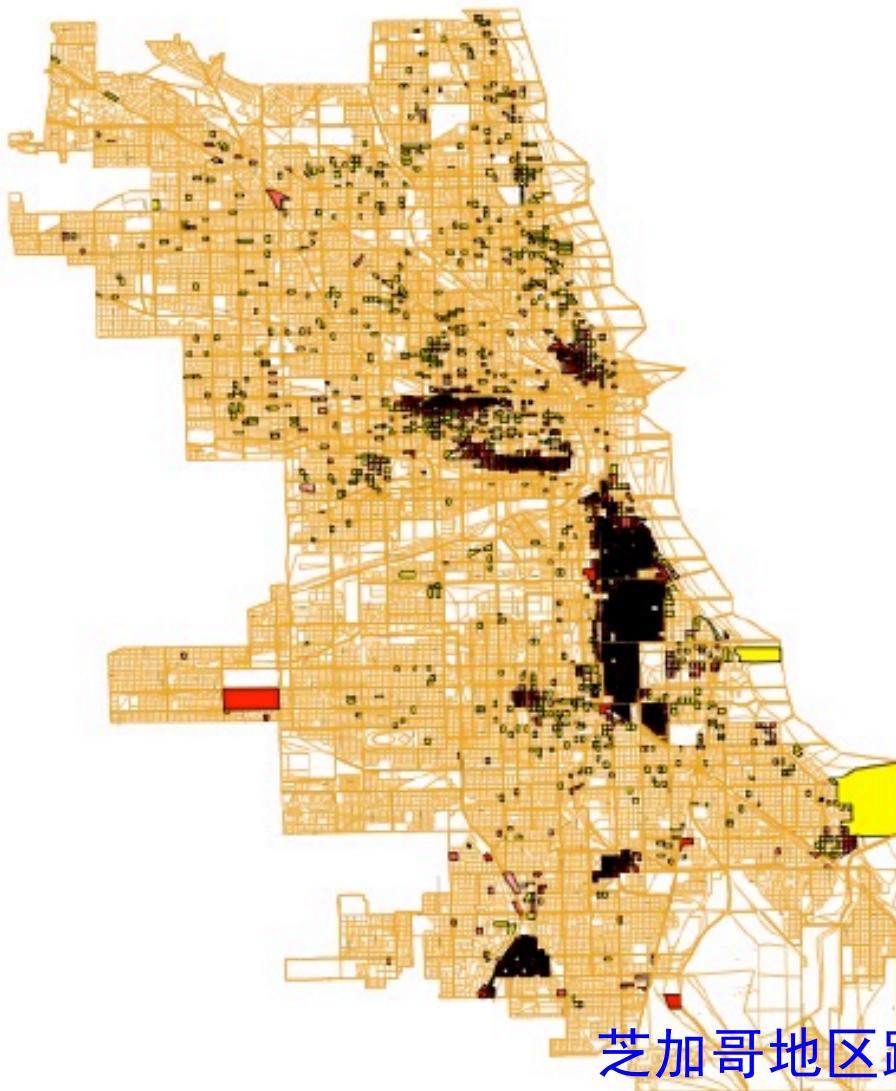
- 两个编辑之间相似性的变化与“选择”和“社会影响”的关系
 - 定义“相似性” — 编辑文章（与 / 或）
- 没有联系（通信）之前，相似主要因为选择；达到足够相似度时则容易发生联系，然后社会影响开始对相似性提高起作用

相似性、选择与社会影响



隔离 (segregation)

- 同质性影响下发生的过程

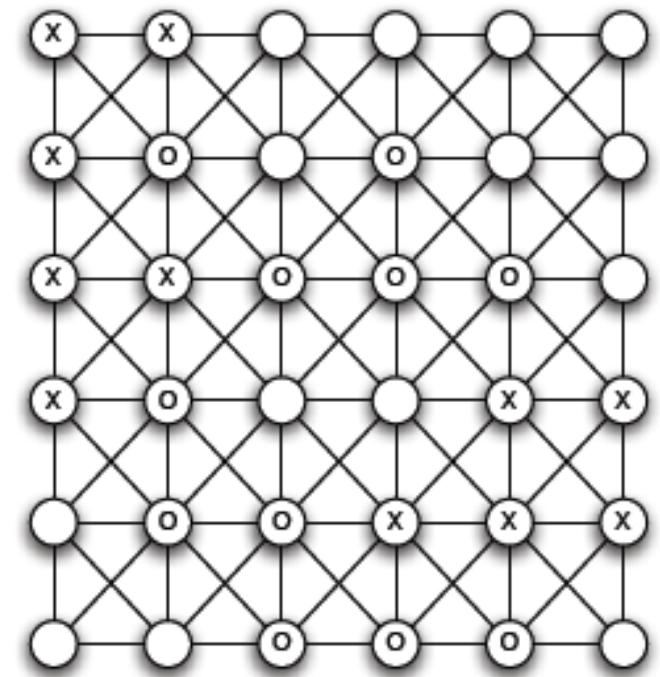


芝加哥地区跨20年间的隔离变化

谢林模型示意

- 解释隔离现象（宏观，全局）并不一定是个体刻意选择（微观，局部）的结果

x	x					
x	o			o		
x	x	o	o	o		
x	o				x	x
	o	o	x	x	x	
		o	o	o		

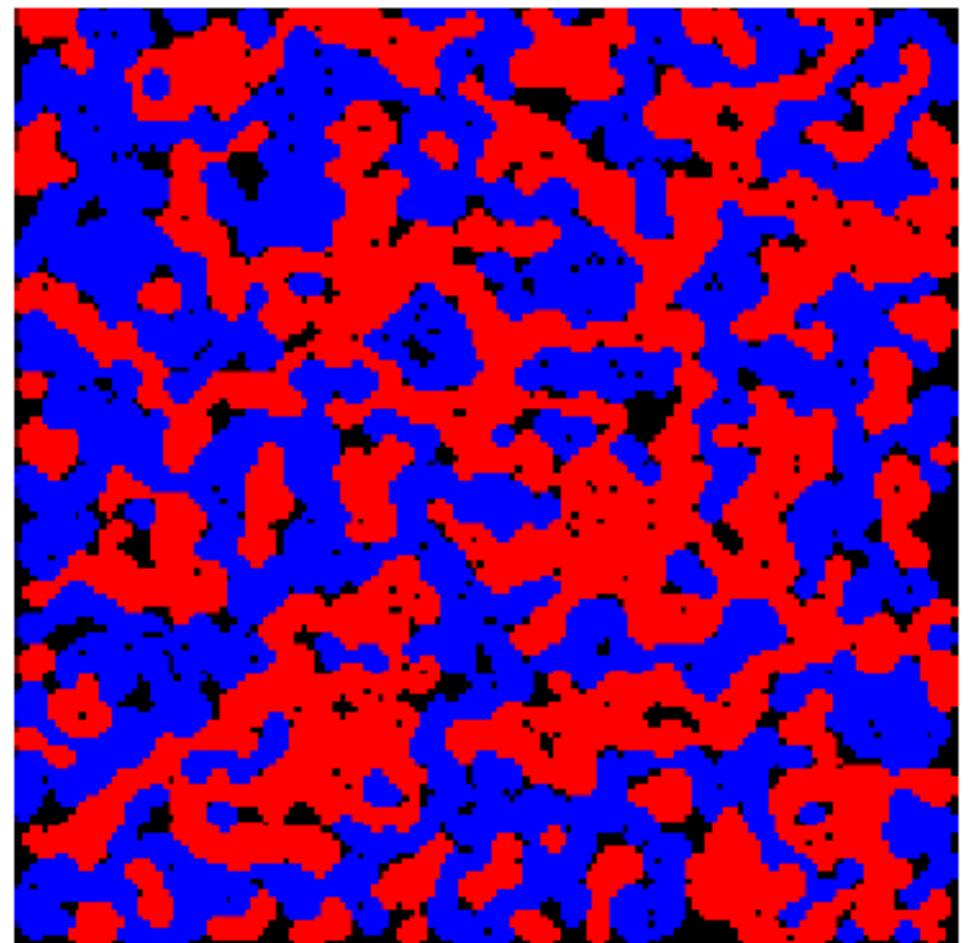
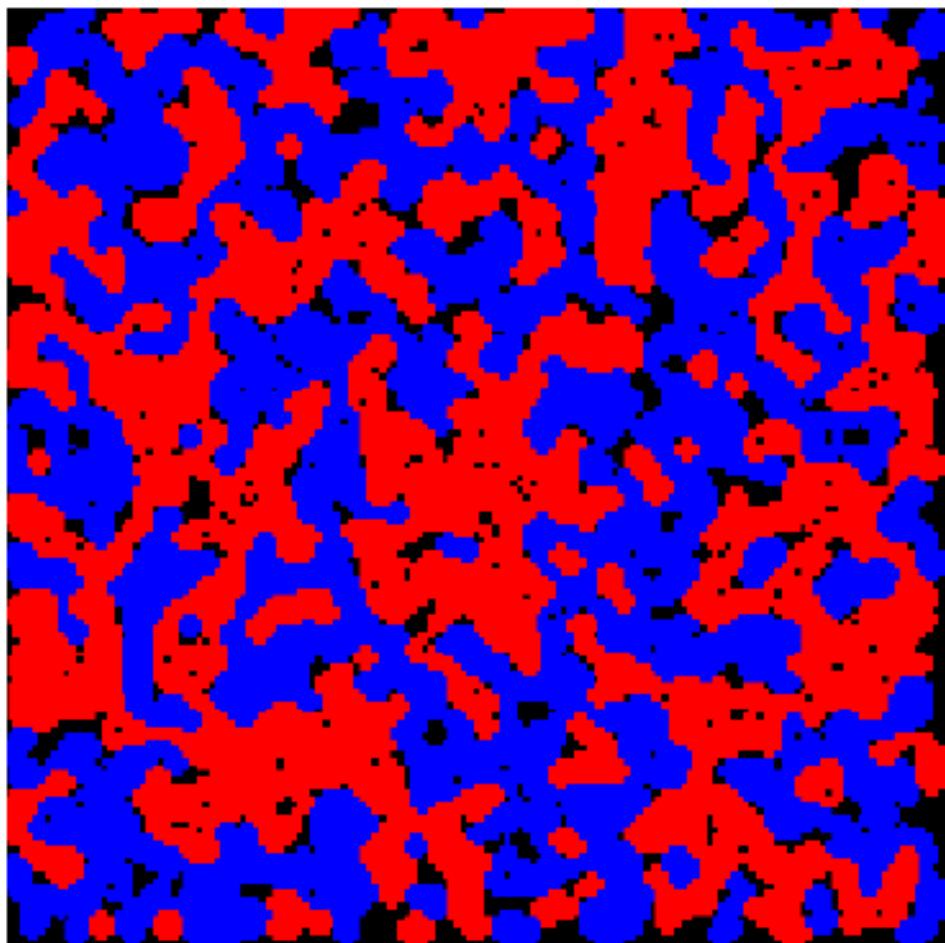


谢林模型 示意

X1*	X2*				
X3	O1*		O2		
X4	X5	O3	O4	O5*	
X6*	O6			X7	X8
	O7	O8	X9*	X10	X11
		O9	O10	O11*	

X3	X6	O1	O2			
X4	X5	O3	O4			
	O6	X2	X1	X7	X8	
O11	O7	O8	X9	X10	X11	
	O5	O9	O10*			

谢林模型的运行（两次）

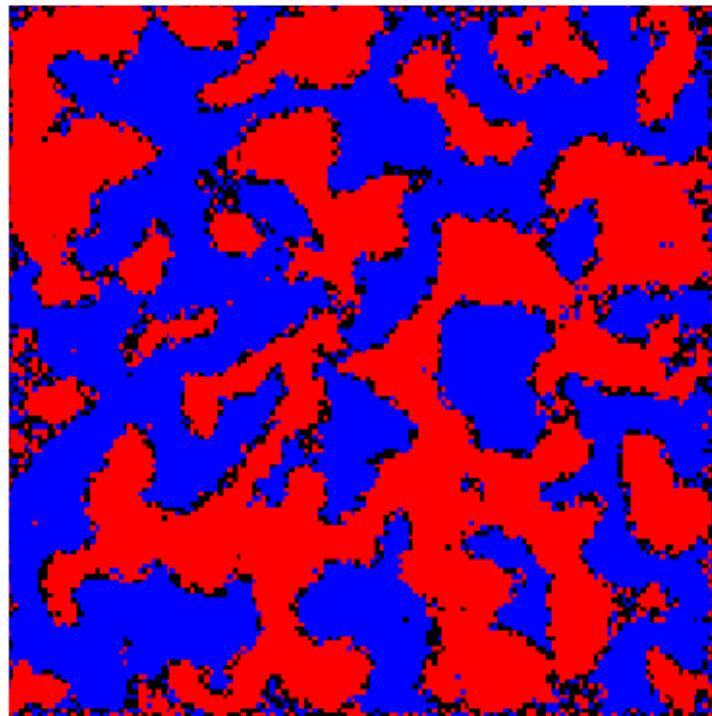


- 150行 * 150列，各10000, $t=3$; 随机初始化

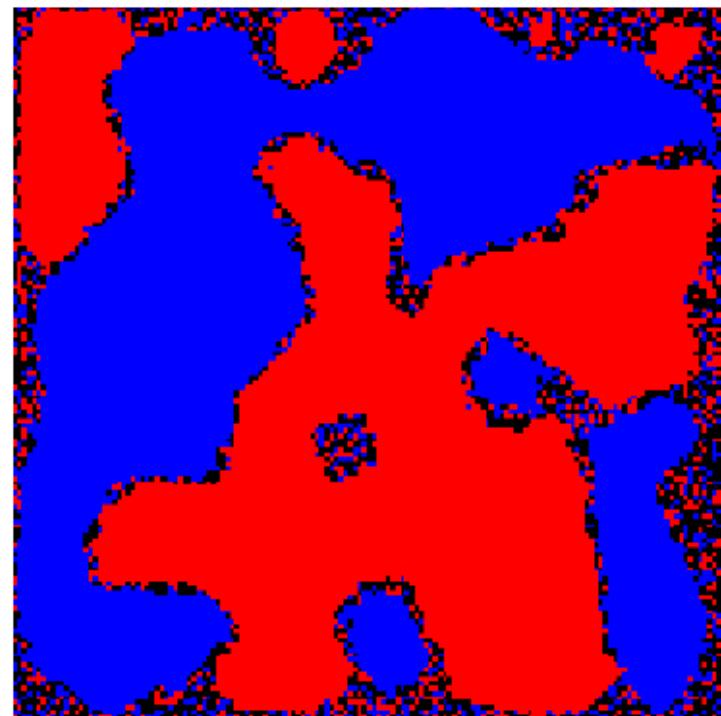
局部要求的一种无隔离的满足

X	X	O	O	X	X
X	X	O	O	X	X
O	O	X	X	O	O
O	O	X	X	O	O
X	X	O	O	X	X
X	X	O	O	X	X

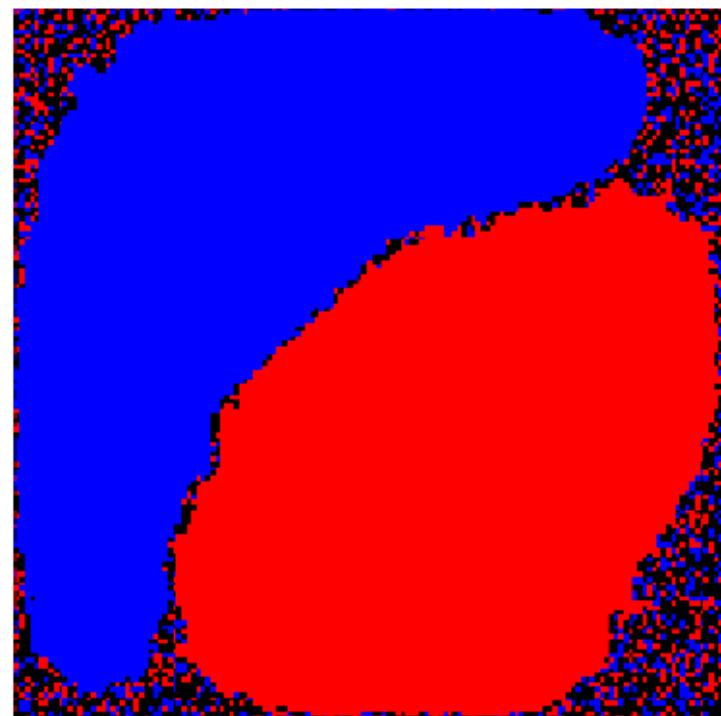
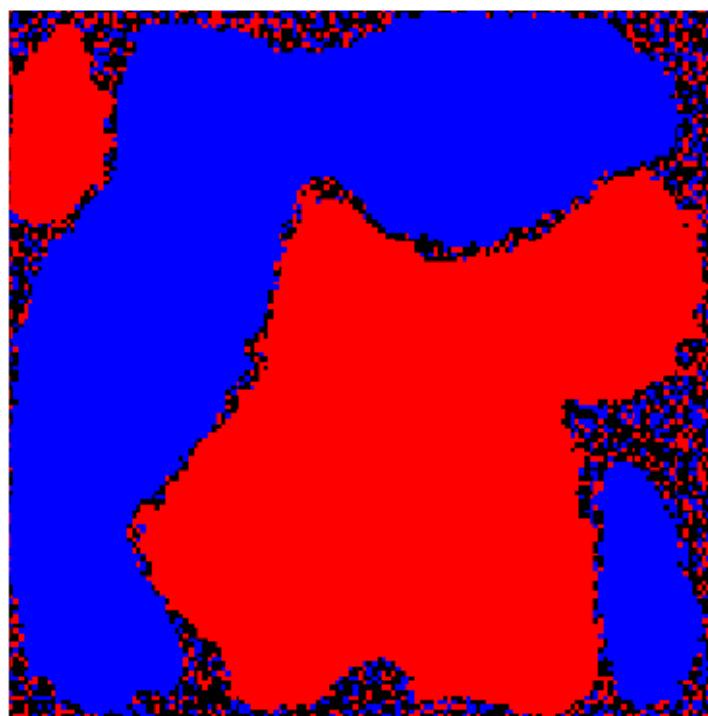
$T=4$



(a) After 20 steps



(b) After 150 steps



小节

- 同质性概念及其对社会网络的影响
- 社会网络中同质性迹象的评估
- 基于社会归属网的概念，结合OSN数据分析，考察同质性现象的方法
- 在同质性研究中的计算机模拟方法（以隔离为例）

作业

- 第4章 3,4
- Or
- 实现谢林模型并提交源代码及运行结果

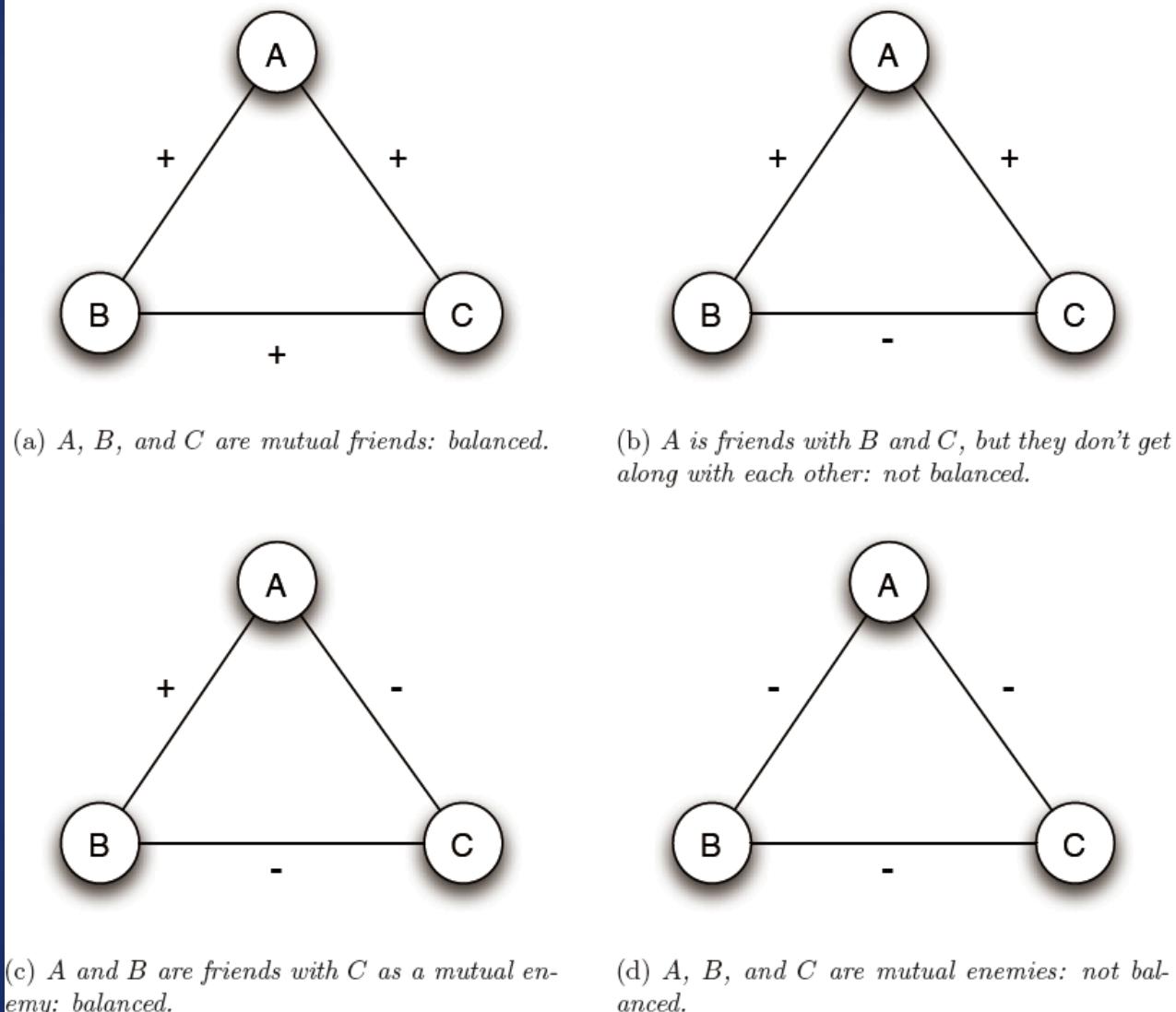
社会网络中关系的 正负及其平衡

边的正负性

- 社会网络中，两个节点之间的关系（边）可能携带着各种各样的社会性含义
- 除了强弱以外，还有支持（+）与反对（-），朋友（+）与敌人（-）等利害关系

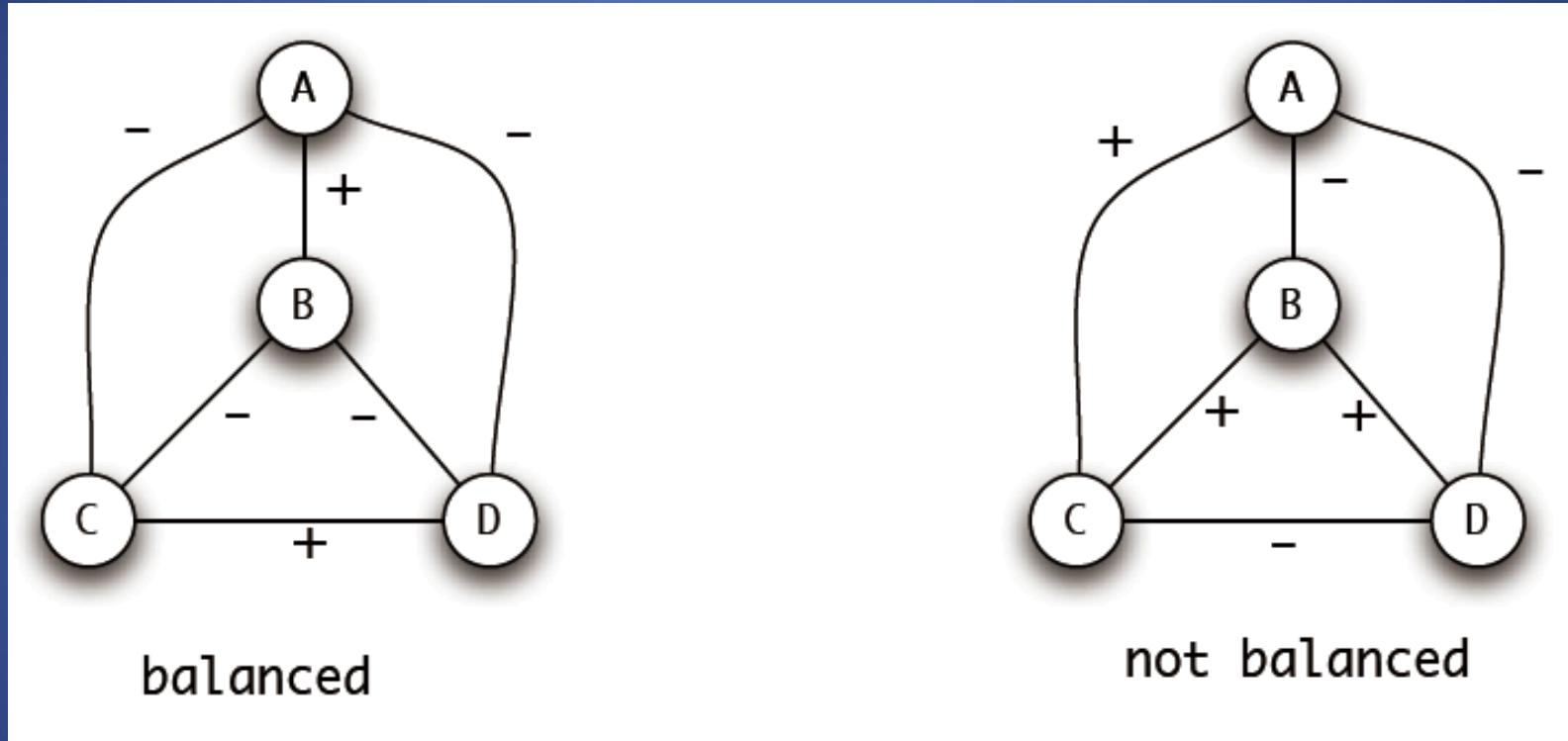
人和人之间如此，国和国之间的外交关系也如此，而且后者常常显现得更加明显（通过同盟条约之类）

三角关系中的结构平衡问题



从社会心理学角度看，一个平衡的三角关系要么3+，要么1+，2-。否则结构不平衡，即隐含有种改变的力量。

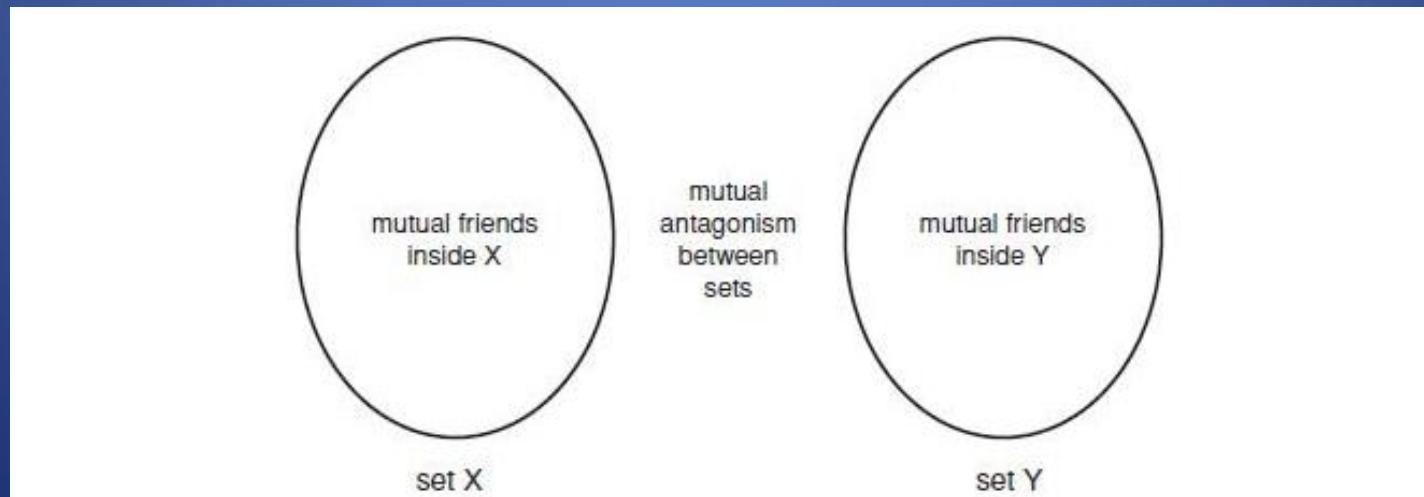
社会网络（图）的结构平衡

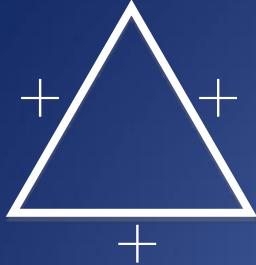


- 定义：（完全）图的结构是平衡的，若其中所有三角关系都是平衡的（即每个三角关系要么3+，要么1+和2-）。

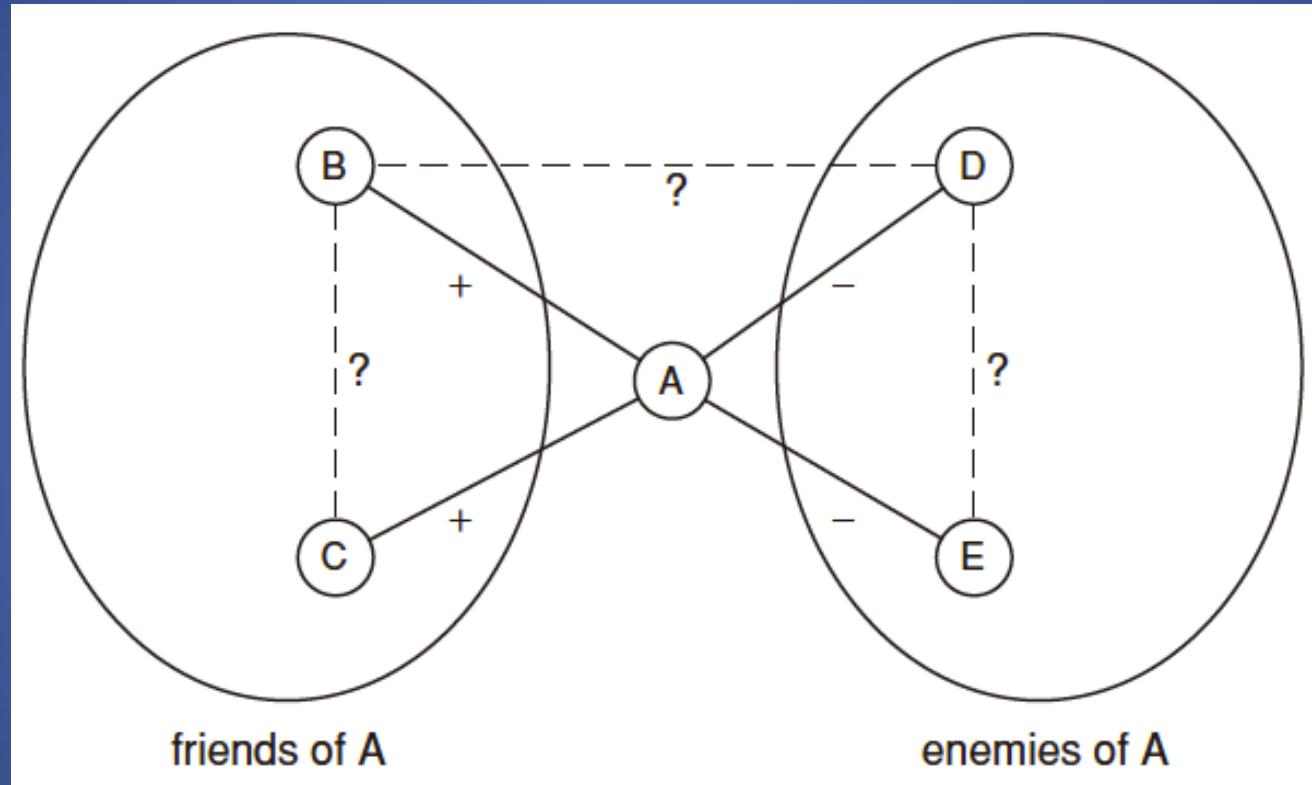
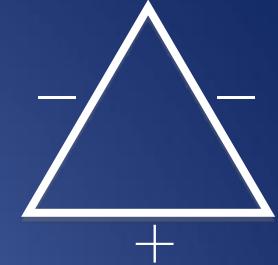
平衡定理

- 如果一个标记（+ / -）的完全图是均衡的，则要么它的所有节点两两都是朋友，要么它的节点可以被分为两组， X 和 Y ，其中 X 组内的节点两两都是朋友， Y 组内的节点两两也都是朋友，而 X 组中的每个节点都是 Y 组中每个节点的敌人



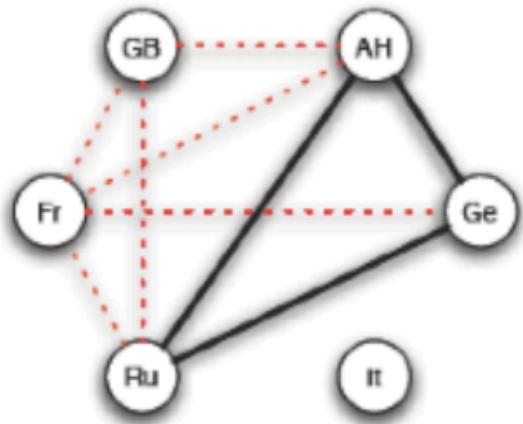


均衡定理的证明

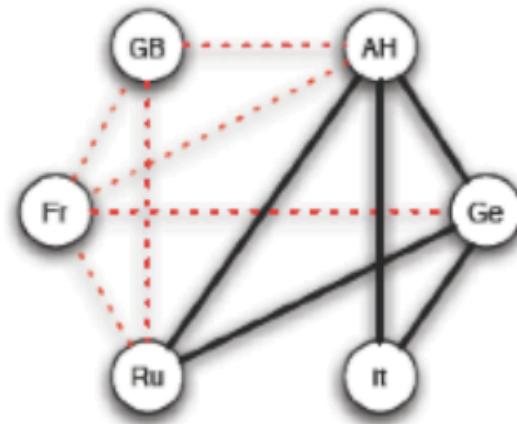


- 满足前述划分性质的图符合平衡定义
- 符合平衡定义的图一定满足前述性质

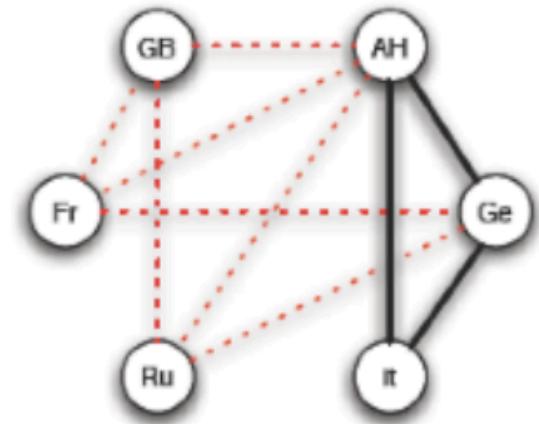
均衡定理在国际关系分析中的应用



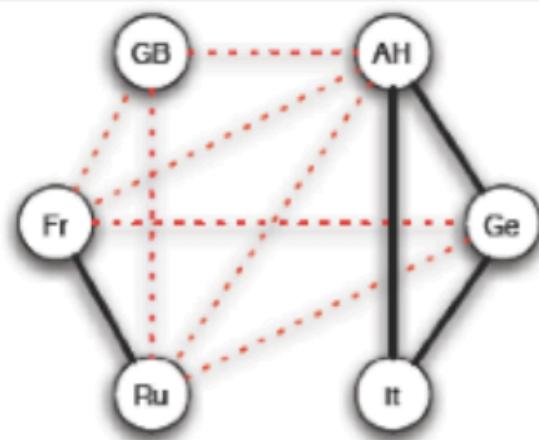
(a) 三国同盟，1872~1881 年



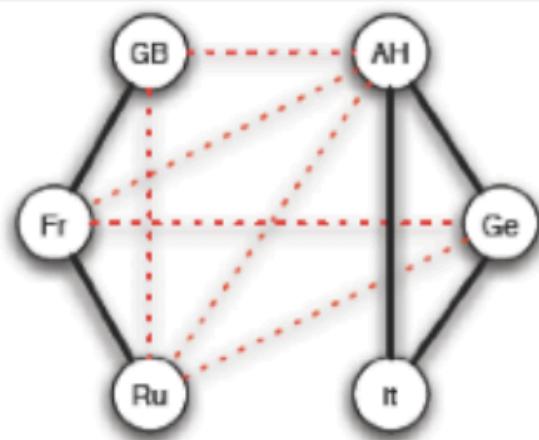
(b) 三国同盟，1882 年



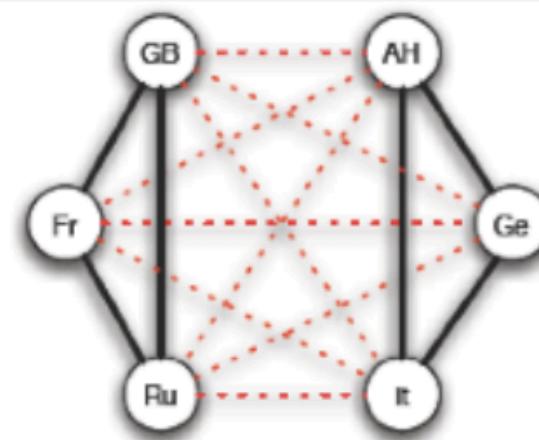
(c) 德国俄罗斯解除联盟关系，1890 年



(d) 法国俄罗斯联盟，1891~1994



(e) 法国英国联盟 1904 年



(f) 英国俄罗斯联盟，1907 年

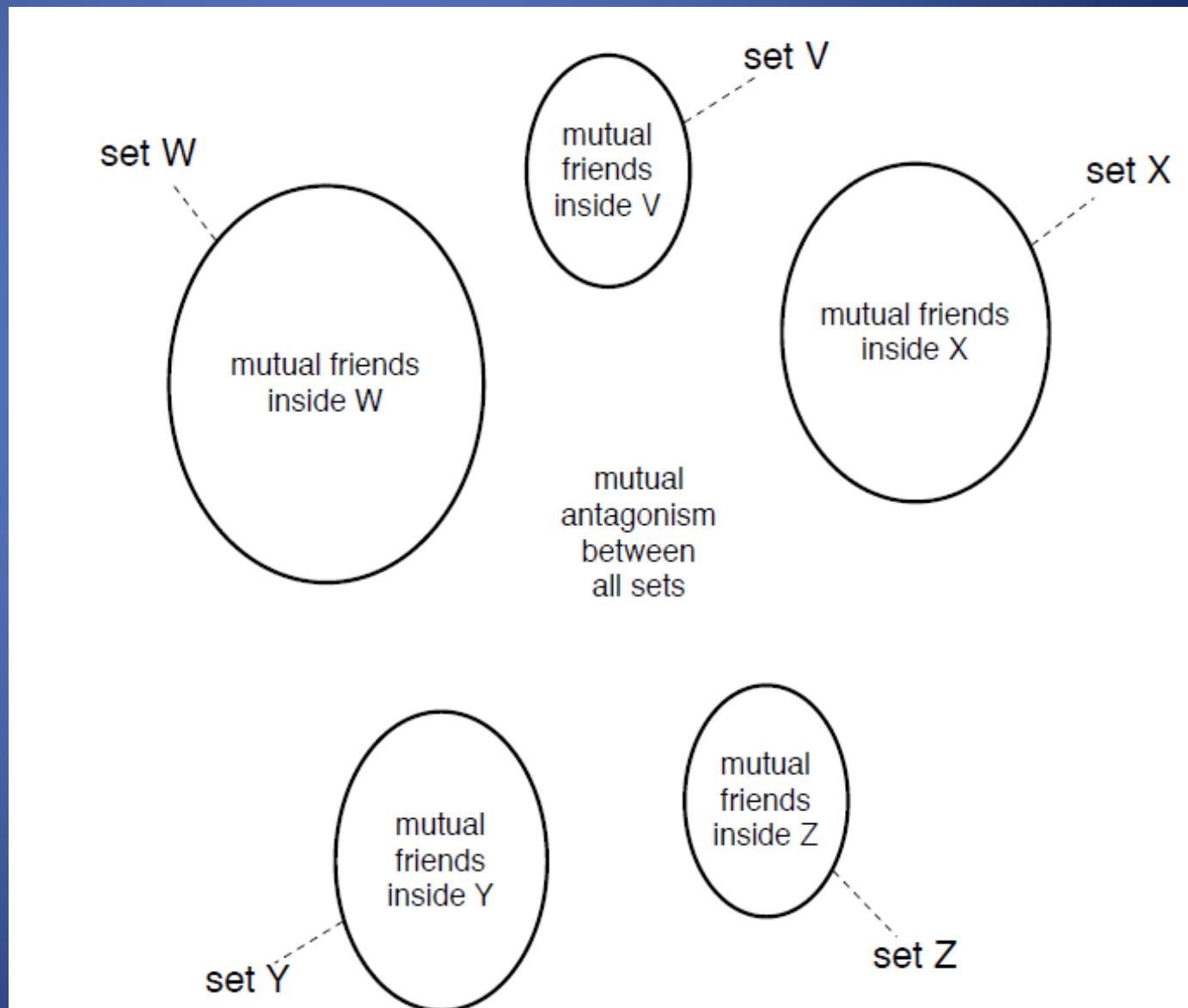
图 5.5 1872 年至 1907 年间，欧洲联盟的演化

弱平衡网络

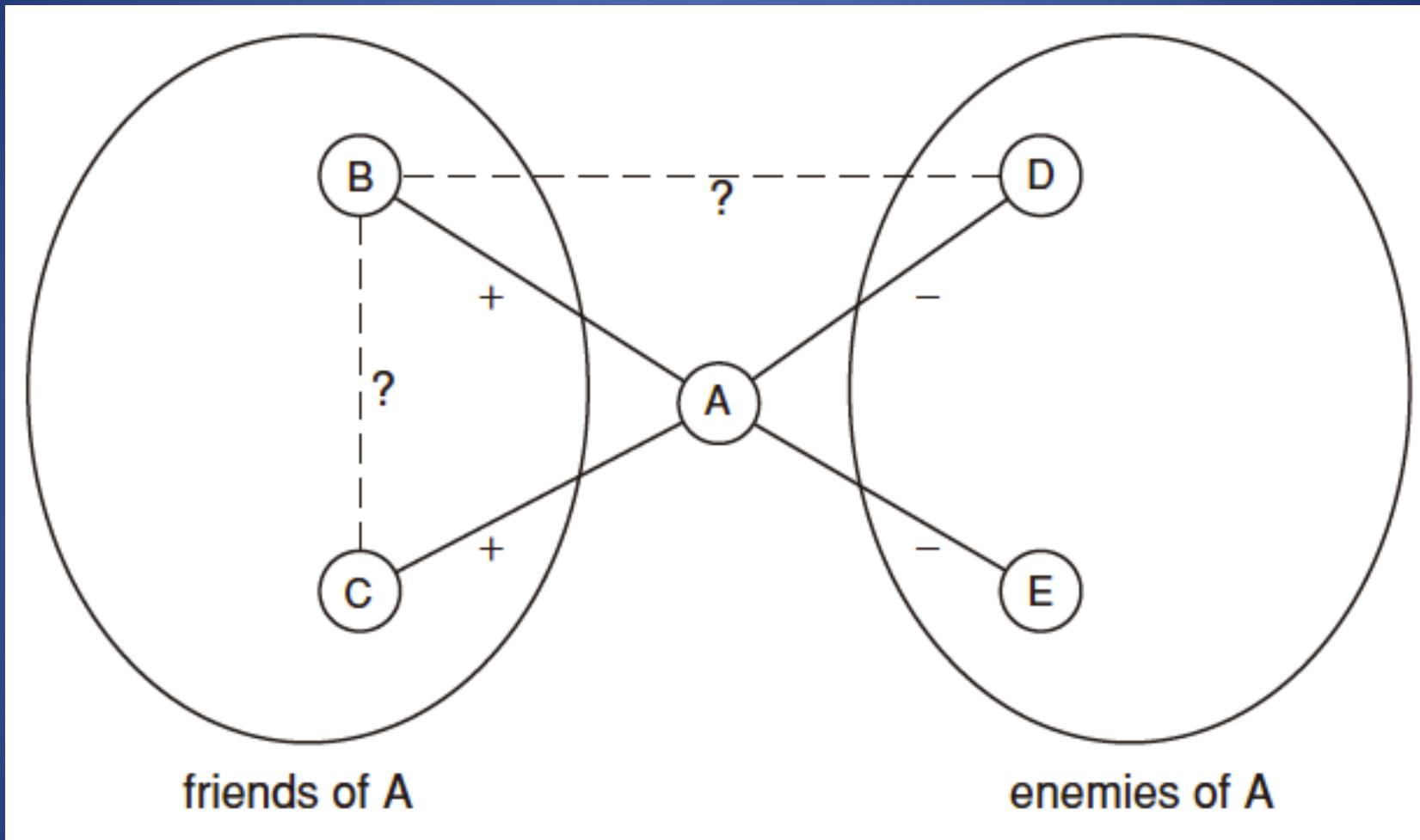
- 注意到在平衡网络中排除的两种三角关系在分量上是有区别的
 - 改变（-、-、-）的动力弱一些
 - 改变（+、+、-）的动力强一些
- 弱平衡网络：不存在（+、+、-）三角关系的标注完全图。
 - 即，我们放松了对平衡的要求
- 弱平衡网络，也有类似于平衡定理那样的性质

弱平衡网络的特性

- 节点可分成若干组，组内均为朋友（+），组间均为敌人（-）



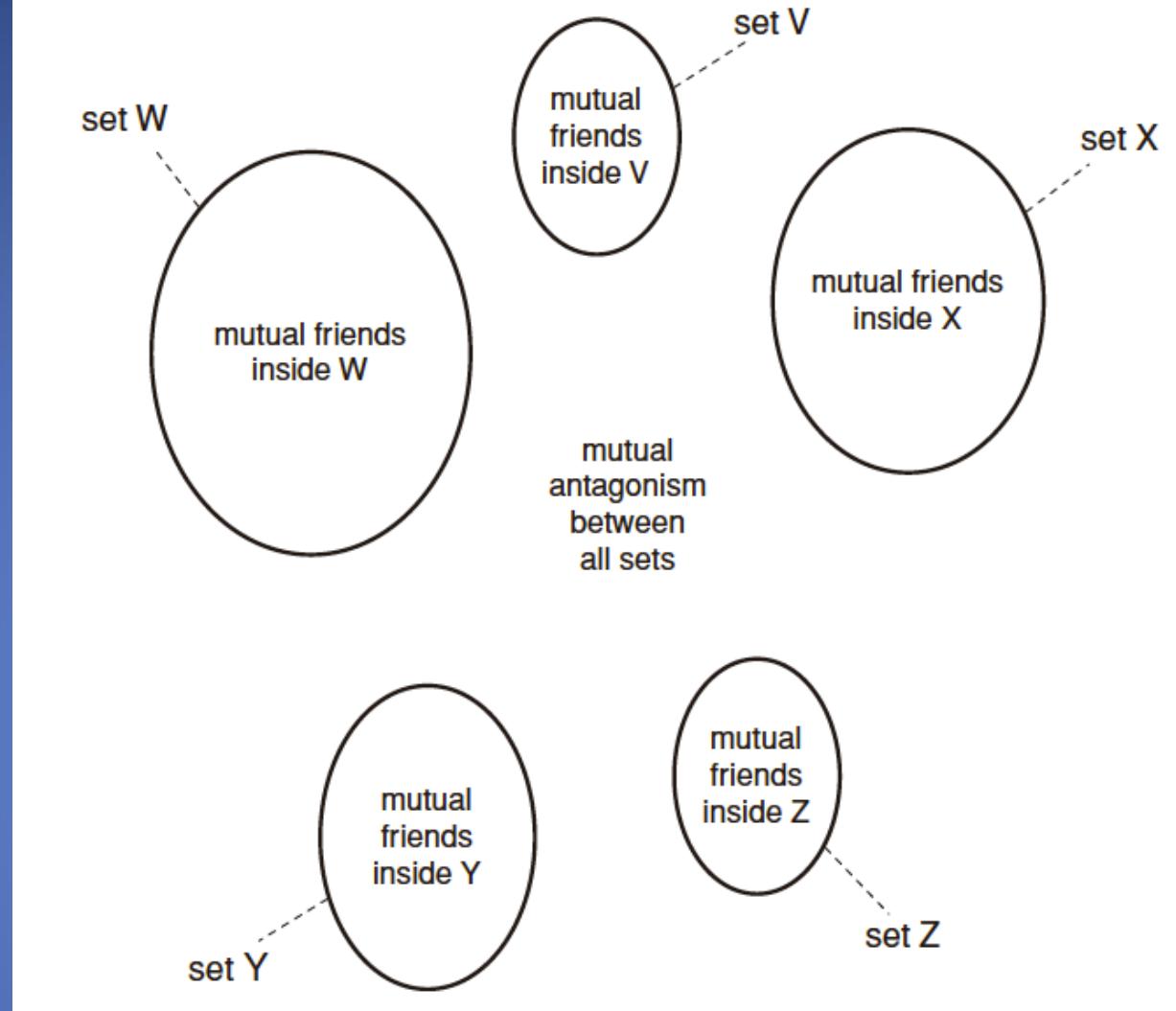
弱平衡性质的证明



- 从一个节点开始，一个个“剥离”满足要求的节点组。

小结

- 现实生活中，人们可能分成多个不同的利益群体
- 平衡网络的概念可用于来分析其中的动态



数学化处理带来了严格漂亮的结果，同时与现实不完全切合，但展示的基本观念“群体内和谐，群体间对抗”对于理解网络的平衡性是有意义的

结构平衡概念的推广

- 任意（非完全）网络中的结构平衡
 - 有些边不存在（表示相关两个节点没关系）
- 近似平衡的网络
 - 完全图，但允许少量（%）的三角形不平衡

作业

- 第5章 3,4
- Or
- 结构平衡概念的推广方面相关的工作（理论推导、证明、说明、验证等）

博弈论基础

[现代博弈论开始于1928年冯诺伊曼的工作]

学习要点

- 理解博弈论的基本概念
 - 参与人, 策略, 收益 (收益矩阵)
 - 最佳应对, 占优策略
 - 纳什均衡
 - 混合策略, 混合策略均衡
 - 帕累托最优, 社会最优
- 几种典型博弈的类型
- 体会 “情景→博弈→求解” 过程中的思想

博弈—从一个例子开始

- “复习考试”还是“准备报告”？
 - 假设在截止日期前一天，你有两件要做的事情：一是复习（为了参加考试），二是准备（给一个报告）。你只能选择做一项。
 - 考试成绩可以预计
 - 如果复习，则考试成绩92分，没复习，则80分
 - 报告需要你和你的拍档合作完成
 - 如果你和拍档都准备报告，则每人都得100分
 - 如果只有一人准备报告，则每人都得92分
 - 如果两人都没准备报告，则每人都得84分
 - 那么你该选择做什么呢？（假设你和拍档各自独立考虑这个问题）

例子：“考试-报告”博弈

- 设你们都追求平均成绩的最大化：
 - 你和搭档都准备报告，则平均成绩均为 $(80+100)/2 = 90$ 分
 - 你和搭档都准备考试，则平均成绩均为： $(92+84)/2 = 88$ 分

- 考试成绩可以预期：
 - 如果复习，则考试成绩92分
 - 如果没复习，则考试成绩80分
- 报告是你和你的拍档合作完成的：
 - 如果你和拍档都准备报告，则每人100分
 - 如果只有一人准备报告，则每人92分
 - 如果两人都没准备报告，则每人84分

若一方复习考试，另一方准备报告：

- ▶ 准备报告一方的得： $(80+92)/2 = 86$ 分
- ▶ 复习的一方得： $(92+92)/2 = 92$ 分

收益矩阵（表达收益的一种直观方式）

		你的拍档	
		准备报告	复习考试
你	准备报告	90, 90	86, 92
	复习考试	92, 86	88, 88

- 其中第一个数字是“你”的收益，第二个是“拍档”的
- 收益（也称“回报”，payoff）

博弈的基本要素

- 一般情况下，博弈具有三个要素：
 - (1) 参与者（至少两个）；
 - (2) 策略集：每个参与者都有一组关于如何行为的备选项，此处备选项指参与者的可能策略。
 - (3) 收益（回报）：每个策略行为的选择，都会使参与人得到一个收益。
 - 这个收益结果还受互动中他人策略选择的影响。
 - 同一组策略，不同参与人的收益可能不同

通常，收益的记号： $P_1(S,T)$, $P_2(S,T)$

博弈行为推理的几点基本假设

- 每个参与人对博弈结构（收益矩阵）有充分了解。
- 参与人都是理性的（rational）
 - 追求自己的收益最大化（尽量大）
 - 也知道其他参与人也是如此
- 决策的独立性
 - 不商量

“考试-报告”博弈中的行为推理

		你的拍档	
		准备报告	复习考试
		准备报告	复习考试
你	准备报告	90, 90	86, 92
	复习考试	92, 86	88, 88

- 严格占优策略 (strictly dominant strategy)：对一个参与人 (A) 来说，若存在一个策略，无论另一个参与人 (B) 选择何种行为策略，该策略都是最佳选择，则这个策略就称为是 A 的严格占优策略。
- 这个例子中，“复习考试”对双方都是严格占优策略。

“囚徒困境”

- 假设有两个疑犯被警察抓住。并且被分开关押在不同的囚室。
- 警察强烈怀疑他们和一场抢劫案有关。但是，没有充足的证据。然而，他们都拒捕的事实也是可判刑的。
- 两个疑犯都被告知以下结果：
 - “如果你坦白，而另外一人抵赖，则你马上释放；另外一人将承担全部罪行，将会被判刑10年
 - 如果你们都坦白，你们的罪行将被证实。但由于你们有认罪的表现——判刑4年。
 - 如果你们都不坦白，那么没有证据证明你们的抢劫罪，我们将以拒捕罪控告你们——判刑1年。
 - 另外一方也正在接受这样的审讯。你是坦白还是抵赖？”

“囚徒困境”的收益矩阵

		疑犯2	
		抵赖	
疑犯1		抵赖	 坦白
 坦白		-1, -1	-10, 0
	 坦白	0, -10	-4, -4

- 疑犯1和疑犯2的严格占优策略都是“坦白”
- 尽管如果两人都抵赖会都判得少些
 - 刻画了“有关个体私利前，建立合作是十分困难”的模型。

“兴奋剂”博弈

		运动员2	
		没服用	服用
运动员1	没服用	3, 3	1, 4
	服用	4, 1	2, 2

- 这种类型通常称为军备竞赛。竞争双方为保持彼此实力相当，都会选择生产更具危险性的武器，尽管对自己内部会有伤害
 - 运动员伤害身体，国家影响民生。

关于“收益”的讨论（收益决定选择）

- “考试-报告”博弈，如果降低考试难度：只要复习了，就会得到100分；否则，也可得到96分。

		你的拍档	
		准备报告	复习考试
你	准备报告	98, 98	94, 96
	复习考试	96, 94	92, 92

囚徒困境类似，如果改变收益矩阵，情况也可不一样

最佳应对与占优策略

- 设 S 是参与人甲的一个选择策略， T 是参与人乙的一个选择策略。在收益矩阵中的某个单元格对应这策略组 (S, T) 。
 - $P_1(S, T)$: 表示参与人甲从这组决策获得的收益
 - $P_2(S, T)$: 表示参与人乙从这组决策获得的收益
- 最佳应对：**针对参与人乙的策略 T ，若参与人甲采用策略 S 产生的收益大于或等于自己的任何其他策略，则称参与人甲的策略 S 是参与人乙的策略 T 的最佳应对。

$$P_1(S, T) \geq P_1(S', T),$$

其中， S' 是参与人甲除 S 外的任何其他策略。

严格最佳应对

- 严格最佳应对：若 S 会产生比任何应对策略 T 的其他策略都更高的收益，则称参与人甲的策略 S 是对于参与人乙的策略 T 的严格最佳应对。

$$P_1(S, T) > P_1(S', T)$$

其中， S' 是参与人甲的所有其他策略。

- 注：最佳应对的概念是针对对方的某一个策略（ T ），相对于自己的所有策略而言的
 - 对于同一个 T ，最多只可能有一个严格最佳应对
 - 对于不同的 T ，最佳应对可能相同，也可能不同

占优策略与严格占优策略

- 定义：（从最佳应对角度给出）
 - 参与人甲的占优策略 s ，是指该策略对于参与人乙的每一策略都是最佳应对。
 - 参与人甲的严格占优策略 s ，是指该占优策略对于参与人乙的每一策略都是严格最佳应对。
- 如果参与人有严格占优策略，则可预期他会采取该策略（与基本假设的一致性）。
- 注：占优策略的概念是相对于对方所有策略而言的。

并不是每人总有严格占优策略

- 例子：“营销战略”博弈
 - 假设有两家公司，分别要规划生产并销售同一种新产品。该产品有两款可能的规格：廉价（低档）或高档。如何决策？
 - 设顾客总体被分成两个市场：一部分消费群体（60%）只购买廉价商品，另一部分消费群体（40%）只购买高档次商品。
 - 每家公司从廉价或高档次商品所得利润是等同的（因此利润仅取决于市场占有率）。
 - 每家公司都追求利润最大化。

“营销战略”博弈

- 假设

- 若两家公司分别定位生产不同类型的产品，则每家公司都会得到该商品市场的全部份额。
- 公司1品牌形象更佳。因此，若这两家公司在同一市场（廉价或高档次）中竞争，则公司1可以得到80%的市场销售量，公司2只能得到20%的市场。

		公司2	
		廉价	高档次
公司1	廉价	0.48, 0.12	0.6, 0.4
	高档次	0.4, 0.6	0.32, 0.08

- 可以预测此博弈的发展趋向。即公司1将会采取廉价策略，公司2将会采取高档次策略。

博弈的行为推理

- 如果参与人都有严格占优策略，则可以预计他们均会采取严格占优策略；
- 如果只有一个参与人有严格占优策略，则这个参与人会采取严格占优策略，而另一方会采取此策略的最佳应对。
- 如果两个参与人都没有严格占优策略呢？

无占优策略例子（三客户博弈）

- 假设有两家公司，都希望和A、B、C三个大客户之一洽谈生意。每家公司都有三种可能的策略：是否找客户A、B或C。
- 他们决策的条件如下所示：
 - 若两家公司都找同一个客户，则该客户会给每个公司一半的业务。
 - 公司1规模太小，以至于不能靠自身找到客户源。所以，只要它和公司2分别寻找不同的客户洽谈生意，则公司1获得的收益将会是0（生意做不成）。
 - 假设公司2单独寻找客户B或C洽谈生意，则会得到客户B或C的全部业务。但是A是一个大客户。寻找客户A洽谈生意时，必须和其它公司合作才能接下业务。
 - 因为A是一个大客户，和它做生意的收益是8（假设两家公司合作，则每家公司会得到收益4）。但是，和B或C做生意的收益价值是2（合作的话，每个公司收益是1）

“三客户”博弈的推理

- 收益矩阵

		公司2		
		A	B	C
公司1		A	4, 4 	0, 2
		B	0, 0 	1, 1
C		0, 0	0, 2 	1, 1

- 两家公司都没有严格占优策略

纳什均衡

- 假定参与人甲选择策略 S ，参与人乙选择策略 T 。若 S 是 T 的最佳应对，且 T 也是 S 的最佳应对，则称策略组 (S, T) 是一个**纳什均衡**。
 - 在均衡状态，任何参与人都没有动机（理性的理由）去换一种策略。
 - 纳什均衡可以被看成是一种**信念上的均衡**
 - 互为最佳应对，谁也不可能通过单方面改变策略而得到额外好处，尽管如果两人都改变可能都会更好（相比都不改变而言）

“三客户”博弈的纳什均衡

		公司2		
		A	B	C
公司1		A	4, 4 	0, 2
		B	0, 0 	1, 1
		C	0, 0 	0, 2 1, 1

- 存在纳什均衡: (A, A)
- 寻找纳什均衡的两种途径:
 - 一是, 检查每一个策略组, 看它们中的每一项是否是彼此间策略的最佳应对策略。
 - 二是, 找出每个参与人对于对方每个策略的最佳应对, 然后发现互为最佳应对的策略组。

多重均衡：协调博弈

- 多重均衡——存在多个均衡
- 例子：协调博弈
 - 假设你和你拍档都为一个合作项目准备幻灯片简报（双方不能通过电话等方式联系商量）。
 - 你必须决定是用微软的PPT或是用苹果的Keynote软件来制作你负责的半份幻灯片。
 - 假设你们使用同样的软件来设计，那就比较容易合并你们的幻灯片。

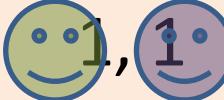
协调博弈的推理

		你的拍档	
		PPT	Keynote
		PPT	0, 0
你	PPT	1, 1	0, 0
	Keynote	0, 0	1, 1

- 存在两个纳什均衡：(PPT, PPT), (Keynote, Keynote)。
- 如何预测协调博弈中参与人的行为?
 - 托马斯·谢林（获得2005年诺贝尔经济学奖）提出一种聚点的想法，利用一些其他外部因素，例如社会习俗。

不对等协调博弈

- 假设你和项目拍档都更喜欢使用苹果软件。

		你的拍档	
		PPT	Keynote
		PPT	Keynote
你	PPT	 1,  1	0, 0
	Keynote	0, 0	 2,  2

- 谢林的聚点理论表明，可以预测到参与人会精选策略，倾向于收益情况更好的均衡。

两人的喜好不同呢

- 假设你和你的拍档喜欢的软件不同。

		你的拍档	
		PPT	Keynote
		PPT	Keynote
你	PPT	 1, 2 	0, 0
	Keynote	0, 0	 2, 1 

- 此时很难预测具体哪种均衡会被采取。
- 可以通过了解他们之间平常发生冲突时解决的惯例来预测。

猎鹿博弈

- 假设两猎人外出猎物。若他们合作，则可以猎到鹿（这可以给猎者带来最高的收益）。
- 猎人若分开单干，都能猎到兔。
- 若一方想单独猎鹿，则收益是0。另一方依然能猎到兔。

		猎人2	
		猎鹿	猎兔
猎人1	猎鹿	 4,  4	0, 3
	猎兔	3, 0	 3,  3

- 选择何种均衡？要在高收益和由于另一方不合作而造成损失之间进行权衡。

多重均衡：鹰鸽博弈

- 假设两只动物要决定一块食物在彼此之间何如分配。
- 每种动物都可以选择争夺行为（鹰派策略）或分享行为（鸽派策略）。
 - 若两种动物都选择分享行为，他们将会均匀的分配食物，各自的收益是3。
 - 若一方行为表现为争夺，另一方行为表现是分享，则争夺方会得到大多数食物，获得收益是5，分享方只能得到收益为1。
 - 当两只动物都表现为争夺行为，由于在争夺中践踏了食物，则它们得到的收益将为0。

鷹鸽博弈推理

		动物2	
		鸽派	鷹派
动物1	鸽派	3, 3	1, 5
	鷹派	5, 1	0, 0

- 很难预测参与者的 behavior
- 纳什均衡概念能有助于缩小合理的预测范围，但它并不能给出唯一的预测。

几种典型多均衡博弈类型对比

你的拍档		你的拍档	
		PPT	Keynote
PPT	1, 1	0, 0	
Keynote	0, 0	2, 2	
猎人2		动物2	
		鸽派	鹰派
猎鹿	猎鹿	4, 4	0, 3
猎兔	猎兔	3, 0	3, 3
鸽派		鹰派	
		3, 3	1, 5
		5, 1	0, 0

简单博弈的推理思路

- 如果双方都有严格占优策略，则都会采用之
- 如果只有一方有严格占优策略，则可以预测另一方会采用此策略的最佳应对
- 如果不存在严格占优策略，则寻找纳什均衡
 - 存在一个纳什均衡，该均衡对应合理结果
 - 存在多个纳什均衡（需要额外信息辅助决策）
 - 协调博弈，鹰鸽博弈
 - 均衡有助于缩小考虑范围，但不保证有效预测
- 如果不存在纳什均衡，该怎么办？

混合策略

- 例子：硬币配对—“零和博弈”（zero sum game）
 - 两个参与人各持一枚硬币，同时选择手中硬币的正反面。
 - 若他们硬币的朝向相同，参与人乙将赢得参与人甲的硬币。反之，则参与人甲将赢得参与人乙的硬币。

		参与人乙	
		正面H	反面T
参与人甲	正面H	-1, +1	+1, -1
	反面T	+1, -1	-1, +1

- 此时，不存在一组互为最佳应对（纳什均衡）

混合策略的引入

- 引入随机性，考虑参与人将以一定的概率分布在不同策略间进行选择，一种分布对应一个“混合策略”（此时，选择策略就是选择分布）
 - 对于双策略（H和T）博弈，混合策略则可简略表示为一个概率。**纯策略**就是概率为（0, 1）的混合策略。
- 通常，我们说
 - 参与人1的策略是概率p，是指参与人1以概率p执行H；以概率 $1-p$ 执行T
 - 参与人2的策略是概率q，是指参与人2以概率q执行H，以概率 $1-q$ 执行T

混合策略的收益

- 采用收益期望作为策略的回报测度
- 设参与人1采用概率 p 执行H, $1-p$ 执行T, 则:
- 若参与人2采用H, 则其收益期望是

$$\bar{P}_2(p, H) = p \cdot P_2(H, H) + (1 - p) \cdot P_2(T, H)$$

- 若参与人2采用T, 则其收益期望是

$$\bar{P}_2(p, T) = p \cdot P_2(H, T) + (1 - p) \cdot P_2(T, T)$$

类似地, 可讨论参与人2采用概率混合策略的情形

混合策略的均衡

- 混合策略的纳什均衡：它是一对混合策略，彼此都是对方的最佳应对（期望收益）
- 纳什的奠基性贡献：证明了具有有限参与者和有限纯策略集的博弈一定存在纳什均衡（包括混合策略均衡）
- 一般来说，找到混合策略的纳什均衡是很困难的，但在某些特定条件下可能有系统的方法。

双人双策略、没有含纯策略均衡的博弈中的混合策略纳什均衡求解

- 给定 H, T : 基本纯策略。按照纳什定理，存在一个混合策略的纳什均衡 (p, q) ，即 p 是 q 的最佳应对， q 也是 p 的最佳应对。如何求 p 和 q ?
 - “没有含纯策略的均衡”的前提意味着 p, q 都是严格在0和1之间。
- 确定参与人2采用的 q 的方法（确定 p 的方法对称）
 - 基于 q 和收益矩阵中的值，分别写出参与人1采用 H 和采用 T 的收益期望（ q 的函数），即 $P1(H, q)$ 和 $P1(T, q)$ ，也就是相当于 $P1(1, q)$ 和 $P1(0, q)$
 - 下面的关键是要认识到此时必定有： $P1(1, q) = P1(0, q)$ ，从而可以借助这等式求出 q

$P1(1,q) = P1(0,q)$: 在两个端点“无差异”原理

- 推理的思路是
 - 若等式不成立，例如 $P1(1,q) > P1(0,q)$ ，则将导致 H （即 $p=1$ ）是参与人1的最佳应对的结论，即 (H,q) 是一个纳什均衡，这与“不含纯策略纳什均衡”的前提矛盾
 - 直观上，若 $P1(T,q) < P1(H,q)$ ，参与人1在应对 q 的时候采用 H 就是最好，将任何机会（概率）分给 T 都只会导致较低收益
 - 数学上就是

$$P1(p,q) = pP1(H,q) + (1-p)P1(T,q) < P1(H,q), \text{ for } p < 1$$

混合策略的收益计算例子

- 用收益期望来表达回报

		参与人2	
		正面H(q)	反面T($1-q$)
参与人1	正面H	-1, +1	+1, -1
	反面T	+1, -1	-1, +1

- 例如，当参与人2采用策略 q 时，若参与人1使用纯策略，则他的回报分别为：
 - 纯策略H的期望收益 = $(-1)(q) + (+1)(1-q) = 1-2q$
 - 纯策略T的期望收益 = $(1)(q) + (-1)(1-q) = 2q-1$

如果系统不存在包含纯策略的均衡，则上述两个表达式必须相等。

硬币配对博弈的混合策略均衡

		参与人2	
		正面H(q)	反面T($1-q$)
参与人1	正面H	-1, +1	+1, -1
	反面T	+1, -1	-1, +1

- 设 (p, q) 是纳什均衡。对参与人2的策略 q ，
 - 参与人1用纯策略H的期望收益 $= (-1)(q) + (+1)(1-q) = 1 - 2q$
 - 参与人1用纯策略T的期望收益 $= (1)(q) + (-1)(1-q) = 2q - 1$
 - 这是一个不存在含有纯策略均衡的博弈，由“无差异”原理，须有 $1 - 2q = 2q - 1$ ，即 $q = 1/2$
- 对称地，可以得到参与人1的最佳应对 $p = 1/2$
- 因此， $(1/2, 1/2)$ 是一个混合策略纳什均衡（合直觉）

混合策略：进一步的例子

- 持球-抛球博弈

- 美式足球比赛：进攻方可以选择持球或者是抛球。防御方可以选择拦断持球或者选择防守抛球。
- 假设正确阻止了进攻方的行为，则进攻方的收益为0。
- 假设进攻方选择持球而防守方却选择防守抛球行为，则进攻方的收益为5（防守方相应损失）。
- 假设进攻方选择抛球，同时防守方却选择拦断持球，则进攻方的收益是10（防守方相应损失）。

		防守方	
		防守抛球	拦断持球
进攻方	抛球	0, 0	10, -10
	持球	5, -5	0, 0

持球抛球博弈的混合策略均衡

- 这是一个没有纯策略纳什均衡的博弈
- 设防守方选择防守抛球的概率为 q

		防守方	
		防守抛球(q)	拦截持球($1-q$)
进攻方	抛球	0, 0	10, -10
	持球	5, -5	0, 0

- 进攻方选择抛球的期望收益: $0*q + 10*(1-q)$
- 进攻方选择持球的期望收益: $5*q + 0*(1-q)$
- 依无差异原理, 令 $10-10q=5q$, 解得 $q=2/3$

持球抛球混合策略均衡（续）

- 进攻方选择抛球的概率为 p

		防守方	
		防守抛球	拦截持球
进攻方	抛球(p)	0, 0	10, -10
	持球($1-p$)	5, -5	0, 0

- 防守方选择防守抛球的期望收益: $-5(1-p)$
- 防守方选择拦截持球的期望收益: $-10p$
- 令 $-10p=-5(1-p)$, 解得 $p=1/3$
- 于是, 这个博弈的混合策略均衡为 $(1/3, 2/3)$

讨论

		防守方	
		防守抛球 <i>(2/3)</i>	拦截持球 <i>(1/3)</i>
进攻方	抛球 <i>(1/3)</i>	0, 0	10, -10
	持球 <i>(2/3)</i>	5, -5	0, 0

- 为什么抛球有可能收益更大，而均衡中进攻方选择抛球的概率只有 $1/3$?
 - 由于防守方高概率防守抛球，若抛球概率 $p>1/3$ ，则损失会比较大
- 为什么进攻方在均衡的抛球概率只有 $p=1/3$ ，但防守方还要更多的防守抛球?
 - 由于抛球对进攻方更有利，需要加大防守力度

例子：罚点球博弈

- 2002年，有人做了一项有关罚点球研究
 - 射手要决定从球门的左侧或是右侧进球。
 - 守门员则是要决定是扑向左侧或是右侧拦断进球。
 - 两人需要同时做选择。

		守门员	
		L	R
射球方		L	0.58, -0.58
		R	0.93, -0.93
守门员		L	0.95, -0.95
		R	0.70, -0.70

统计数据。可以看到，罚球方总是有赢头（符合实际）。

混合策略均衡

		守门员	
		L(q)	R
射球方		L(p)	0.58, -0.58 0.95, -0.95
		R	0.93, -0.93 0.70, -0.70

- 计算得到的均衡：
 $0.58q + 0.95(1-q) = 0.93q + 0.70(1-q)$, $q=0.42$
 $-0.58p - 0.93(1-p) = -0.95p - 0.70(1-p)$, $p=0.39$
- 实战统计得到的数据: $q=0.42, p=0.40$

兼具纯策略和混合策略均衡的博弈

- 例子：不平衡的协调博弈

		你的拍档	
		PPT(q)	Keynote
		PPT(p)	1, 1 0, 0
你	PPT	0, 0	2, 2
	Keynote		

- 除了两个纯策略均衡 (PPT,PPT) 和 (Keynote,Keynote) 外，还存在一个混合策略均衡： $q=2(1-q)$, $q=2/3$; $p=2(1-p)$, $p=2/3$

关于混合策略均衡的一般性推理

- 若双人双策略博奕存在混合策略均衡 (p,q) , $0 < p, q < 1$
 - 甲所选择的 p , 应该使乙在他的两个纯策略上无差异;
否则乙就会用优势策略（例如1）作为应对，收益期望
 $(p,1) >$ 收益期望 (p,q) , 即没有 $q < 1$ 是 p 的最佳应对，从而
与存在上述混合策略均衡矛盾
 - 于是可利用这无差异原则，**试求** p ; 类似地，求 q 。（
注意，并不保证总能求出在 $(0,1)$ 上的 p 和 q ）
- 若分别求得了 $(0,1)$ 区间中的 p 和 q ，则说明
 - 对于甲的策略 p ，乙采用 $[0,1]$ 上的任何策略都是一样的，即都是最佳应对，**后来求出的** q **当然也是**。对于 p 也有对称的认识，亦即 p 和 q 互为最佳应对

考试一报告博奕没有混合策略

		你的拍档	
		准备报告	复习考试
你	准备报告	90, 90	86, 92
	复习考试	92, 86	88, 88

- $P1(1,q) = q * 90 + (1-q) * 86; P1(0,q) = q * 92 + (1-q) * 88$
- 容易检查，不存在 q ，使 $P1(1,q) = P1(0,q)$

混合策略均衡的概率也可能恰好取在端点

		你的拍档	
		PPT	Keynote
你	PPT	1, 2	0, 2
	Keynote	0, 0	2, 1

这个例子中，纯策略均衡有两个，
(PPT, PPT) 和
(KT, KT)
PPT → PPT, KT
KT → KT
PPT ← PPT
KT ← KT

$$q * 1 + 0 = 0 + (1-q) * 2, q = 2/3$$

$$p * 2 + 0 = p * 2 + (1-p) * 1, p = 1$$

(1, 2/3) 为混合均衡。

从收益矩阵也能看出，当你采取PPT，对手有什么都无差异

纯策略均衡不是概率恰好取端点的混合策略均衡！

帕累托最优和社会最优

- “个体最优”与“整体最优”
- 帕累托（Pareto）最优
 - 一个策略组：每个参与者对应其中一个策略选择。
 - 一个策略组被称为帕累托最优，若不存在其他策略组满足：所有参与者得到至少和目前一样高的回报，且至少有一个参与者会得到严格较高的回报。

		你的拍档	
		准备报告	复习考试
你	准备报告	90, 90	86, 92
	复习考试	92, 86	88, 88

这个例子中，存在三个Pareto最优，但都不是均衡。

社会最优

- 定义：一组策略选择是社会最优（或社会福利最大化），若它使参与者的回报之和（总收益）最大。

		你的拍档	
		准备报告	复习考试
你	准备报告	90, 90	86, 92
	复习考试	92, 86	88, 88

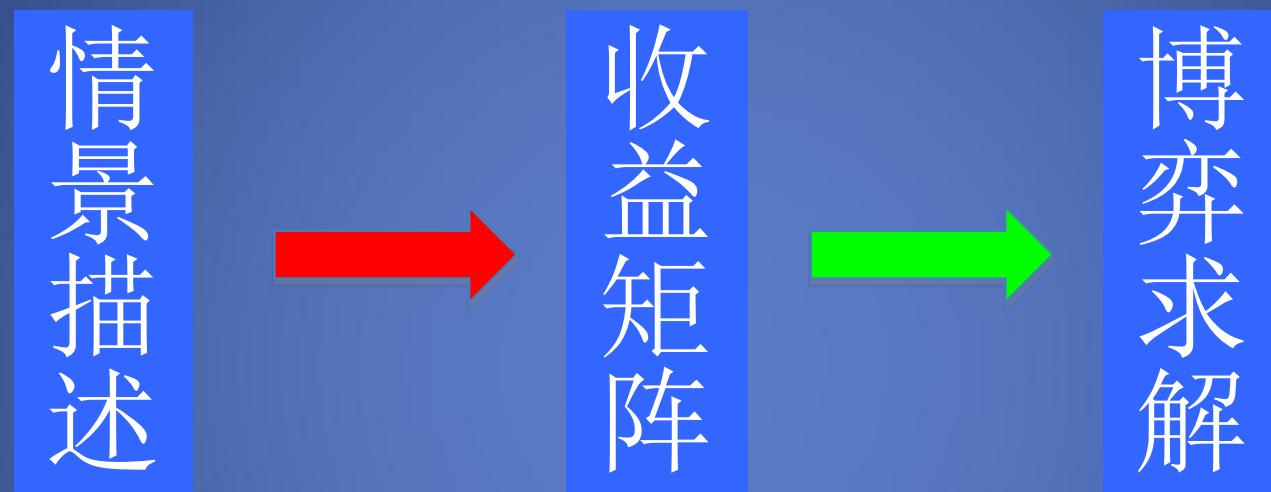
- （报告， 报告）是社会最优。
- 社会最优也一定会是帕累托最优。

社会最优和纳什均衡有可能一致

- 按照下面的收益矩阵，（报告，报告）既是社会最优也是纳什均衡

		你的拍档	
		准备报告	复习考试
你	准备报告	98, 98	94, 96
	复习考试	96, 94	92, 92

用博弈论思想分析问题



- 理解不同博弈的类型，以及求解的基本方法
重要（science）。均衡是一个基本目标。
- 将问题（情景）要求准确抽象成收益矩阵至
少同样重要（art）。

作业

- 第6章，第11，13题
- 第13题图6.44有错，请参考英文版书籍的图

课程报告说明

- 这个作业占成绩的**20%**，最后1周前提交网盘
- 基本目的是希望学生根据自己的兴趣和以往知识积累，对与课程内容相关的某一方面问题进行比较深入地探讨
- 具体结果是一篇报告，形式和内容自己决定。下面是几点具体指南意见
 - 长度，4页以上，正常的字号和排版形式，包括相关引用的文章，书籍，网站等。
 - 文章应该体现对某一主题创造性地探讨，可以参考其他文献，但不能只是文献综述。文章应该体现对该主题某种新颖的讨论或分析。

课程报告说明（续）

- 指南意见（续）
 - “新颖的讨论或分析”的含义可以从多方面考虑，例如定量或者定性；结合课程内容对现实社会问题的新理解；也可以是结合本课程内容讨论其他课程中所学的知识；还可以是对我们学的数学模型进行评述和推广，例如结合某些具体的应用进行讨论；如果你有兴趣，还可以是收集一些数据进行分析或者用计算机来进行模拟
 - 引用别人的内容（包括网上）需要标注
- 欢迎课间讨论你的构思

网络关系中的权力

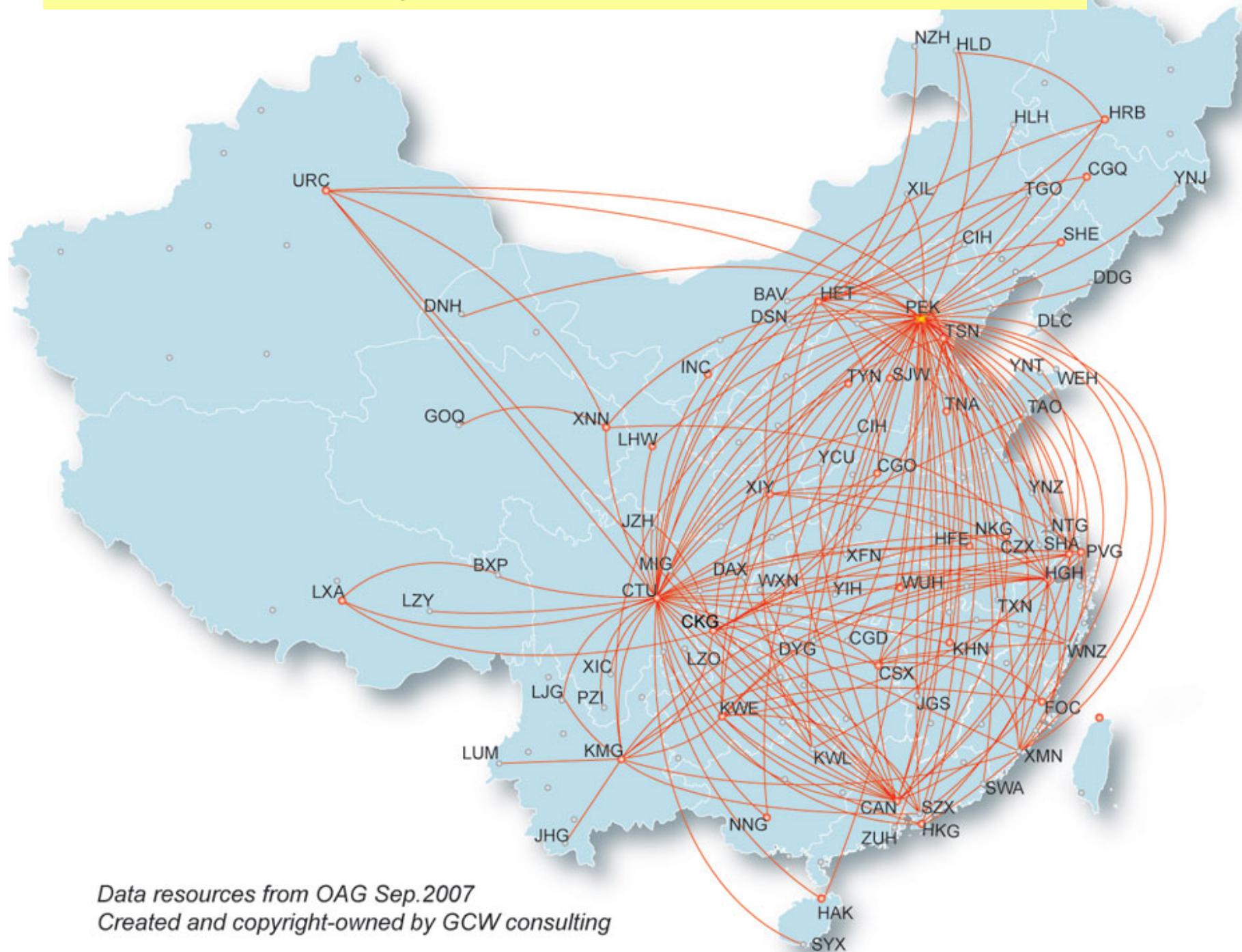
提要

- 问题：如何判别网络中一个节点的权力？
- 社会网络：节点的权力 **vs** 关系上的利益
- 课堂实验
- 网络交换实验的一般性描述
- 从若干例子看相关概念和认识
- 理论基础：纳什议价解与终极博弈
- 追求理论与实践的吻合
 - 稳定性：不可能有人来“破坏”已有的结果
 - 平衡性：关系本身也达到均衡

基本问题

- 给定一个网络（结构），其中的节点在“重要性”、“权力”、“地位”（或其他什么意义）上是否有不同？
- 如果有，如何确定（估计，预测）？
 - 计算机科学的问题：能否计算？如何计算？
- 容易预想到，不同类型的网络对这问题的回答会不同

节点的度数（degree）是重要性的直接指标



航线网

Data resources from OAG Sep.2007
Created and copyright-owned by GCW consulting

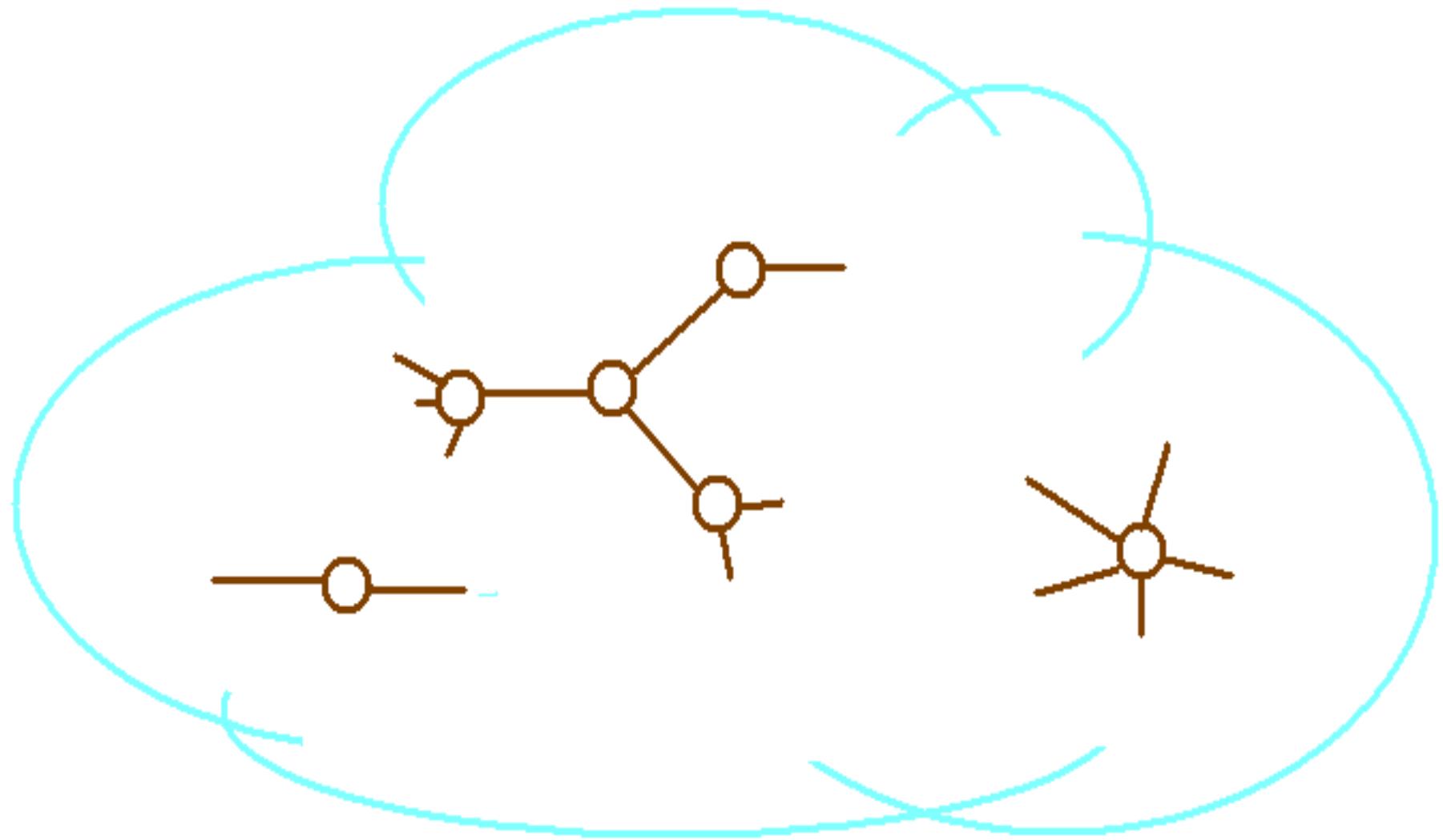
信息网 (The Web)

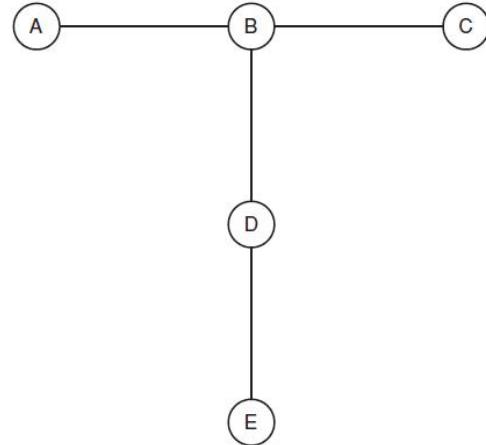
- 节点是网页
- 边是网页之间的超链 (**hyper link**)
- 每个节点的“出度” (**outdegree**)：对应网页上的超链数
- 每个节点的“入度” (**indegree**)：指向对应网页的链接数
- 入度高是重要性的一种体现，但还不够具有区分性，于是人们发明了 **PageRank**（成为**Google**最初的核心技术）

社会网络

- 节点：人
- 边：朋友关系、经济关系、同事关系等等
- Richard Emerson认为
 - 在社会网络上，与其讨论一个节点的权力（power），不如讨论在一个关系中的两个人哪一个人更有权力
 - 于是，同一个节点，与不同的邻居，所表现出来的权力就可能不一样。（现实生活不就是如此吗？）
 - 社会交换（social exchange）框架的著名学者，曾任华盛顿大学社会学系系主任，1982去世

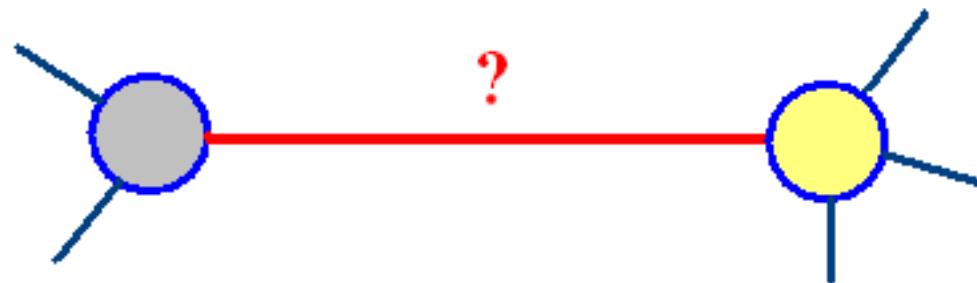
从“边”上看社会网络中节点的权力





- 出5个有手机的同学
 - 分别代表图中的节点
- 假设每条边上有10元钱， 相关的两个人要商议如何进行划分（取整数单位）
 - 商议的过程通过短信（微信）进行（可以来回多次）
 - 每个人只能和自己的邻居商议（为区别：可发“**A4**”等）
 - 每人最终只能最多与一个邻居达成协议；因此一旦在某个边上达成了协议，就通知另外的邻居，例如“**B-end**”
 - 整个过程进行5分钟，时间到则强行终止，没达成协议的节点收益为**0**（无论是时间到终止还是被邻居“抛弃”）
- 每人追求得到尽量多的收益

关系上的价值



- 谁在这个关系中具有较大的权力?
 - 如果有利益冲突，谁将比较有优势（优越）？
- 利益的“均衡”≈关系的稳定
 - 双方共识认定的，不是单方面强加的
- “均衡” ≠ “均等”

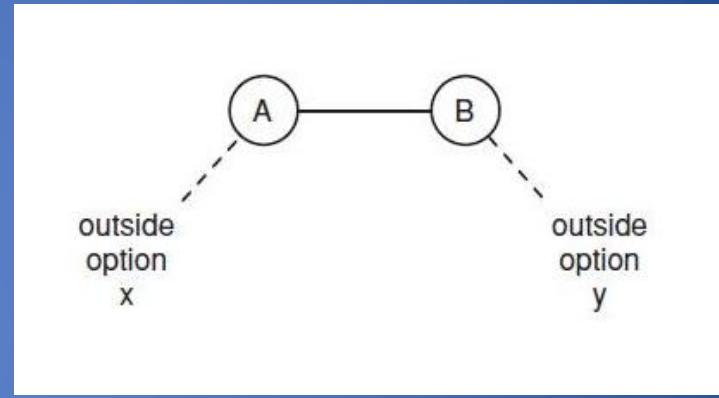
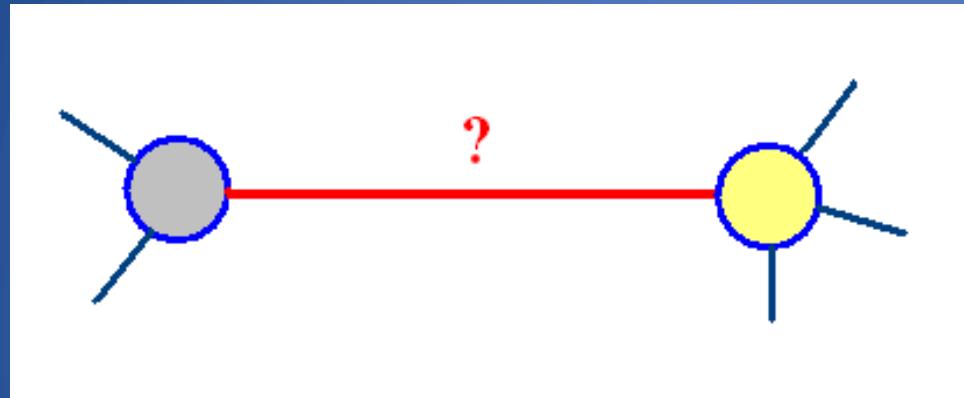
网络交换实验的一般性描述

- 通过一种有经济象征意义的操作，考察社会网络结构对权力的影响
- 选择有代表性的网络
 - 针对一个原理，一种认识等
 - 依赖性、排他性、饱和性、中心性，等等
- 同一网络，多次重复实验，得统计性结果
 - “高信息”和“低信息”的差别不大
 - 1—交换，2—交换，等等，影响复杂
- 如果一种实验结果一致性强，问：为什么会如此？
 - 背后当然会涉及人们的心理因素，但明显是由网络结构调动出来的心理因素。

若干例子与其中蕴含的概念

- 2—节点路径：完全对称性
- 3—节点路径：权力的极端不平衡
- 4—节点路径：权力的弱不平衡
- 5—节点路径：中心不一定有强权
- 柄图：较强，但不极端的权力
- 三角图：不稳定性
- 一般地，任意给定一个网络，若执行网络交换实验，我们能预期什么结果？

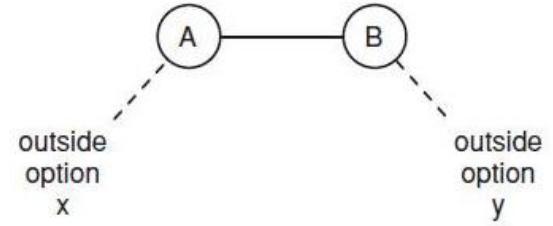
理论基础：纳什议价解与终极博弈



- 讨论两个节点之间的权力关系，可将网络中其他部分的影响归结为一个“外部选项”
 - 类似于“谈判底线”，即人们不会同意达不到外部选项的谈判结果，或者说有“退路”。
 - 是节点在网络中地位的一种集中体现

纳什议价解

(Nash Bargaining Solution)



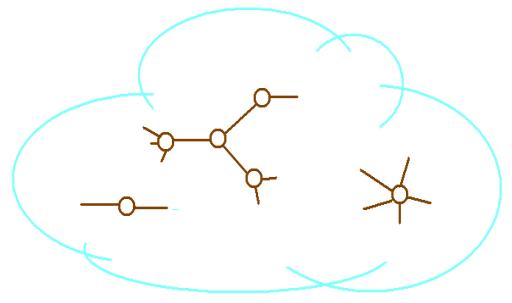
- 假设网络中两个节点的外部选项可以量化为 x 和 y , 在关系上 \neq 划分的预期结果如何?
 - 规格化, $0 \leq x < 1$, $0 \leq y < 1$, $x+y < 1$
- 纳什的理论结果: 均分 $s = 1-x-y$
 - 对于A, $(x+1-y) / 2$
 - 对于B, $(y+1-x) / 2$
- 证明不简单, 但与直觉(平分剩余)相符
- 因此, 我们将用这个结果作为(理论上)分析任意网络中节点关系价值划分的基础

终极博弈



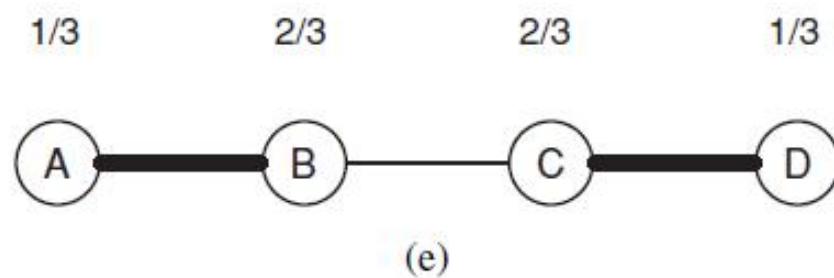
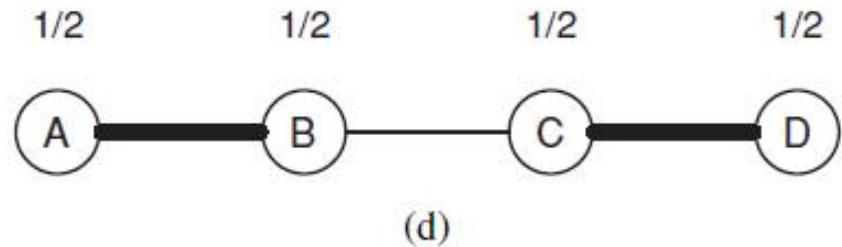
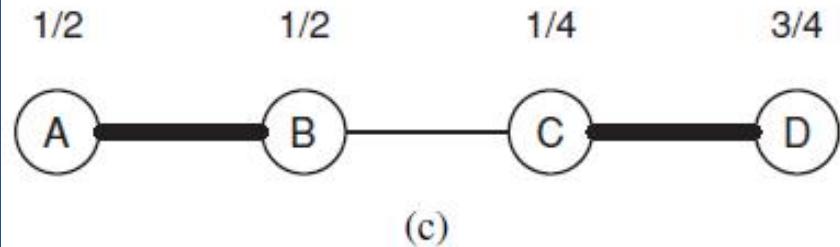
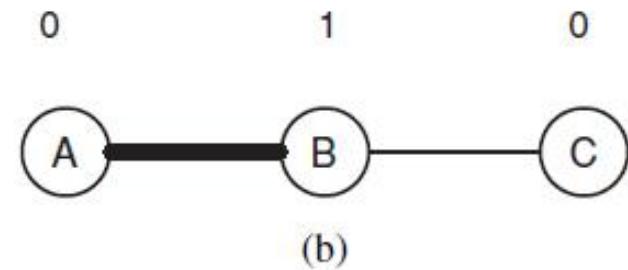
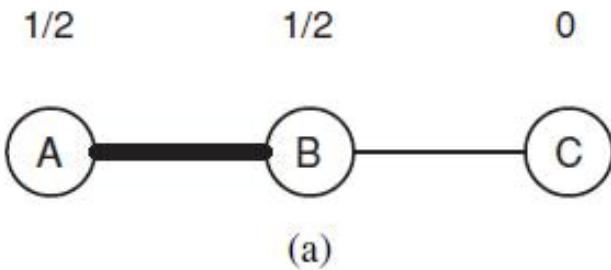
- 解释纳什议价解与实验结果的出入
- 3-节点路径，两边节点（A）别无选择（ $x=0$ ），中间节点（B）具有绝对支配地位，似乎可以使A的份额压缩到趋于0
- 但实验结果不支持这一点，而是给出类似于（ $1/6, 5/6$ ）的分配关系。也就是说，这种相对“温和”的结果是自然的
- 什么原因？
 - 在趋于极端的情况下，“金钱至上”不是人类的典型行为
 - “回报”在均衡中的综合含义
- 结论：现实中得到 $1/6-5/6$ 之类就可以认为达到理论极端结果

追求理论与实践的吻合



- 任意给定一个网络，若执行网络交换实验，我们能（从理论上）预期什么结果？
- 定义“结果”
 1. 图的一个**匹配**（一组边的集合，任何节点最多只出现一次）
 2. 匹配中每个节点在交换中得到的价值（0和1之间），匹配中一条边上两节点价值之和为**1**
 3. （不在匹配中的节点价值为**0**）

哪些结果是可能的？

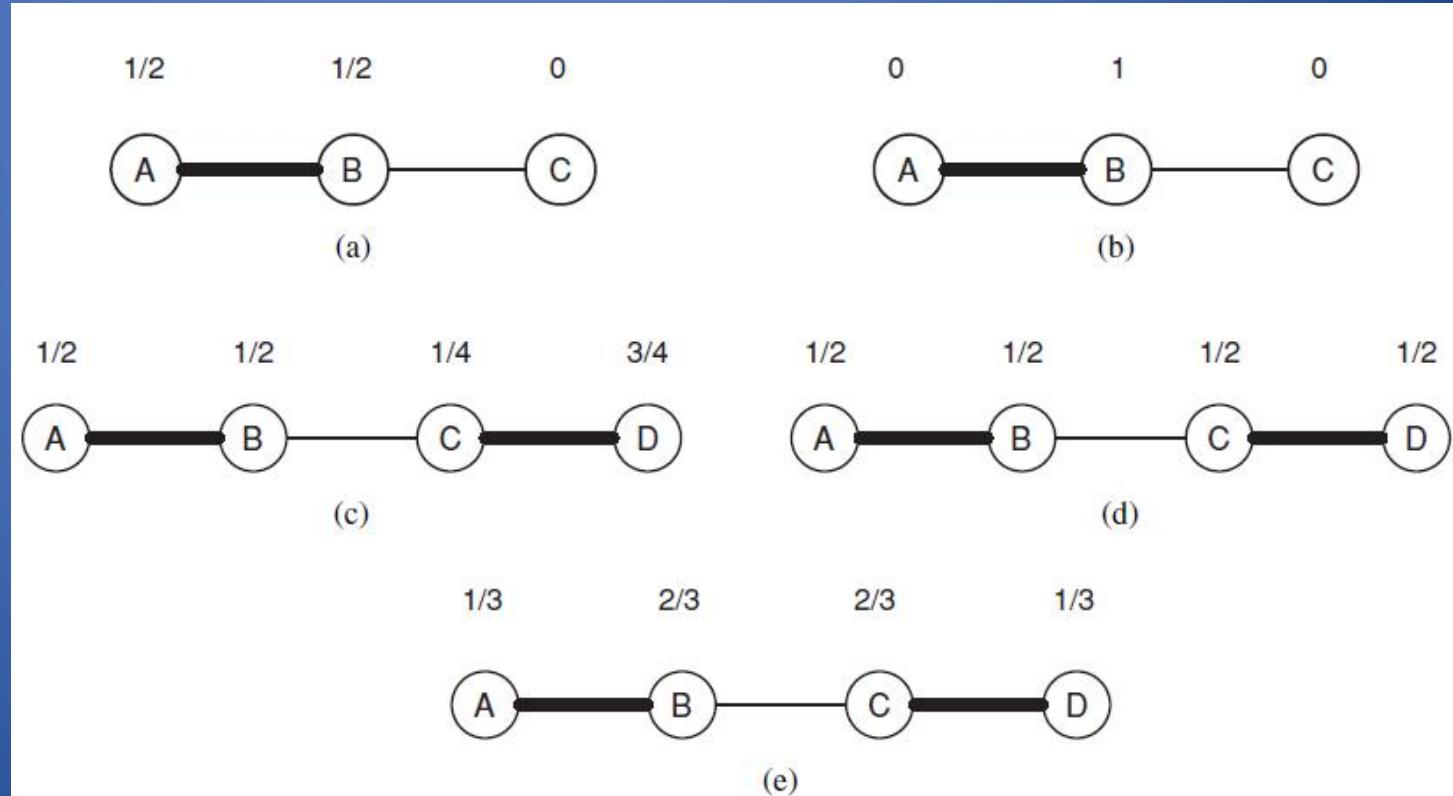


提问题的角度

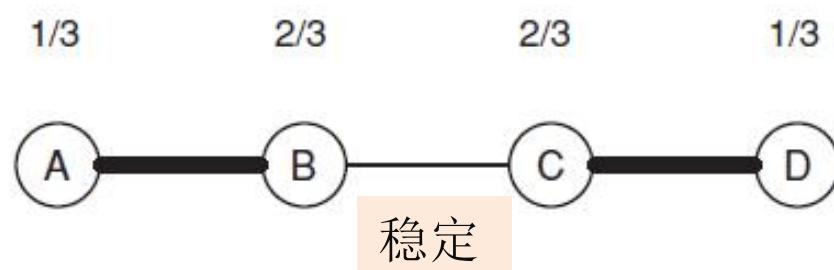
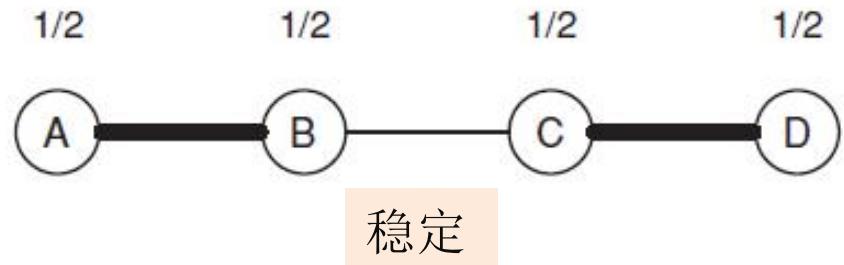
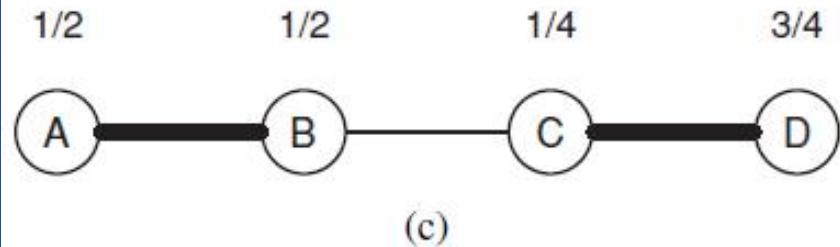
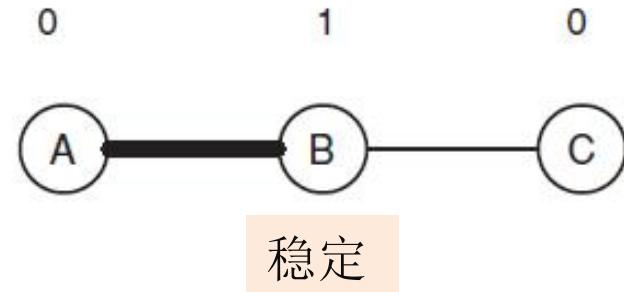
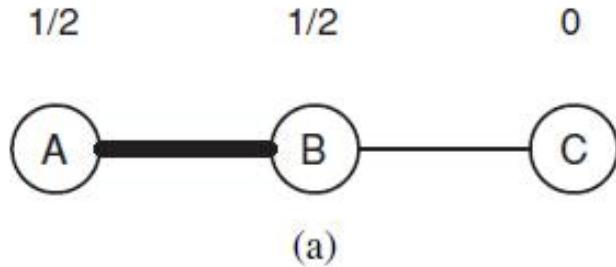
- 给定一个网络，若按照关于结果的定义，显然能有许多种结果出现，其中大概可分三类
 - 第一，实验中不太可能看到的
 - 第二，实验中很可能看到的（包括终极博弈体现的近似）
 - 第三，说不清楚的
- 理工背景的考虑：能否有一些判别准则（计算方法），来将第一类排除，将第二类识别？
- 社会学背景的考虑
 - 判断一个结果稳定还是不稳定
 - 判断一个关系应该对等还是不对等
 - 判断结构微小变化对权力的影响

稳定结果 (stable outcome)

- 不稳定因素: 不在结果中的一条边, 其两端节点的价值之和小于1
- 稳定结果: 不存在不稳定因素的结果
- 考察例子



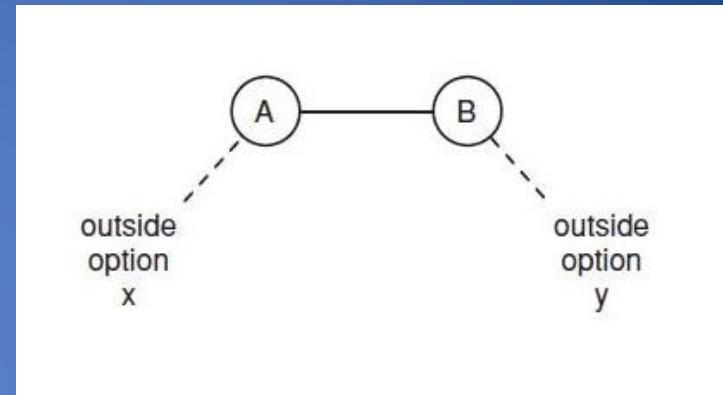
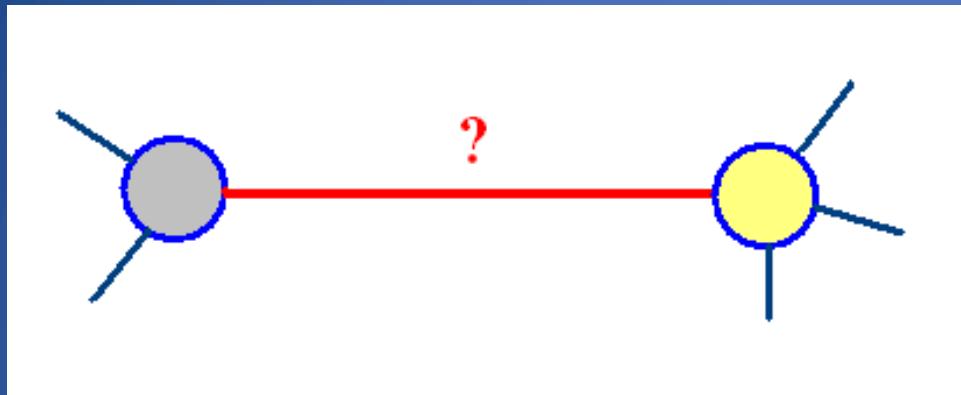
稳定结果都是“很可能出现的结果”吗？



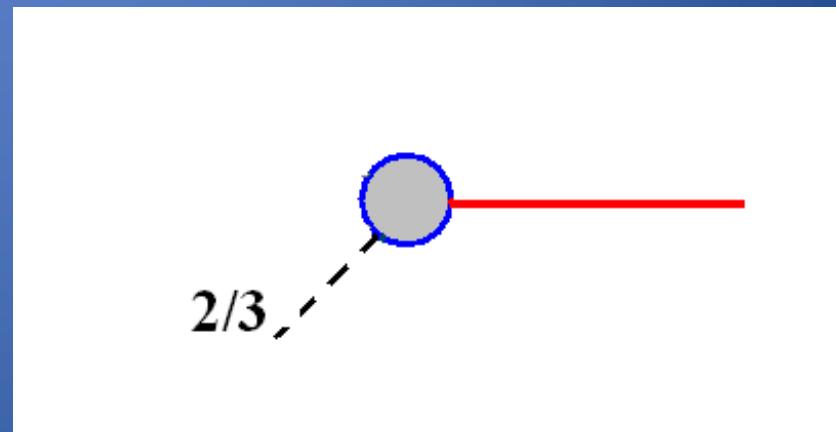
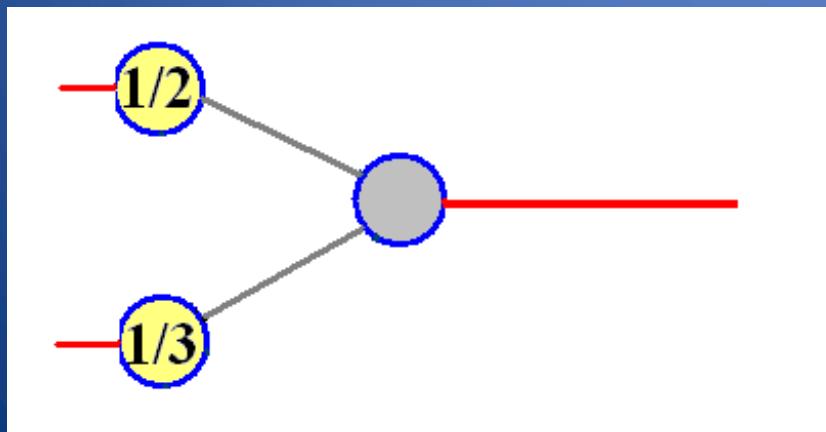
平衡结果 (balanced outcome)

- 稳定结果意味着一个“外人”不可能通过提供好处来破坏一个已有的价值分配关系
- 但那个分配关系本身是不是合理呢?
 - 即关系中的两个人是不是都满意?
- 前面明确了, 若一个价值分配对应纳什议价解, 则认为双方应该是满意的 (反过来就不一定满意, 即存在议价空间)
 - 但纳什议价解涉及到“外部选项”, 如何得知?

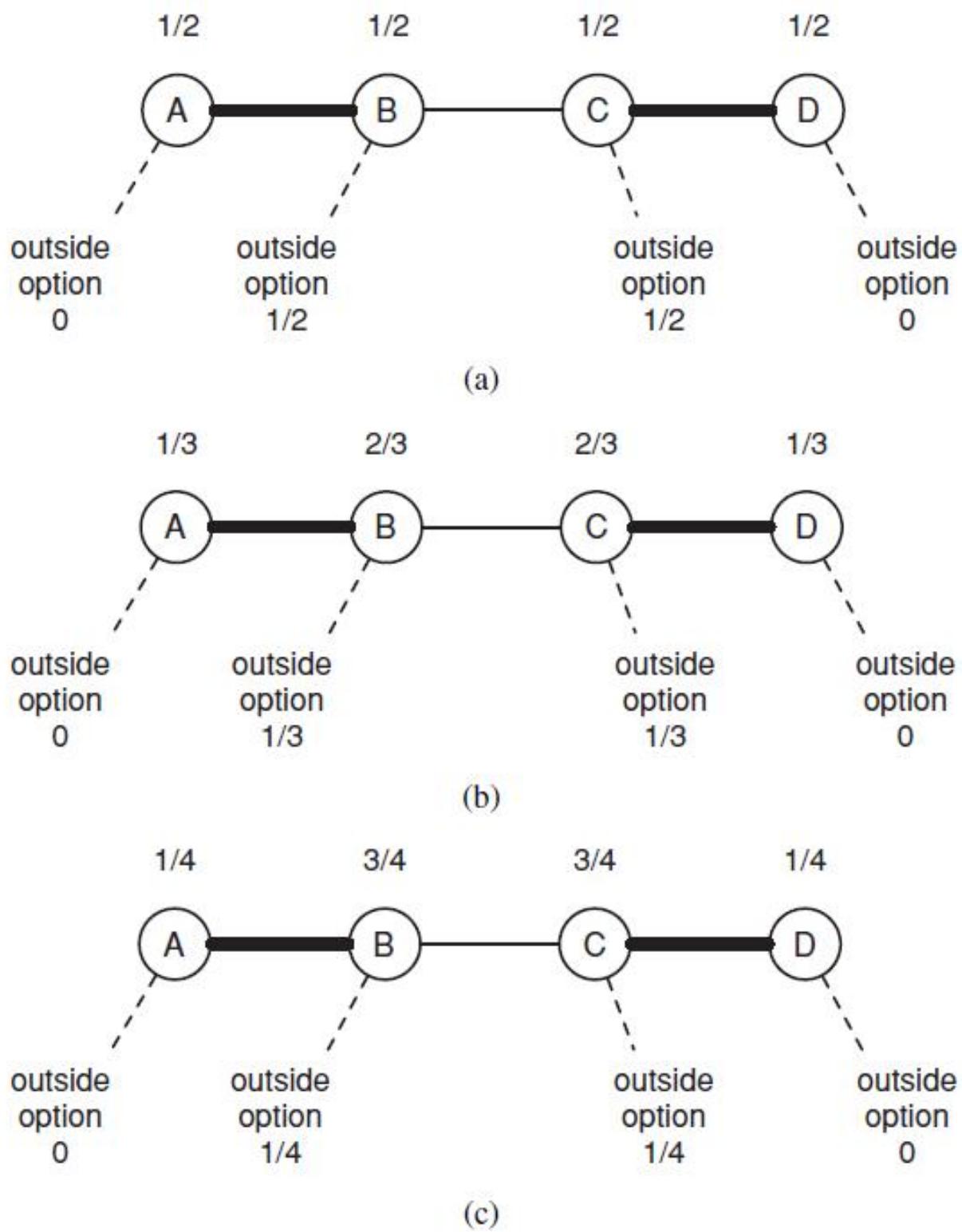
外部选项：网络其他部分的影响



- 放弃当前的匹配关系后所能得到的最大好处
– 即“退路”的价值

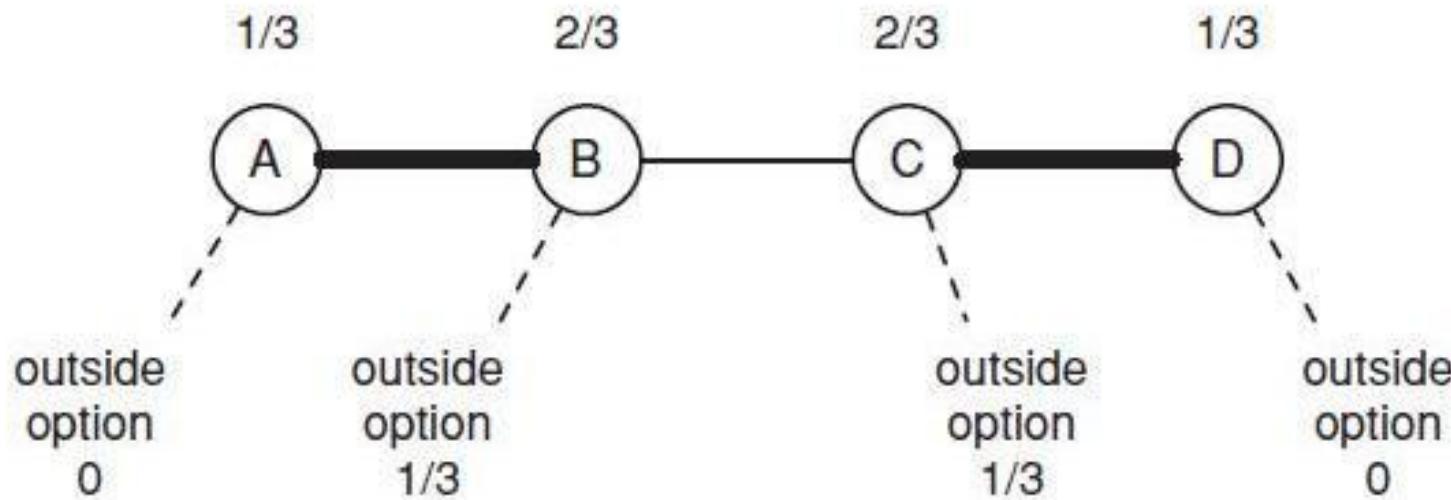


结果中节点的外部选项

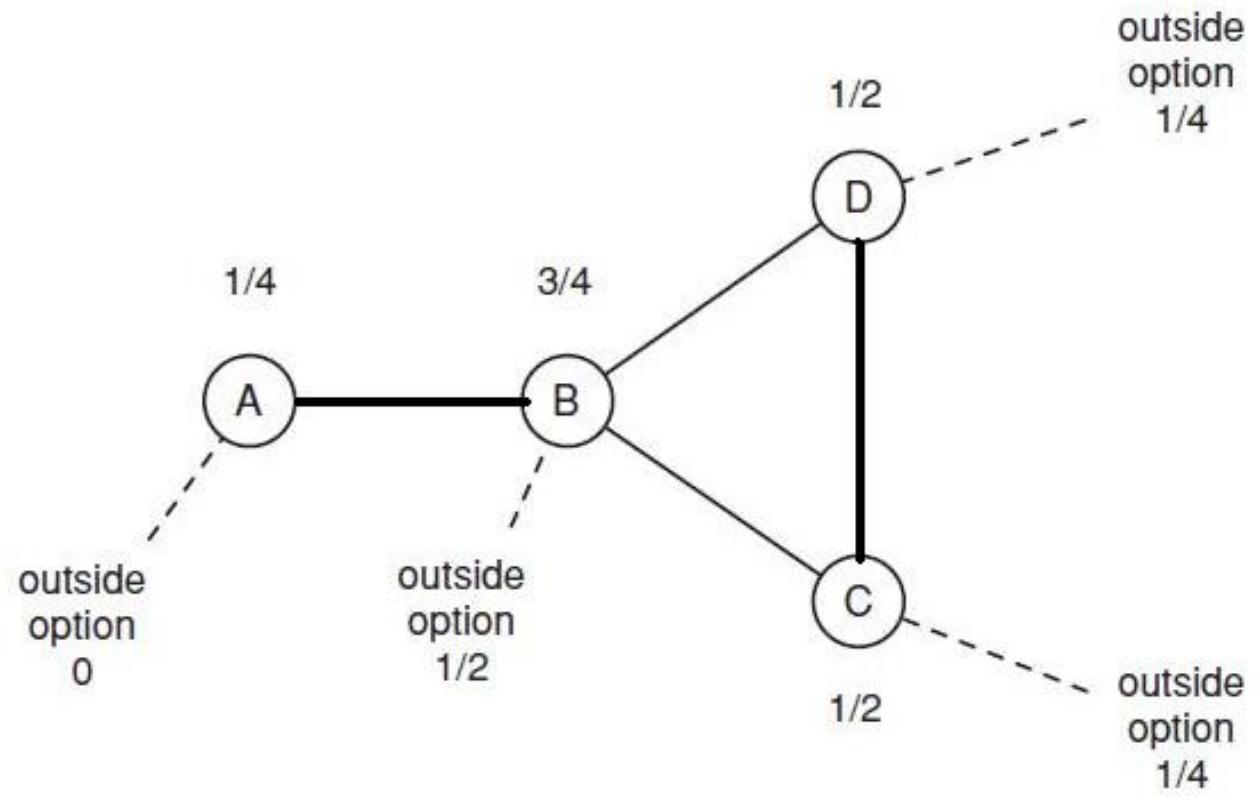


平衡结果（续）

- 于是，给定一个结果，我们可得每一个节点的外部选项，进而可以根据结果中节点的价值算得匹配中每条边上的交换是否为纳什议价解
- 平衡结果（定义）：结果中匹配的每条边上的价值划分都满足纳什议价解
- 平衡结果一定是稳定结果（证明）
- 社会网络中一定有平衡结果？



平衡结果
概念的区
分能力



小结

- 在社会网络中，讨论节点权力的差别在关系上的体现比讨论节点的权力更有意义
- 网络交换实验是社会学中用于理解关系上权力差别的有效手段
- 纳什议价解与终极博弈是理解关系上权力差别的两个基础原理
- 追求理论与实践的吻合：合理结果的性质
 - 稳定性：不可能有人来破坏已有的结果
 - 平衡性：关系本身达到均衡，与实验结果更接近

进一步可能关心的问题

- 还有没有比“平衡结果”更精细反映实验结果的概念？
- 1—交换，2—交换，多交换
- 计算机专业人员本能会问：给定社会网络，能否（如何）求得所有平衡结果？
 - Balanced Outcomes in Social Exchange Networks, Symposium on Theory of Computing, 2008
-

作业

- 第12章 5,6

万维网结构

提要

- 信息的组织
- 超文本（**hypertext**），超链（**hyperlink**）
- 网页、网站
 - 导航性（**navigational**）、事务性（**transactional**）
、主题性（**topical**）
- 万维网结构
 - 有向图（层次观）
 - 强连通分量
 - 领结结构的概念
 - 领结结构的计算

万维网 (World Wide Web)

- 有多大? (**size**)
 - 每个人从浏览器中看到的都是其中很小很小的一部分
 - 大型搜索引擎试图覆盖 (**index**) 其中一大部分
- 长什么样? (**shape**)
- 成长规律?
 - 规模
 - 形状
- 对国家和地区性Web可问同样问题

“size of the web”→google



WorldWideWebSize.com
DAILY ESTIMATED SIZE OF THE WORLD WIDE WEB

The size of the World Wide Web (The Internet)

[Tweet](#)

The Indexed Web contains at least **4.64 billion pages** (Monday, 13 April, 2015).

The Dutch Indexed Web contains at least **230.46 million pages** (Monday, 13 April, 2015).

[The Indexed Web](#) | [The Dutch Indexed Web](#)



WorldWideWebSize.com
DAILY ESTIMATED SIZE OF THE WORLD WIDE WEB

The size of the World Wide Web (The Internet)

The Indexed Web contains at least **4.51 billion pages** (Wednesday, 05 April, 2017).

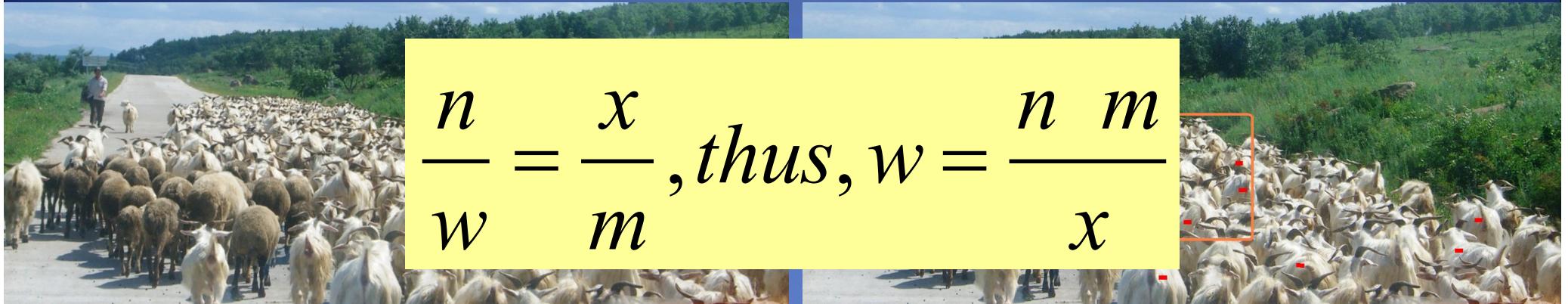
The Dutch Indexed Web contains at least **187.91 million pages** (Wednesday, 05 April, 2017).

人们关心Web规模问题

- People seriously report it from time to time
 - 1997, ~200 million, K. Bharat and A. Broder
 - 1998, ~800 million, S. Lawrence and C. Giles
 - 2000, ~2.1 billion, Shayna Keces
 - 2005, ~11.7 billion, A. Gulli and A. Signorini
- People maintain websites to talk about it
 - <http://www.worldwidewebsize.com>
 - <http://www.boutell.com>
 - <http://www.pandia.com>
- In China, CNNIC annually reports it since 2002
 - CNNIC – China Internet Network Information Center
 - 阎宏飞, 李晓明, “关于中国Web的大小、形状和结构”, 《计算机研究与发展》, 第39卷, 第8期, 2002年8月, 第958-967页。
 - 李晓明, “对中国曾有过静态网页数的一种估计”, 《北京大学学报》(自然科学版), 第39卷, 第3期, 2003年5月, 394-398。

如何估计Web
的规模?

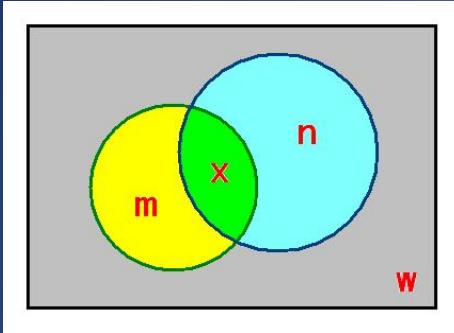
种群规模估计: capture/recapture模型



$$\frac{n}{w} = \frac{x}{m}, \text{ thus, } w = \frac{n \cdot m}{x}$$

- How many, **w**, sheep are there ?
- Capture a sub set **N**, count them, **n**, and release them **tagged**.
- Recapture a sub set **M**, count them, **m**, and count how many of them were in **N**, **x**.
- Web size is estimated essentially the same way.

Estimation: capture/recapture



$$\frac{n}{w} = \frac{x}{m}, \text{ thus, } w = \frac{n \cdot m}{x}$$

- How do we get n, m, and x for the Web?
- In practice, researchers explore multiple search engines (via queries) to estimate those numbers.
- We may consider each search engine has a random subset of the Web, different SEs have different subsets. Through a set of queries, x can be figured out from the returned urls (pages) by each search engine.

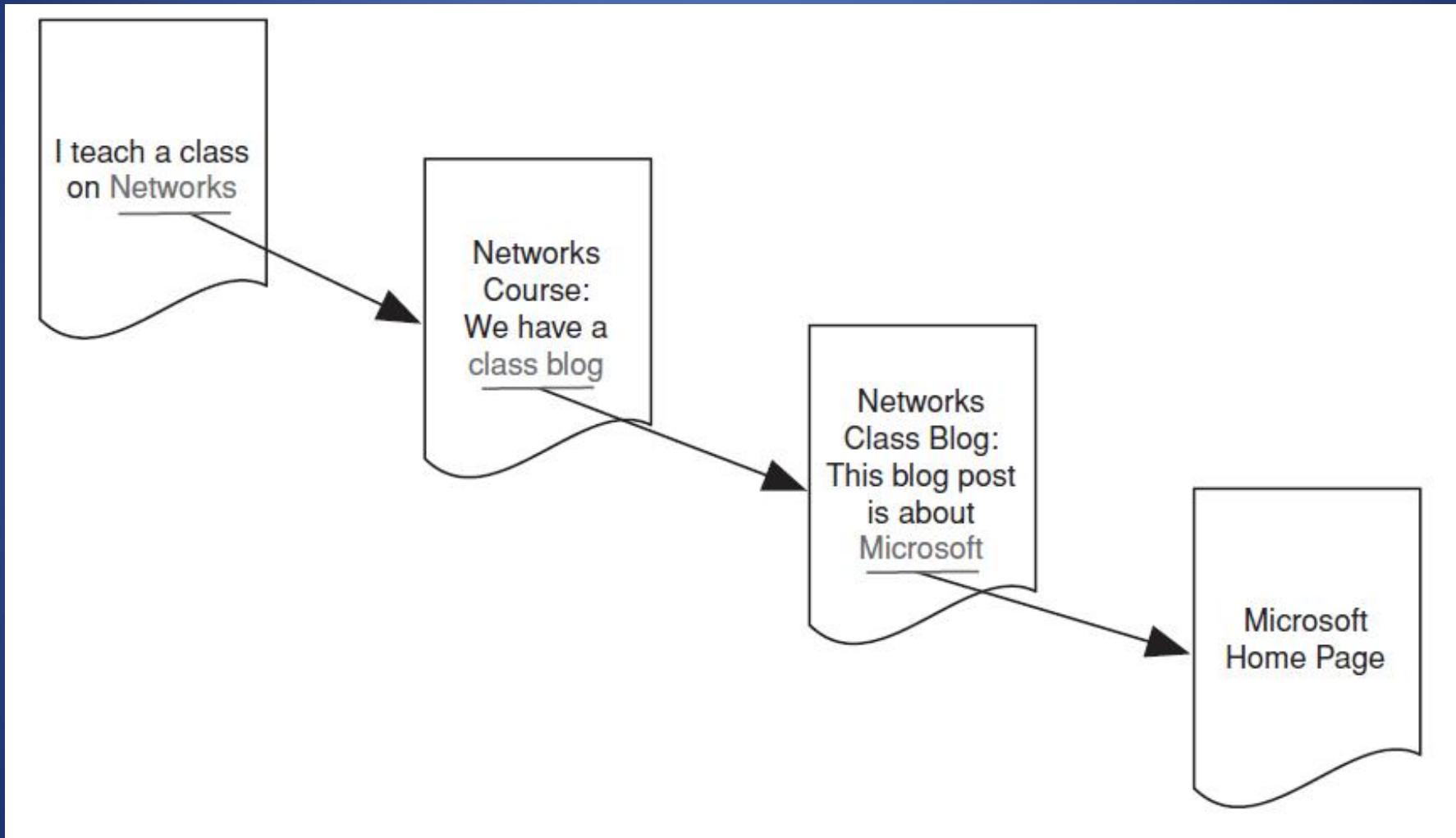
信息的组织（一般意义的）

- 信息单元（元素）：书籍、文件、网页等
- 动态变化的信息单元的集合
 - 一个图书馆的书，一个人计算机中的文件等
- 如何将集合中的信息“组织”起来
 - 便于利用（查找，使用）
 - 索引、目录、关于信息的信息（元信息）
 - 便于维护（信息单元的加入和删除）
 - 维护的方式？集中式 **vs** 分布式
 - 杜威分类体系，目录树，文件夹，关联链接

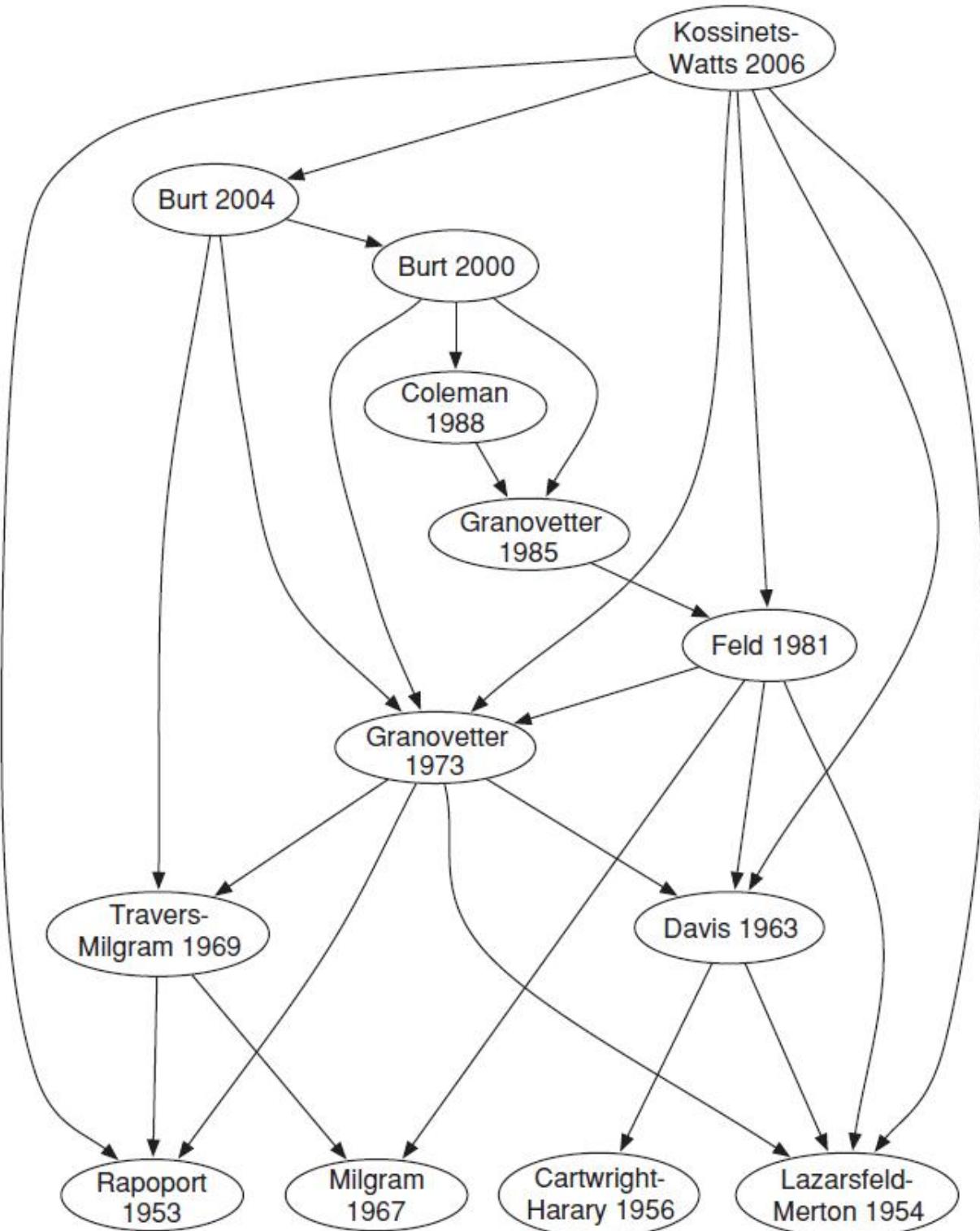
Web信息的组织方式

- 信息单元：网页
 - 准确定义的困难。按“地址”（URL）？
 - 从浏览器看到的不同于搜索引擎搜到的
 - 同一个地址，不同的人看到的可能不同
 - 从浏览器中看到的有些网页（例如一个电子商务活动的一个收据）并没有像样的地址
- 网页之间的关系
 - 超链接（hyperlink）
 - A包含一个指向B的超链通常意味着A对B的一种“认可”
- 基本优势：可扩展性

几篇网页之间的链接关系



- 注意，不仅信息所处的位置可以相距很远，其中的主题也可能“漂移”很远；不奇怪，人的思维也如此。

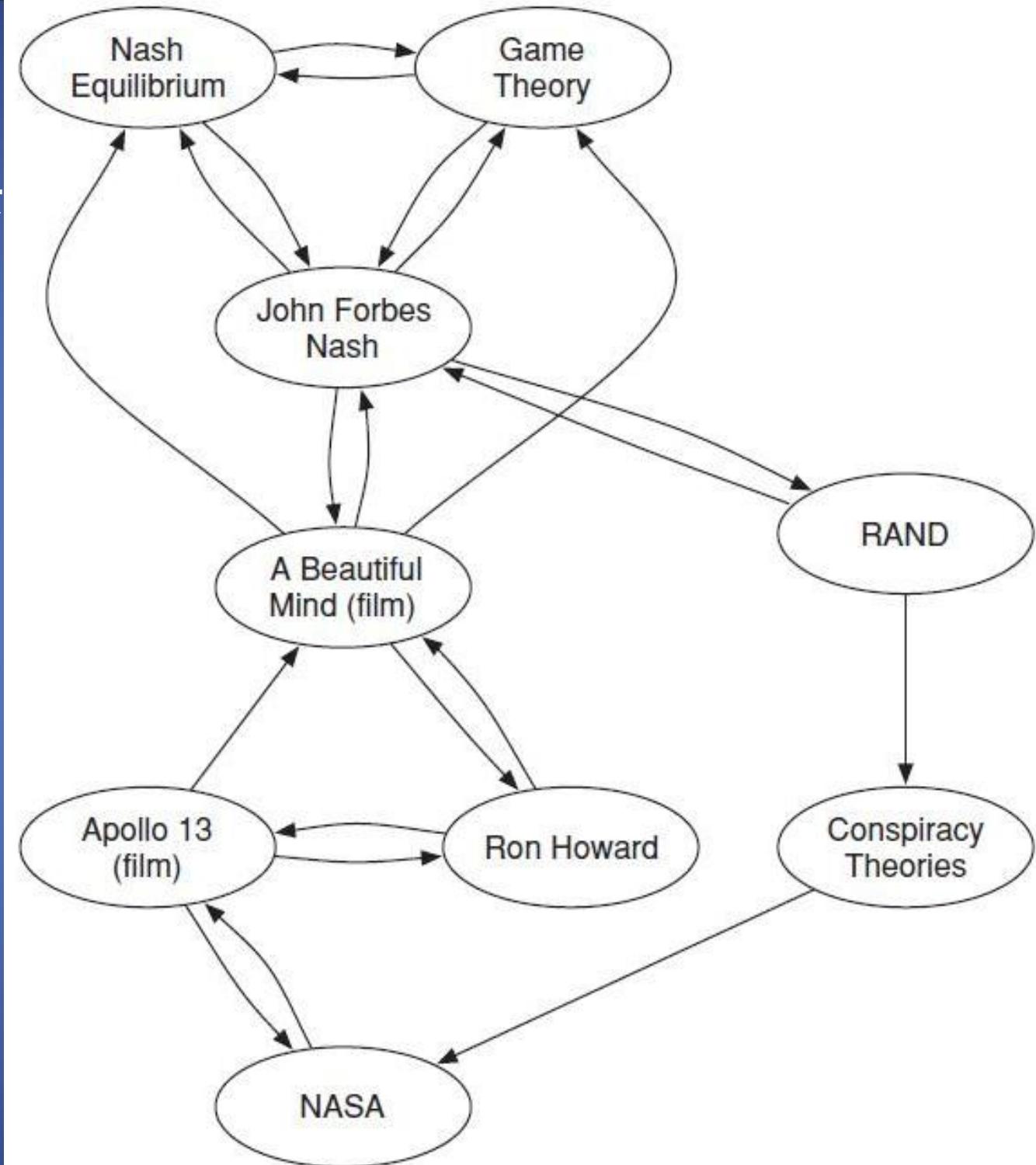


“链接”不仅
用于表达网页
之间的关系

- 文献引用关
系之间具有
时间流向性

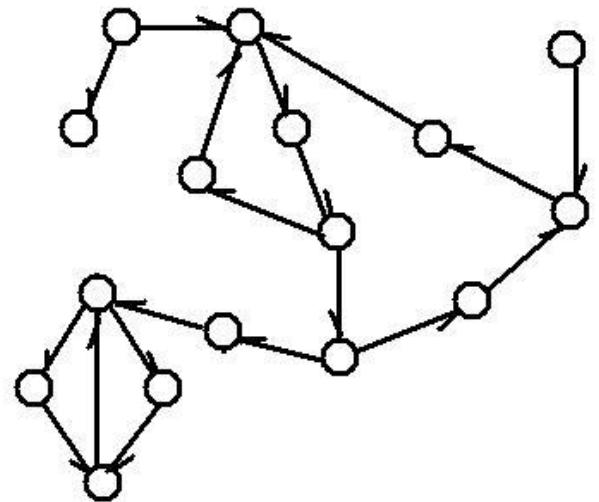
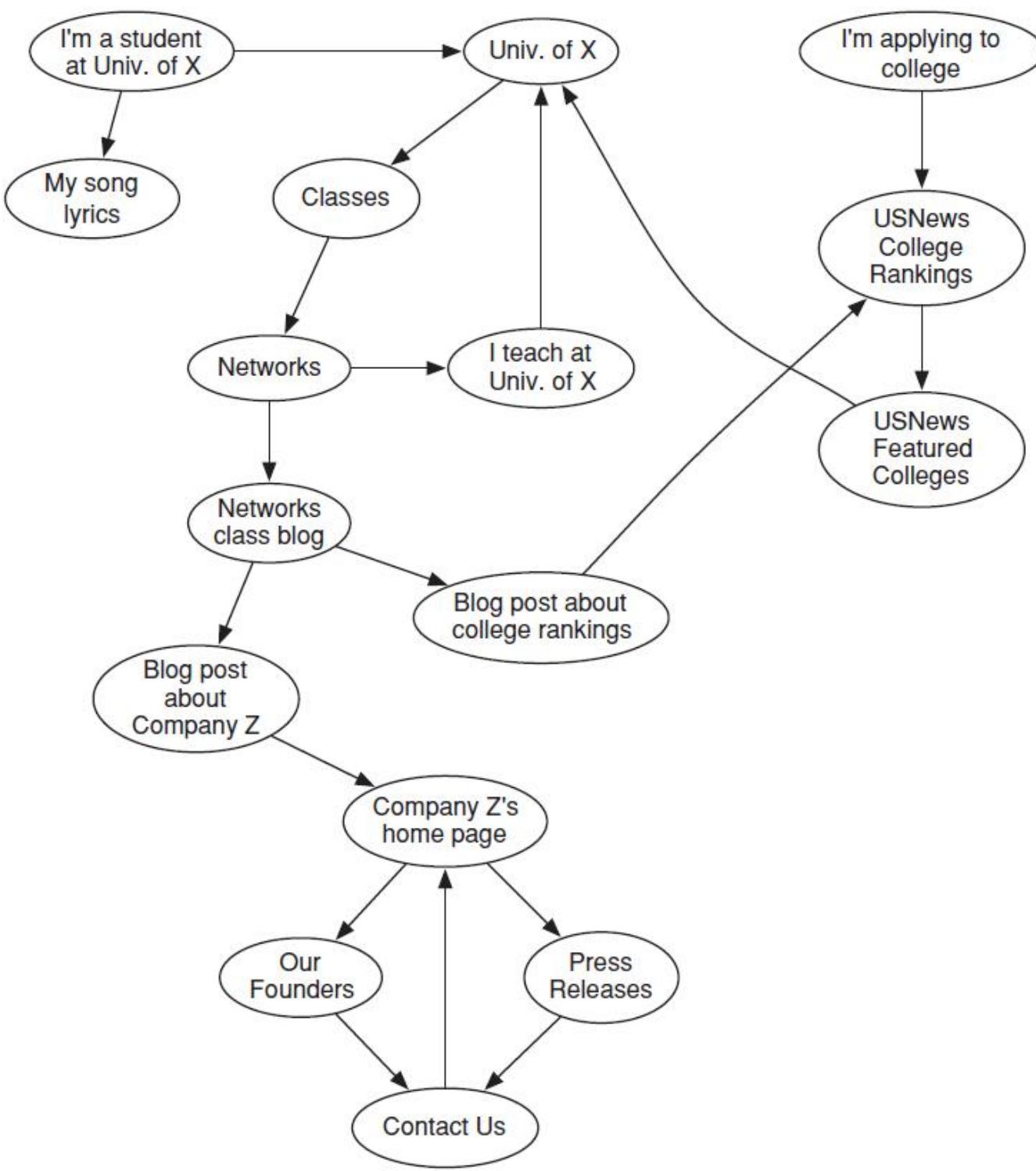
信息网络中链接两端无隐含时间序的其他例子

- 维基百科文章的交叉参考关系构成信息网络
- 大百科全书的词条之间的引用网络



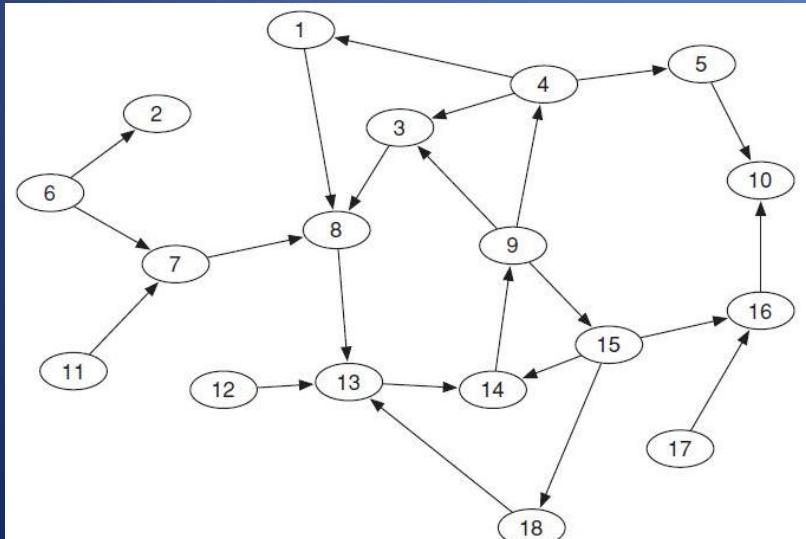
一组网页之间
构成的一个有
向图示例

* 具体与抽象



有向图的几个关键概念

- 有向图
 - 有向路径：两节点之间边的方向一致的路径
 - 强连通有向图：任何两节点之间都存在两个方向的有向路径（不一定经过相同节点）
 - 强连通分量：尽可能大的节点子集，其中每个节点都有到其中任何另一节点的有向路径



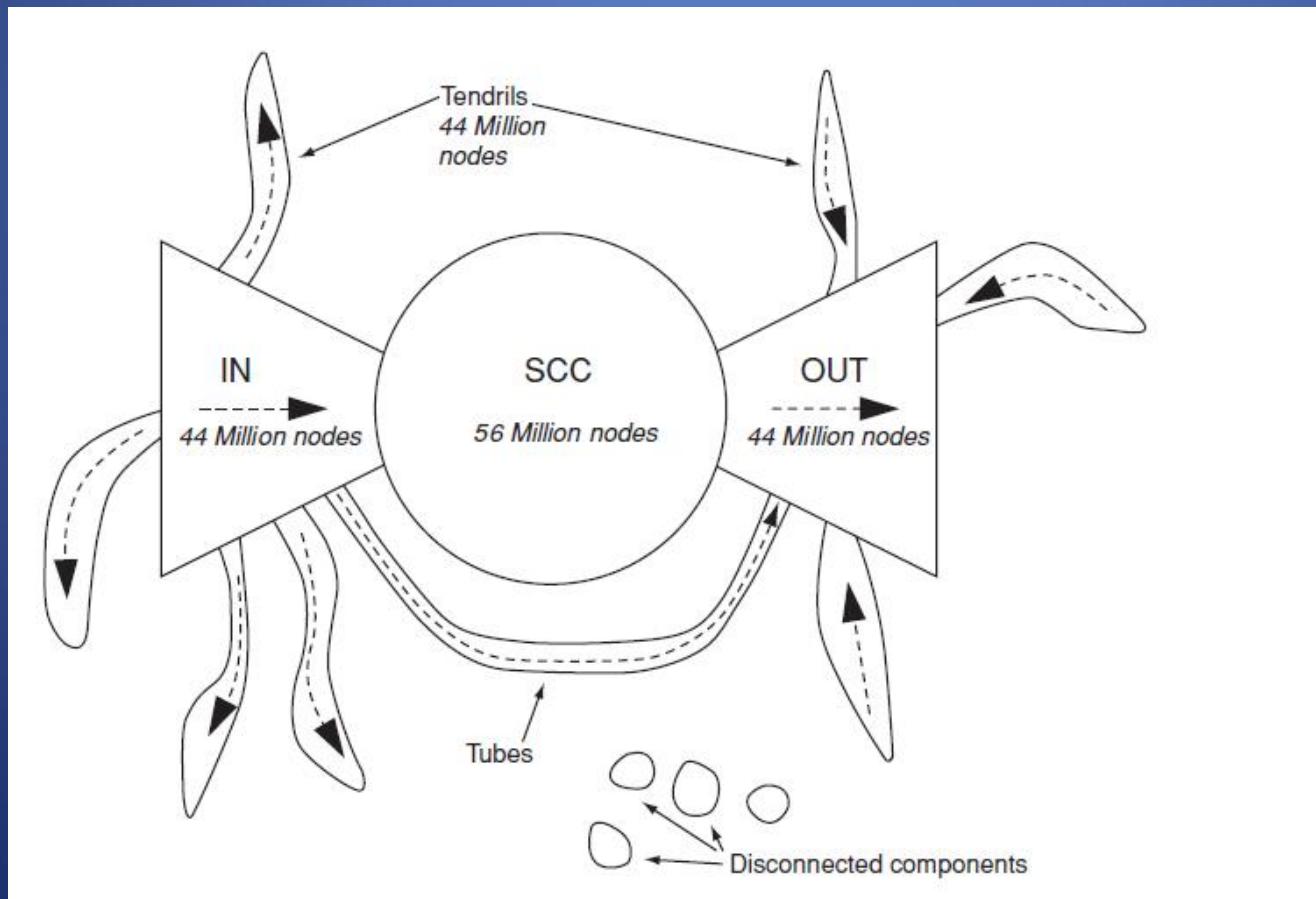
- 与图（无向图）对比
 - 路径
 - 连通
 - 分量
- 易见，强连通概念视觉上不如连通概念直观

万维网的结构模型：有向图

- 根据用途，可在**不同层次**定义图的节点和边的含义
 - 网页层次：网页
 - 网站层次：网站（例如系统结构所的网站）
 - 机构层次：机构网站总体（例如北航各院系）
- 还可以按行政层级分
 - 县、地市、省
- 网页层次是基础

“领结”：万维网图结构的一种概貌

- 1999, Andrei Broder等发现万维网包含一个超大强连通分量SCC，加上其他部分，显示出一种形象的结构
 - 链入，链出，卷须（管道），游离



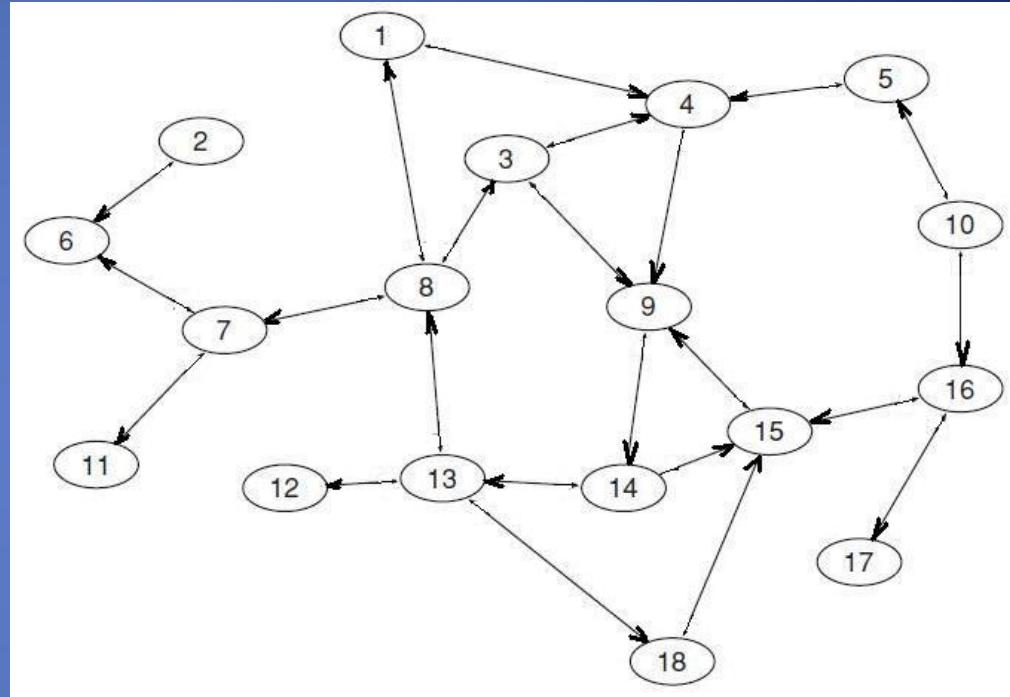
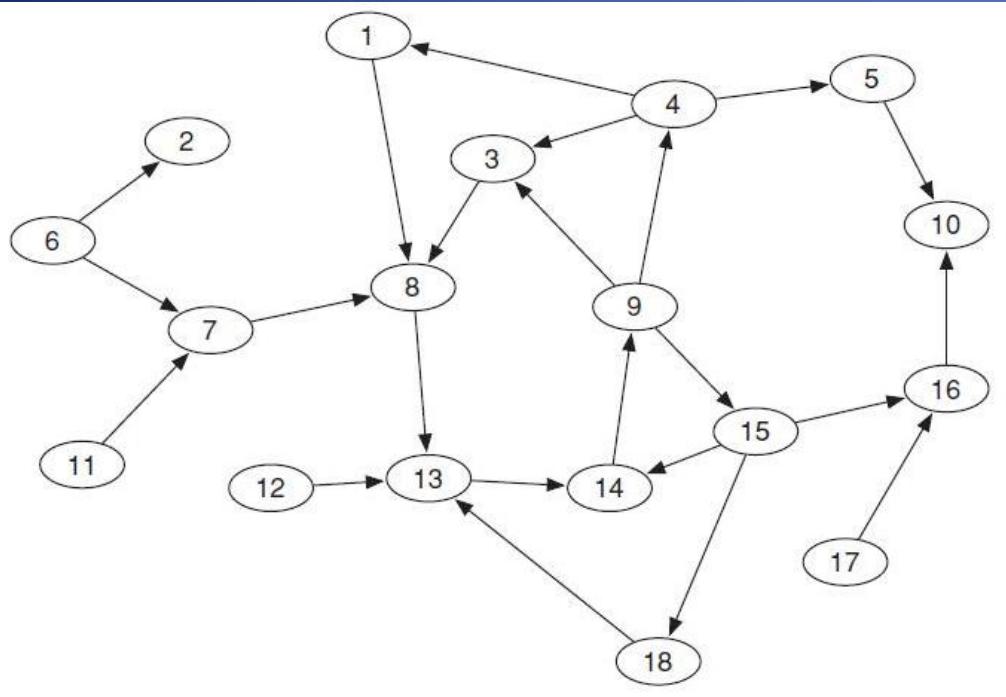
这是怎么知道的？

基本问题

- 给定一个有向图，如何得到其中的强连通分量？
 - 显然不一定就一个。强连通分量的划分性。
- 以最大的强连通分量为基础，如何描述其他部分与它的关系？
 - 链入，链出，卷须（管道），游离；这四个概念是否足够？

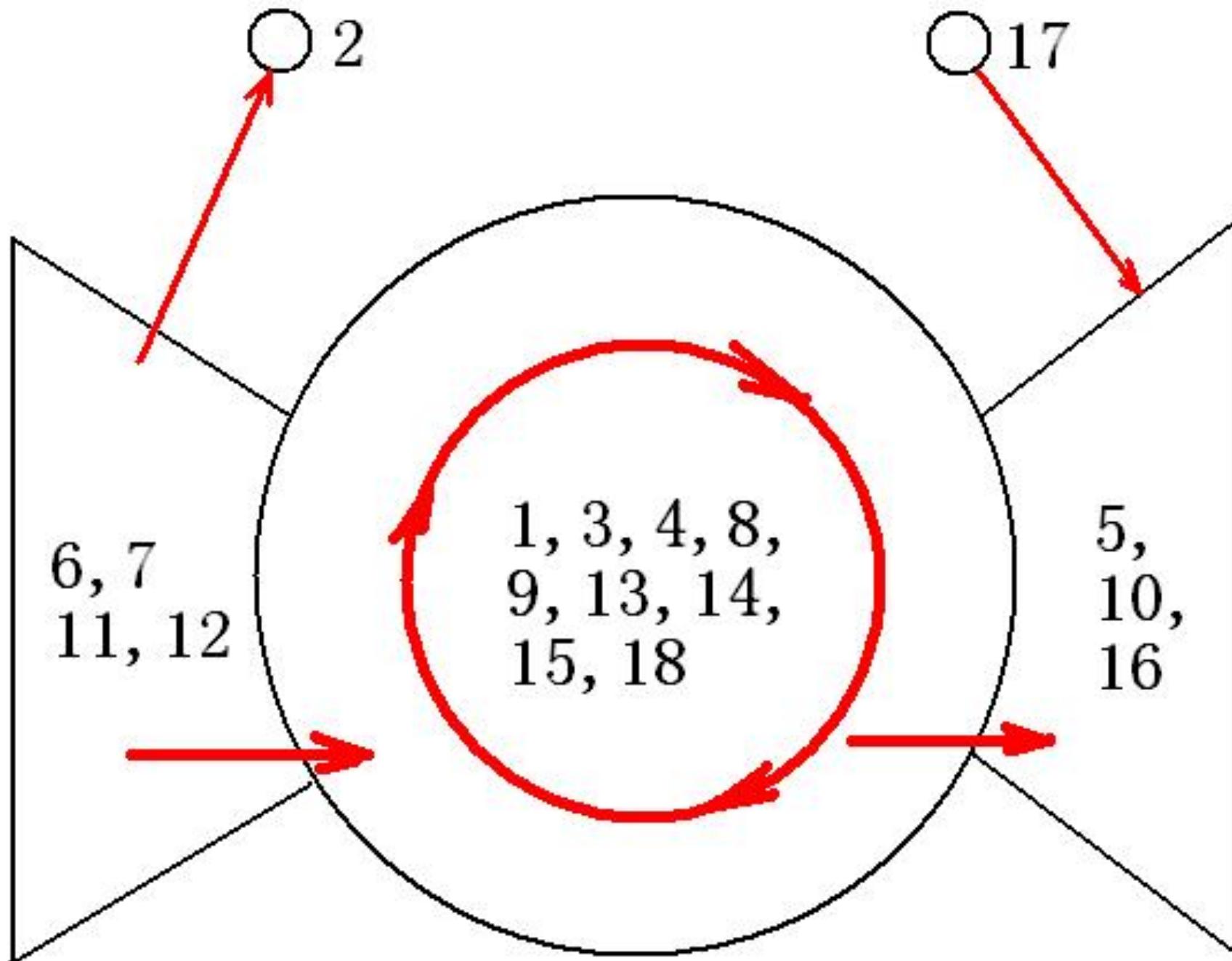
为了回答第一个问题，我们问一个更具体些的问题：
给定一个节点，如何确定包含它的强连通分量？

从一个具体例子入手



- $\{1, 3, 4, 8, 9, 13, 14, 15, 18\}$;
- $\{2\}$, $\{5\}$, $\{6\}$, $\{7\}$, $\{10\}$, $\{11\}$, $\{12\}$, $\{16\}$, $\{17\}$

有向图的“领结”表示



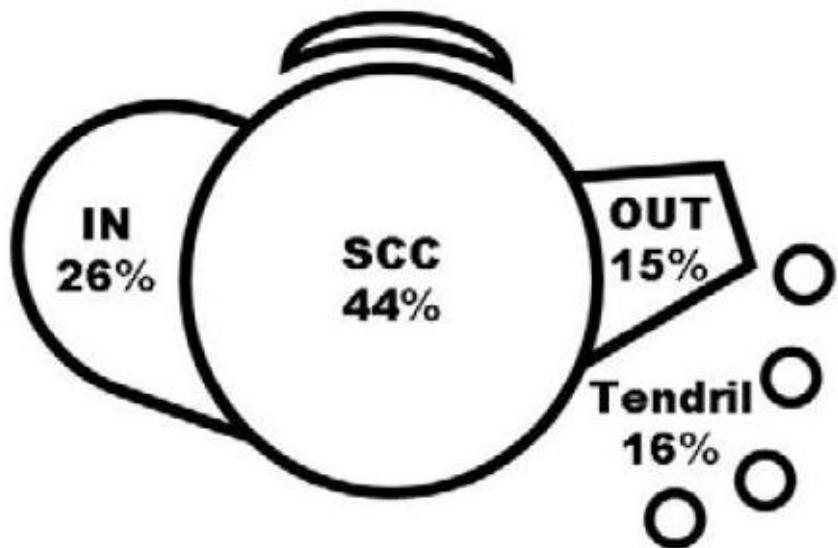
计算领结结构的方法（算法）

- 输入：有向图 G
- 第一步：生成图 G 的“反向图” G'
- 第二步：选择一个在最大强连通子图中的节点 A （tricky?）
- 第三步：以 A 为出发节点，在图 G 中宽度优先搜索直到没有新的节点发现，得节点集合 FS
- 第四步：以 A 为出发节点，在图 G' 中宽度优先搜索直到没有新的节点发现，得节点集合 BS
- 结果
 - $SCC = FS$ 和 BS 的交集，即共同元素
 - IN （链入） = $BS - SCC$
 - OUT （链出） = $FS - SCC$
- 在 FS 和 BS 基础上进一步操作可给出卷须和游离（细节略）

一个计算实例

- From Jan-Feb, 2006, PKU conducted a relatively thorough crawl of Chinese web, 830 million pages were collected
- As a result, PKU constructed a huge directed graph of 830 million nodes, summing to 400GB+ data
- A program ran one week on a 16 nodes cluster and generated the shape parameters

Figure 1. A Teapot Graph of Chinese Web



- 网页: <http://.../....html>, (完整地址)
- 网站: http://.../*, 对应例如大学的一个系
- 机构: http://*..../*, 对应例如一所大学所有院系网站的集合

Table 1. Components of Chinese Web Graph

	Italy ¹	UK ¹	Indochina ¹	China (page- level)	China (host- level)	China (domain- level)
SCC	72.3%	65.3%	51.4%	44.1%	50.7%	63.3%
IN	0.03%	1.7%	0.7%	25.5%	1.4%	0.7%
OUT	27.6%	31.8%	45.9%	14.6%	47.4%	34.9%
DISC /Tendrils	0.01%	1.2%	2.1%	15.8%	0.5%	1.1%
Total	100%	100%	100%	100%	100%	100%
N of Pages	41.3M	18.5M	7.4M	836.7M (hosts)	16.9M (domains)	0.79M
N of Links	1.15G	194.1M	298.1M	43.28G	43.28G	43.28G

¹ Taken from [2].

结果：
自相似、层次性

作业

- 第13章 1,3

链接分析与网络搜索

提要

- 网页排序（**ranking**、排名、排位）
 - 搜索服务的基本问题，传统信息检索技术的不足
- 中枢与权威
 - 一篇网页的两面性；有向图的启示
- 中枢值与权威值的计算（**HITS算法**）
- **PageRank**（含义）
- **PageRank**（计算）
- 退化图结构带来的问题
- 随机游走及其与**PageRank**定义的等价关系

搜索引擎关心的基本问题

- 计算机显示屏一次只能显示5-6个结果，典型搜索引擎掌握的网页超过10亿
- 对用户提交的一个查询，如何从这种海量网页集合中将最可能满足用户需求的少数几个结果找出来，展现在计算机显示屏上？
 - “最可能满足”的多义性
 - 同一个查询，不同的需求（苹果，病毒等）；
 - 不同的查询，相同的需求（电脑，计算机等）

传统信息检索技术的要点 (information retrieval, IR)

- 基于词语之间的相关性 (relevance)
 - $\text{similarity}(q, d) \approx \sum \text{score}(d, \text{term})$
- 传统应用背景
 - 文档集合: 图书, 规范的文献
 - 查 询: 主题词, 关键词
 - 查询意图: 获取与查询词有关的书籍和文章
 - 用 户: 图书管理人员
- “查询目标包含查询词” 是一个合理假设
 - 在形成查询词的时候就有这样的潜意识

现在查找学术文献有类似预期

- 但人们在网络上不光是要找“文献”，而是多方面意义的“信息”
- 例如，人们给出“北京航空航天大学”查询词，多数会有什么预期？
- 查询“大学”呢？（意图会相当多样化）

The strength of weak ties – Google Search

http://www.google.com/#hl=en&source=hp&q=+The+strength+of+weak+ties

Apple中国 中国雅虎 维基百科 新闻 热门

The strength of weak ties – Goog...

Sign in | More

Google The strength of weak ties

Search About 2,950,000 results (0.26 seconds)

Everything Scholarly articles for **The strength of weak ties**

Images  [The strength of weak ties](#) - Granovetter - Cited by 17579

Maps [The strength in weak ties](#) - Liu - Cited by 81

Videos [The strength of weak ties: A network theory revisited](#) - Granovetter - Cited by 3331

News [PDF] [The Strength of Weak Ties - Stanford Sociology](#)
sociology.stanford.edu/people/.../granstrengthweakties.pdf
File Format: PDF/Adobe Acrobat - Quick View
by MS Granovetter - 1973 - Cited by 17543 - Related articles

Shopping [The Strength of Weak Ties](#). Mark S. Granovetter. American Journal of Sociology, Volume 78, Issue 6 (May, 1973), 1360-1380. Your use of the JSTOR database ...

Books

More

Show search tools

[Mark Granovetter - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Mark_Granovetter - Cached
Jump to [The strength of weak ties](#): Granovetter's paper "The Strength of Weak Ties" is a highly influential sociology paper, with nearly 15000 citations ...
Background - Major ideas - Bibliography (selected) - See also

[PDF] [THE STRENGTH OF WEAK TIES: A NETWORK THEORY ...](#)
citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128...
File Format: PDF/Adobe Acrobat - Quick View
by M Granovetter - 1983 - Cited by 3321 - Related articles
hypotheses of my 1973 paper "The Strength of Weak Ties". (hereafter "SWT") and work that elaborates those hypotheses theoretically or uses them to suggest ...

[The Strength of Weak Ties - Sociable Media Group](#)
smg.media.mit.edu/classes/library/...weak.ties/granovetter.htm... - Cached
by MS Granovetter - 1973 - Cited by 3 - Related articles

为什么能恰到好处？

- 主页放在最前面，一定不是因为其中包含许多“北京航空航天大学”字样
- 很可能是由于许多包含“北京航空航天大学”字样的网页指向它
 - 利用链接中隐含的信息

Baidu 百度 北京航空航天大学 百度一下

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约7,280,000个 搜索工具

为您推荐 北京理工大学 南京航空航天大学 北京邮电大学

北京航空航天大学 官网

北京航空航天大学(简称北航)成立于1952年,是一所具有航空航天特色和工程技术优势的多科性、开放式、研究型大学,肩负着高层次人才培养和基础性、前瞻性科学研究,以及...
www.buaa.edu.cn/ - 百度快照 - 125条评论

北京航空航天大学高考分数线_招生信息_中国教育在线

办学类型: 211高校 985高校 普通本科
院校类型: 理工类
高校地址: 北京市海淀区学院路37号 [校园全景](#)
相关信息: [学校官网](#) [招生计划](#) [招生章程](#) [专业设置](#)

选择生源地: [北京](#) 选择科属: [理科](#)

年份	最高分	平均分	省控线	录取批次	专业分数线
2015	689	678	548	一批	各专业录取分数线
2014	--	658	543	一批	各专业录取分数线
2013	722	668	550	一批	各专业录取分数线

[查看更多北京航空航天大学信息»](#)

gkcx.eol.cn 2017-03-07

北京航空航天大学_百度百科

北京航空航天大学

基础信誉积累, 可接洽商谈 累计时间: 23个月
网民评价: 86%好评 125评价

网站地址: bnd.buaa.edu.cn
工商地址: 北京市海淀区学院路37号
经营范围: 培养高等学历航空航天人才、促进科技发展。力学、机械、材料、仪器仪表、信息通信、能源动力、电气、自...
[查看更多>>](#) 由百度信誉提供

211工程大学 展开 ▾

南京航空航天大学 对外经济贸易大学 厦门大学 中央财经大学
第一批次航空高等院校 教育部直属211综合类国家重点大学 第一所高等财经院校

I ❤ DHU
官方微博

北京理工大学 北京邮电大学 四川农业大学 东华大学
国内一流理工科院校 通信领域的领头军 国家211工程建
教育部直属的7

这个两个结果哪一个较好？

百度搜索_大学

http://www.baidu.com/s?wd=%B4%F3%D1%A7&rsv_bp=0&rsv_spt=3&in

Apple中国 中国雅虎 维基百科 新闻 热门

百度搜索_大学

Baidu 百度 新闻 网页 贴吧 知道 MP3 图片 视频 地图 更多

大学

大学_百度百科

大学，指提供教学和研究条件和授权颁发学位的高等教育机关。分为综合大学、专科大学或学院。它选拔具有高中以上学历者进行教育和培训，并以考试考核的方式检验其所学知识和技能。另有，儒家基本经典之一《大学》，也指聚集在特定地...共226次编辑

baike.baidu.com/view/4410.htm 2011-11-1

欢迎访问北京大学主页

北京大学作为国内前茅的文理医工综合性大学，在培养高素质创新型人才、取得突破性科研进展，以及为国民经济发展和社会进步提供智力支持等方面都发挥着极其重要的作用。

www.pku.edu.cn/ 2011-10-12 - 百度快照

北京大学_百度百科

北京大学，简称北大，创办于1898年，初名京师大学堂，是中国近代第一所国立大学，被公认为中国的最高学府，也是亚洲和世界最重要的大学之一。在中国现代史上，...

baike.baidu.com/view/1471.htm 2011-10-20 - 百度快照

hao123网址之家 -- 大学

把hao123设为首页 网友反馈 首页>大学 分数线查询 地区批次线正在加载，请稍候...211高校名单 985 高校名单 校园社区 大学排名(仅供参考) 阳光高考 大学专业介绍 ...

www.hao123.com/edu.htm 2011-10-21 - 百度快照

在北京市搜索大学_百度地图



- A. 北京大学 - (010)62752114
海淀区颐和园路5号
- B. 清华大学
北京市海淀区
- C. 清华大学 - (010)62782165
清华园
- D. 西南财经大学 - (010)82380594
北京市海淀区学院路30号北京科技大学...

大学 - Google Search

http://www.google.com/#hl=en&newwindow=1&q=大学

Apple中国 中国雅虎 维基百科 新闻 热门

大学 - Google Search

Google 大学

Search

About 5,030,000,000 results (0.11 seconds)

Everything

Images

Maps

Videos

News

Shopping

Books

Blogs

More

All results

Sites with images

More search tools

hao123网址之家 -- 大学

www.hao123.com/edu.htm - Cached - Translate this page

30+ items - 把hao123设为首页 · 网友反馈 · 首页>大学. 分数线查询. 地区批次 ...
· 地区 - 普通本科院校(820所 - 高职院校(1228所 - 独立学院(311所
· 北京 - 58所[点击查看 - 25所[点击查看 - 5所[点击查看
· 天津 - 19所[点击查看 - 26所[点击查看 - 10所[点击查看

211高校名单 - 北京普通本科 - 山东普通本科 - 江西普通本科

搜狐教育高校信息库-高校、专业一站查询

daxue.learning.sohu.com/ - Cached - Translate this page

中国大学信息查询系统所有信息、数据均来源于高校网站或相关出版物，仅供考生参考，请以官方公布信息为准。建议考生综合考虑各校公布的信息及各种因素填报 ...

大学-精品学习网

www.51edu.com/daxue/ - Cached - Translate this page

大学排行：精品学习网大学频道为广大网友搜集了2011三本大学排名汇总，供各位考试参考。[详细] · 2011中国三星级私立大学学科专业名单（港澳台） · 2011年全国二 ...

News for 大学

寒门子弟如何上大学？

新华网 - 5 hours ago

网上舆情要览：在大学教育阶段再去争取教育平权，其实为时已晚，教育的公平首先应当是起跑线上的公平，中小学的招生同样应当注重教育平权，向农村孩子作出一定倾斜，...

1225 related articles

郎朗哈佛大学开课 与刘翔相约明年奥运见(图)

腾讯网 - 57 related articles

张尧学任中南大学校长 黄伯云不再任校长职务

红网 - 47 related articles

有效利用链接关系蕴含的信息，是
搜索引擎超越传统信息检索系统、
技术进步的最重要标志

餐馆推荐问题

	甲	乙	丙	丁	
新辣道	*		*	*	21
海底捞	*	*	*		20
麦当劳		*			6
五方院	*			*	15
俏江南		*		*	13

8 6 6 7

看推荐人的“水平”

3
3
1
2
2

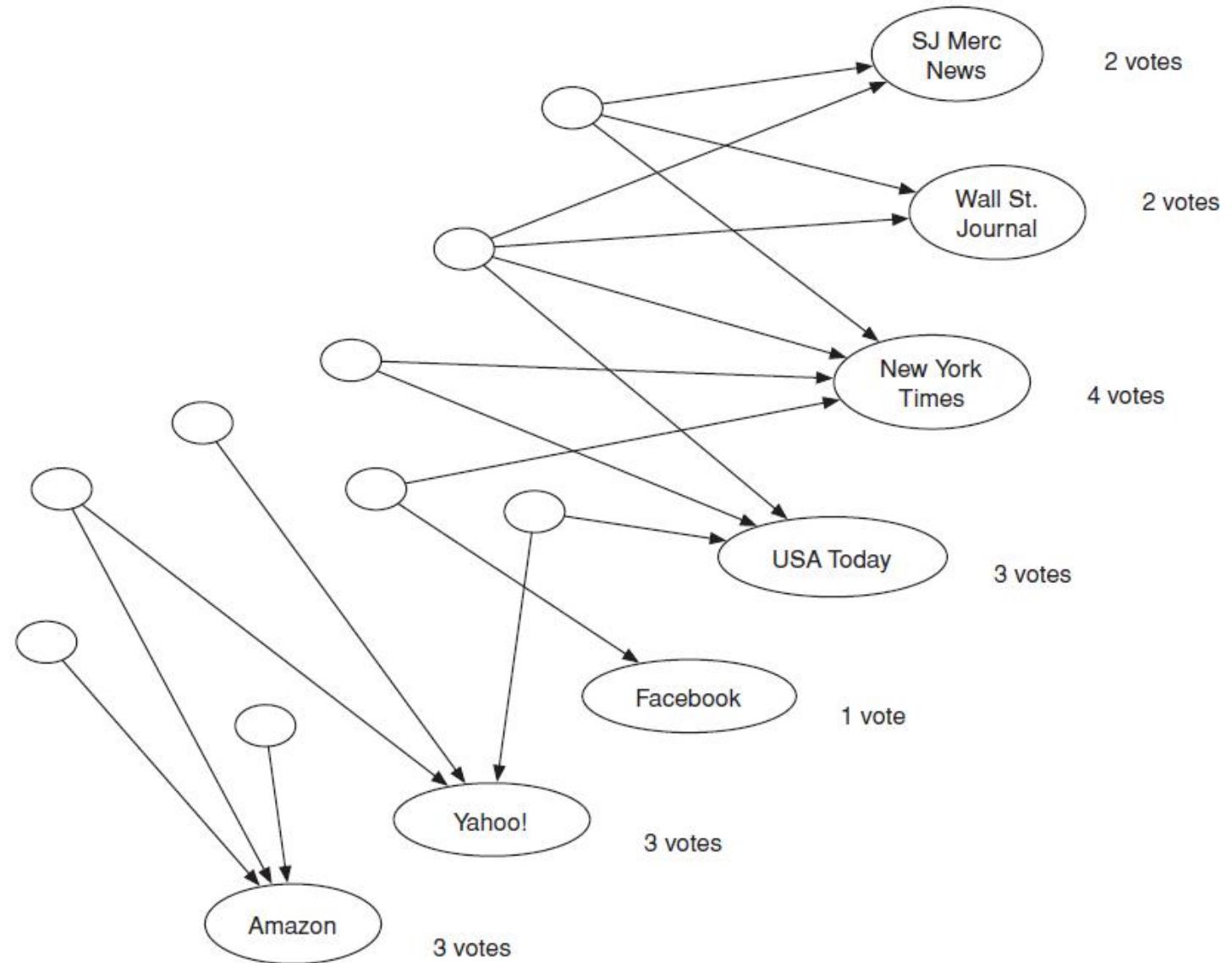
不能完全区分

完全区分开来

反复改进原理（例）

假设查询词“newspaper”

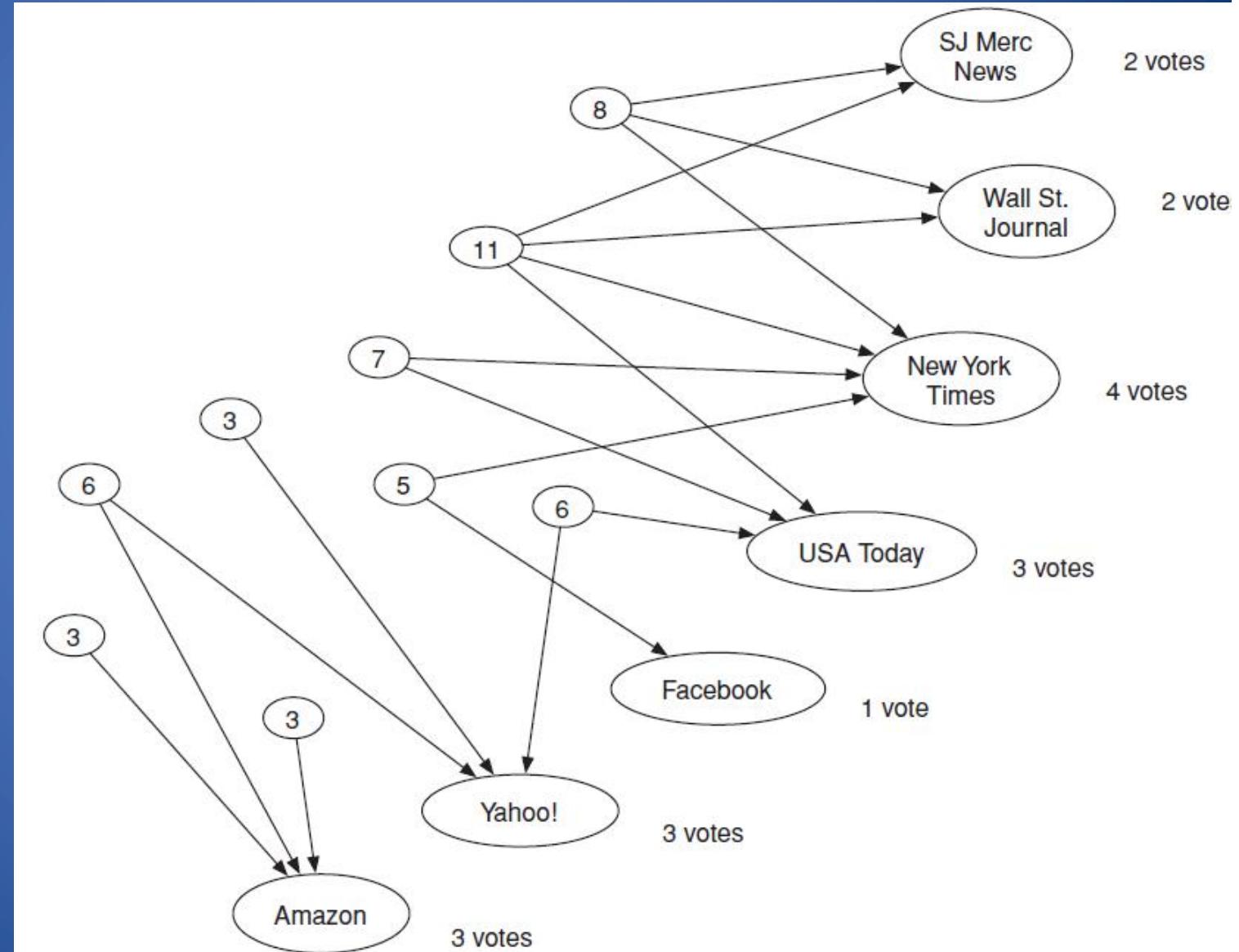
- 左边是与“newspaper”字面上相关的网页。
 - 右边是它们所指向的网页，得到的“票数”表示一定的认可度



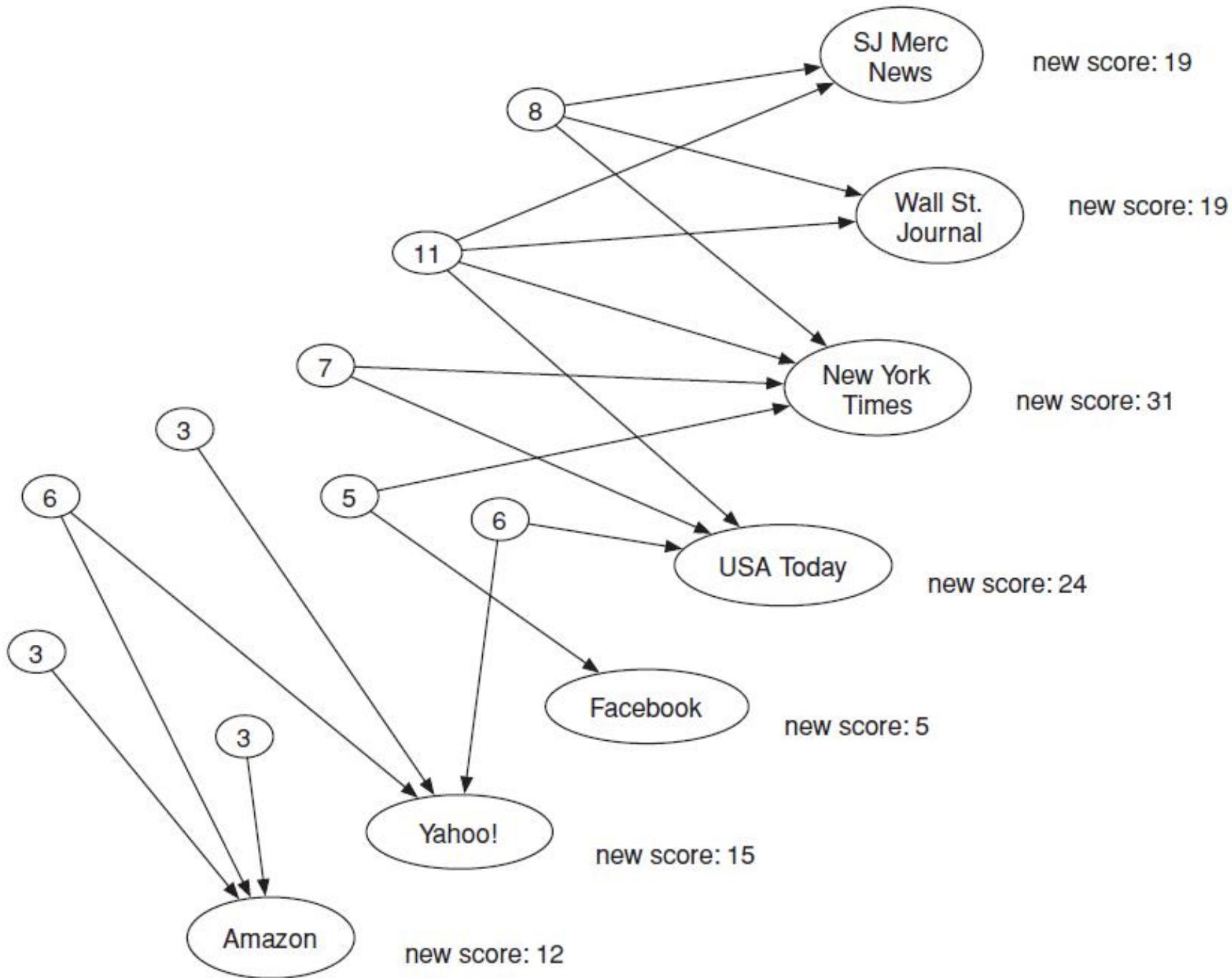
反复改进原理（续）

(principle of repeated improvement)

- 也可以反过来评估“推荐者”的份量
- 然后可以在考虑推荐者份量的情况下重新评估网站相对于“newspaper”的重要性



反复改进原理



“中枢”（hub）与“权威”（authority）

- 万维网中一篇网页的两面属性。观念：
 - 被很多网页指向：权威性高
 - 指向很多网页：中枢性强
- **HITS算法**：计算网页的权威值（auth）和中枢值（hub）
 - Hyperlink-Induced Topic Search
- 在实际中算法实施的针对性：相关网页集合
 - 不是全部网页集合。为避免赘述，这一点后面不再总强调，认为所讨论的就是相关网页集合。

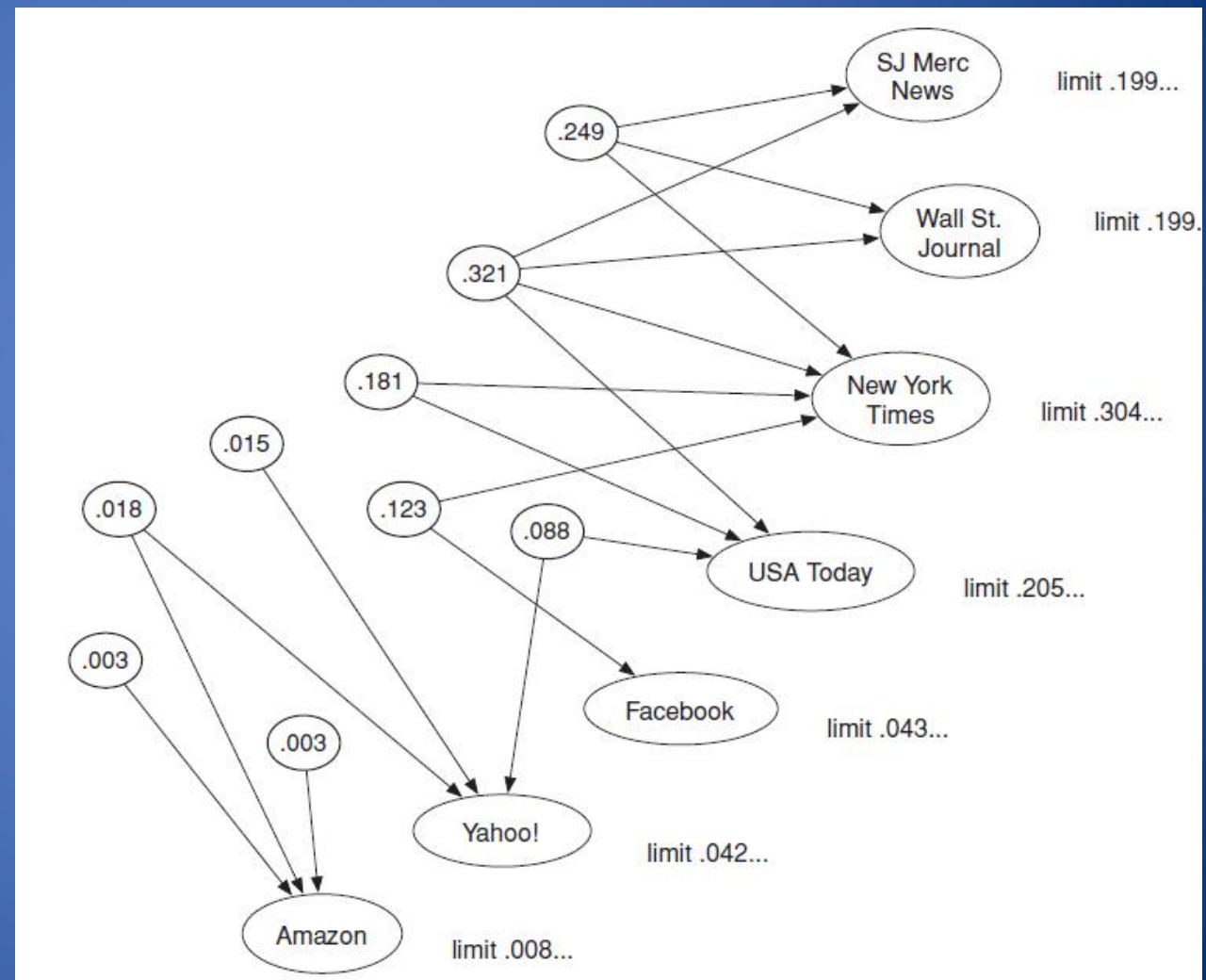
auth(p) 和 hub(p) 的计算方法

- 输入：一个有向图
- 初始化：对于每一个节点 p , $auth(p)=1$, $hub(p)=1$
- 利用中枢值更新权威值
 - 对于每一个节点 p , 让 $auth(p)$ 等于指向 p 的所有节点 q 的 $hub(q)$ 之和
- 利用权威值更新中枢值
 - 对于每一个节点 p , 让 $hub(p)$ 等于 p 指向的所有节点 q 的 $auth(q)$ 之和
- 重复上述两步若干 (k) 次

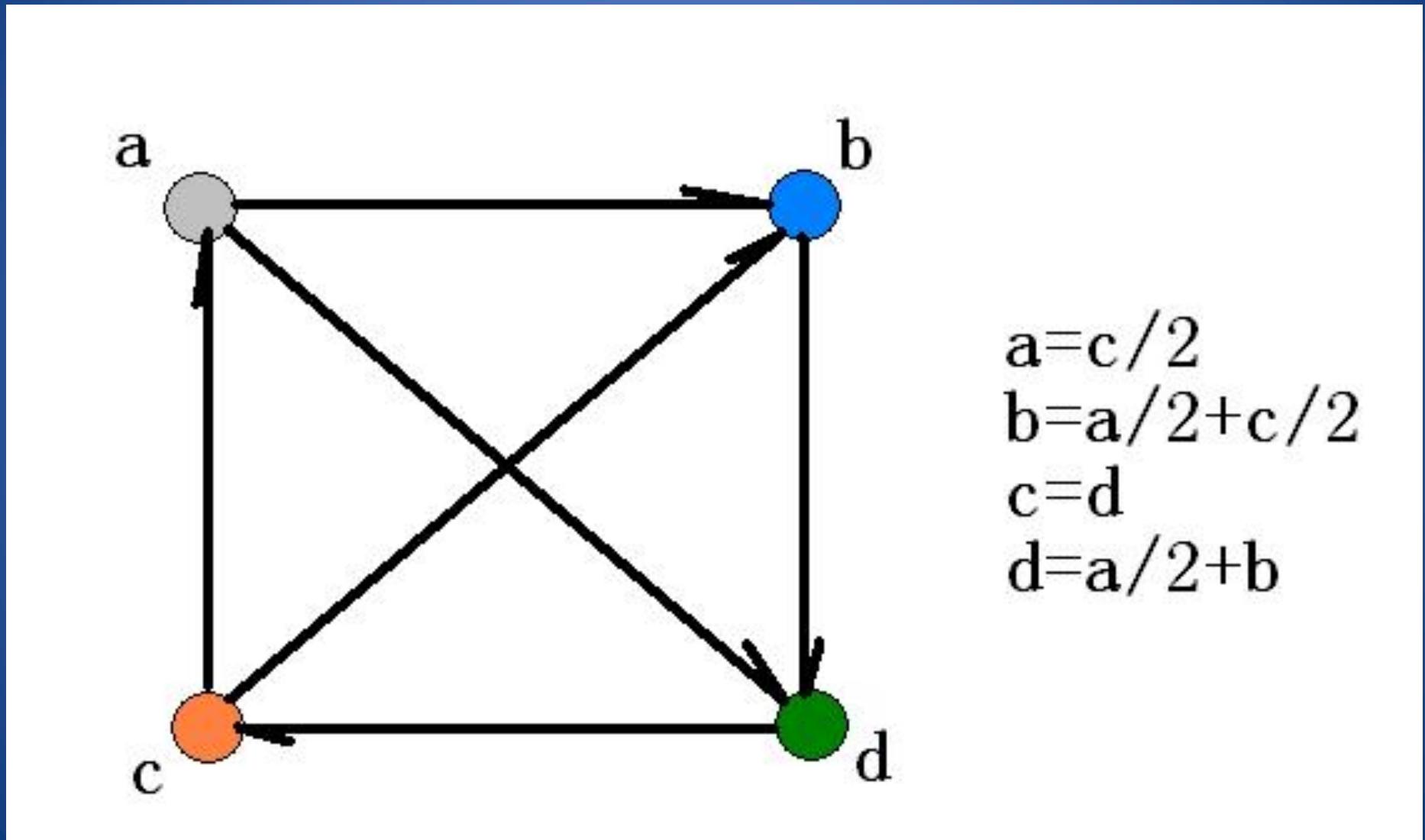
在搜索引擎领域, **auth**值或**hub**值高的网页, 有时分别称为“权威网页”和“中枢网页”。一篇网页可以兼具二者。

归一化与极限

- 数值随迭代次数递增
- Auth和hub值的意义在于相对大小
- 在每一轮结束后做归一化：值 / 总和
- 归一化结果随迭代次数趋于一个极限
 - 相继两次迭代的值不变
 - 极限与初值无关，即存在“均衡”



PageRank：节点的一种重要性测度

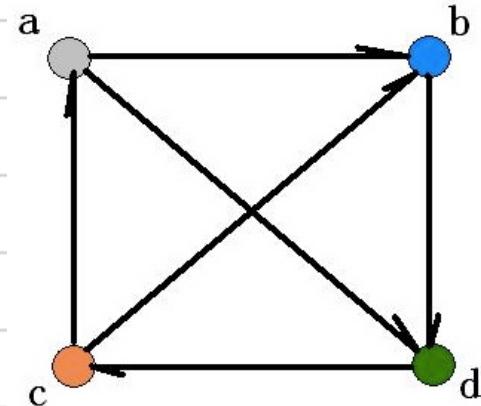


搜索引擎形成查询结果网页排序的重要参数

上图的算例

经过约70次迭代，最后收敛到：
 $A=0.615, B=0.923, C=D=1.231$

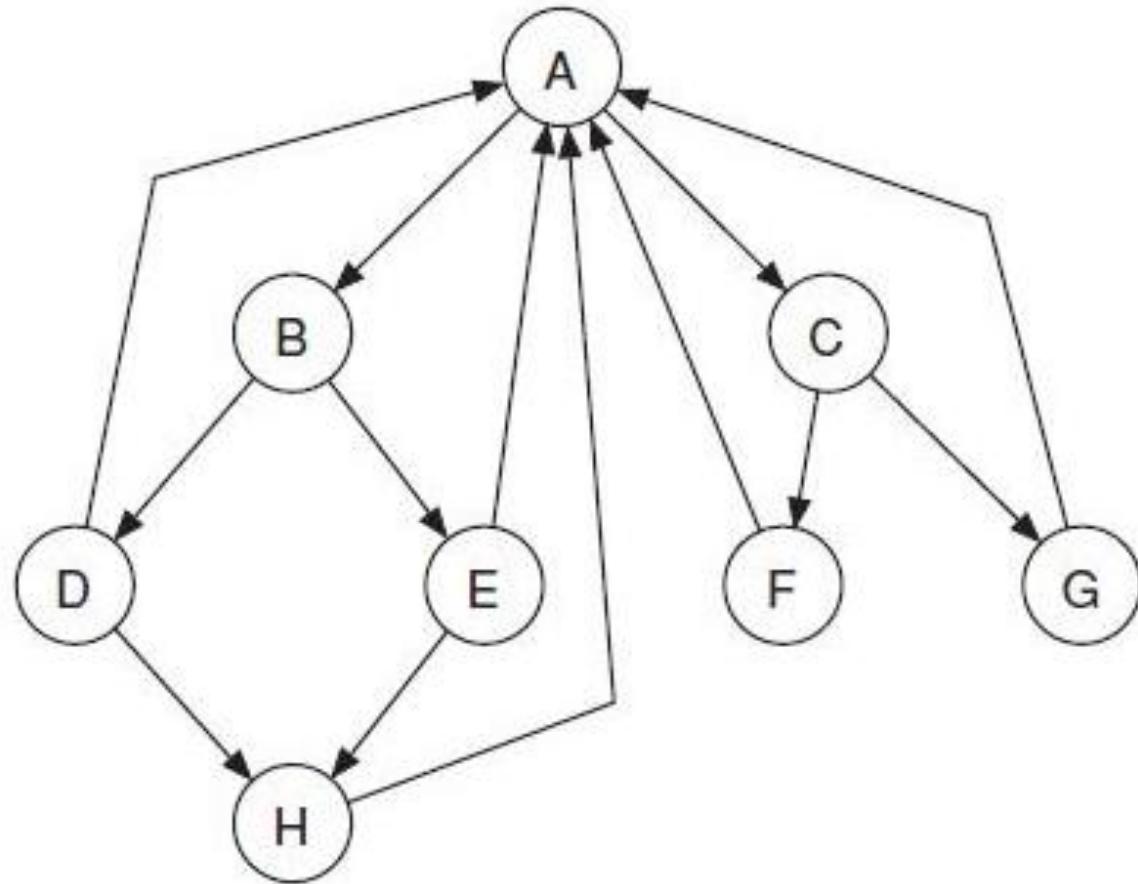
	A	B	C	D	E
1	$a=c/2$	$b=a/2+c/2$	$c=d$	$d=a/2+b$	传递关系
2	a	b	c	d	
3	1	1	1	1	初值
4	0.5	1	1	1.5	按传递关系计算
5	0.5	0.75	1.5	1.25	
6	0.75	1	1.25	1	
7	0.625	1	1	1.375	
8	0.5	0.8125	1.375	1.3125	
9	0.6875	0.9375	1.3125	1.0625	
10	0.65625	1	1.0625	1.28125	
11	0.53125	0.859375	1.28125	1.328125	
12	0.640625	0.90625	1.328125	1.125	
13	0.6640625	0.984375	1.125	1.2265625	



PageRank基本算法描述

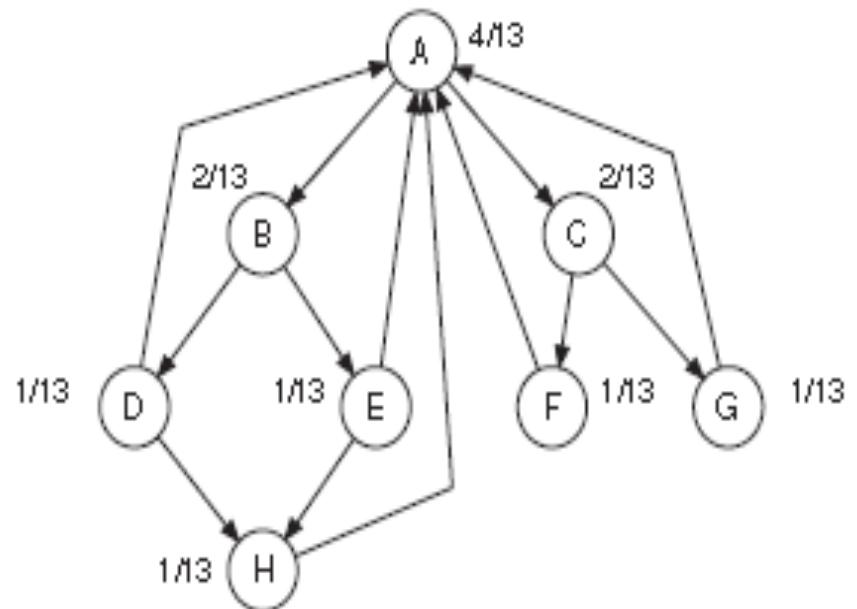
- 输入：一个有 n 个节点的网络（有向图），设所有节点的PageRank初始值为 $1/n$ 。
- 选择操作的步骤数 k
- 对PageRank做 k 次更新操作，每次使用以下规则：
 - 每个节点将自己当前的PageRank值通过出向链接均分传递给所指向的节点
 - 若没有出向链接，则认为传递给自己
 - 每个节点以从入向链接获得的（包括可能自传的）所有值之和更新它的PageRank

一个计算网页排名的实例



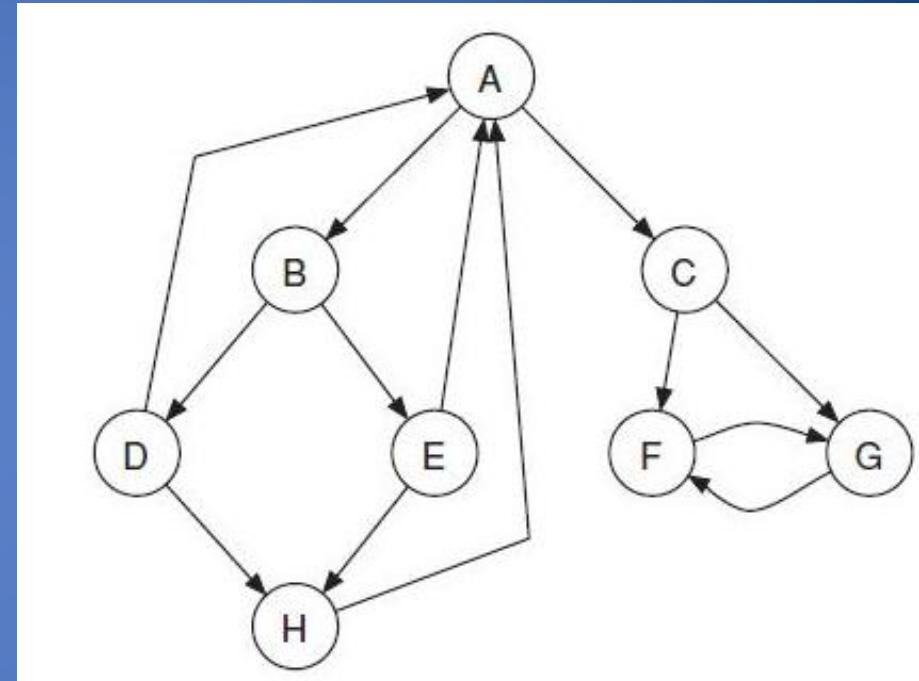
Step	A	B	C	D	E	F	G	H
1	$\frac{1}{2}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$
2	$\frac{3}{16}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$

- 每个节点的初值都是 $1/8$
- 最后收敛结果见下图



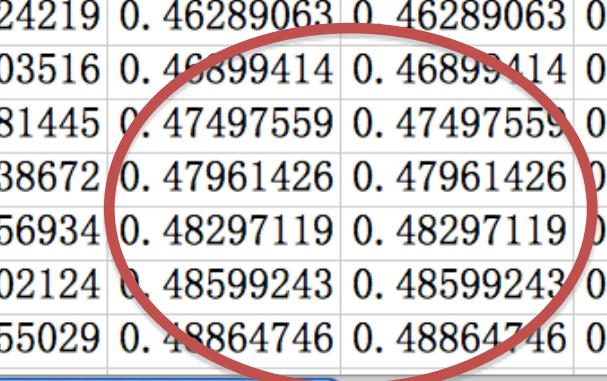
PageRank基本算法在某些网络结构上表现不好

- PageRank算法不象HITS 算法那样需要归一化问题，但有新问题
- F和G两个节点显得很“**自私**”：吸收别人的价 值，但不向外传
 - 导致它们最后各自 $1/2$ ， 其他人都 0



这也显示了共谋（**colluding**）制造垃圾网页的一个原理

PageRank值很快集中到F和G



	A	B	C	D	E	F	G	H
62	$a=d/2+e/2$	$b=a/2$	$c=a/2$	$d=b/2$	$e=b/2$	$f=c/2+g$	$g=c/2+f$	$h=d/2+e/2$
63	a	b	c	d	e	f	g	h
64	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
65	0.25	0.0625	0.0625	0.0625	0.0625	0.0625	0.1875	0.1875
66	0.1875	0.125	0.125	0.03125	0.03125	0.21875	0.21875	0.0625
67	0.09375	0.09375	0.09375	0.0625	0.0625	0.28125	0.28125	0.03125
68	0.09375	0.046875	0.046875	0.046875	0.046875	0.328125	0.328125	0.0625
69	0.109375	0.046875	0.046875	0.0234375	0.0234375	0.3515625	0.3515625	0.046875
70	0.0703125	0.0546875	0.0546875	0.0234375	0.0234375	0.375	0.375	0.0234375
71	0.046875	0.03515625	0.03515625	0.02734375	0.02734375	0.40234375	0.40234375	0.0234375
72	0.05078125	0.0234375	0.0234375	0.01757813	0.017578125	0.41992188	0.41992188	0.02734375
73	0.04492188	0.02539063	0.02539063	0.01171875	0.01171875	0.43164063	0.43164063	0.01757813
74	0.02929688	0.02246094	0.02246094	0.01269531	0.012695313	0.44433594	0.44433594	0.01171875
75	0.02441406	0.01464844	0.01464844	0.01123047	0.011230469	0.45556641	0.45556641	0.01269531
76	0.02392578	0.01220703	0.01220703	0.00732422	0.007324219	0.46289063	0.46289063	0.01123047
77	0.01855469	0.01196289	0.01196289	0.00610352	0.006103516	0.46899414	0.46899414	0.00732422
78	0.01342773	0.00927734	0.00927734	0.00598145	0.005981445	0.47497559	0.47497559	0.00610352
79	0.01208496	0.00671387	0.00671387	0.00463867	0.004638672	0.47961426	0.47961426	0.00598145
80	0.01062012	0.00604248	0.00604248	0.00335693	0.003356934	0.48297119	0.48297119	0.00463867
81	0.00799561	0.00531006	0.00531006	0.00302124	0.00302124	0.48599243	0.48599243	0.00335693
82	0.00637817	0.0039978	0.0039978	0.00265503	0.002655029	0.48864746	0.48864746	0.00302124

PageRank的同比缩减与统一补偿规则

- 同比缩减
 - 在每次运行基本PageRank更新规则后，将每一节点的PageRank值都乘以一个比例因子 s ， $0 < s < 1$ ，经验值在0.8-0.9之间。
- 统一补偿
 - 在每一节点的PageRank值上统一加上 $(1-s)/n$ 。

这样，既维持了所有PageRank之和等于1的性质，也防止了PageRank值不恰当地集中到个别节点。

pagerank.xlsx

G86

	A	B	C	D	E	F	G	H
62	a=(d/2+e/2+f/2)*0.8+0.2/8			取s=0.8	其他类推			
63	a	b	c	d	e	f	g	h
64	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
65	0.225	0.075	0.075	0.075	0.075	0.175	0.175	0.125
66	0.185	0.115	0.115	0.055	0.055	0.195	0.195	0.085
67	0.137	0.099	0.099	0.071	0.071	0.227	0.227	0.069
68	0.137	0.0798	0.0798	0.0646	0.0646	0.2462	0.2462	0.0818
69	0.14212	0.0798	0.0798	0.05692	0.05692	0.25388	0.25388	0.07668
70	0.13188	0.081848	0.081848	0.05692	0.05692	0.260024	0.260024	0.070536
71	0.1269648	0.077752	0.077752	0.0577392	0.0577392	0.2657584	0.2657584	0.070536
72	0.12762016	0.07578592	0.07578592	0.0561008	0.0561008	0.26870752	0.26870752	0.07119136
73	0.12683373	0.07604806	0.07604806	0.05531437	0.055314368	0.27028038	0.27028038	0.06988064
74	0.12515601	0.07573349	0.07573349	0.05541923	0.055419226	0.27164353	0.27164353	0.06925149
75	0.12473658	0.0750624	0.0750624	0.0552934	0.055293396	0.27260822	0.27260822	0.06933538
76	0.12470302	0.07489463	0.07489463	0.05502496	0.055024961	0.27311154	0.27311154	0.06923472
77	0.12440774	0.07488121	0.07488121	0.05495785	0.054957852	0.27344708	0.27344708	0.06901997
78	0.12418226	0.0747631	0.0747631	0.05495248	0.054952483	0.27371015	0.27371015	0.06896628
79	0.12413501	0.0746729	0.0746729	0.05490524	0.054905239	0.27387336	0.27387336	0.06896199
80	0.12409378	0.074654	0.074654	0.05486916	0.054869161	0.27396785	0.27396785	0.06892419
81	0.12403468	0.07463751	0.07463751	0.0548616	0.054861602	0.27403588	0.27403588	0.06889533
82	0.12400554	0.07461387	0.07461387	0.054855	0.054855005	0.27408371	0.27408371	0.06888928

随机游走：PageRank的另一种等价理解

- 想象一个人从一篇随机选择的网页开始，随机选择其中的链接浏览到下一篇网页，并不断如此进行，称为“随机游走”。
- 考虑一篇网页 x ，问：经过 k 步随机游走到达 x 的概率是多少？
- 可以证明：到达 x 的概率等于运行PageRank基本算法 k 步得到的值。
- 随机游走概念稍加修改也可以和同比缩减统一补偿的PageRank等价。

链接分析小结

- 有向图作为万维网的结构模型
 - 概貌形状：领结
 - 得到领结概貌形状的方法：宽度优先搜索
- 万维网信息节点的（结构性）重要程度
 - 权威性，中枢性：一个节点的双重作用
 - 反复改进原理（交叉支持原理）
 - 相对查询需求或者某个主题
 - **PageRank**（网页排名）：节点的全局相对重要性
 - 局部结构信息在全局扩散达到的均衡

作业

- 第14章 2,4

小世界现象

小世界现象

- Stanley Milgram, 1933-1984
 - The Small-world Problem, Psychology Today, 1967
- 基本实验
- 两点启示
 - “小世界”现象
 - “自我寻的”现象
- 体现小世界现象的模型
- 体现自我寻的现象的模型



- 模型在实践中的验证与推广

Stanley Milgram 的实验

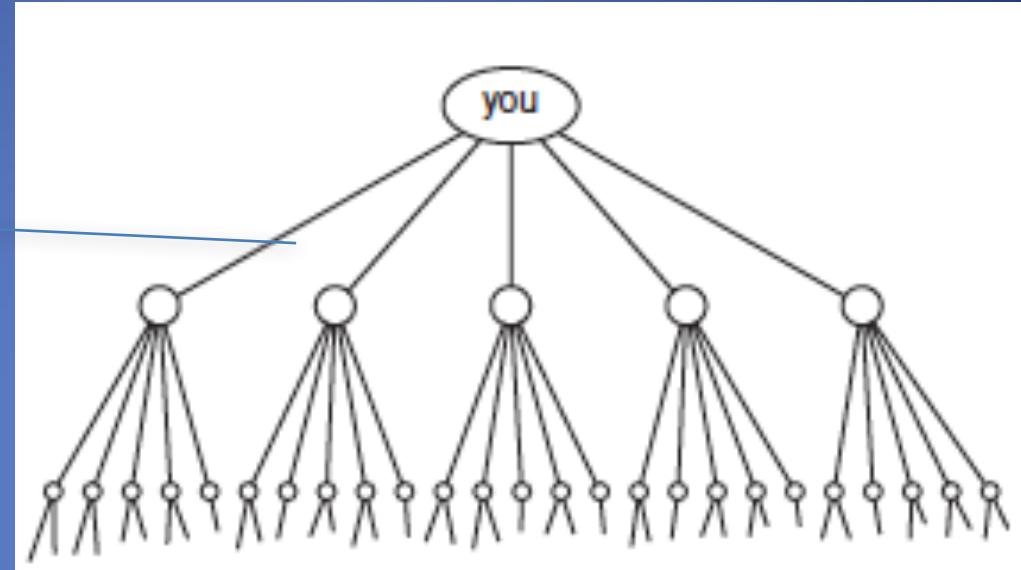
- 几百名“初始者”，要求每人尝试通过转发让一个指定的人收到一封信；
- 向每个初始者提供了目标收信人的姓名、地址、职业等个人信息；
- 规定：参与者只能将信件转发给能直呼其名的熟人，并请他继续转发。因此，如果一个参与者不认识目标收信人，则他不能直接将信寄给他；
- 要求参与者力争让这信件能尽早达到目的地；
- 结果，约三分之一的信件经过平均六次转发到达了目标

实验结果带给人们两点惊奇

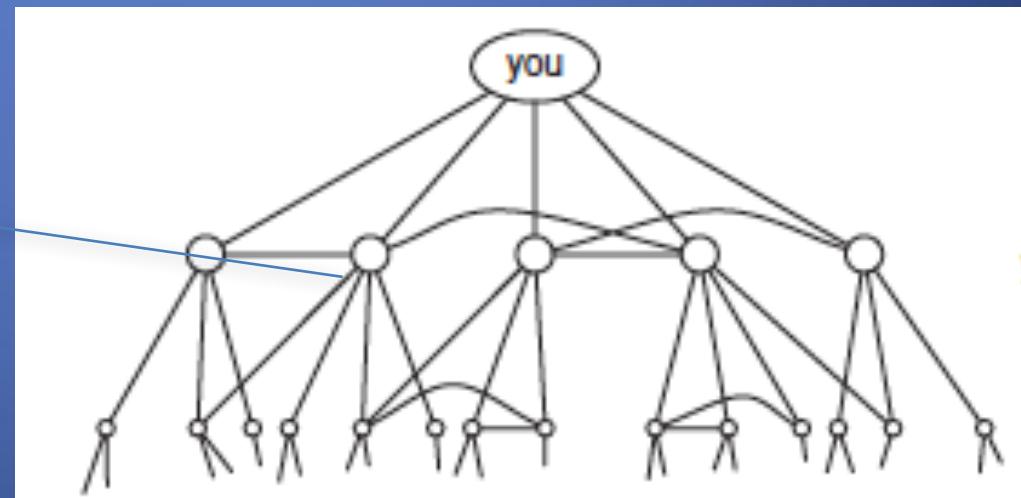
1. 社会网络中包含丰富的短路径
 2. 通过“有意识的转发”能够“自动地”找到这些短路径
- 问题
 - 为什么社会网络具有这样的性质？它们源于社会网络的哪些基本原理？
 - 反过来说，能否依据社会网络的某些基本原理，构建出反映这种性质的网络模型？
 - （完全随机的网络不具备这样的性质）

短路径的存在性初探

没有三元闭包的朋友关系。
可通过很短路径到达任何人。



体现三元闭包的朋友关系。
是否一定有短路径就不明显了。



形成社会网络的两种基本力量

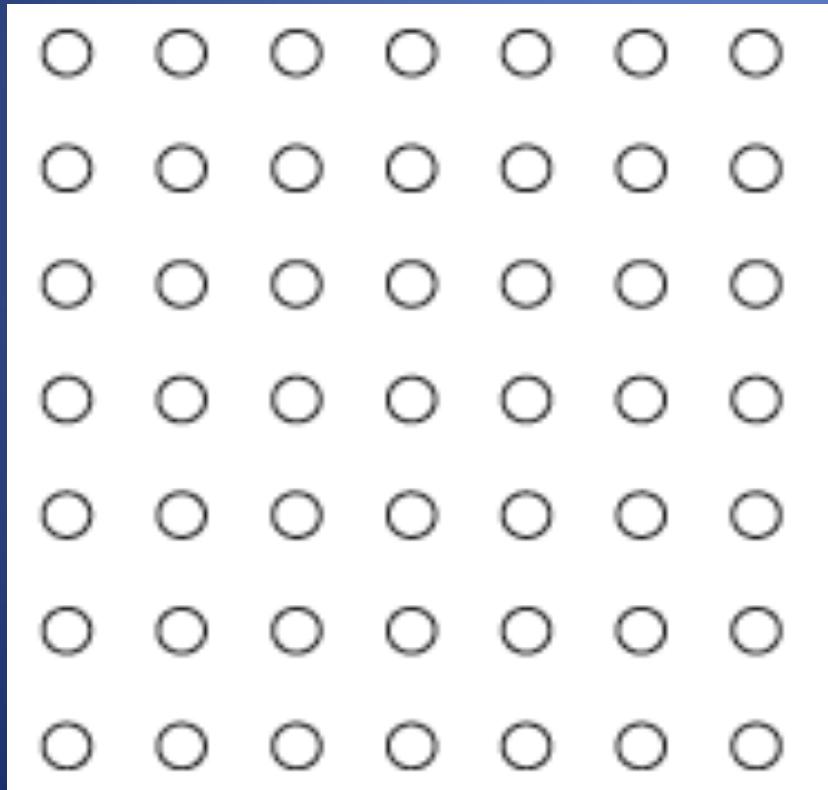
- 同质性（选择，社会影响），三元闭包
 - 家庭成员，邻里关系，同学，同事
 - 对应社会网络中的大量的“三角形”
 - 体现某种“亲近”（例如地理范围的）
- 弱联系
 - 偶然的原因，认识的“远程”朋友
 - 对其所在的圈子并不一定熟悉

一种什么样的形式化网络，既体现这两种力量的作用，也便于我们分析其中是否有小世界现象？

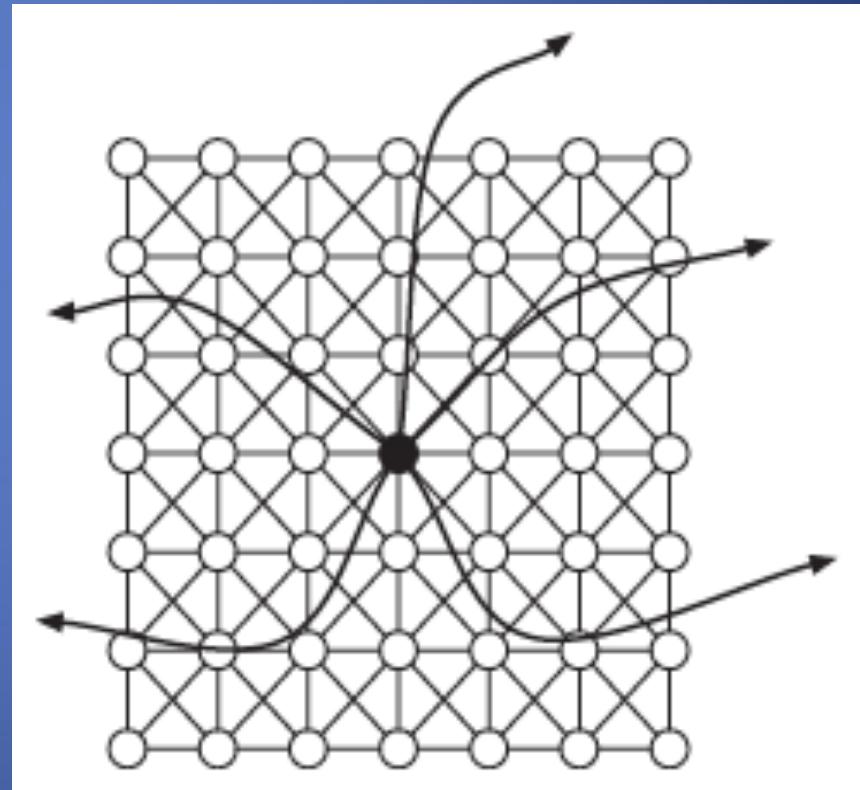
Watts-Strogatz模型 [1998]



- 定义一种图（网络），它体现这两种力量
 - 存在许多“三角形” 和少数随机的“远程边”



想象**大量**节点排布成均匀网格
状

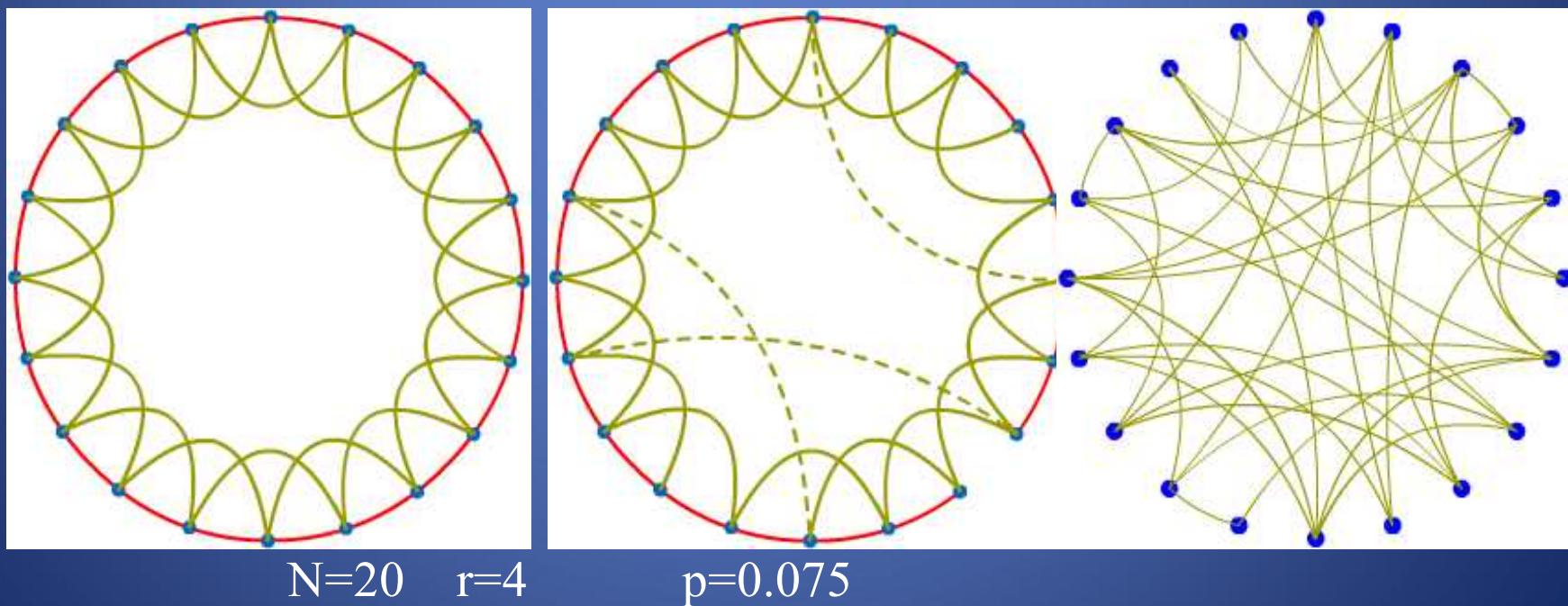


连接近邻:确定性，连接远程:随机性

WS小世界网络模型（其他论文提出）

小世界网络指聚类系数较大且平均路径长度较小的网络 Watts-Strogatz(W-S) 模型

W-S模型的构建： (1)从一个有N个节点的一维规则圆环开始，每一个点与它的r个最近邻居连结。 (2)以概率P按顺时针方向对规则圆环中每一条边进行重新连结。重连生成k条边。在重连过程中，不能进行自我连结和重复连结。

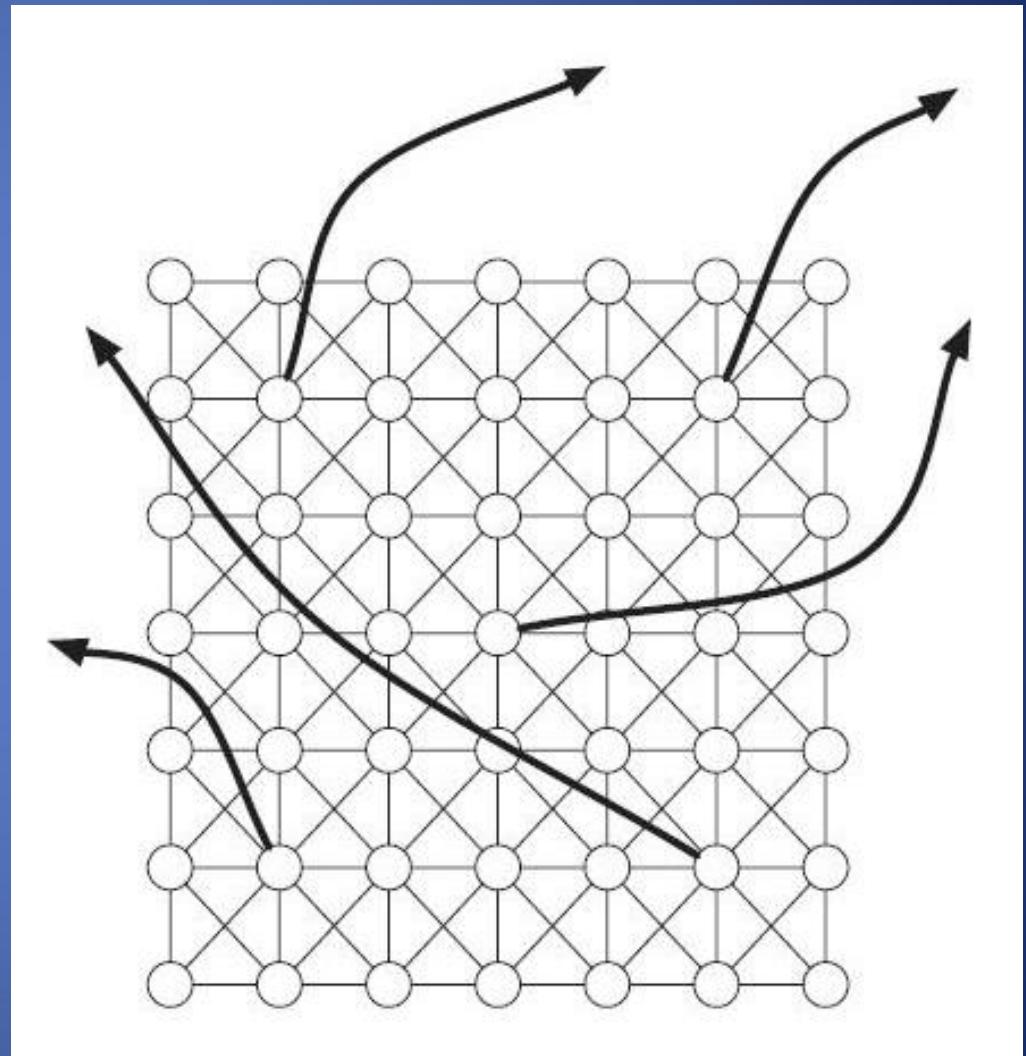


Watts-Strogatz模型 (r,k)

- 体现了同质连接和弱关系连接的概念（利用 r 和 k 控制相关的强度），于是可以看成是现实社会网络的一个合理近似
- Watts-Strogatz 证明了：在这样的网络中，任意两点之间存在短路径的概率很高
 - 基本思路：弱连接的随机性使其形成闭包的概率很小
- Ballobas 和 Chung 进一步从数学上确定了典型路径的长度

古代的社会网络是不是也是小世界？

- 那时许多人实际上没有“远程联系”
- 回答是肯定的
 - 设 $k \times k$ 的子网格中只有 k 个远程随机连接
 - 将这子网格看成是一个节点就回到了最初的Watts-Strogatz模型形式
 - 所谓“短路径”此时只不过增加了在子网格中的长度($2k$)



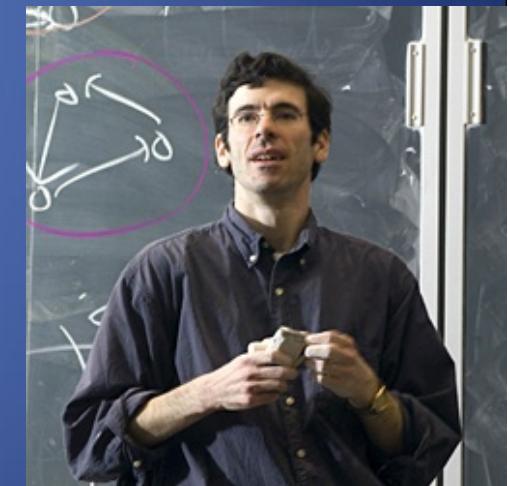
亦即，只要在基础网络结构上存在很少的全局随机性，就可以看到小世界现象

分散搜索 (decentralized search)

- 如果当年 Milgram 的实验经费很充足
 - 让每个参与者可以将信件复印任意多份，寄给所有能直呼其名的朋友 (flooding, BFS)
- 就不会给我们带来第二个惊奇了：那些信怎么就能真的到达目的地？
 - 虽然人们是有意识地希望信能送到，但她若不认识目标人（他），就只能“估计”她的哪个朋友可能“更接近于”认识他。
 - 因此，我们没有理由认为信件会以很高的概率送到，更没有理由认为它会走一条短路径

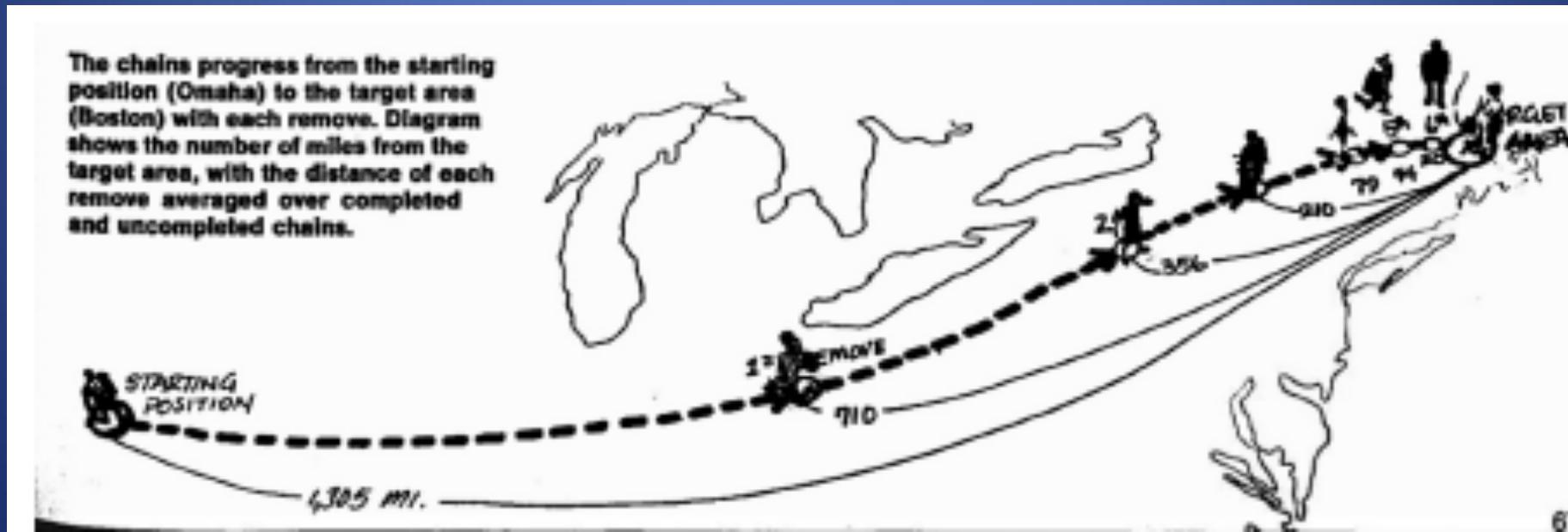
探索一种社会网络模型

- 既反映任何节点对之间短路径的存在性，也反映这种转发方式达到目的的可能性
- 可以证明，Watts-Strogatz模型不能很好地体现第二个要求（搜索路径太长）
- 重新构造一个模型？
- 扩展Watts-Strogatz模型？
 - Jon Kleinberg, Navigation in a small world. Nature, 2000



思路

- 为什么Watts-Strogatz模型不奏效?
 - 体现弱关系的边太过于随机，不支持现实中人们会有意识向比较接近目标的朋友转发信件的行为的表达



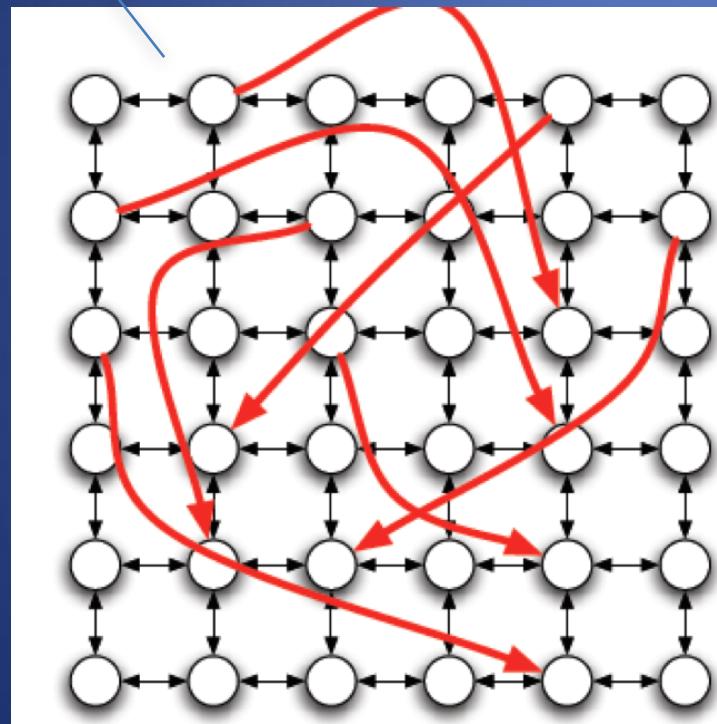
每经过一步就向目标更接近

扩展的Watts-Strogatz网络模型

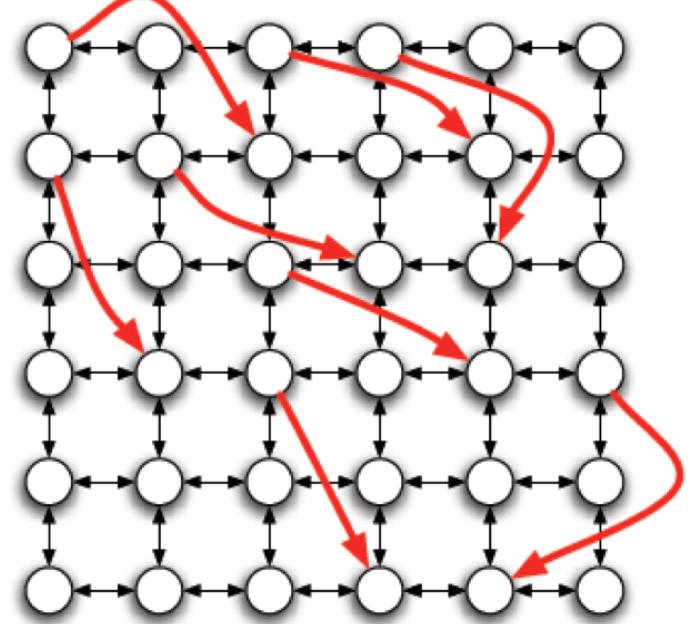
- 节点依然在网格点上排布
- 同质性局部连接相同：每个节点与在 r 个网格步内的节点都有直接的边
- 对随机远程弱连接加以一定的控制：让两个节点之间连接的概率与它们的网格距离的幂次成反比关系（即随距离增加而衰减）
 - 记 $d(v,w)$ 为 v 到 w 的距离（网格步数），则产生一条从 v 到 w 的随机边的概率与 $d(v,w)^{-q}$ 成正比。
- 理论上可以证明，对于适当的 q ，在此形成的网络中的分散搜索有很高的效率（平均步数）
- “适当的 q ” 是多少？

不同q值对随机连接长度的影响

q值较小，随机边倾向于较远



q值较大，随机边倾向于较近

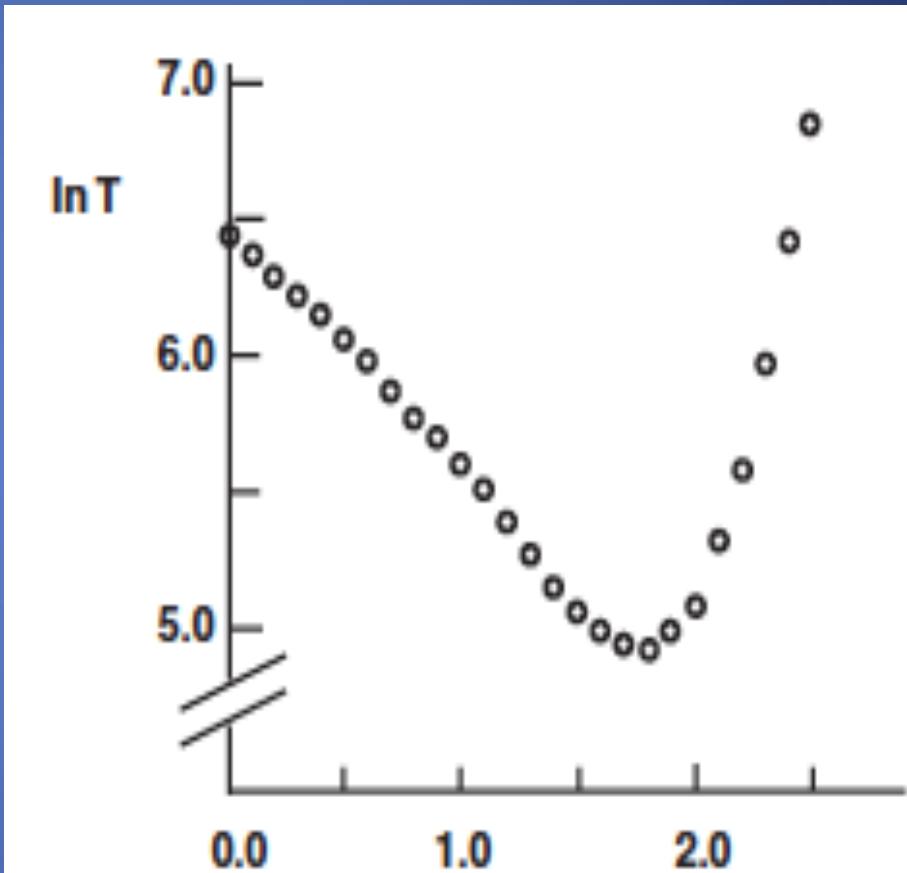


$$p \propto \frac{1}{d^q}$$

- Watts-Strogatz模型对应于 $q=0$

该模型的最佳工作参数

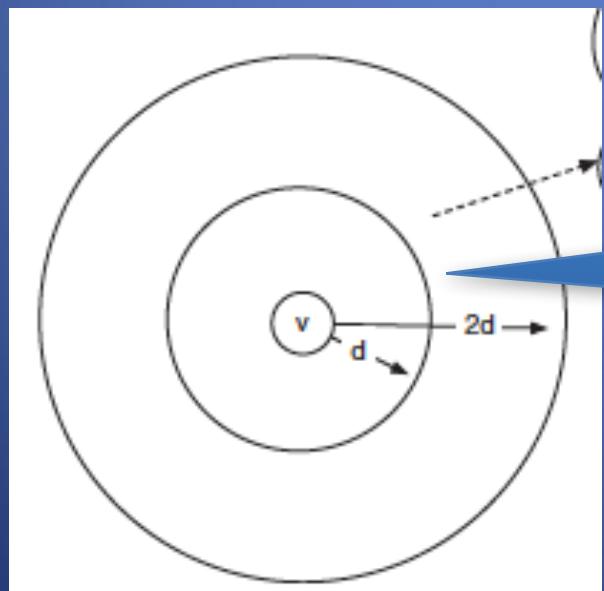
- 理论结果：当 $q=2$ 时，分散搜索达到最佳效果
- 仿真实验：由几个节点组成的网络中，针对不同的 q 值而实施的分散搜索
 - 对于这种规模的网络，在指数 q 介于1.5和2.0之间时搜索效果最佳
 - 随着网络规模的扩大，最佳的性能发生在指数 q 越来越接近2



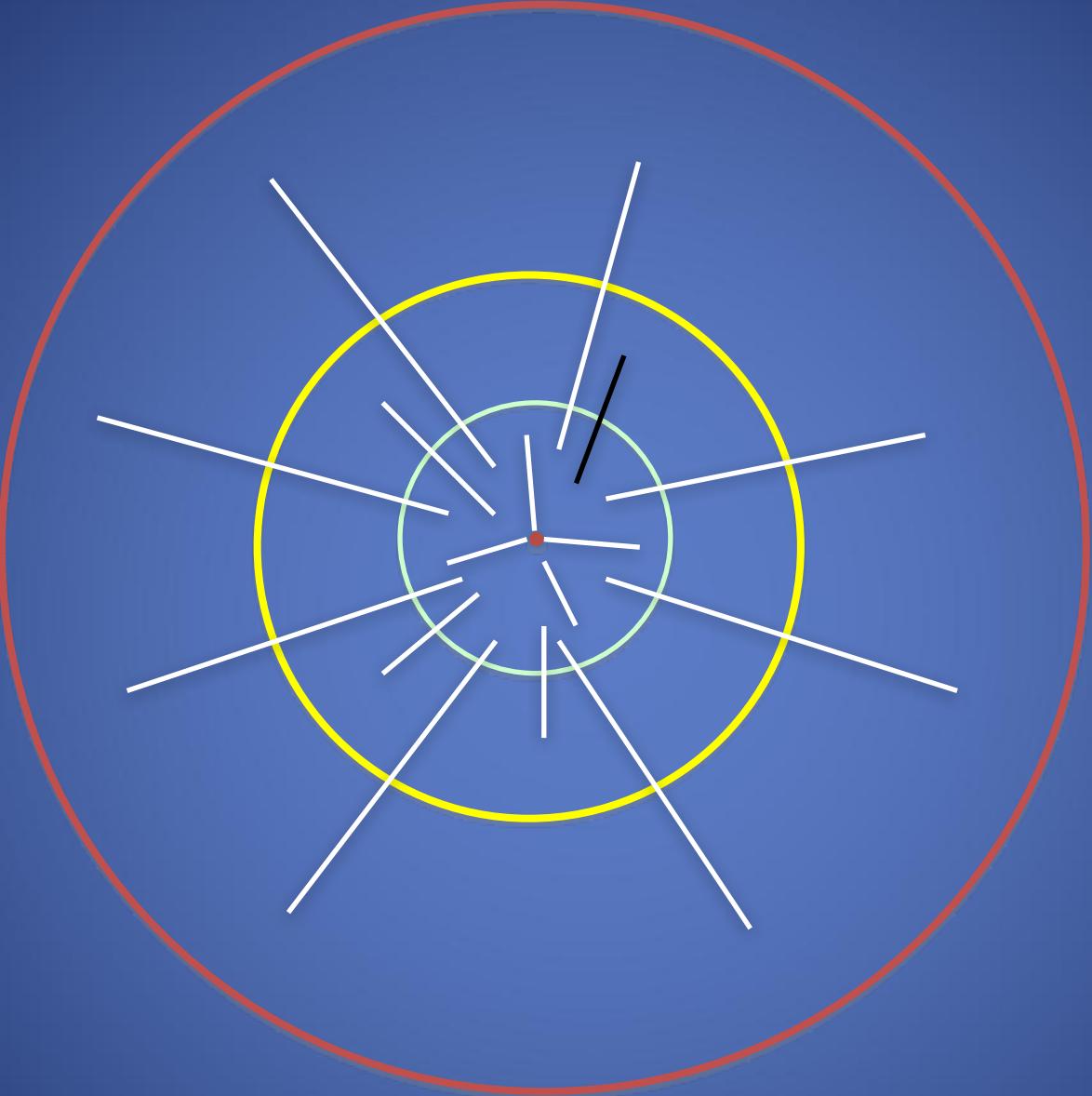
横轴为参数 q ，纵轴为从一个节点到达另一个节点所需的平均时间（跳步）

d^{-q} : 为什么 $q = 2$ 时效果最佳?

- 考虑节点 v 随机连接到距离在 d 到 $2d$ 环内的概率
 - 节点数与 d^2 成正比 (均匀分布, 节点数与面积成比例)
 - 随机连接到其中一个节点的概率与 d^{-2} 成比例
- 因此从 v 发出的随机连接落到该区域的概率与 d 无关
 - 意味着 (1) 无论一个转发节点与目标节点相距多远, 他都很可能有一个离目标距离近一半的朋友; (2) 信件离目标越近, 与目标有边的可能性就越大



节点数与 d^2 成正比,
连接到每个节点的概
率与 d^{-2} 成正比



模型的验证与推广

- 大规模在线社会网络是否（如何）体现了这个（Watts-Strogatz-Kleinberg）网络的特点？
 - 如果没有体现，则可能说明这个网络作为社会网络的模型有问题
- 什么是需要体现的特点？
 - 从地理上看，一个节点在任一距离上的朋友数量在同等距离上网络节点总数中的占比（对应于概率）随距离按反平方递减
- 在线社会网络如何体现地理位置关系？
 - 同一“距离”所“包含”的节点数相同

LiveJournal (LJ) , 1999年建立的社交网站

LiveJournal.com - Start a Free Blog / Journal Today

http://www.livejournal.com/ Google

Apple中国 中国雅虎 维基百科 新闻 热门

LiveJournal.com - Start a Free Blo... ☀

LIVEJOURNAL™

Home Create an account Explore Shop LJ Extras Games

用户名: 开设一个账户
密码: 登录
 记住我 [f](#) [t](#) [d](#)

English • Español • Deutsch • Русский...

Writer's Block

Frozen delights

What is your favorite ice cream flavor?

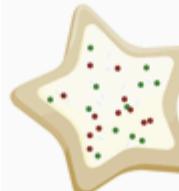
[Answer](#)

Nov, 29(1/2)

[» View Answers](#)

LiveJournal Shop

 Snow Globe \$0.99

 \$0.99

 Colorful Lights \$0.99

Sign up for LiveJournal now: It's fast, fun, and free!

Spotlight

 **anythingdisney**
Miley Cyrus, Selena Gomez, the Jonas Brothers, and the current crop of Disney Channel stars... Love 'em or hate 'em, you can still get in on all their latest news and gossip!

 **aramatheydidnt**
Take one part **ohnotheydidnt** and one part j-pop and what do you get? **aramatheydidnt** is the premiere community to fulfill all your japanese pop culture needs.

 **vaginapagina**
(vaginapagina.com) **vaginapagina** is an online community that offers a supportive, progressive, body- and sex-positive environment in which to discuss issues related to female sexual and reproductive health and wellness. It is a unique empowerment based safe space that is GLBTQIA+ and kink friendly. And boy, there really is no

About LiveJournal

44.3 million journals and communities

130.2 thousand posts in the last 24 hours

Open Minded and Open Source.
Volunteer supported by people who love LJ.

Read about our latest features in [News](#).

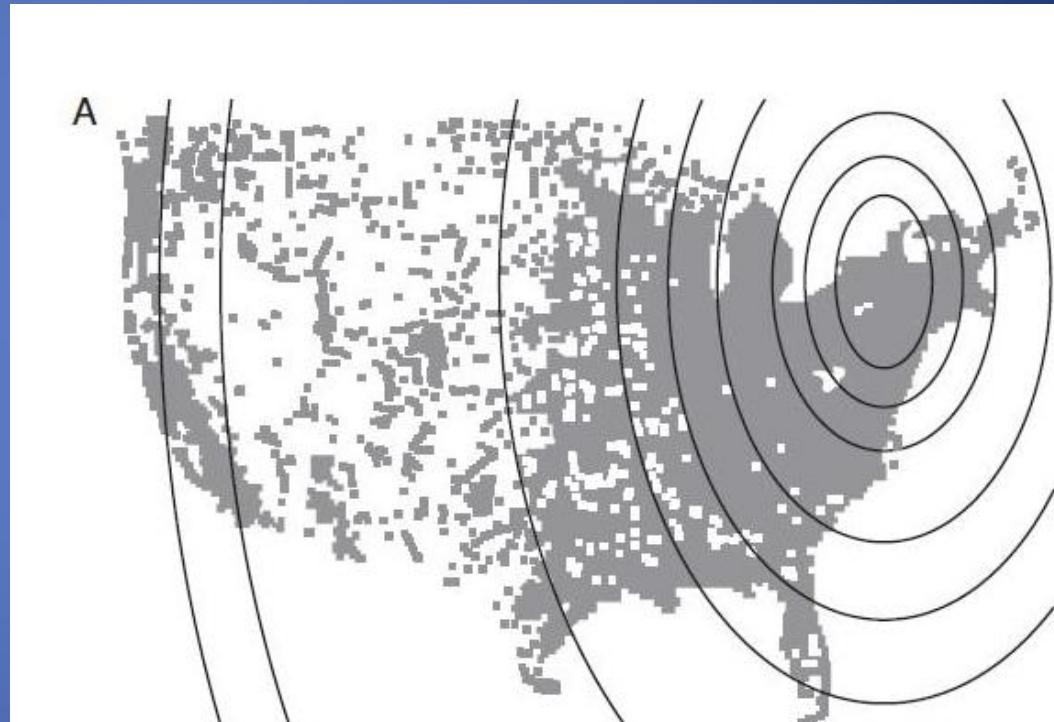
Creative Expression

 **ashleybouvier** for **lj_photophile**

正在载入“<http://www.livejournal.com/>”, 已完成 89 项 (共 93 项, 1 个错误)

来自LiveJournal的实验数据

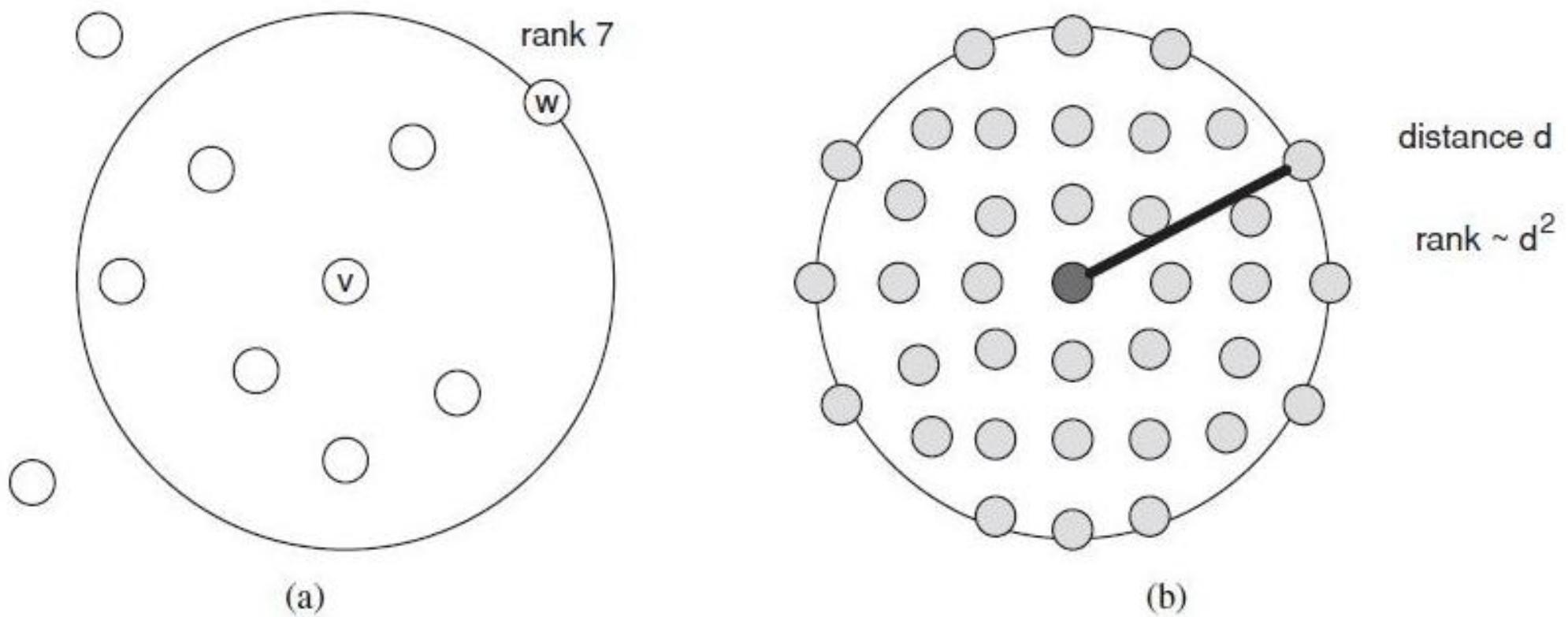
- 50万用户，含邮政编码信息（地理信息）
- 但他们是不均匀分布的，不符合模型的假设，需要做一些“适配性”工作
- 依据地理信息定义一个节点相对于另一个节点的**排名**
 - 节点 w 在节点 v 眼里的排名， $\text{rank}(w)$ 等于网络中**比 w 离 v 近**的节点的个数。



LiveJournal中用户的地理位置分布

同一排名含同样多节点，统一了不同密度的区域

社会网络中结合地理位置信息的节点排名

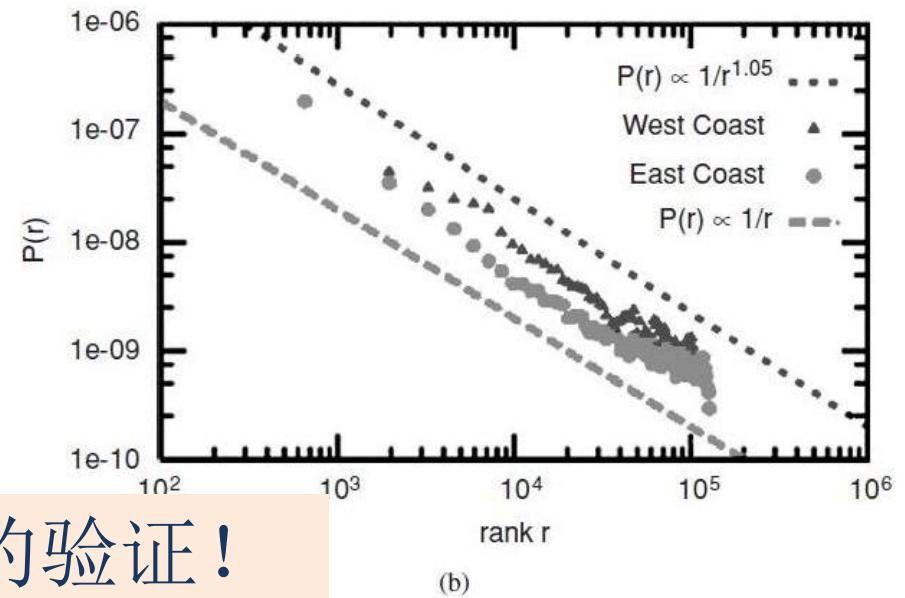
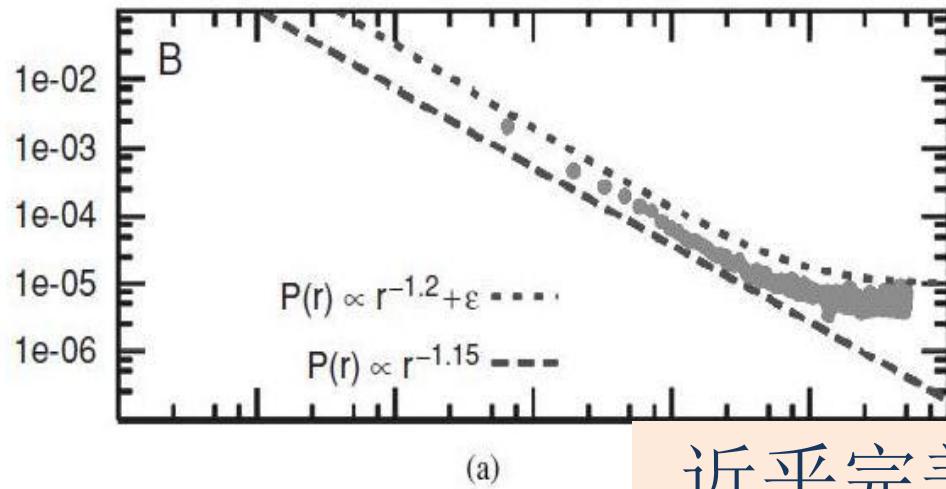


- 可以看成是节点在地理上均匀分布时区域范围的一种推广， “排名”与“距离”有对应关系
- 这就使我们能一般性地处理节点分布不均匀的问题了

这样

- 要验证的是
 - 从地理上看，一个节点在任一距离上的朋友数量在同等距离节点总数中的占比随距离按反平方递减 ($1/d^2$)
- 此时等价于看

- 一个节点在任一排名上的朋友（即有连接）数量在同等排名节点总数中的占比随排名成反比 ($1/r$)



近乎完美的验证！

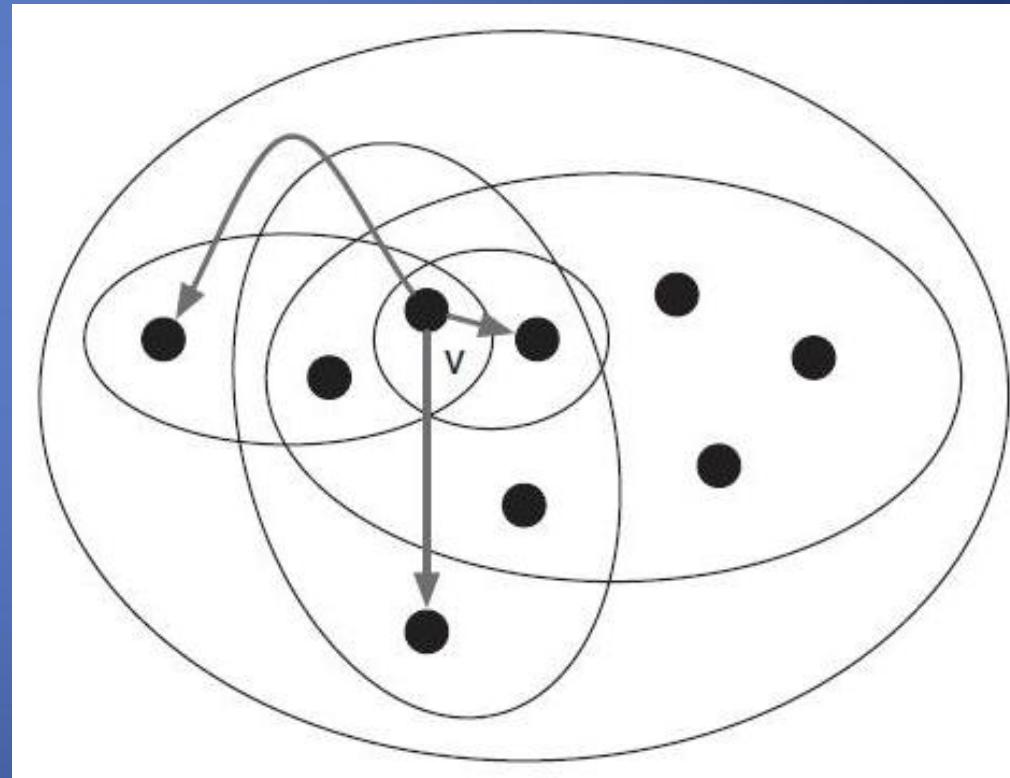
真实社会网络的测量参数与模型最优参数相当吻合！

进一步考察网络模型与社会网络的关系

- Watts-Strogatz-Kleinbeig网络模型试图反映的社会网络特征
 - 由某种“亲近”（相似性）形成的近距离边
 - 随机的远距离弱联系边
 - 远程相邻节点的数量在同距离节点数中的占比随距离的平方减少
- 距离，是其中一个关键概念
 - 地理空间位置的关系体现最直接的距离概念
 - 节点的相对排名也体现了一种距离概念，克服了处理节点在地理空间上分布不均匀情形的困难
 - 在社会网络中，还有什么有意义的距离概念？

社团 (social foci, 社交聚点) 与距离

- 一个人可能参加多种社团 (组织机构, 兴趣爱好群体等), 社团是两人建立关系的一个基础 (社团闭包)
- 可以想象, 两人的亲近程度 (“距离”) 与社团的规模有关, 越小越近
- 定义 “社会距离”: 两人同属最小社团的规模。

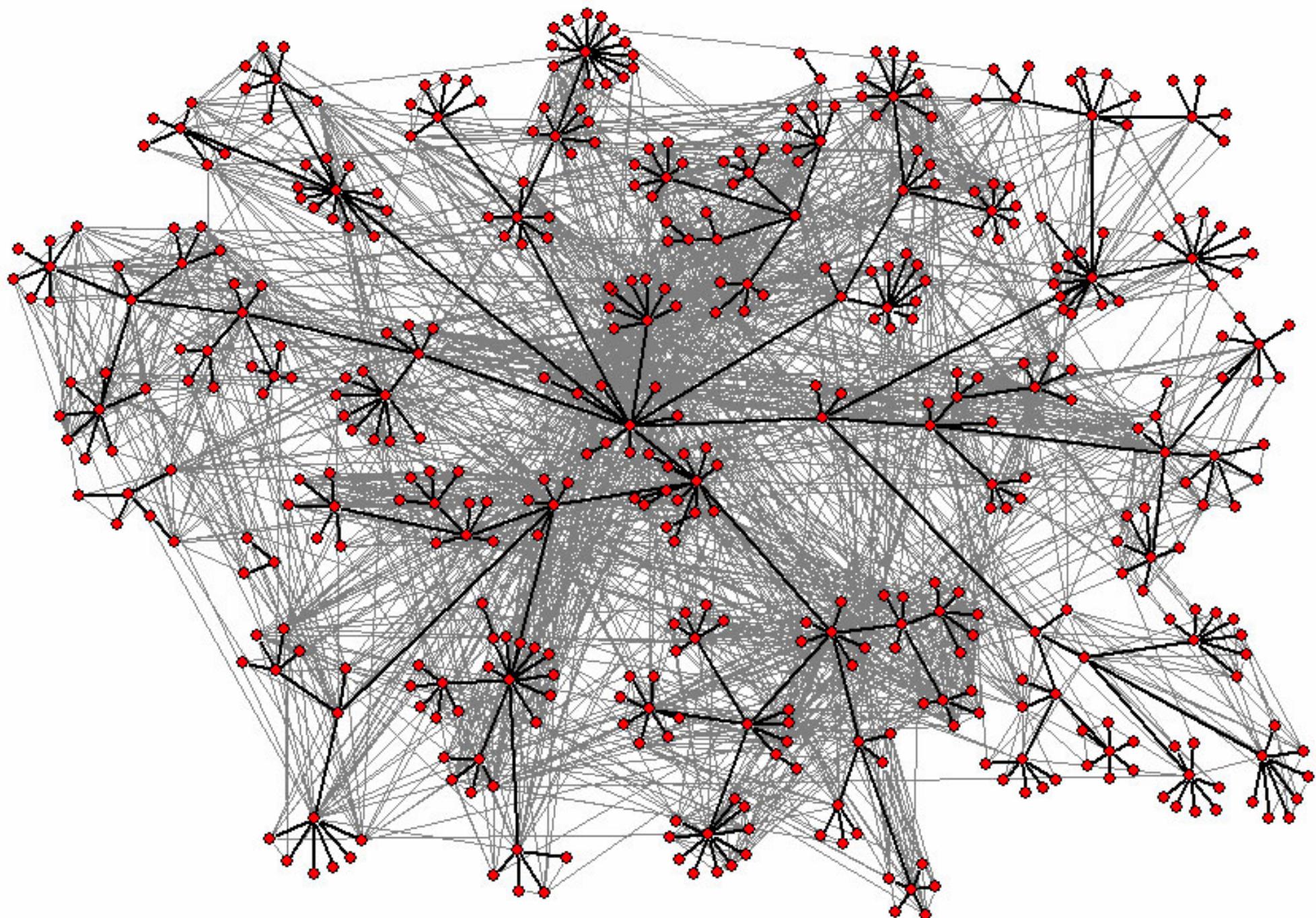


v属于5个大小分别为2, 3, 5, 7, 9的社团

进一步

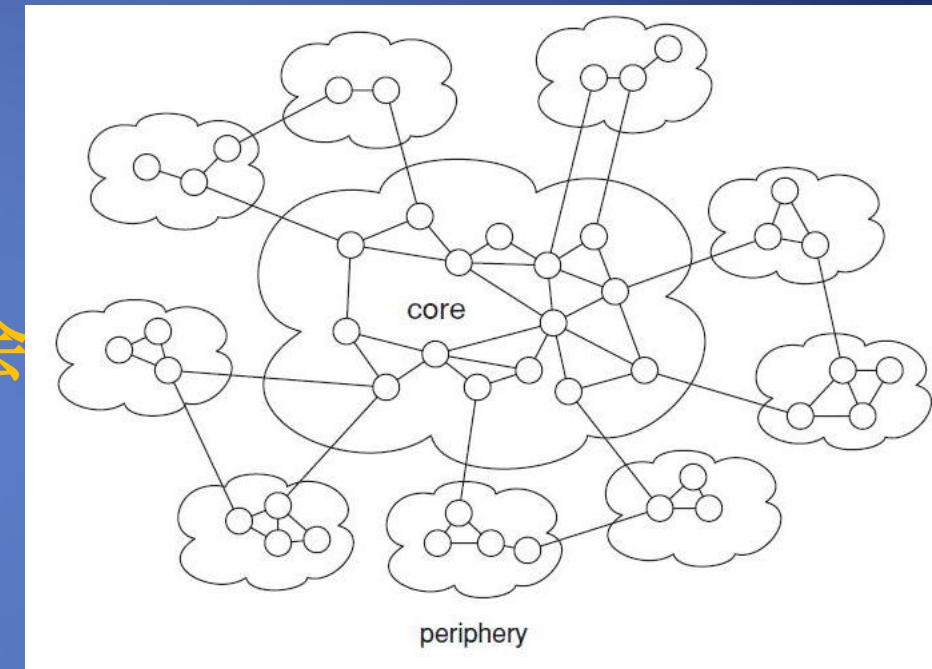
- 如同排名，可以建立（现实社会网络）中的社会距离与Watts-Strogatz-Kleinbeig网络空间距离之间的对应关系
- 从而我们可以来考察以社会距离度量人们之间亲近关系的真实社会网络中是否也体现了Watts-Strogatz-Kleinbeig网络的第三条性质
- Adamic和Adar以惠普实验室人员的email数据为对象进行的分析表明，该性质体现明显：社会距离为 d 的员工之间产生连接的概率与 $d^{-3/4}$ 成比例
 - 节点：员工；边：一段时间内交换过邮件
 - 社团：惠普实验室的管理层次结构

Email联系与组织机构层次的叠加分析



社会网络的“核心—外围”结构

- 大量类似于Milgram的实验表明，找到路径不容易
 - 人们的参与度是原因之一
- 发现相对容易找到通向社会地位状态高的人的路径
 - 不仅是人们可能更加愿意参与找通向他们的路径
- 还有社会网络的结构性原因！（他们相对容易找到）
 - 同质性：人们大致上按照类似社会地位状态相关联
 - 社会地位高的人倾向于有更多、更广的连接（关系）
- 这就促成了所谓“核心—外围”结构（研究前沿）



本章要点

- 围绕小世界问题体现出的研究范式
 - 实验现象 → 理论模型 + 完善 → 实际验证与推广
(这可能需要几十年的时间！)

“图20.10以及随后的相关结果标志着典型科学研究中心一系列环节的终结，(1) 实验观察 (Milgram) , (2) 建立基于观察到的现象的数学模型, (3) 基于这个模型进行预测 (最优参数) , (4) 根据实际数据验证这个预测 (模型概念的“适应性处理”)。这就是人们所希望看到的实验、理论、测量相结合的研究方法论的实践。”

作业

- 第20章 1,3题

反转课堂

- 任选第7、8、9、15、22、23、24章节
- 自行制作PPT
- 预计12、13周开始
- 最多2人一个小组，每人限参加1次
- 根据课堂讲解情况，给期末成绩直接加分（最高10分。若小组为2人，则平分所得分数）

幂率与富者更富现象 及其与长尾、齐普夫定律等的关系

流行度 (popularity)

- 同一类事物的不同实例被关注（认知、偏爱）的程度
 - 人（明星），书籍，歌曲，某一类产品（例如软饮料），某一类服务（例如提供同一种服务的网站），微博主
- 为什么会有差别？
- 这种差别有没有什么规律？
- 有没有办法增进某些实例在这种差别中的优势？

以Web上网页得到的链接数量为例

- 得到一个链接，意味着得到某种“认可”。于是可认为得到的链接越多，流行度越高
- 问
 - 给定一个国家（地区）的网页集合 (S)，其中一个网页的入向链接数为 k 的概率 $f(k)$ 是多少？
 - 这个概率函数是否反映了一种规律，即普适于任何大规模搜集的网页集合？
 - 如果这是反映网页集合的一种规律，为什么会有这规律？它是否也适合其他具有流行性的事物？

先看如何回答第一个问题

- 给定一个国家（地区）的网页集合 (S)，发现其中一个网页的入向链接数为 k 的概率 $f(k)$ 是多少？

$$S = \{x_1^{(p_1)}, x_2^{(p_2)}, \dots, x_i^{(p_i)}, \dots, x_n^{(p_n)}\}$$

n是网页总数
p_i表示x_i的入向链接数

$$f(k) = \frac{\sum_{i=1}^n equal(p_i, k)}{n}$$

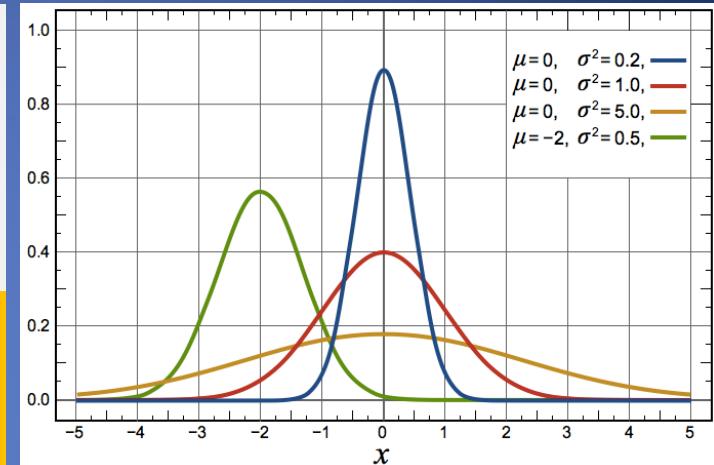
什么性质?
曲线是什么形状?

正态分布—随机量的一种规律

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2}$$

概率密度函数

μ : 均值; σ^2 : 方差; σ : 标准差



中心极限定理: 大量独立同分布的随机变量之和（均值）是正态分布的随机变量；与原始分布是什么无关。

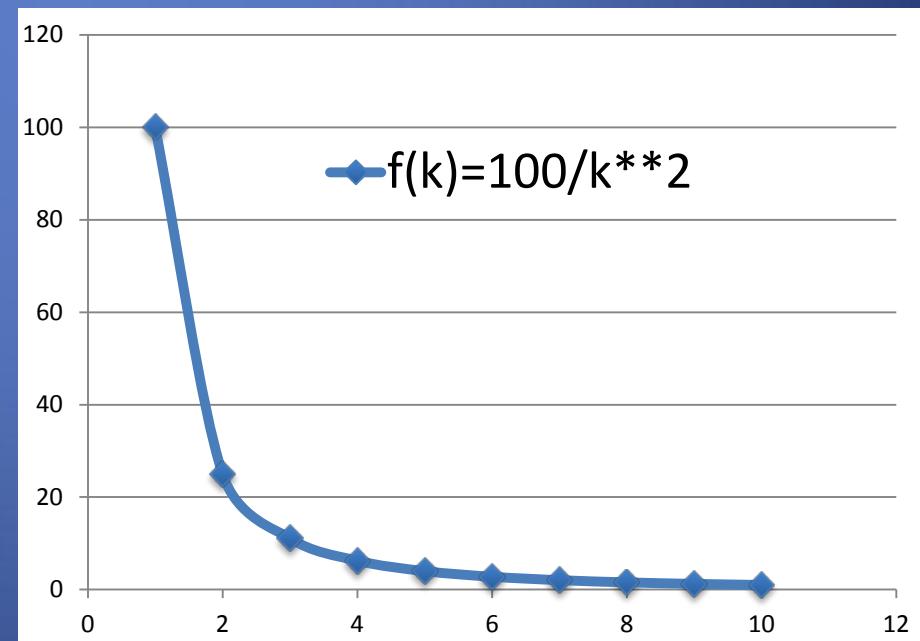
网页入向链接的个数（随机量）应该是什么分布？
如果想象：网页A是否给网页B链接是一个随机变量；那么，B得到的入链个数就是大量随机变量之和。于是，正态分布？

数据实验表明：

$$f(k) = \frac{a}{k^c} = a \cdot k^{-c}$$

- 大量各种不同的数据集都显现出这种性态
- 因此，我们说这就是反映网页入度分布的规律——“幂率”

k	$f(k) = 1/k^{**2}$	$g(k) = 1/2^{**k}$
1	1	0.5
2	0.25	0.25
3	0.111111111	0.125
4	0.0625	0.0625
5	0.04	0.03125
6	0.027777778	0.015625
7	0.020408163	0.0078125
8	0.015625	0.00390625
9	0.012345679	0.001953125
10	0.01	0.000976563



幂率的习惯（图形）表示

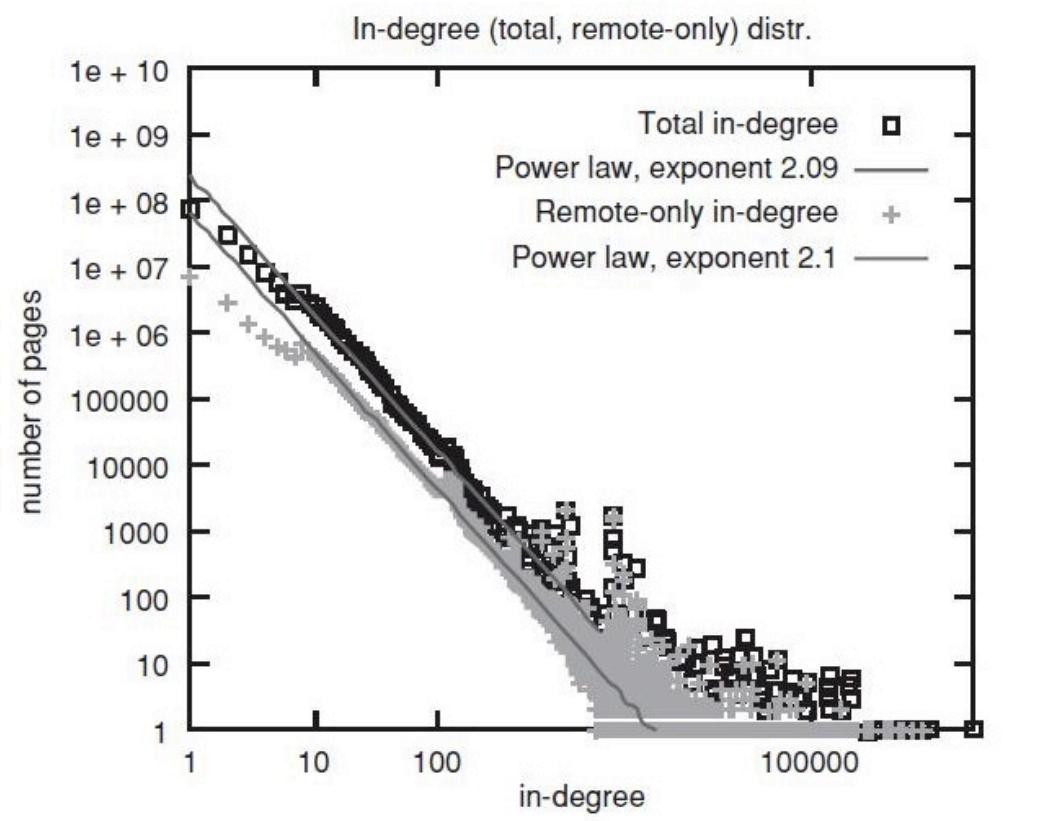
$$f(k) = \frac{a}{k^c} = a \cdot k^{-c}; \quad \log(f(k)) = \log(a) - c \cdot \log(k)$$

- $\log(f(k))$ 是关于 $\log(k)$ 的线性函数
 - 以 $\log(k)$ 为横轴, $\log(f(k))$ 为纵轴的图像是一条直线
- 这等价于说
 - 以 k 为指数标度的横轴, $f(k)$ 为指数标度的纵轴的图像是一条直线



因此，给定一
组原始数据

$$\begin{aligned} k &: 1, 2, 3, \dots \\ f(k) &: \dots \end{aligned}$$



- 为查看 $f(k)$ 是否幂率，一种做法就是取 $\log(k)$ 和对应的 $\log(f(k))$ ，然后用得到的数据值在常规坐标下绘制曲线图形，观察结果看起来像不像一条直线。
- 在数据量很大的时候（流行度数据常常如此），这种方式很有效。许多绘图工具直接支持对数坐标。

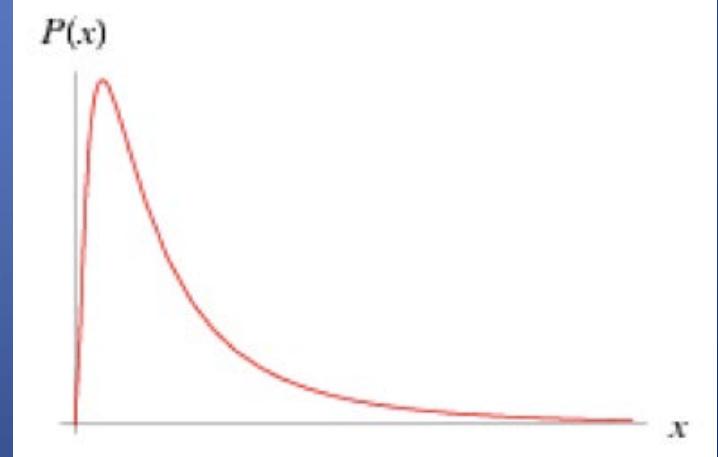


幂率：流行度的一种主导规律

- 网页（网站）的入度，网站的出度
- 网站的规模（其中网页的数量）
- 每天能接到k个电话的电话
- 一种书籍的销量
- ...

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\ln x - \mu\right)^2/\sigma^2}$$

但不是完全普适的规律。
对数正态分布（**log normal**）
也反映某些事物流行的现象。





幂率的基本特征

- Scale free (不受尺度影响的)
 - 数学上，一个函数 $f(x)$ 称为“scale free”，若 $f(a*x)=b*f(x)$
 - Scale free函数隐含着自相似 (self similarity)
- 平均行为不反映典型行为
 - “典型行为” – 经常遇到的；
 - “平均行为” – 总和 / 个数
 - 正态分布的“平均行为”反映“典型行为”
 - 典型看到“中等个子”，大个子很稀少

比较容易看到
“个大的”



$$f(x) = \frac{a}{x^2} = ax^{-2}, \quad x \in [1, n]$$

一个算例

To determine the normalizing factor a, set

$$\int_1^n f(x) dx = 1, \text{ i.e.}$$

$$\int_1^n ax^{-2} dx = -ax^{-1} \Big|_1^n = a - an^{-1} = 1$$

$$a = \frac{n}{n-1}, \text{ then, figure out the mean}$$

$$\int_1^n xf(x) dx = \int_1^n ax^{-1} dx = a \ln x \Big|_1^n = a \ln n = \frac{n \ln n}{n-1}$$

suppose $n=100$, we have:

$$\int_1^n xf(x) dx = \frac{n \ln n}{n-1} = \frac{200 \ln 10}{99} \approx \frac{200 \times 2.3}{99} = 4.65$$

see the probability observing larger than mean

$$-ax^{-1} \Big|_{4.65}^{100} = \frac{100}{99} \times \left(\frac{1}{4.65} - \frac{1}{100} \right) \approx 0.207, \text{ also}$$

$$-ax^{-1} \Big|_{9.3}^{100} = \frac{100}{99} \times \left(\frac{1}{9.3} - \frac{1}{100} \right) \approx 0.1$$

取值范围

$n=1, \dots, 100$

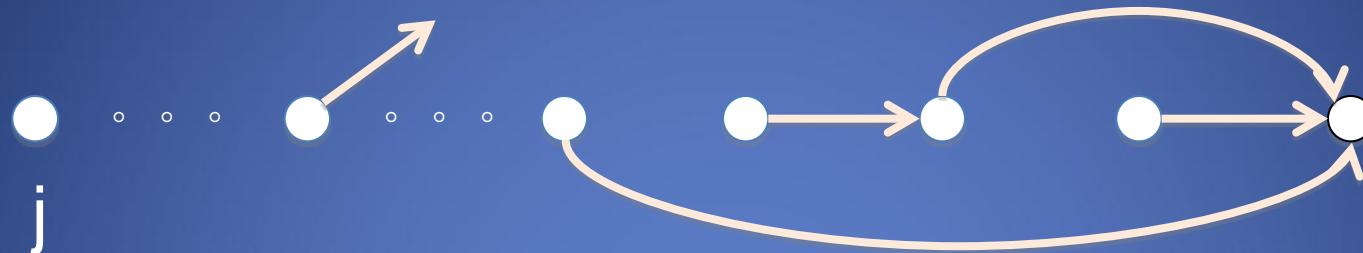
均值 = 4.65, 相对
比较小

意味着：看到比均
值大的对象的可能
性很高

具体算出来，看到
较大对象的概率约
为 0.2

最后这个计算表明
看到比均值大一倍
对象的概率约为 0.1

幂率的成因（“富者更富”模型）



- 网页按照顺序创建: 1, 2, 3, ..., j, ...
- 当创建网页 **j** 时, 以概率 p 或 $1-p$ 选择如下(a)或(b)执行
 - (a) 以概率 **p**, 均匀地、随机地选择一个早先创建的网页 **i**, 建立一个从 **j** 到 **i** 的链接
 - (b) 以 **$1-p$** 的概率, 均匀地、随机地选择一个早先创建的网页 **i**, 建立一个从 **j** 到 **i** 所指向的网页的链接。

此模型产生幂率, 其中的指数 c 取决于概率 p

为什么说这体现了“富者更富”

- 网页按照顺序创建: 1, 2, 3, ..., j, ...
- 当创建网页 **j** 时, 以概率 **p** 或 **1-p** 选择如下(a)或(b)执行
 - (a) 以概率 **p**, 均匀地、随机地选择一个早先创建的网页 **i**, 建立一个从 **j** 到 **i** 的链接
 - (b) 以**1-p** 的概率, 均匀地、随机地选择一个早先创建的网页 **i**, 建立一个从 **j** 到 **i** 所指向的网页的链接。
- 等价于:
-
- (b) 以**1-p** 的概率, 按照与已有入度成比例的概率, 选择一个早先创建的网页 **i**, 建立一个从 **j** 到 **i** 的链接。

BA无标度网络模型

Barabasi-Albert (B-A) 模型

无尺度网络形成的两个基本机制：

- (1)增长。
- (2)优先连接。

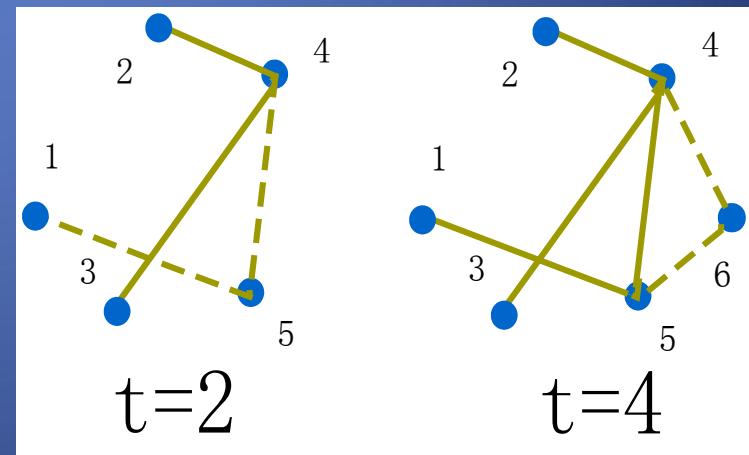
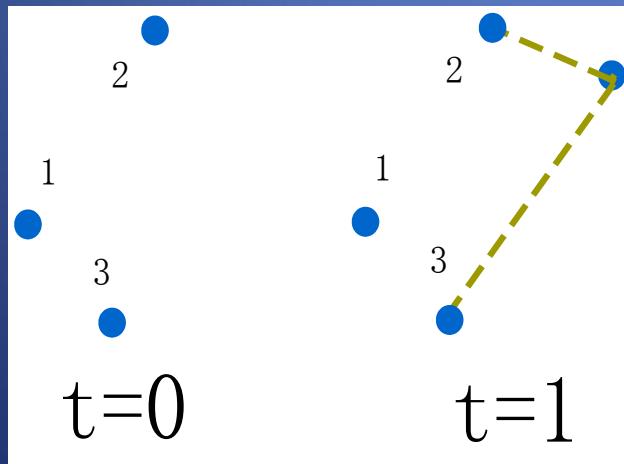
$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

B-A 模型的构建

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

(1) 增长：在初始时刻，假定系统中已有 m_0 个点，在以后的每一个时间步长中，我们新增一个度为 m 的点($m \leq m_0$)，这 m 条边连向网络中已经存在的 m 个不同的点。

(2) 优先连接：当我们在原来网络中选择一些点被新增加的边连结时，这些点被连结的概率与这些点自身的度的大小成正比。比如度为 K_i 的点 i 被新增点连结的概率为：



适应度模型

BA模型中老节点具有较高的度，增加适应度

- (1) 增长：在初始时刻，假定系统中已有 m_0 个点，在以后的每一个时间步长中，我们新增一个度为 m 的点($m \leq m_0$)，这 m 条边连向网络中已经存在的 m 个不同的点。
- (2) 优先连接：

$$\Pi(k_i) = \frac{\eta_i k_i}{\sum_j \eta_j k_j}$$

局域世界演化模型

增加局域世界

(1) **增长**: 在初始时刻, 假定系统中已有 m_0 个点, 在以后的每一个时间步长中, 我们新增一个度为 m 的点($m \leq m_0$) ,

(2) **局域世界优先连接**: 随机从网络已有的节点中选取 M 个节点($M \geq m$), 作为新加入节点的局域世界。第 t 步新加入的节点根据优先连接概率

$$\Pi_{\text{Local}}(k_i) = \frac{M}{m_0 + t} \sum_j \text{Local } k_j$$

富者更富效应的不可预测性

- “富者更富”也具有级联的意味，现实生活中有不少体现这种情形的现象
- 最初阶段充满不确定性，“富”到一定程度后就开始“起飞”
 - 与《哈利波特》同样质量的小说在同一时期其实很多，但真正流行起来的很少
 - 同样水平的歌星在同一时期其实很多，但真正出名的很少
- 一类事物流行史的细节不可能重演，但历史的结果宏观上总是如此

“长尾” (long tail) 又是什么?

- 一类产品（例如书籍，个人音乐专辑）各个品种的销售量（流行度）常符合幂率

$$f(x) = \frac{a}{x^c}, \quad c \geq 2$$

发现销量为x的品种的概率

- 人们更方便直接谈销量（而不是概率），设该类产品的总销量为n，于是

$$n \cdot f(k) = \frac{n \cdot a}{k^c}, \quad c \geq 2$$

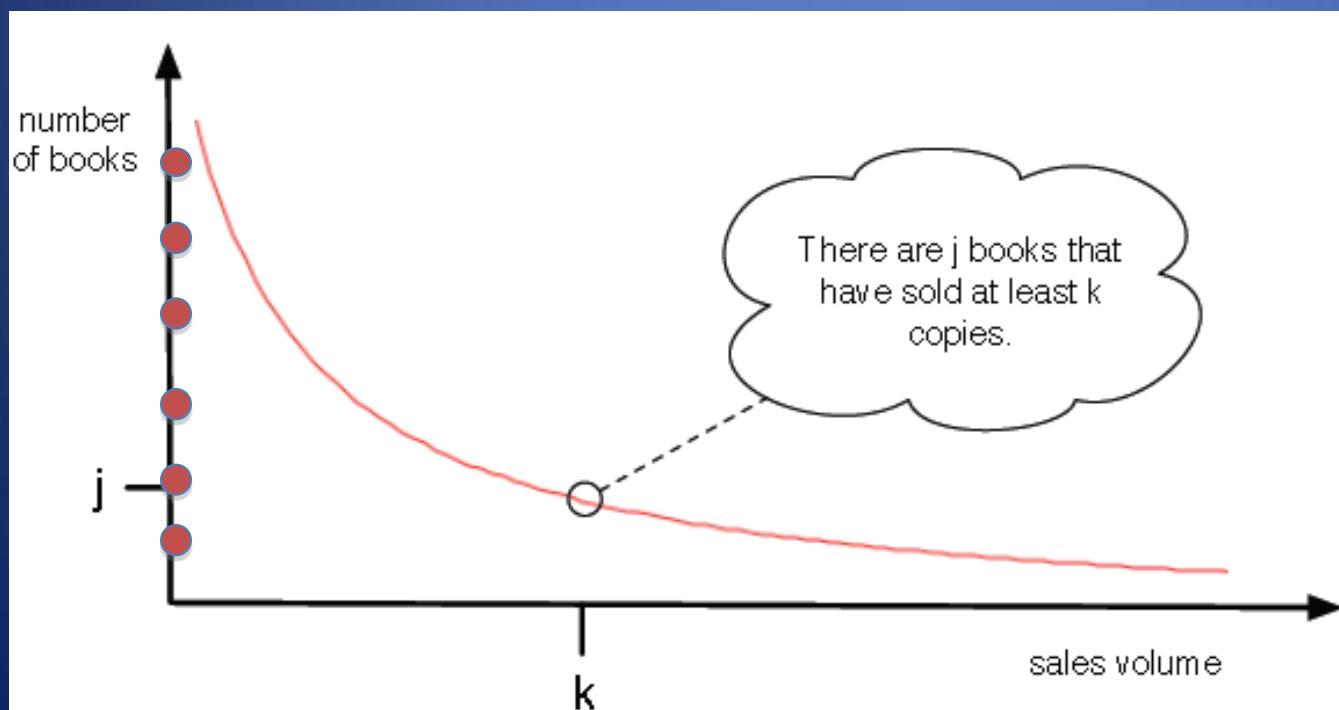
销量为k的品种的个数

“长尾”（进一步）

- 关心“销量至少为K的品种数”

也是“幂率”
(但幂次变了)

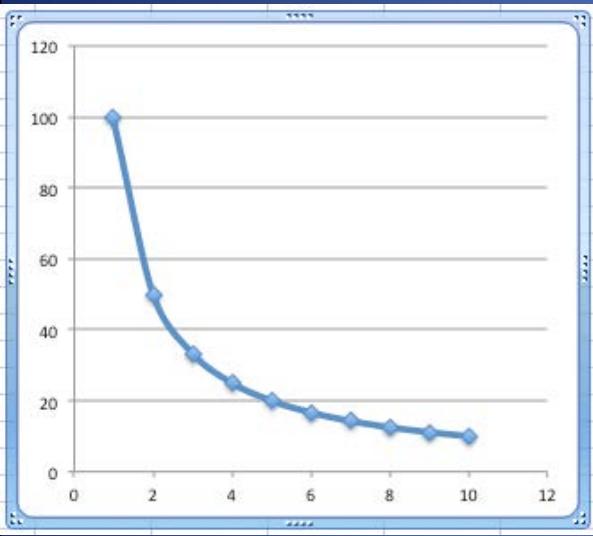
$$\int_K^{\infty} n \cdot f(k) dk = \int_K^{\infty} \frac{n \cdot a}{k^c} dk = -\frac{n \cdot a \cdot k^{-c+1}}{c-1} \Big|_K^{\infty} = \frac{na/(c-1)}{K^{c-1}}, \quad c \geq 2$$



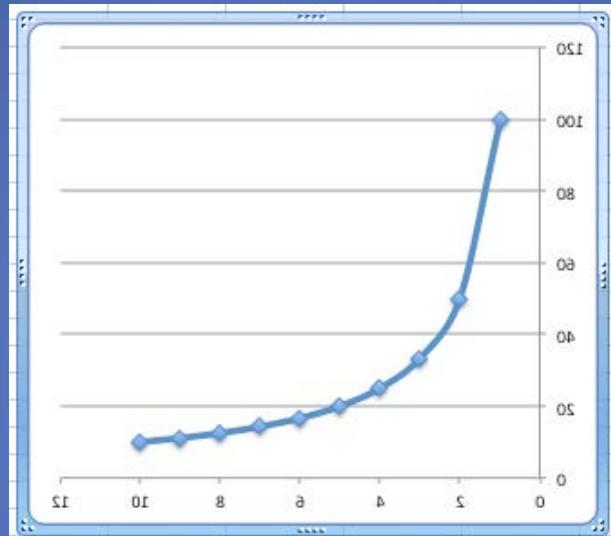
长尾的典型图示。由于降了一个幂次，尾巴显得更加明显

齐普夫定律 (Zipf's Law)

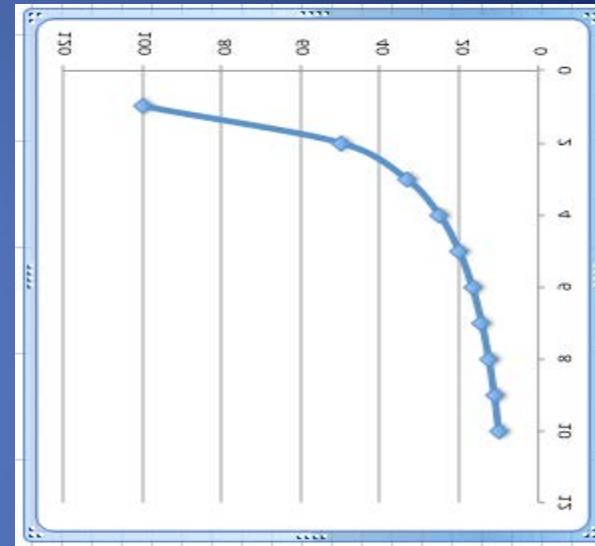
—— 另一个视角看“长尾”



销量至少为k的品种数



“向左翻转”



“顺时针旋转”

- 横轴此时可看成“销量排名位次”，纵轴则是对应位次的销量。从函数关系看：

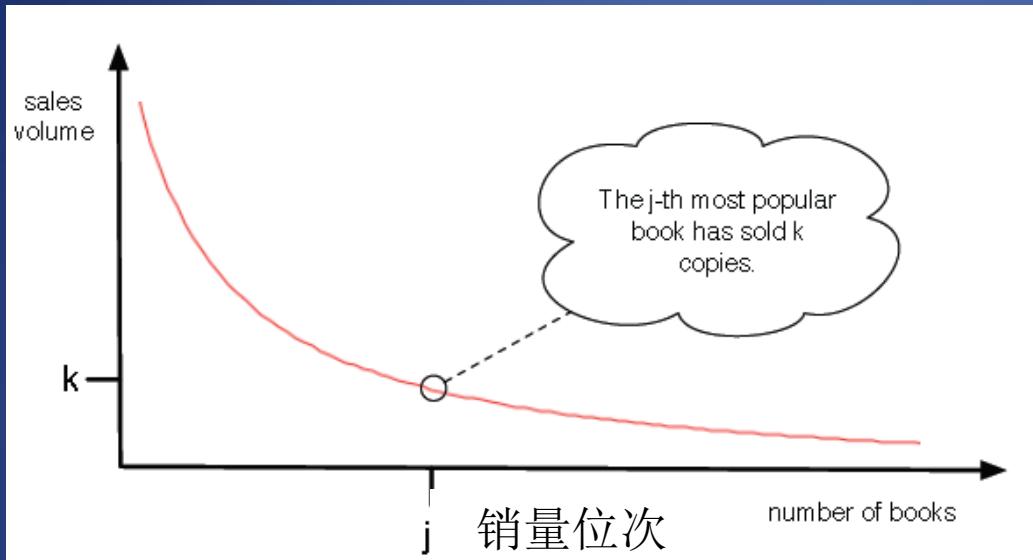
$$y = \frac{a}{x^c}, \quad c \geq 1$$

$$x^c = \frac{a}{y}, \quad c \geq 1$$

$$x = \frac{a^{1/c}}{y^{1/c}} = \frac{b}{y^d}, \quad d \leq 1$$

也是幂率！

长尾效应与营销策略



结论是：在长尾规律下，如果品种足够多（即 \max 很大），经营利基产品也能够获得很大利益。

$x = \frac{b}{y^d}$, consider all "non hits" sales

考虑top-100之后

assume $d = \frac{1}{2}$, which correspond to $c=2$

$$\int_{100}^{\max} b \cdot y^{-d} dy = -\frac{by^{-(d-1)}}{d-1} \Big|_{100}^{\max} = 2b(\sqrt{\max} - 10)$$

if $d=1$, which correspond to $c=1$

$$\int_{100}^{\max} b \cdot y^{-d} dy = b \ln y \Big|_{100}^{\max} = b(\ln(\max) - \ln 100)$$

对应概率意义
幂率中的幂次3

对应概率意义
幂率中的幂次2

但有两个前提
* 降低库存成本
* 让顾客容易发现那些产品

销售排行榜、推荐、搜索

- 是促进“畅销产品”还是促进“利基产品”的销售？
- 排行板：推动富者更富
- 推荐（相关推荐）
 - 取决于“相关”的含义，若是“买了这产品的其他人通常也买了...”，则倾向于富者更富；
 - 若是按照某种“内容相关性”，则可起到推动利基产品销售的作用
- 搜索：也是有两面性

要点小结

- 幂率是流行现象的主导规律
 - 但不是普适规律
 - “富者更富”是幂率的一种成因。发现一种流行现象的规律有意义，理解其成因更重要
- 符合幂率的流行现象也可以通过“长尾”或齐普夫定律来刻画
 - 它们本身也满足幂函数关系（但幂次不同）
 - 不仅幂率是“长尾”，还有许多长尾分布
- 对营销策略的启示

作业

- 第18章 2,3题

信息级联

信息级联（连锁反应）

(information cascades)

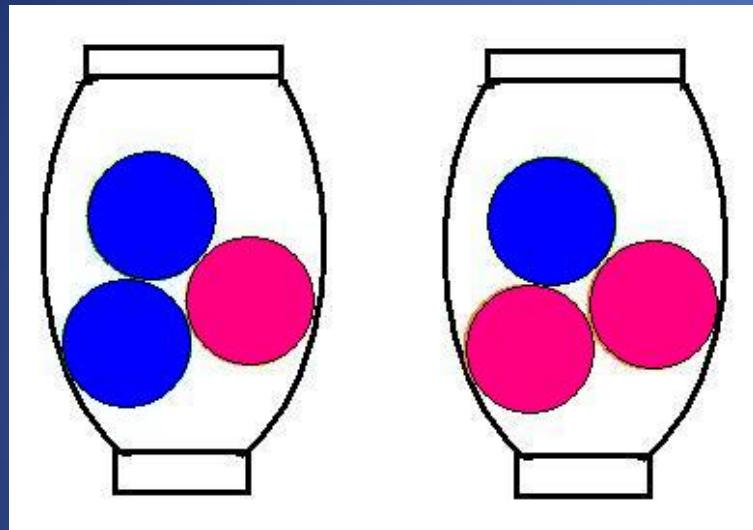
- “随大流”（从众）现象及其理性
 - 常见的社会现象：依据信息的决策 vs 依据结果（反映某种信息）的决策
- 分析随大流现象的一个工具：贝叶斯规则
 - 状态 → 引起状态出现的原因的可能性
- 一种通用的（简单）信息级联模型
 - 对象：一个可能处于“利好”和“利坏”状态的事物
 - 决策：对该事物表示“接受”还是“拒绝”的选择
 - 回报：
 - (1) “接受”的回报取决于状态，在没附加信息情况下，期望为0；
 - (2) “拒绝”的回报总为0
 - 信号（私有，附加信息）：有可能辅助决策以争取较大的回报期望
 - 依次决策过程分析

“随大流”（从众、模仿）

(following the crowd)

- 常见的社会现象
 - 产品的选择，委员会的抉择，政治观念的采纳
- 为什么会模仿别人？
 - (1) 好奇（信息），(2) 直接获益
- 感性的还是理性的？
 - 信息（私有信息，群体行为背后的信息）、推理与结果
- 压力下的顺从还是自觉的选择？

一个简单的群集 (herding) 实验



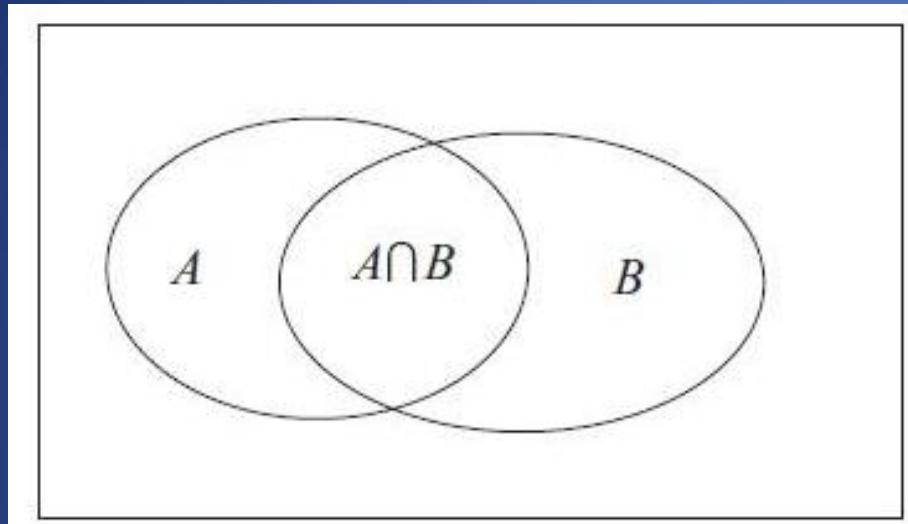
- 两个坛子，以 $p=0.5$ 的概率拿出一个来进行实验
- 参与人（1, 2, 3, ..., k）顺序来到跟前，随机摸出一个球看看，然后大声宣布他认为坛子是“蓝多”还是“红多”
- 放回小球，离开
- 注意，每个人只公开宣布自己的判断，不告诉他看到的颜色（即不揭示自己的私有“信号”）。
- 最后，判断对了的人有奖。
- 在理性的分析下，我们会看到什么现象？

当一个人连续听到前面是一串“蓝多”的时候，他知道并不是每一个人都真的看到蓝球了，但他自己逻辑上的最好选择也是宣布“蓝多” — 尽管他可能看到了红球！而且他也知道有可能前面的人多数看到的是红球！

贝叶斯规则：从一个例子开始

- 设一个城市，出租车中有 80% 是绿色， 20% 是黄色。
- 出现了一个交通事故，肇事出租车逃离，现场目击者说是“黄色”，但他可能看错了
 - 假设出错概率0.2（即绿说成了黄），亦即真是黄色也说黄色的概率为0.8
- 问，那辆车是黄色的可能性（概率）有多大？

贝叶斯规则 (Bayes's Rule)



- 总体
- A, B两个事件（对应的情况）
- 事件可能同时出现 $A \wedge B$
- 面积的相对大小表示“可能性”（概率）

- $\Pr[A]$, $\Pr[B]$: 对应单个事件的概率
- $\Pr[A|B]$, 在观察到B的情况下出现A的（条件）概率
 - 直观上，就是“ $A \wedge B$ 的面积”与“B的面积”之比，即

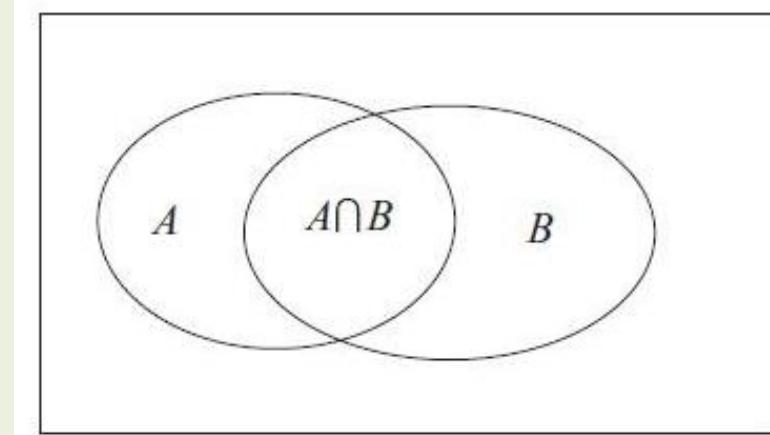
$$\Pr[A|B] = \Pr[A \wedge B] / \Pr[B]$$

$$\Pr[A | B] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

$$\Pr[B | A] = \frac{\Pr[A \cap B]}{\Pr[A]}$$

$$\Pr[A | B] \cdot \Pr[B] = \Pr[B | A] \cdot \Pr[A]$$

$$\Pr[A | B] = \frac{\Pr[B | A] \cdot \Pr[A]}{\Pr[B]}$$



$$\Pr[A|B] = \Pr[B|A] * \Pr[A] / \Pr[B]$$

- 判断肇事车的颜色
 - $\Pr[\text{是黄色}|\text{报告黄色}]?$
 - $\Pr[\text{报告黄色}|\text{是黄色}]$ =目击者正确率=0.8
 - $\Pr[\text{是黄色}]$ =城市中黄车的百分比=0.2
 - 于是, 分子= $0.8 * 0.2 = 0.16$
- 分母, $\Pr[\text{报告黄色}]$ 可表为下面两个量之和
 - $\Pr[\text{报告黄色}|\text{是黄色}] * \Pr[\text{是黄色}] = 0.2 * 0.8 = 0.16$
 - $\Pr[\text{报告黄色}|\text{是绿色}] * \Pr[\text{是绿色}] = 0.8 * 0.2 = 0.16$
- 结果, 肇事车是黄色的概率为0.5

贝叶斯公式在一种特殊情况下的方便表示

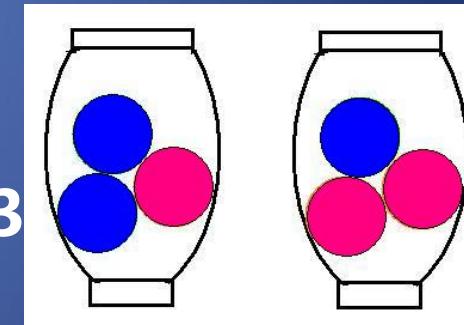
$$\Pr[A | B] = \frac{\Pr[B | A] \cdot \Pr[A]}{\Pr[B]}$$
$$= \frac{\Pr[B | A] \cdot \Pr[A]}{\Pr[B | A] \cdot \Pr[A] + \Pr[B | \bar{A}] \cdot \Pr[\bar{A}]}$$

if $\Pr[A] = 0.5$, we then have

$$\Pr[A | B] = \frac{\Pr[B | A]}{\Pr[B | A] + \Pr[B | \bar{A}]}$$

用贝叶斯规则分析群集实验

- 本质上，每个人是在听到前面人的宣布结果和自己看到球的颜色基础上，对坛子类型进行条件概率的估计
- $\Pr[\text{蓝多} | \text{已有信息}]$ ？（若大于 $1/2$ ，则接受）
 - 第一个人？ $\Pr[\text{蓝多} | \text{blue}] = \dots = 2/3$
 - 第二个人？ $\Pr[\text{蓝多} | \text{blue, blue}] =$
 - 第三个人？ $\Pr[\text{蓝多} | \text{blue, blue, red}] = 2/3$
 - 第四个人？ $\Pr[\text{蓝多} | \text{blue, blue, *, red}] =$
 $= \Pr[\text{蓝多} | \text{blue, blue, red}] =$
- （具体推理计算过程见书）



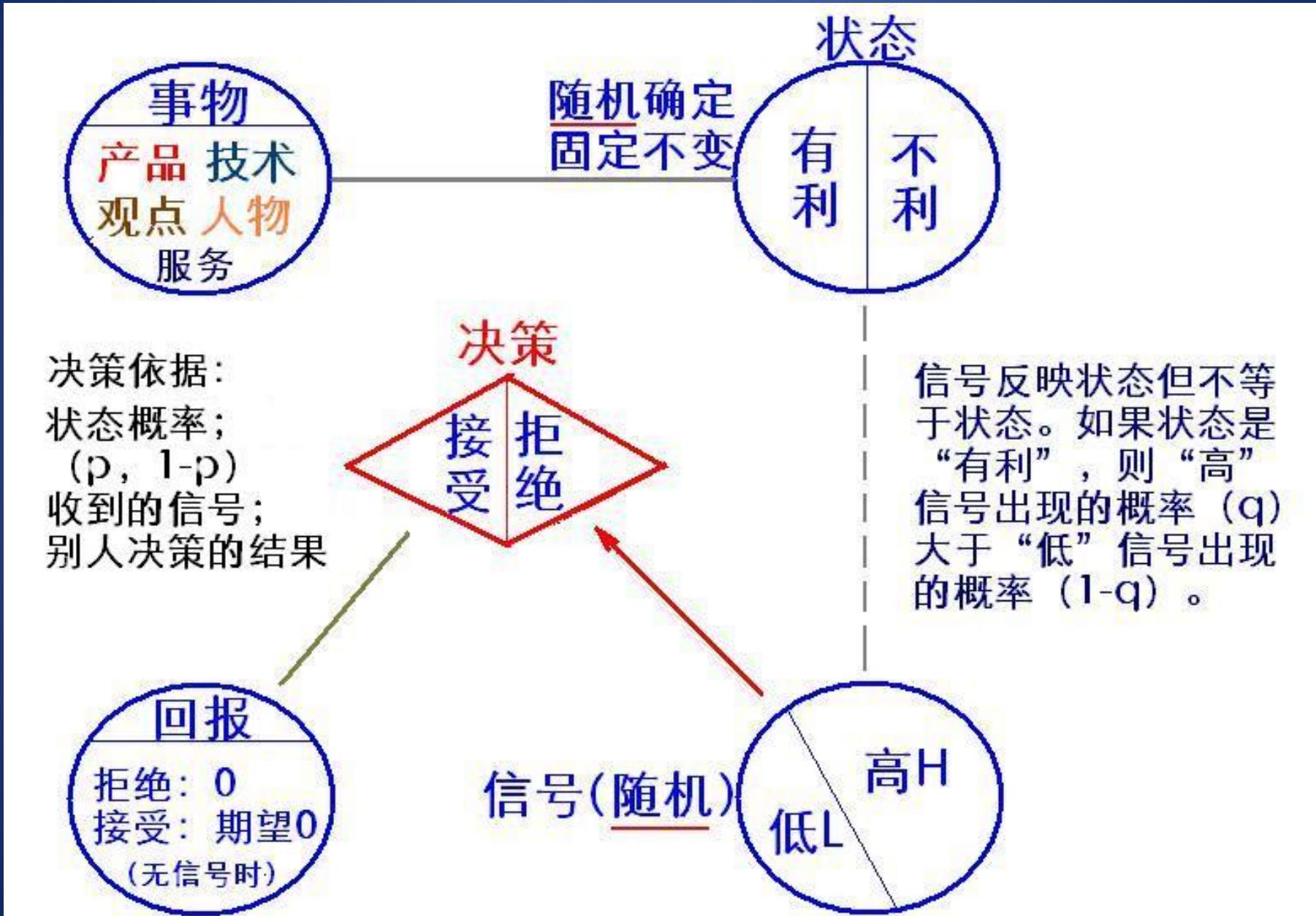
第三个人的推导

Using ($\Pr[A]=0.5$): $\Pr[A | B] = \frac{\Pr[B | A]}{\Pr[B | A] + \Pr[B | \bar{A}]}$, we have

$$\Pr[maj-blue | blue, blue, red]$$

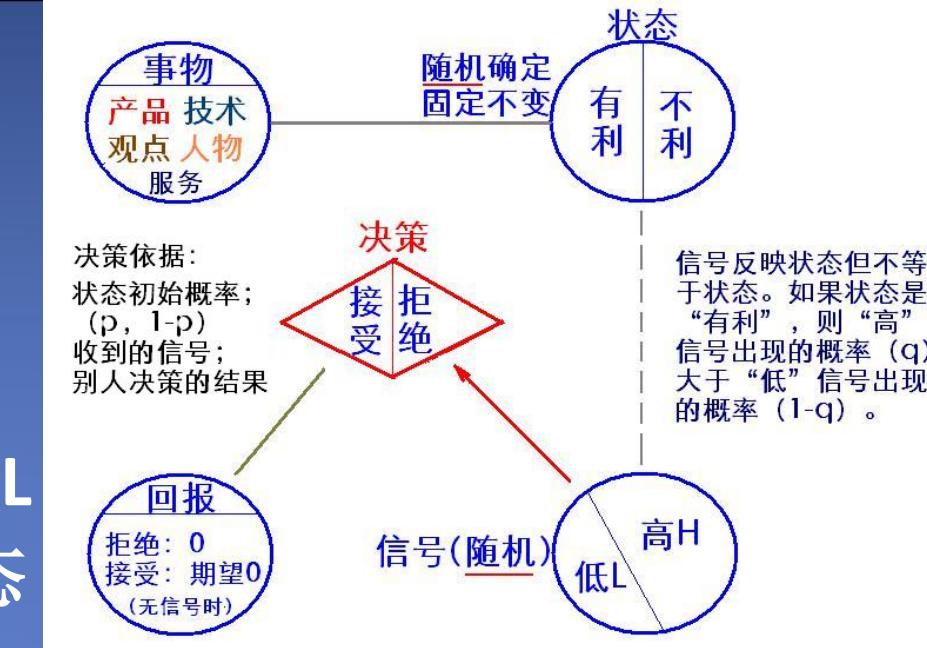
$$= \frac{\Pr[blue, blue, red | maj-blue]}{\Pr[blue, blue, red | maj-blue] + \Pr[blue, blue, red | maj-red]}$$
$$= \frac{(2/3) \cdot (2/3) \cdot (1/3)}{(2/3) \cdot (2/3) \cdot (1/3) + (1/3) \cdot (1/3) \cdot (2/3)} = \frac{2}{3}$$

一种简单通用的级联模型



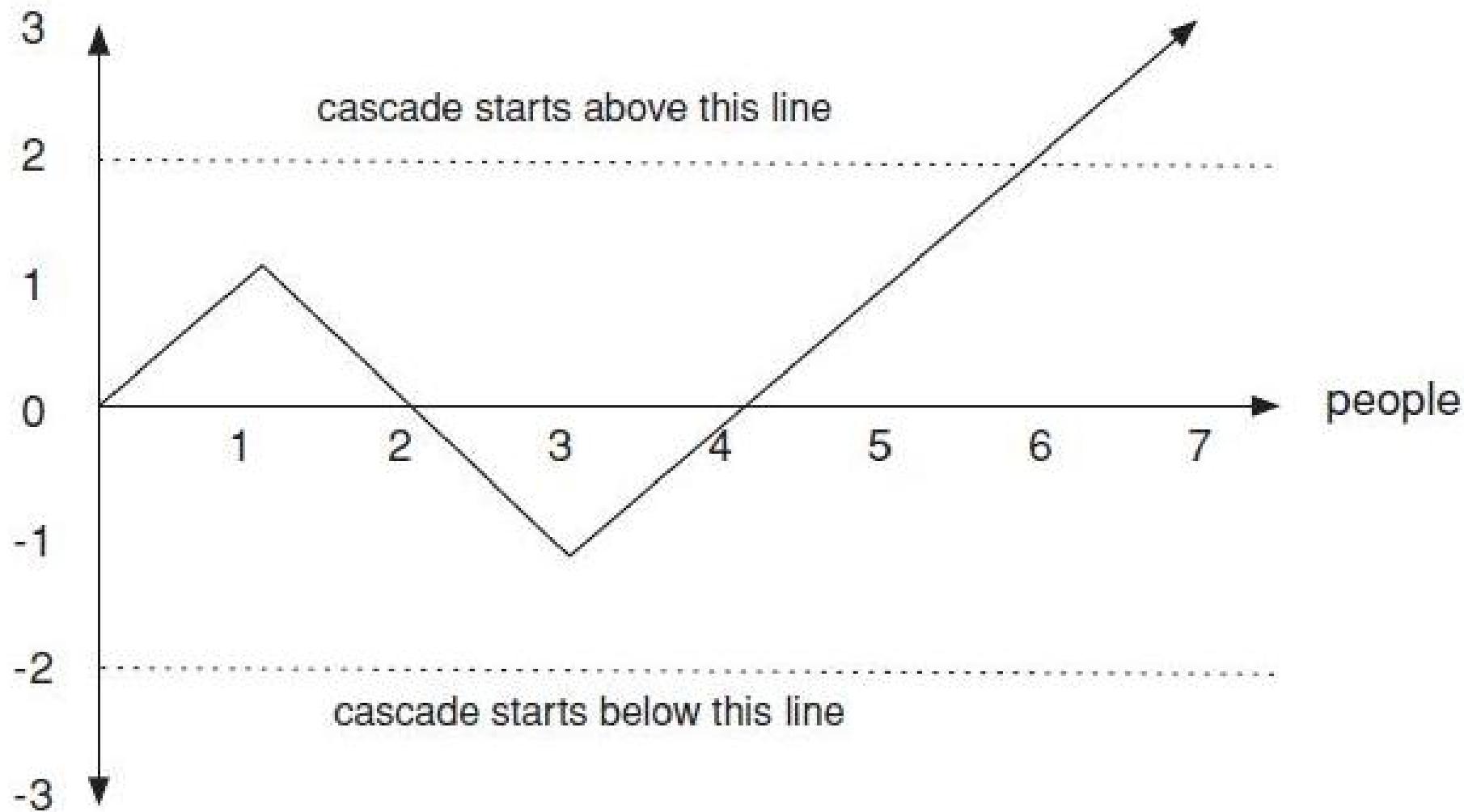
简单通用级联模型

- 试图回答的问题
 - 当收到一个信号（H或L），该如何决策（状态是否有利）？
 - 当若干人一起参与，依次决策，第N个人可见前N-1人的信号时，该如何决策？
 - 当若干人一起参与，依次决策，第N个人只可见前N-1人的决策结果时，该如何决策？
- 假设每一个人都是理性的
- （有些情况下，容易有正确的直觉，但我们要模型下的推理结果，从而也可能得到不容易从直觉中得出的结论）



级联开始的条件

#acc - #rej



信息级联现象

- Signal: H, L, H, L, H, L, H, L
- Decision: A, R, A, R, A, R, A, R

- Signal: H, H, L, L, H, L, L, L
- Decision: A, A, A, A, A, A, A, A

- Signal: H, L, H, L, L, L, H, H
- Decision: A, R, A, R, R, R, R, R

关于信息级联的进一步认识

- 级联可能是错误的
- 基于很少的信息，级联也可能开始（级联效应，连锁反应——“多米诺骨牌效应”）
- 级联是脆弱的，中间信息的微小扰动就可能终止甚至改变级联方向
- 级联现象与“群体智慧”不矛盾
- 级联现象的防止和利用
 - 独立决策与商讨决策的平衡
 - 新产品的推广，虚假火爆的终止

作业

- 第16章，第2，5题

网络中的级联

- 想像一个社会关系网（例如由同一个专业圈子的人构成的“同事网”），其中大家本来都用着PC Notebook（windows）
- Mac出来了，并且迷上了一些人，他们决定抛弃PC，改用MacBook，而且很坚定，尽管与同事之间交换文件不再那么顺畅
- 其他同事看到了Mac的一些优势，也看到了由此带来的一些不便
- 因此，有些人也开始从PC换到Mac，但另一些还在掂量、观望
- 问：Mac能在这个社会网络中“走多远”？

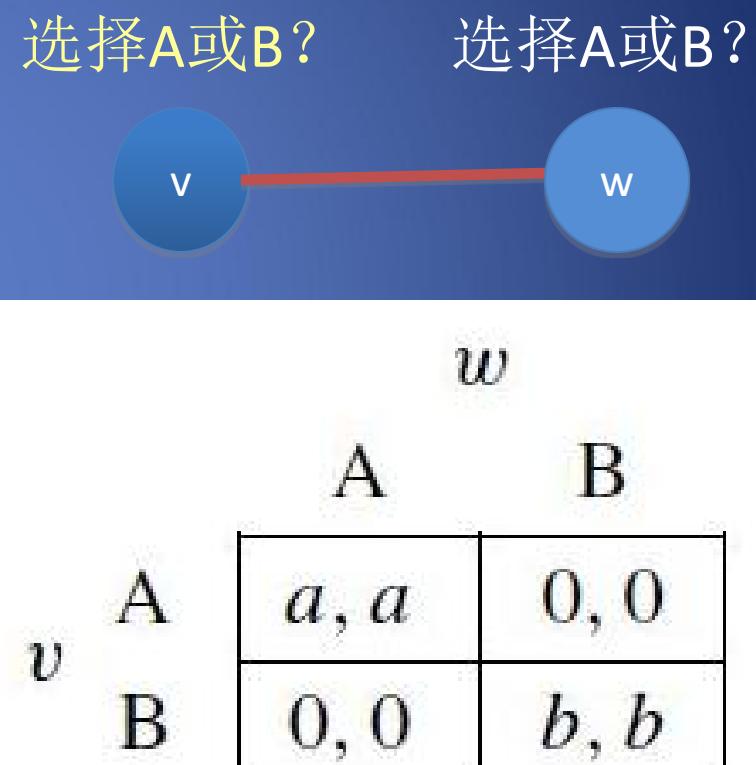
- 想像你是某公司的副总裁之一。副总裁之间的关系有亲疏，因此也有一个“亲密关系网”
- 在工作中，你感到总裁工作有问题，不应该再当下去，而且与亲密的朋友之间也有些非正式的议论
- 明天又要开总裁办公会了，你考虑向总裁发难，同时知道若有足够多的副总裁响应就会成功，否则对自己就会是灾难
- 这事还不能和别人商量。能否从关系网结构中得到些判断？

网络结构对新事物传播影响的模型

- 场景
 - 一个社会网络；A，B两类事物要在其中流行；
 - B是“旧的”，一直以来大家都采用B
 - A是“新的”，开始吸引了几个坚定份子
- 假设
 - 每个人只能采纳A或B之一
 - 两个相邻的人若都采用A，则得回报a；若都采用B，则得回报b；若采用不一样的，则回报0
 - 在从一种选择换到另一种过程中没有其他成本

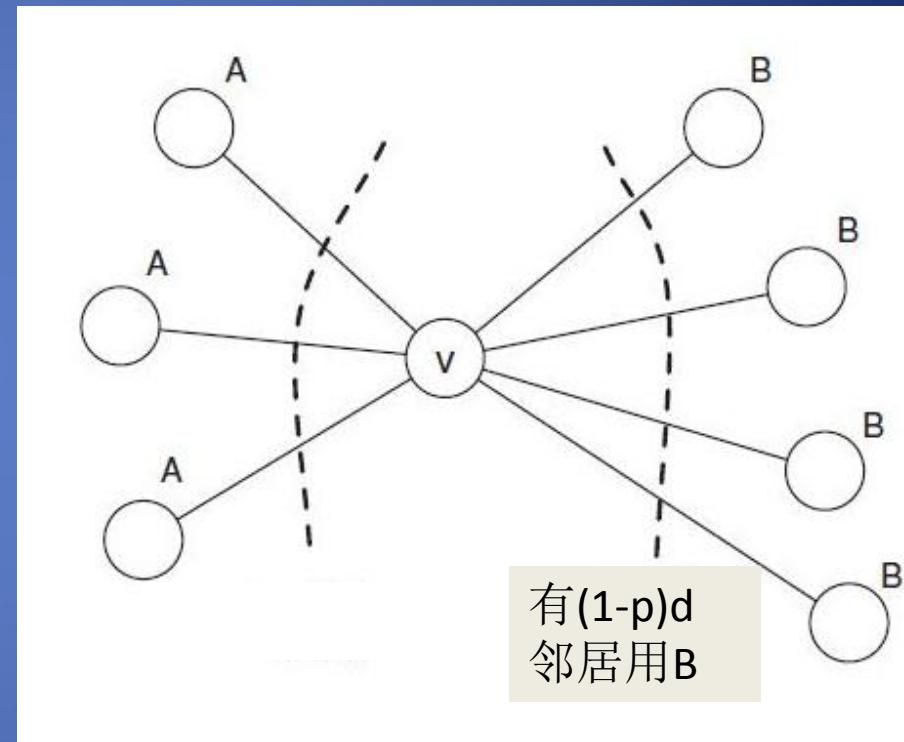
表达为一个协调博弈

- 在一条边上的博弈
 - 如果 v 和 w 都选择A，它们分别得到回报 $a > 0$ ；
 - 如果它们都选择B，分别得到回报 $b > 0$ ；
 - 如果它们选择不同的选项，那么都得到回报为0
- 一个节点 v 需要考虑其所有邻居选择的综合结果



网络节点v的决策门槛

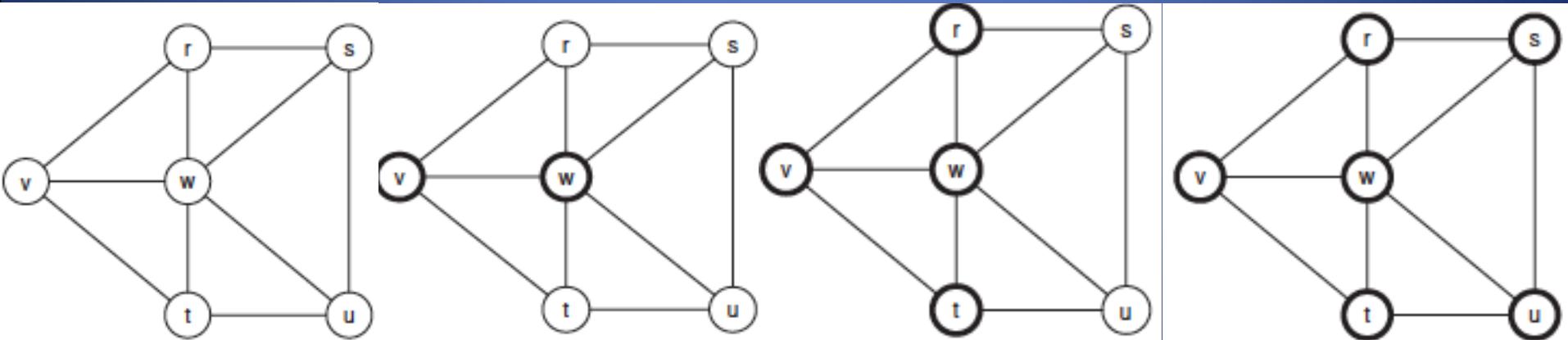
- 设 v 有 d 个邻居，在某一时刻，若比例 p 的邻居选A，比例 $(1-p)$ 的邻居选B
- v 选A的回报： pda
选B回报： $(1-p)db$
- 如果 $pda \geq (1-p)db$ ，即
 $p \geq b/(a+b)$ ，选A好。
否则，选B更好。



$$p \geq \frac{b}{a+b} = q \quad \text{门槛}$$

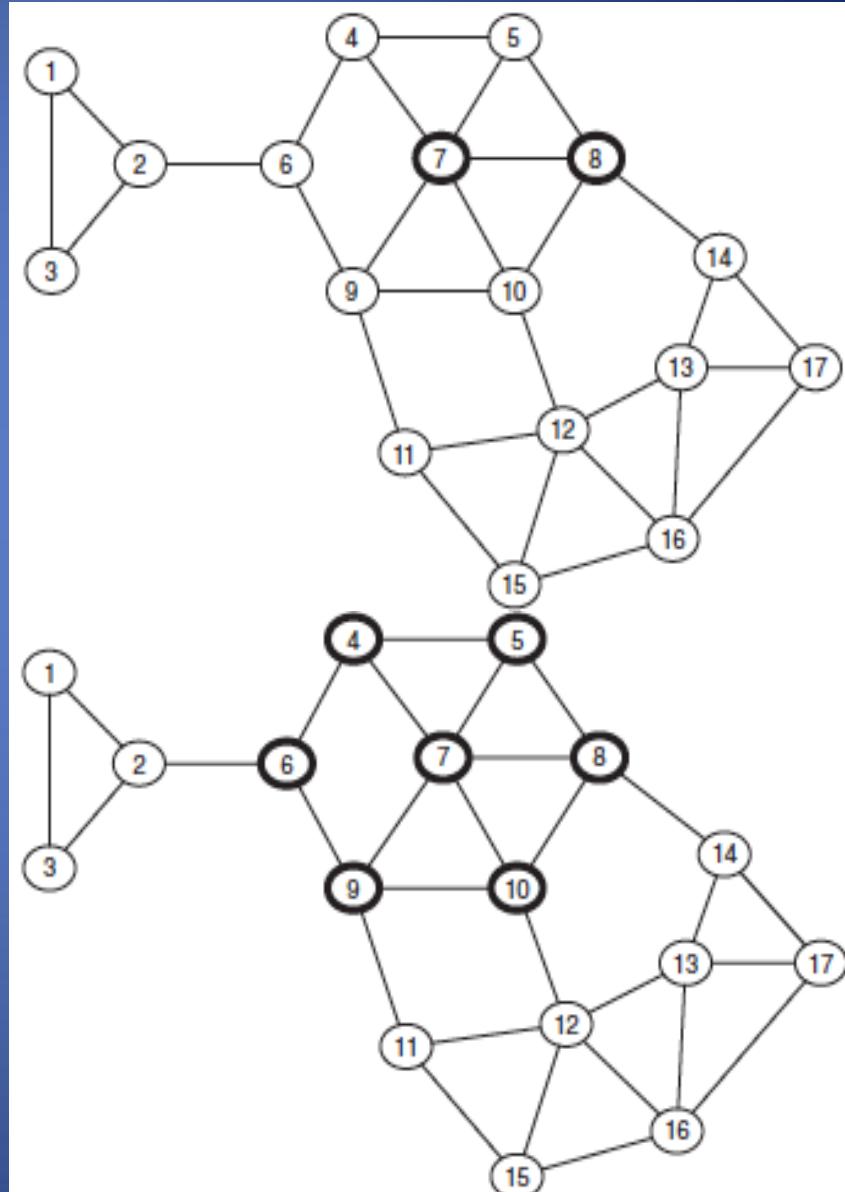
新生事物在网络中的传播

- 同时考察每一个采用B的节点采用A的邻居比例是否达到门槛 $q=b/(a+b)$
 - 是，则节点放弃B，转用A，否则继续采用B
- 重复这个过程，直到网络中采用A的节点集合不再变化
- 例子： $a=3$; $b=2$; $q=2/5$ 。此例实现了完全级联。



一个不能形成完全级联的例子

- $a=3, b=2, q=2/5$
- 最初，7、8是A的初用节点，其他所有节点均采用B
- 一步后，5、10
- 两步后，4，9
- 三步后，6
- 在那之后，级联停止
- （如何促进级联？）



问题

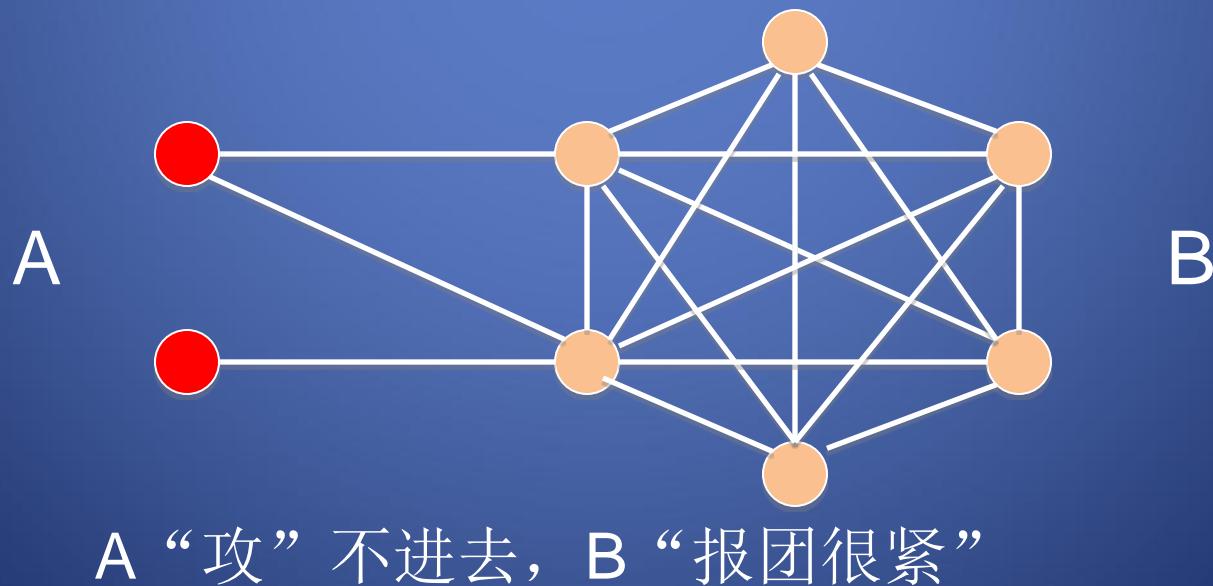
- 什么条件下，网络中节点将会全部放弃B，逐步转而选择A，实现采用A的完全级联？
- 什么情况下，A在网络中停止了传播？

可以进一步思考：

- 假若不仅是当前采用B的节点在判断是否转换，所有节点都同时考虑用什么对自己更好，有没有可能节点在某一步决定采用A，在后面某一步又决定退回B，从而造成“震荡”？
- 如果 $a < b$ ，有可能发生传播吗？

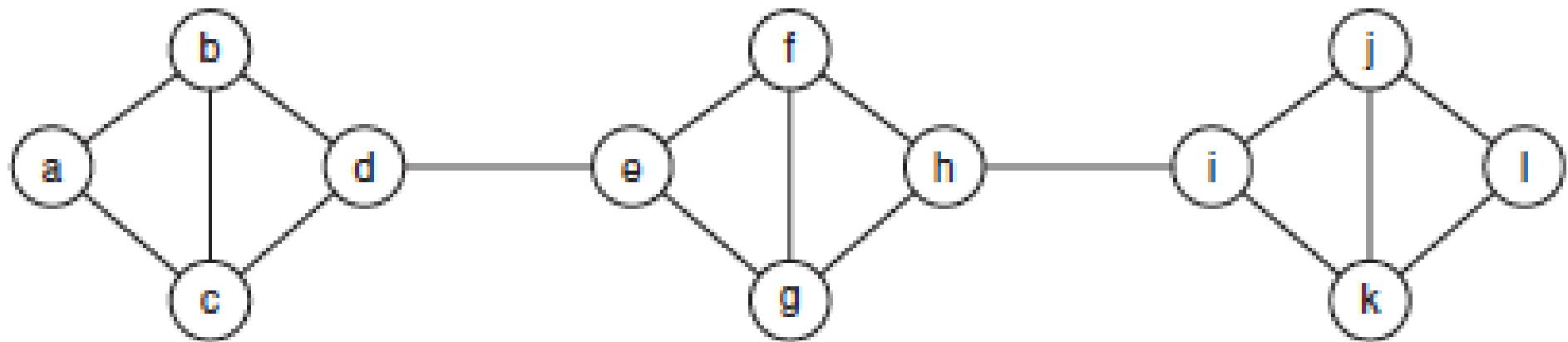
级联进行不下去了？

- 也就是剩下的节点（采用B）的A邻居数占比小于门槛 $q = b/(a+b)$
- 换句话说，也就是它们各自的B邻居数占比都大于 $1-q$



阻挡级联的因素——聚簇

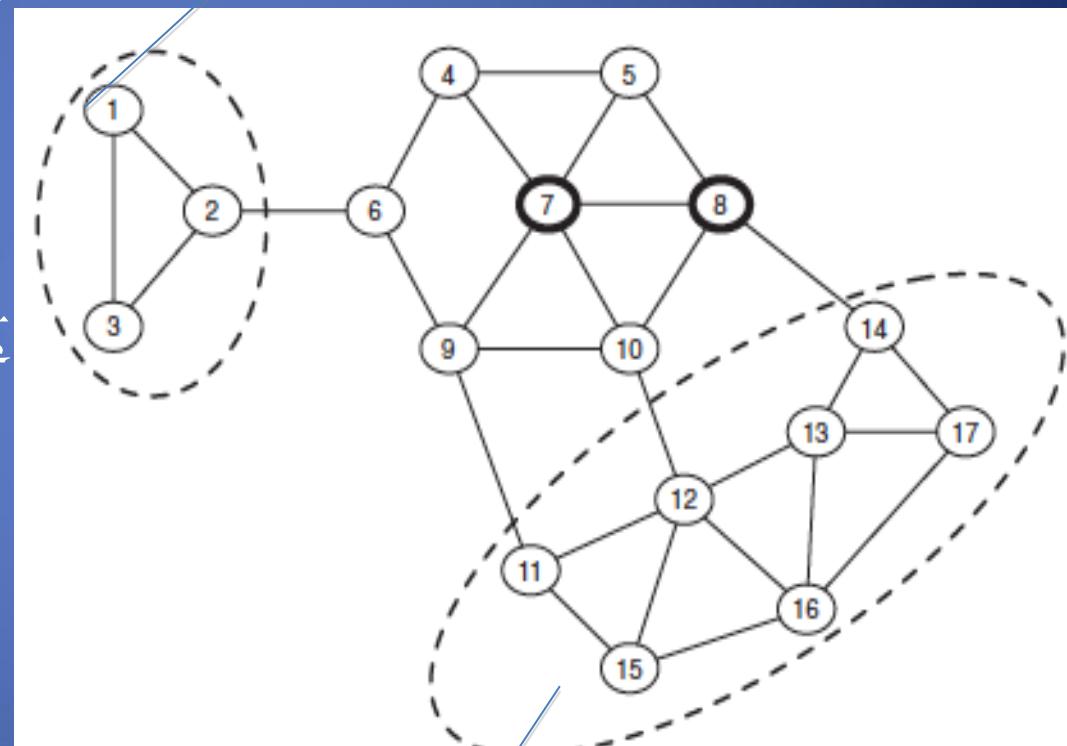
- 聚簇（“抱团”）：称一个节点集为密度为 r 的聚簇，则其中每个节点至少有比例为 r 的网络邻居也属于这个节点集合
- 下图有3个密度为 $2/3$ 的聚簇



聚簇和级联的关系（定理）

- 设网络中一个初用节点集采用 A，剩余网络的其他节点采用 B，且它们改用A的门槛值为 q
- 如果剩余网络中包含一个密度大于 $1-q$ 的聚簇，则这个初用节点集不能形成A的完全级联
- 而且，如果一个初用节点集不能形成一个完全级联，则剩余网络一定包含一个密度大于 $1-q$ 的聚簇

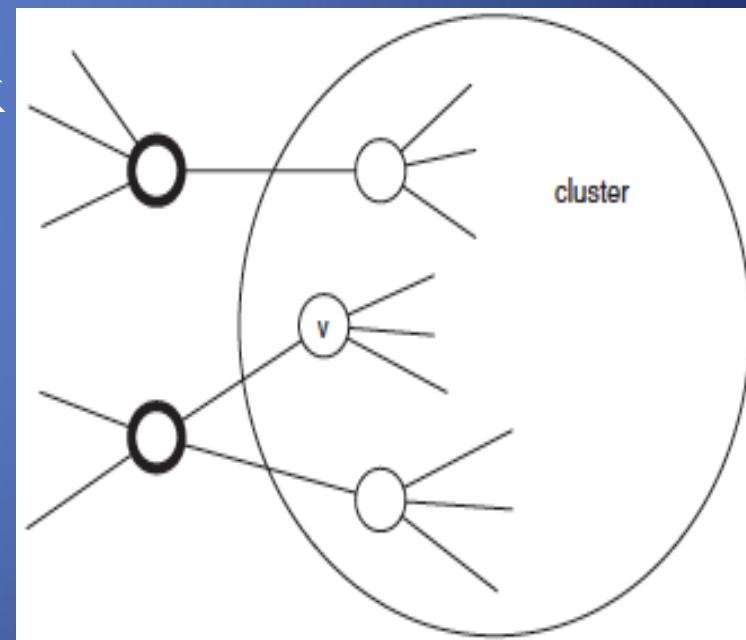
$$\text{密度} = 2/3 > 1-q = 3/5$$



$$\text{密度} = 2/3 > 1-q = 3/5$$

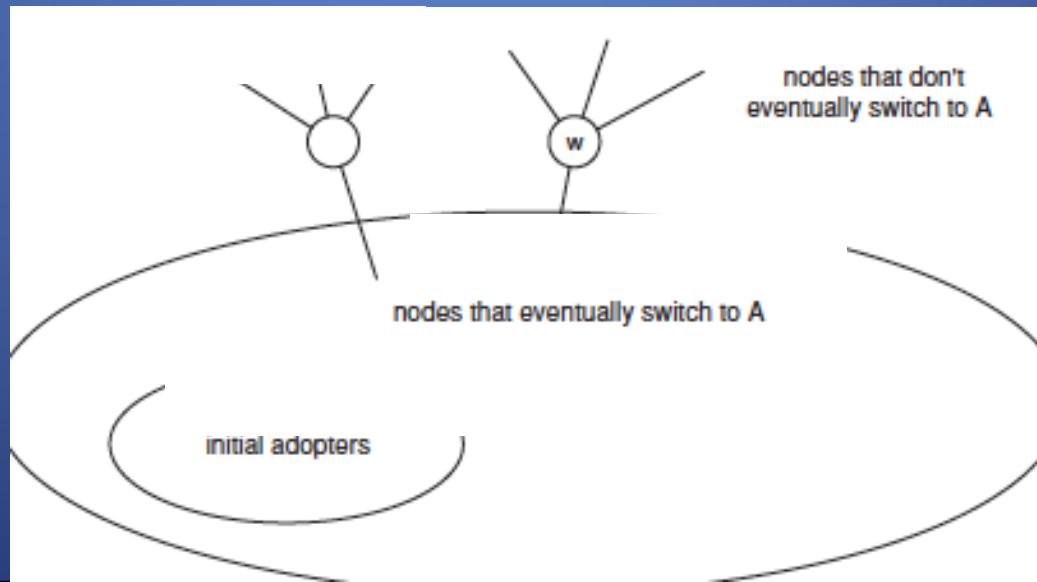
证明——聚簇是级联的障碍

- 剩余网络中，节点改用A的门槛值为 q ，设包含一个密度大于 $1-q$ 的聚簇，证明该聚簇中没有节点会采用A。
- 反证法，设该聚簇中第一个改用A的节点为 v ，则在 v 决定采用A之前，其邻居中有大于 q 占比的已经采用了A。
- 根据定义， v 所在聚簇密度大于 $1-q$ ，即它的邻居中至少有 $1-q$ 占比的邻居也在该聚簇中（采用的B），这说明它的邻居中用A占比不可能大于 q 。
- 矛盾，假设不成立。



证明——聚簇是级联的唯一障碍

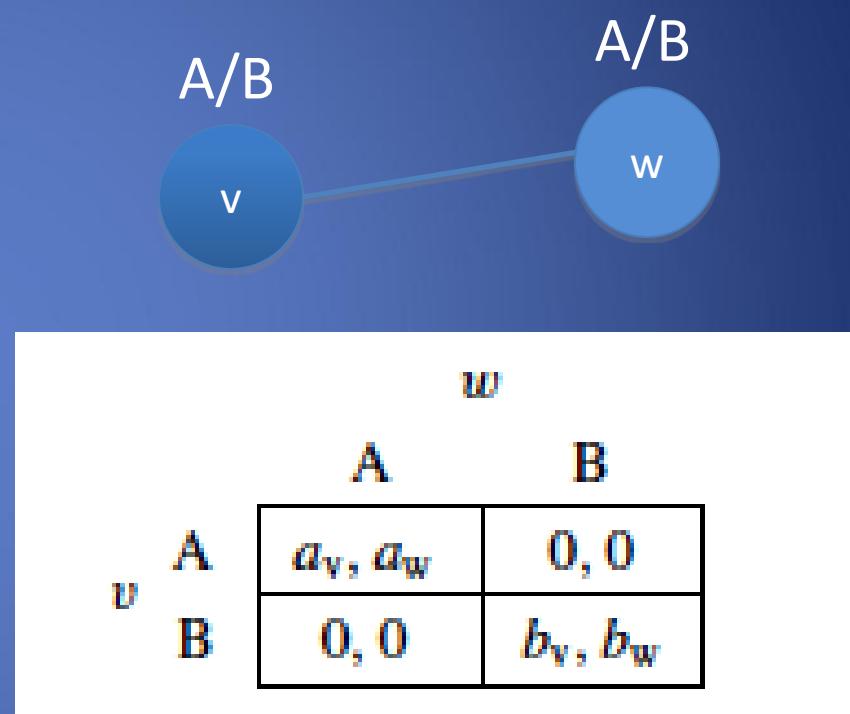
- 需要证明只要一个初用集不能形成完全级联，剩余网络中必然存在一个密度大于 $1-q$ 的聚簇
- 设S是最终未转用A的节点集，证明S密度大于 $1-q$ 。考虑S中任何一个节点w，最终没有转向A，说明它邻居中只有占比小于q的采用A，即大于 $1-q$ 的邻居使用B，因为整个网络中所有使用B的节点都在S中，因此S是一个密度大于 $1-q$ 的聚簇



基本级联模型的扩展—异值门槛

- 异值门槛：假设网络中每个人对行为A和B的估值不同，节点v采用A的回报为 a_v ，采用B的回报 b_v 则：

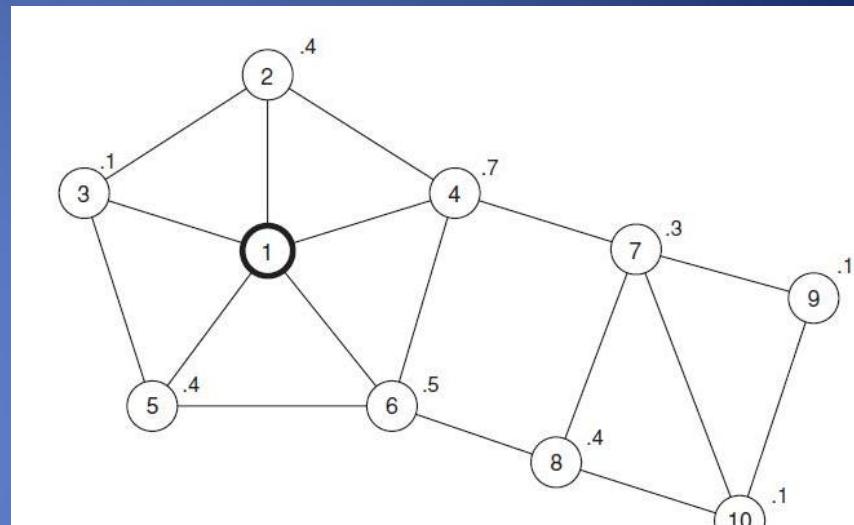
$$q_v = \frac{b_v}{a_v + b_v}$$



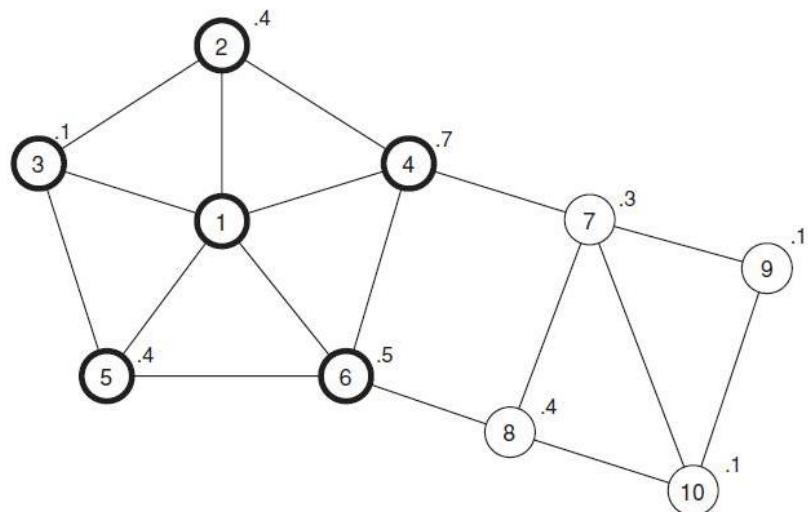
- 传播过程中每个节点根据自身的门槛值决定是否采纳新事物， $p_v \geq q_v$ ？

异值门槛：同样简洁的结论？

- 异值门槛阻塞聚簇
 - 节点集，其中任何节点 v 至少有 $1-q_v$ 占比的邻居也在该集合中
- 直觉上，这就是阻止传播（或者防止一个节点被感染）的条件
- 在此定义下，也有类似的形成一个完全级联的充要条件。



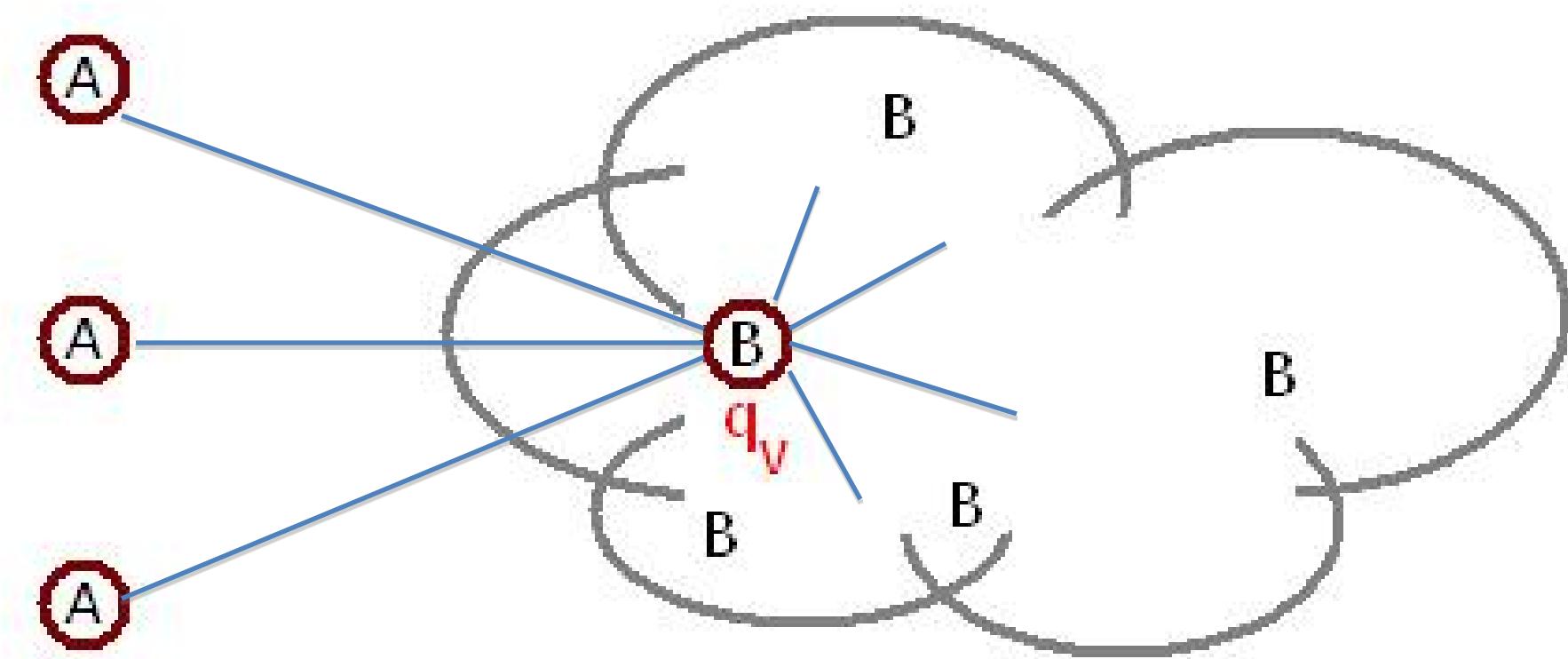
(a)



(b)

每个节点对诱惑的抵御能力不同

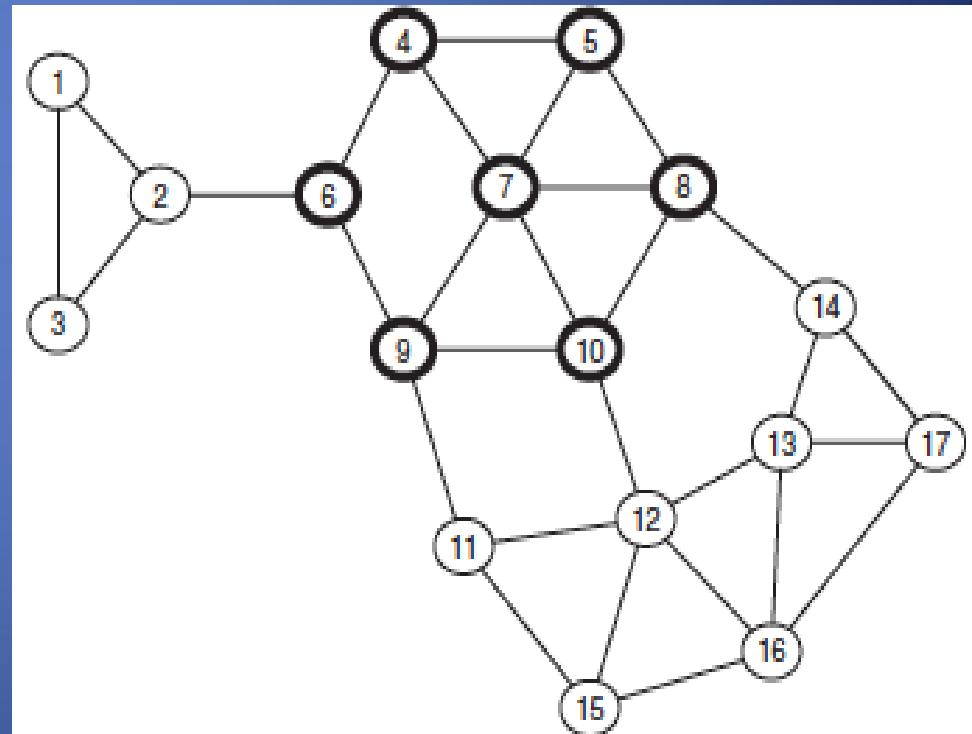
(注: q_v 只是基于对两种事物优势的判断)



- 进一步考虑不同的节点对邻居的不同影响力?

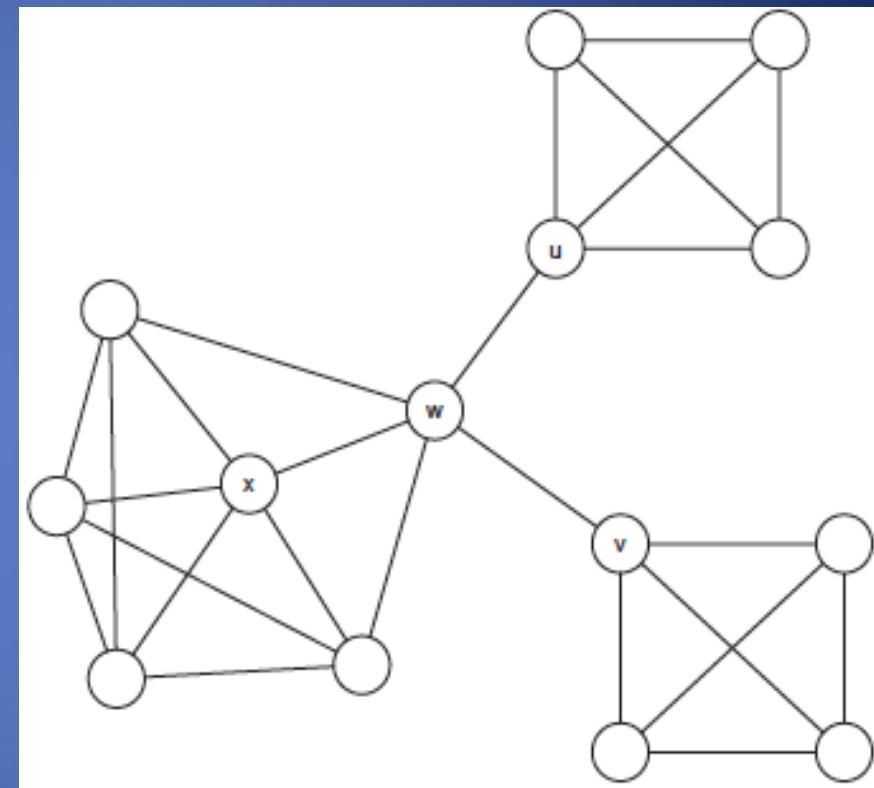
一些思考：病毒式营销 (viral market)

- 如果A是一种新的产品，如何突破聚簇的阻碍，形成A的完全级联？
 - 提高 A 的质量，从而减少 q 值，如 q 减至 $1/5$ ，A形成完全级联
 - 在密度大于 $1-q$ 的聚簇中选择关键人物，利用其他外部因素使其转向 A，致使 A 在这些区域继续传播



一些思考：弱关系的作用

- 弱关系：对于信息传播有很重要的作用，如在线视频、各种开放信息.....
- 但对于一个行为的传播，特别是风险较高的行为（例如罢工、示威），弱关系作用较小（一端聚簇的阻碍）



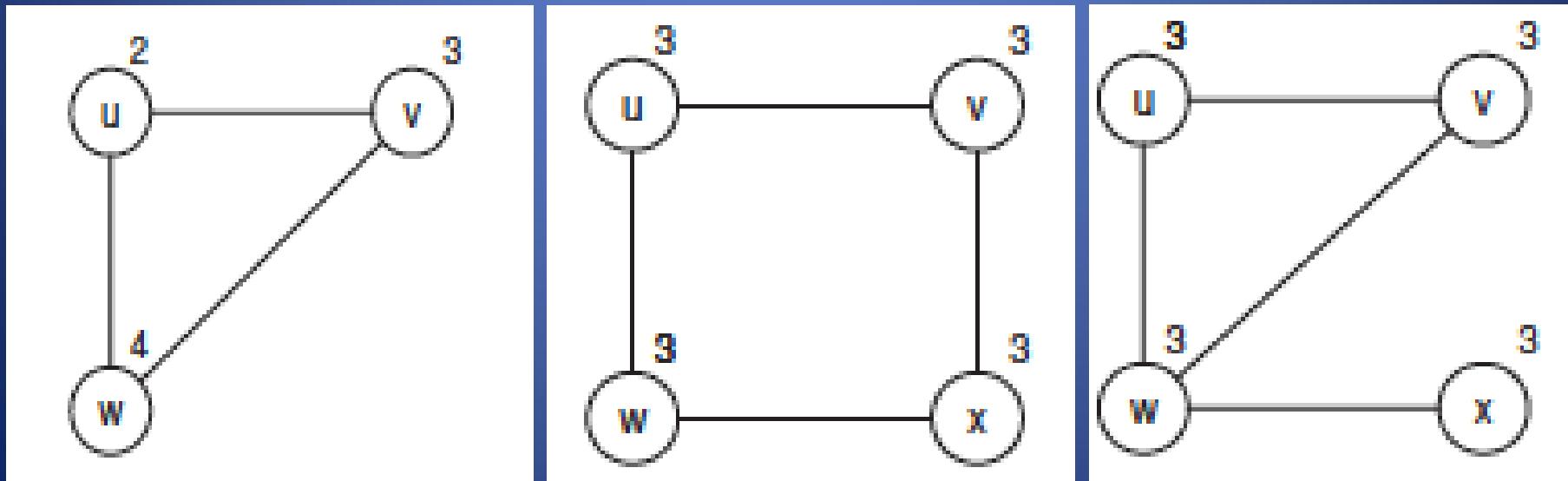
人们倾向于在肯定有足够的其他人参加的情况下才去参加。

一些思考：集体行动中知识的作用

- 集体行动：一项活动只有足够多的人参加才会得益
- 每个个体只能了解其邻居的门槛值，并不能完全了解网络中其他人的情况
- 知识在集体行动中的作用：分析社会网络结构如何影响人们对集体行动的个体决策
 - 每个节点都了解网络的整体结构
 - 每个节点都了解邻居（强关系）的决策门槛值
 - (1) 我是否参加？(2) 行动会不会发生？

一个个体在集体行动中决策的例子

- 每个节点代表公司的一个副总裁，每个人需要决定在第二天的董事会上是否参与对总裁发难的行动
 - 每人都了解网络结构
 - 每人了解邻居的门槛值，但不了解非邻居的门槛值
 - 节点上的数字表示各自的门槛值



宣传作用的双重性

- 大规模的宣传（营销活动等），对于一个人来说，不仅使他对内容有所了解，而且使他潜意识地认为还有许多其他人也了解这件事了
- 于是，当他倾向于接受这件事，就会认为还有许多人也会接受这件事；如果他对这件事持反对态度，也会认为还有许多人也持反对态度
- 当这种态度涉及到行动（涉及利益或者代价），上述认识就会被考量到其中。

要点小结

- 在网络中，新生事物的扩散（被接受的程度）受三种因素的影响
 - 新生事物的优势
 - 网络结构
 - 初用者的选择
- 个体决策与总体状态关系的又一实例
 - 每个节点独立、并行决策
 - 总体达到某种状态的条件
- 公共知识、门槛值、集体行动

作业

- 第19章 3,5题

传染病的网络传播模型

(第21章)

疾病传播模型：基本要素

接触网络

传染概率

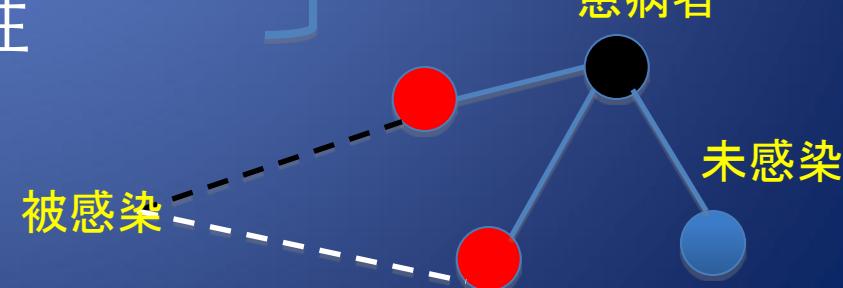
疾病特征

关心的问题：

- 多大范围（规模）？要不要采取措施？
- 许多流行病会自生自灭，例如季节性的
- 多长时间？会不会反复（“死灰复燃”）
- 对患者影响的严重性（生命、损伤、不适）

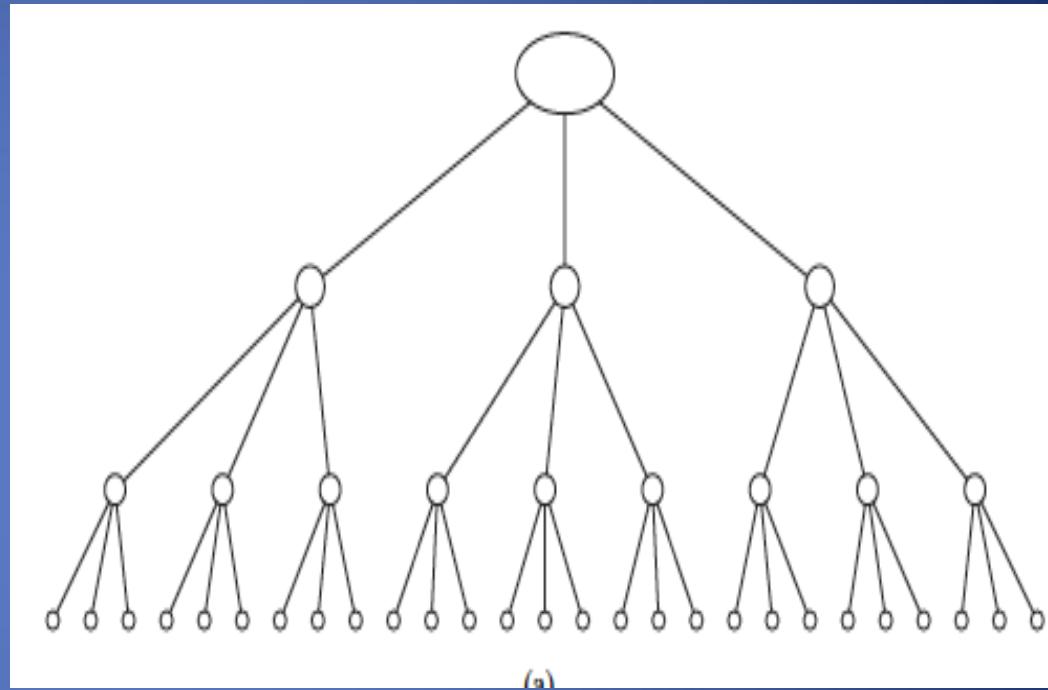
流行病传播与社会行为传播的差别

- 已经讨论过的传播（影响）模型
 - 总体模型——总体效应
 - 网络模型——朋友关系网络结构
 - 基于门槛值的模型
- 流行病传播特点
 - 传播原因：传染
 - 传播途径：接触关系网络结构
 - 传播方式：具有随机特性



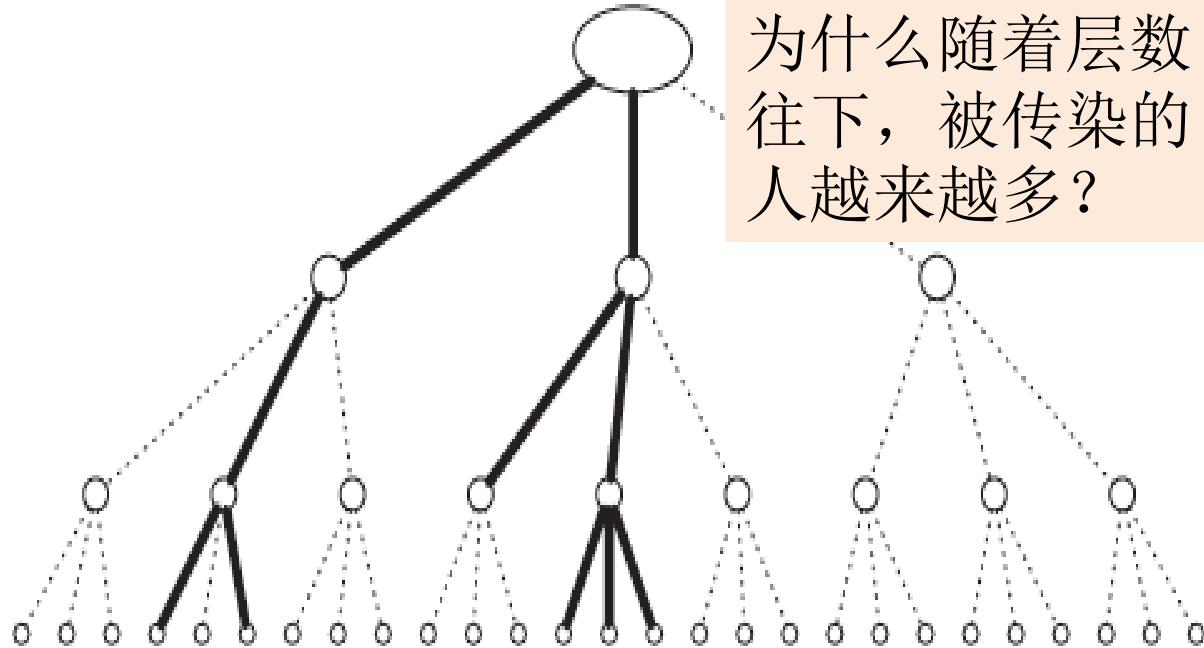
分支过程：一个最简单的疾病传播模型

- 分支过程形成树结构接触网络，最初一个人携带病菌进入人群，以一个独立的概率 p 传染给遇到的每个人。假设疾病感染期间每个人遇到 k 个其他人



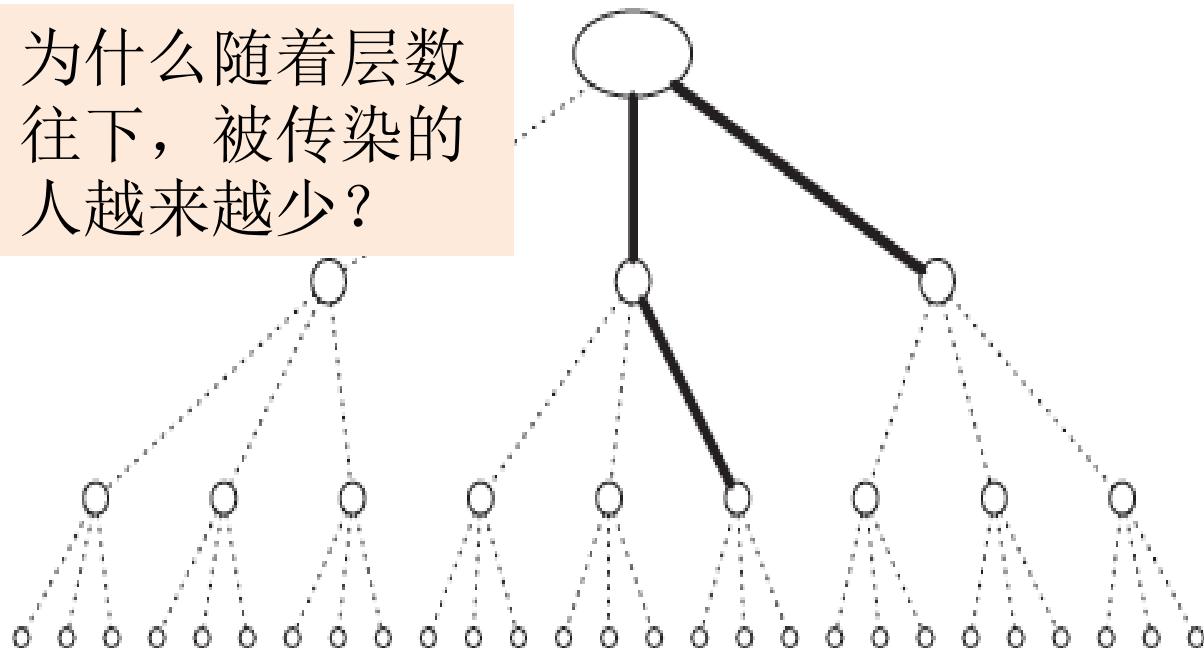
如此简单、“不实际”的模型有什么用？

传染过程停止的条件？



(b)

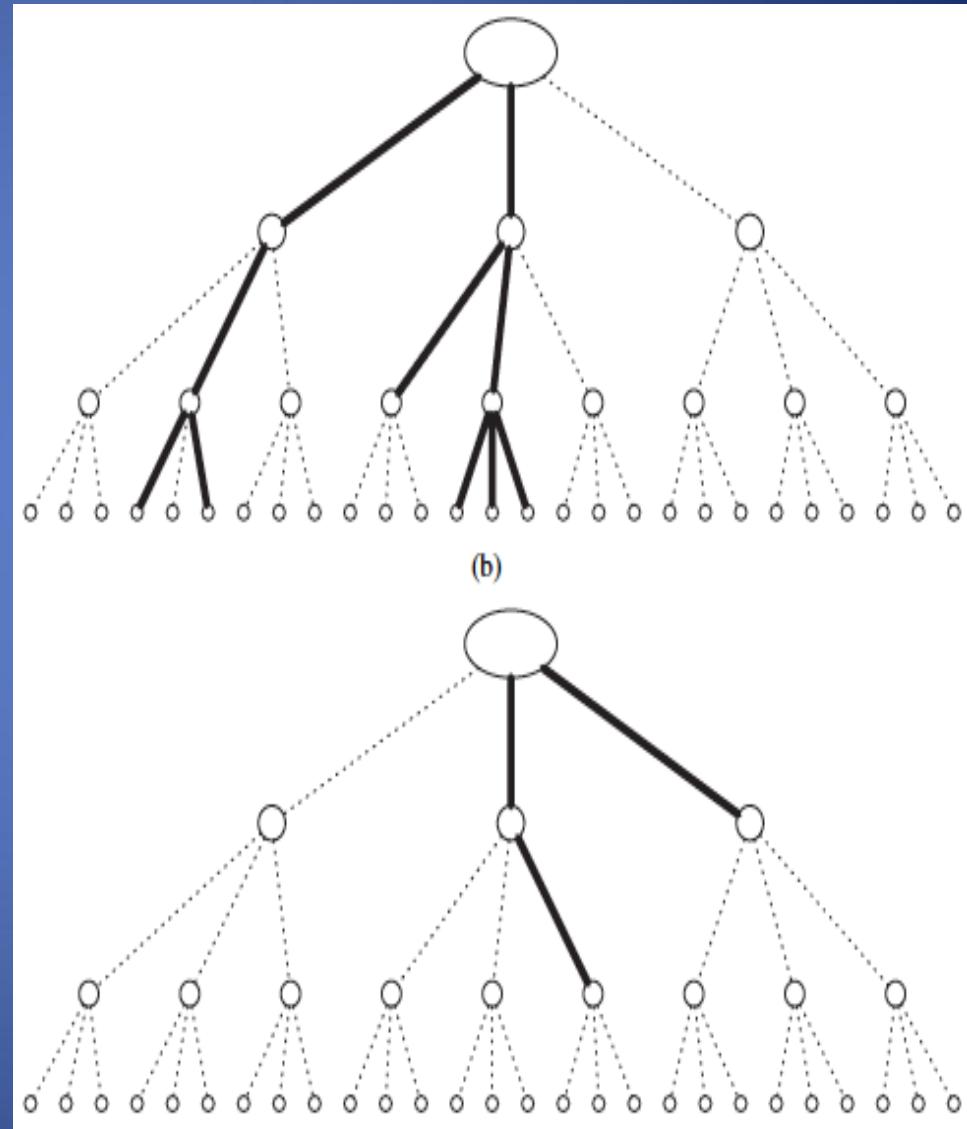
为什么随着层数
往下，被传染的人
越来越多？



为什么随着层数
往下，被传染的人
越来越多？

基本再生数 (reproductive number)

- 基本再生数：由单一个体引起的新发病例数期望值，记为 R_0 , $R_0=pk$
- 断言：如果 $R_0 < 1$ ，则疾病将以概率1在有限的过程后消失。如果 $R_0 > 1$ ，则疾病持续在每一波以 >0 的概率至少感染一个人



基本再生数的特点

- 因为 $R_0=p*k$, 当 R_0 接近于1时, 稍微改变接触人数k或传染概率会对结果产生很大影响。 R_0 接近于1时有一个“刀刃”特性
 - 如果 R_0 略低于1, 稍微增加传染概率p; 结果可能会使 R_0 最终高于1, 造成一个突然的疾病爆发
 - 略微减少疾病的传染性, 可以将 R_0 减小到1以下, 消除疾病大范围流行的风险
- 结论: 当门槛值 $R_0=1$ 附近, 社会应该付出努力减小基本再生数(隔离, 减少接触就是措施之一)
- 实际的疾病传播要复杂得多, 分支结构没有考虑三角关系, 弱连接……

SIR疾病传播模型

- SIR模型可以描述更一般性网络结构的传播情况，接触网络中的节点在流行病传播过程中经历三个状态(susceptible,infectious,removed)：
 - 易感状态(S)：易感者，容易被一个患病邻居节点感染
 - 传染状态(I)：患病者，以一定的概率将疾病传染给处于易感状态的邻居
 - 移出状态(R)：痊愈并免疫者，当一个节点经历了完整的病期康复后，就不再会被传染，也不会再对其他节点造成威胁，相当与从接触网络中移出了

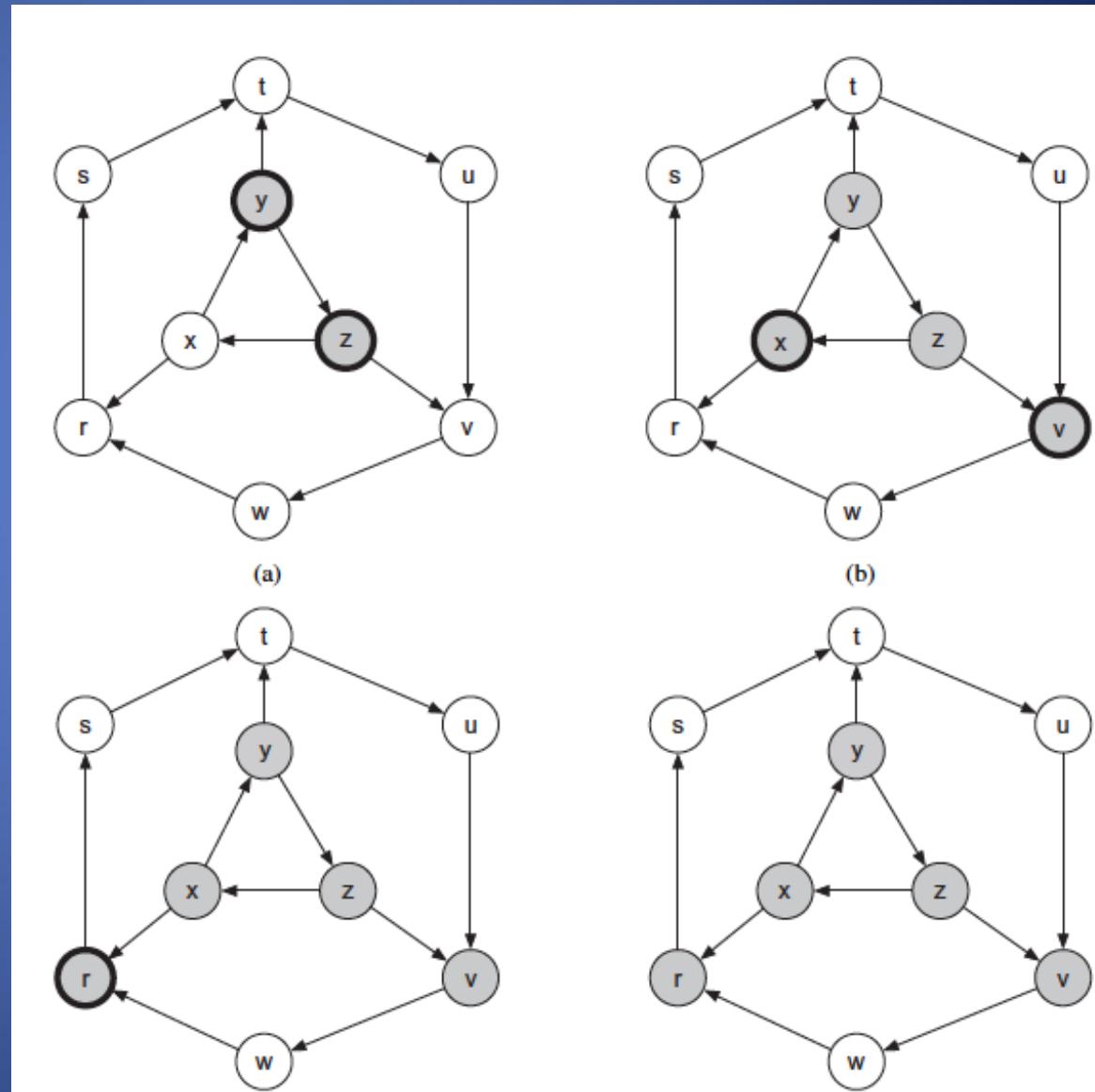
SIR模型描述

- SIR模型流行病的传播过程取决于接触网络结构，以及两个量值： p （传染的概率）和 t_I （传染期的长度）
 - 最初，一些节点处于状态 I，所有其他节点处在状态 S。
 - 每个进入状态 I 的节点 v 在固定的步骤 t_I 期间具有传染性。**在 t_I 的每一步**， v 以概率 p 将疾病传染给它的处于易感状态的邻居
 - 经过 t_I 步后，节点 v 进入移出状态 (R)，成为网络中的惰性节点，不会再被传染，也不会向其他人传播



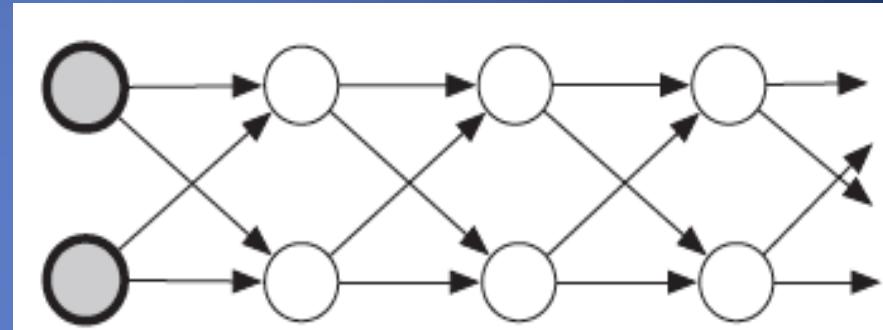
一个 $t_i=1$ 的SIR模型示例

- 接触网是一个有向图
- 图中有三种不同的节点
 - 白色（易感）
 - 粗边界的灰色（传染）
 - 细边界的灰色（移出）



非树型网络的 R_0 不再有“刀刃”结论

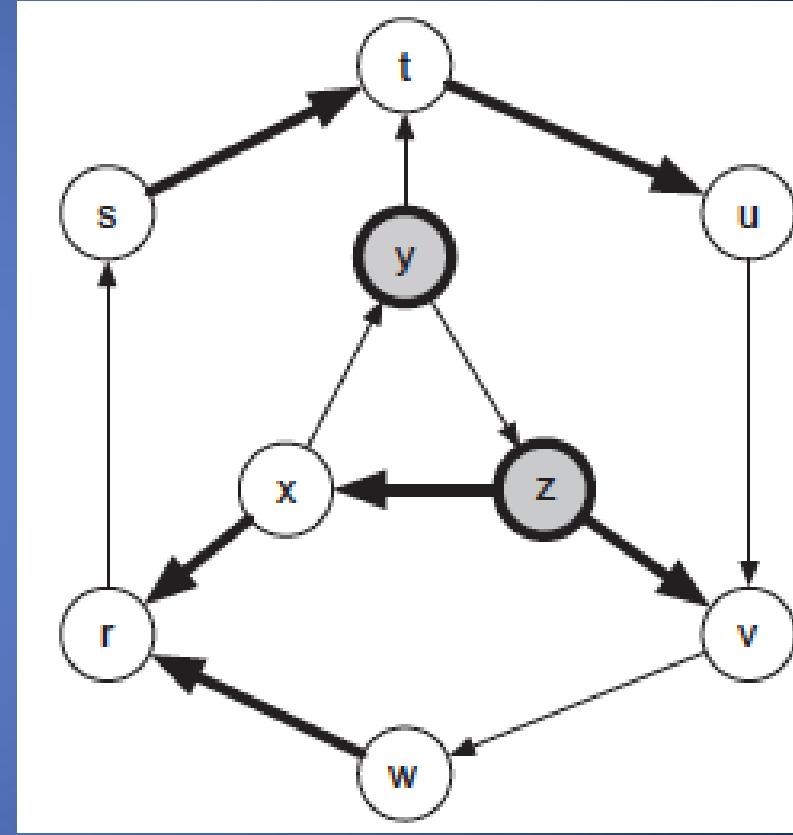
- 网络由每层两个节点组成向右无限伸延, $t_l=1$, 传染概率 p 为 $2/3$, 最左边的两个节点是最初传染者
- R_0 是由一个节点造成的新病例, 预期为 $4/3 > 1$
- 每条边不传播疾病的概率为 $1/3$, 所有四条边都不会传播疾病的概率为 $(1/3)^4$, 每一层成为最后一层(疾病传染终止层)的概率至少是 $1/81$



这种结构相当于形成了一个路障, 疾病无法超越它向远方传播

SIR模型（运行）的静态图示

- 接触网络每条边预先投掷硬币（按照给定传染概率 p ），投掷硬币成功的边定义为 **开放边**；其余的边定义为 **阻塞边**
- 利用开放边和阻塞边描述流行病传播过程：最终受感染的节点正是那些能够从最初感染节点沿着在网络中随机选定的 **开放边**能够达到的节点
- 意义：有利于对随机现象的观察



图中，粗黑线为开放边。注意到不是所有开放边都起作用的（例如 $w \rightarrow r$ ）

SIR模型的形式化描述

- 易感人数、传染人数、移出人数分别为 $S(t), I(t), R(t)$, 占比分别为 $s(t), i(t), r(i)$
- β : 个人单位时间与其他人平均有效接触次数
- γ : 单位时间治愈人数
- SIR模型描述为:

$$\frac{ds(t)}{dt} = -\beta \cdot i(t) \cdot s(t)$$

$$\frac{di(t)}{dt} = \beta \cdot i(t) \cdot s(t) - \gamma \cdot i(t)$$

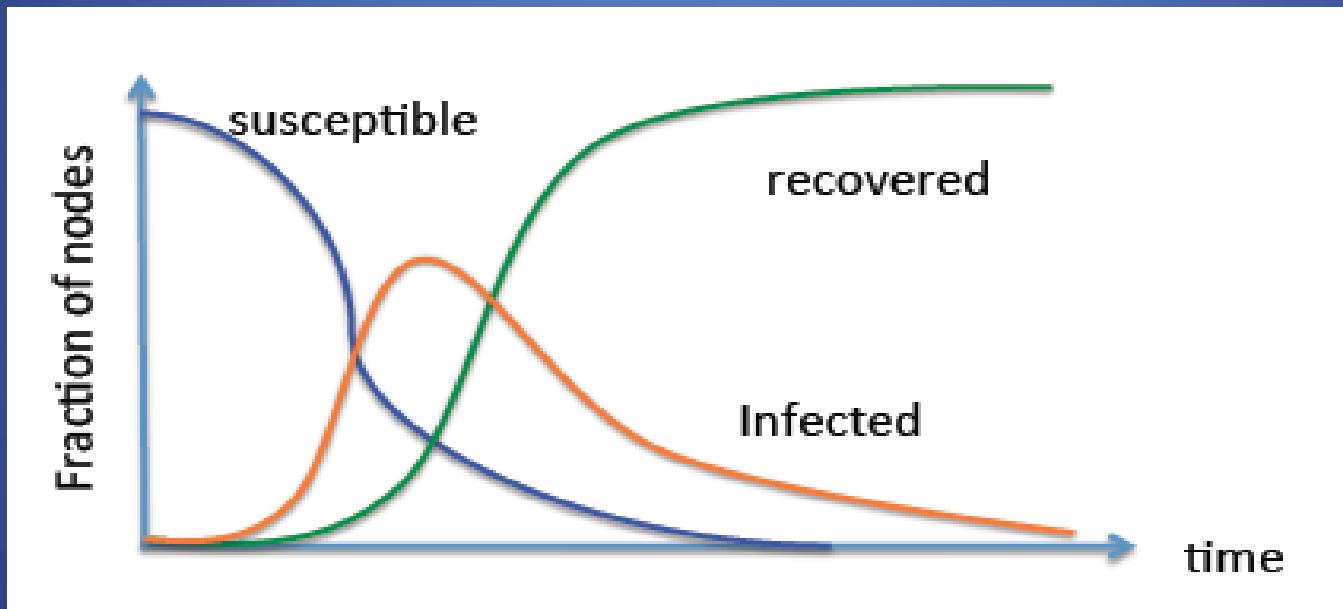
$$\frac{dr(t)}{dt} = \gamma \cdot i(t)$$

$$s(t) + i(t) + r(t) = 1$$

来自最早提出SIR模型的文献:
[Kermack W O, McKendrick A G. Contributions to the mathematical theory of epidemics, part I[J]. Proceedings of the Royal Society of London A, 1927, 115: 700–721]

一个给定 β 和 γ 的SIR时间曲线

- 设 $\beta=1, \gamma=0.4, s(\text{at start})=0.99, i(\text{at start})=0.01, r(\text{at start})=0$

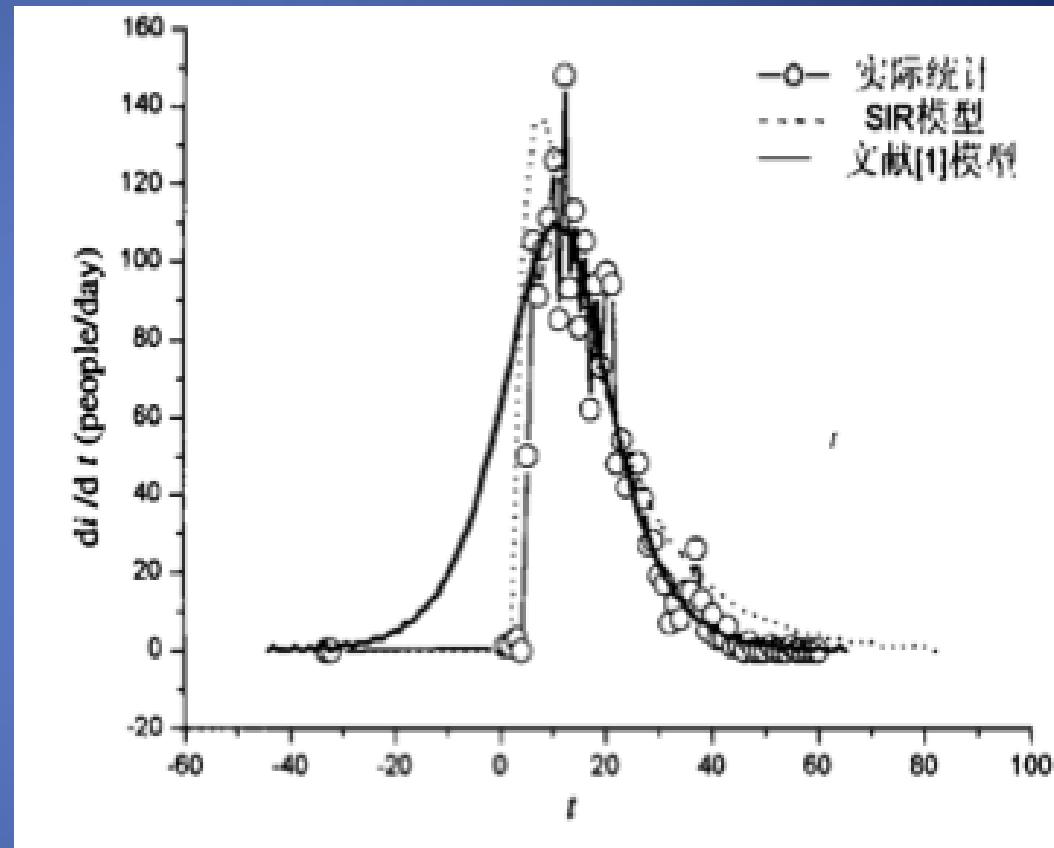


- 疾病什么时候消失呢？

取决于： β/γ 。如果发病率高于痊愈率，则疾病会持续，否则疾病将消失

利用SIR模型分析SARS传播

- 利用SIR模型中 di/dt 方程模拟北京在2003年4月18日到6月16日的每天新增病例
- 模拟中使用的参数是：
 $\beta = 0.148$, $\gamma = 0.85$, 初始值为: $s(0) = 1199$,
 $i(0) = 1$ (这里总人数应该看作是接触到病源的易感、传染、免疫的人数之和,而不是北京市总人口)



SIS流行病传播模型

- SIR模型只适用于一次患病终身免疫的情况。SIS模型描述了更一般的疾病传播。节点结束传染状态后再回到易感状态，节点在 S 和 I 两种状态中交替，因此得名**SIS模型**
 - 最初，一些节点处于状态 I，其余节点处于状态 S
 - 每个进入状态 I 的节点 v 在 t_i 期间内具有传染性，在 t_i 的每一步，节点以概率 p 将疾病传染给其处在易感状态的邻居
 - 经过 t_i 步后，感染节点不再具有传染性，返回状态 S

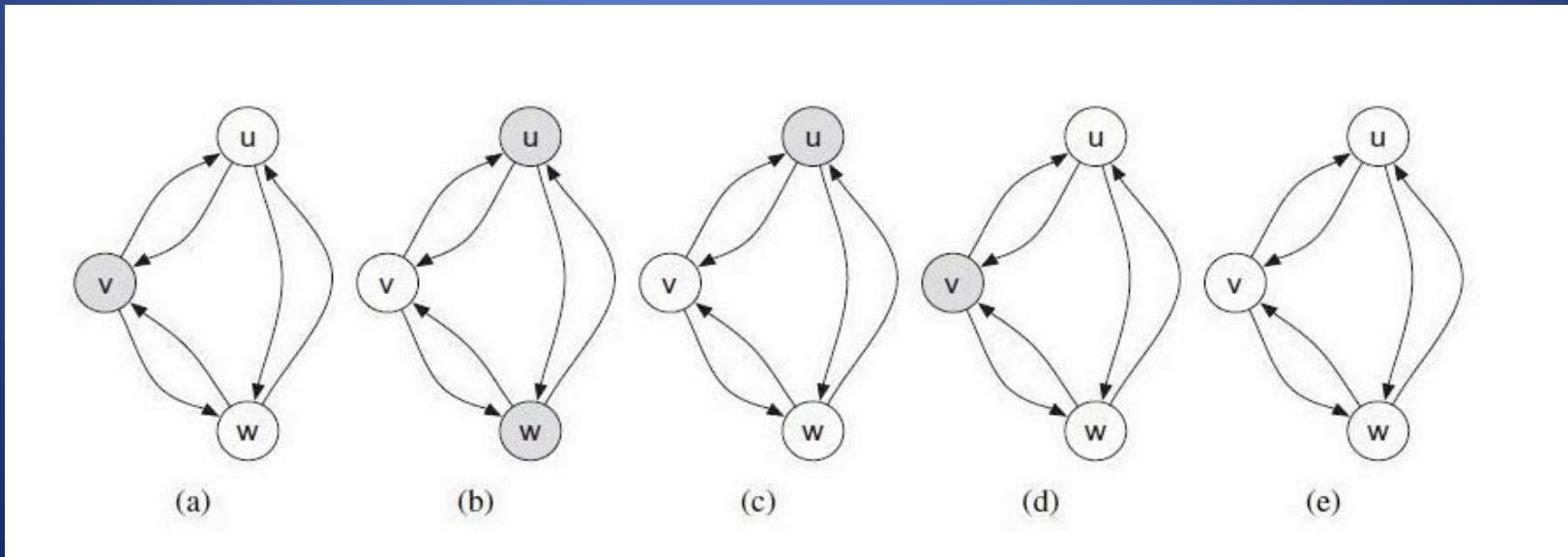


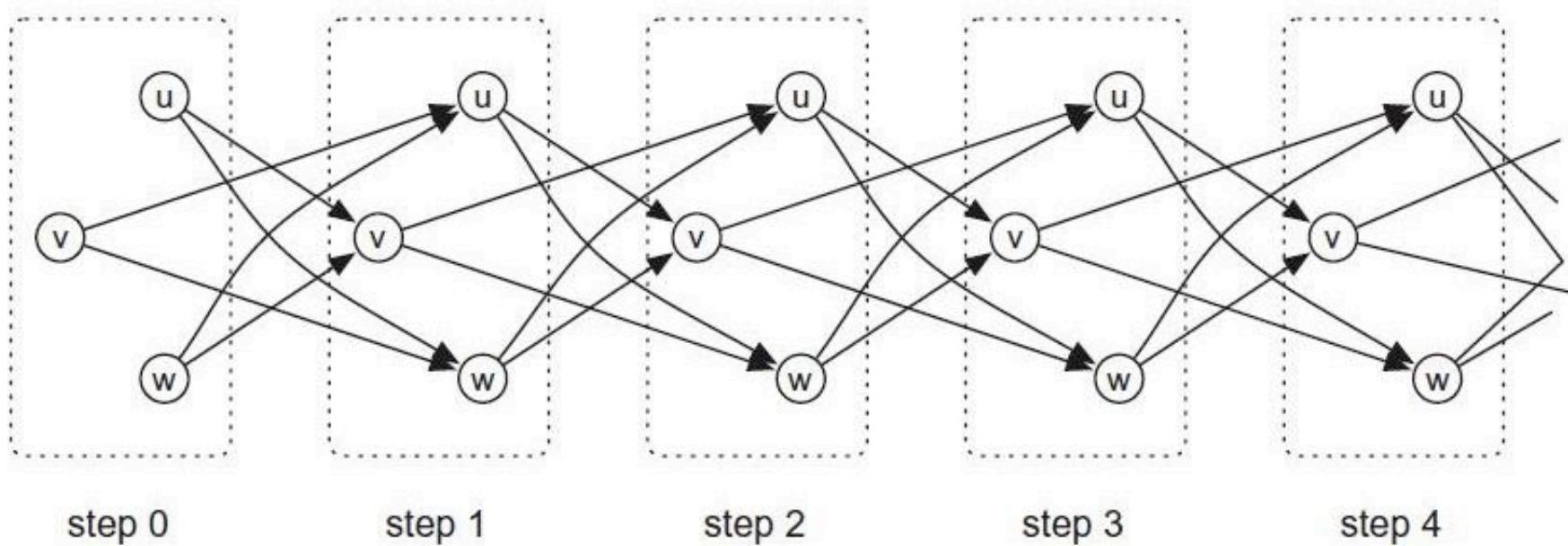
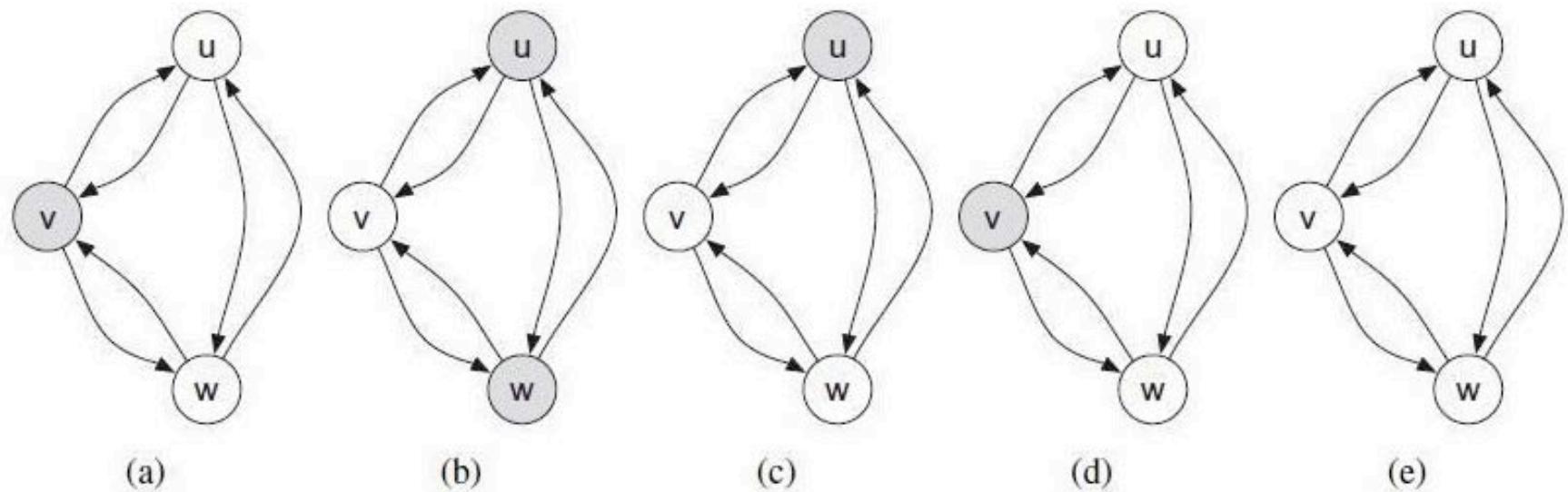
SIS模型的特点

- SIS模型刻画的流行病的生命周期会比SIR模型的要长
- 基本再生数具有刀刃特点：
 - 可以证明，对于某些类型的接触网络，存在特定传染概率 p ，在其附近，网络中的SIS流行病在一边表现为“快速消失”，在另一边则会“持续一个很长的时间”。该传染概率临界值以微妙的方式依赖于网络结构。

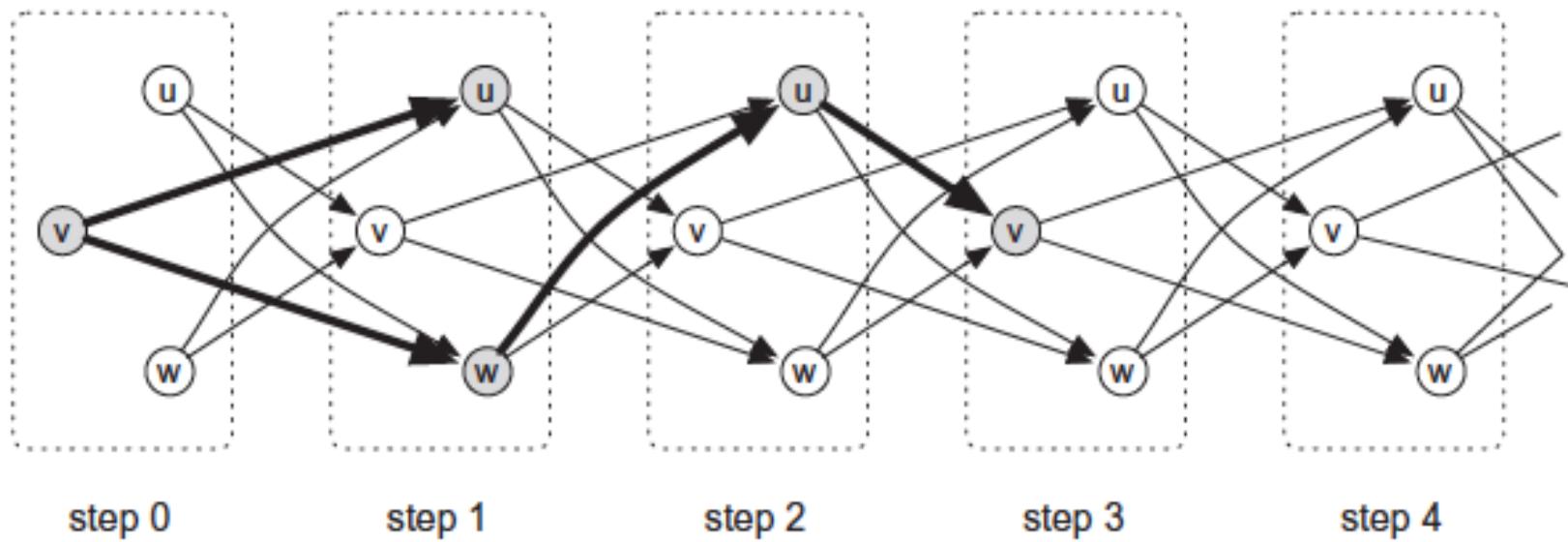
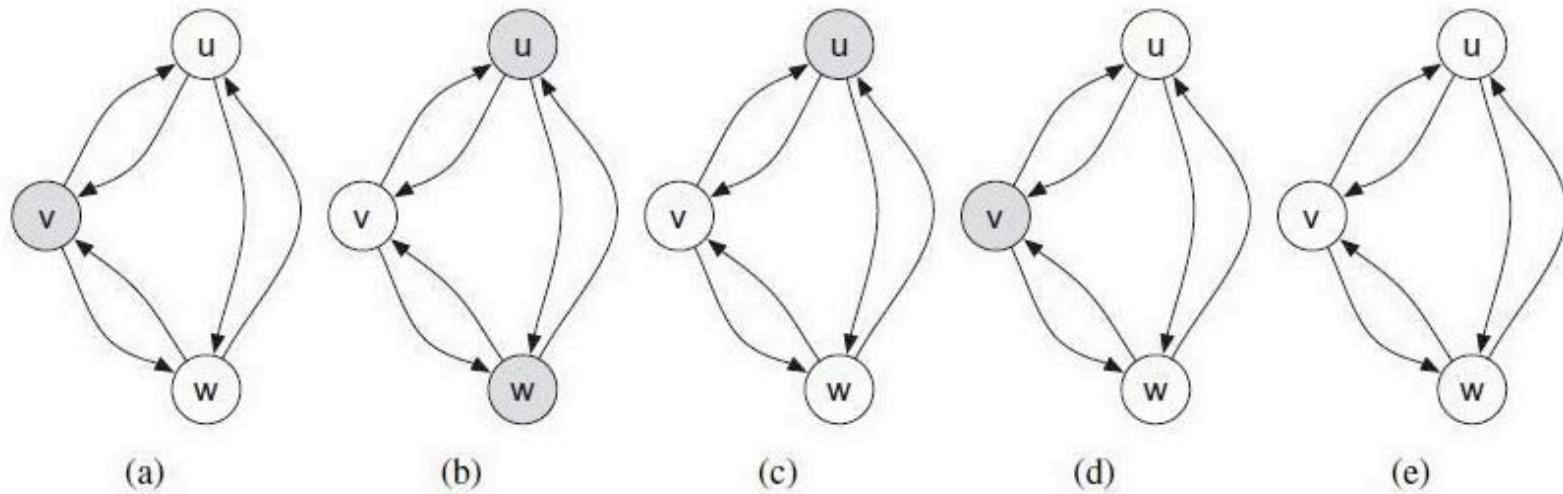
SIR与SIS的关系

- SIS模型可以看成是SIR模型的特例
- 设 $t_i=1$ ，将患病节点 v 在每一个时间步视为“不同的个体”，可以将SIS模型表示为SIR模型
- 在每个时间周期 $t=0、1、2、3……$ ，为每个节点分别创建一个节点副本，对于原网络中 v 连接到 w 的边，复制 t 时刻扩展网络中 v 到 $t+1$ 时刻 w 的边



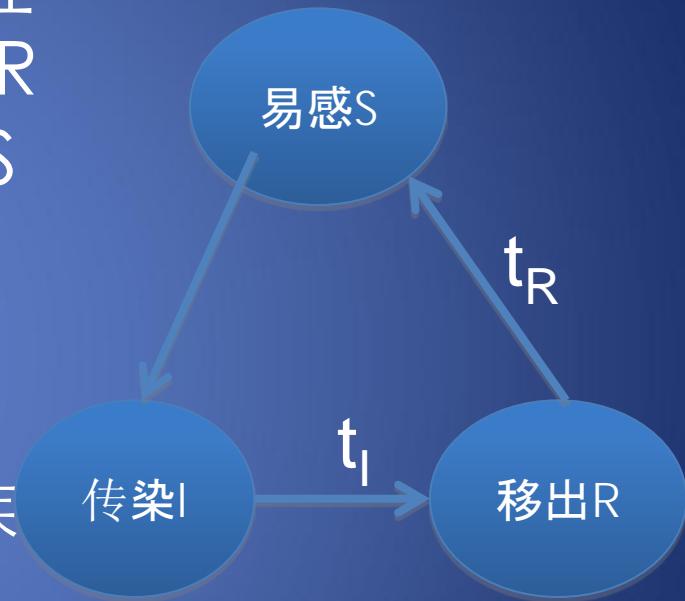


SIR与SIS的关系



SIRS 流行病传播模型

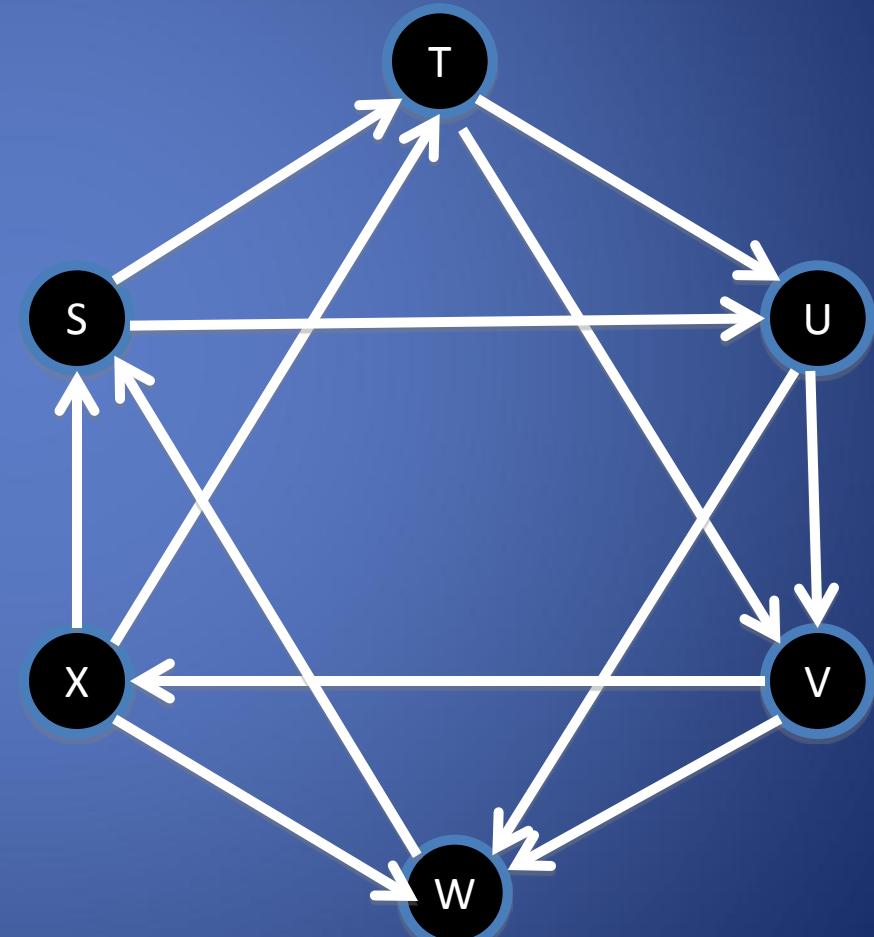
- 结合SIR和SIS，疾病有暂时免疫特性
：受传染节点恢复后回短暂地经过R状态，然后再次进入S状态——SIRS模型
 - 开始某些节点在I状态，其他在S状态
 - 进入I状态的节点 v 在固定时间 t_I 内具有传染性，在 t_I 的每一步以概率 p 将疾病传播给易感状态的邻居
 - 经过 t_I 步后，节点进入状态R，经过 t_R 后再次回到S状态



疾病传播过程受量值 p 、 t_I 以及 t_R 的影响，同时受到网络结构的影响

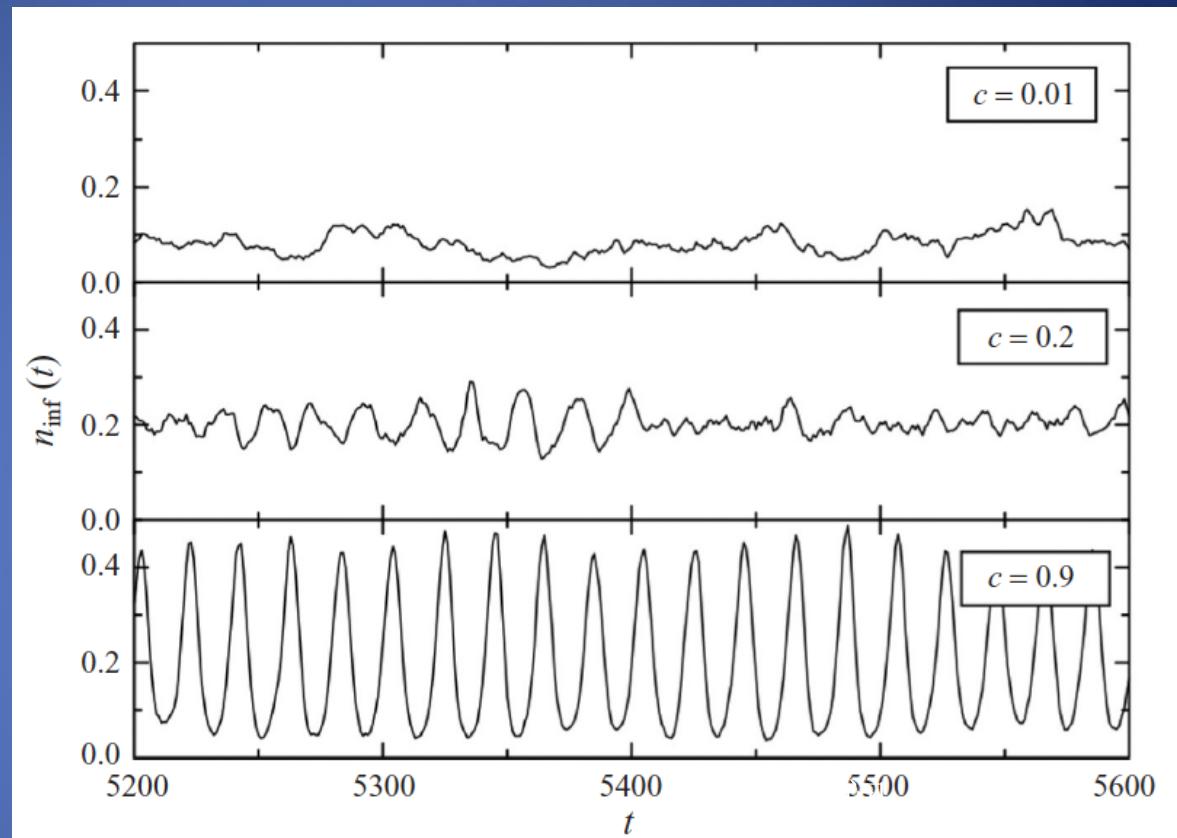
SIRS模型用于研究疾病传播的震荡效应

S	T	U	V	W	X
S	I	S	S	S	S
..	R	I
..	S	R	I
..	..	S	R	I	..
..	S	R	I
I
R	I
S	R	I
..	S	R	I
..	..	S	R	I	..
..	S	R	I



Watts-Strogatz小世界模型上运行SIRS模型

- $T_I=4, T_R=9$
- $N_{\text{inf}}(0)=0.1$
- 模拟 $c=0.01, 0.2, 0.9$ 三种情况



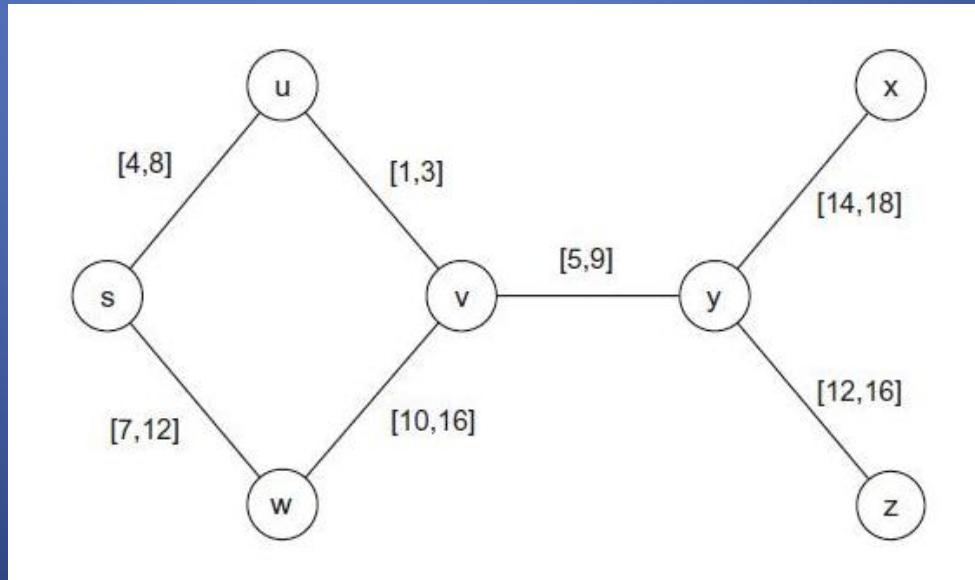
结论：当弱连接概率较高时，所有节点的感染周期趋于同步。同样的时间点被感染，持续相同的时间，进入隔离，之后又同时被感染

为什么会有同步震荡？

- 当 p 较小时： $0 \sim 0.5$ ，网络呈现较高的有序的聚集特点，疾病周期会在一个局部区域内完成，之后进入免疫期（不活跃期）
- 当 p 值较大时： >0.5 ，弱连接打破了有序的聚集区域，刚刚痊愈的个体可能又通过一个弱连接被再次感染。因此整个网络成同步状态

模型中可以考慮的一些其他因素

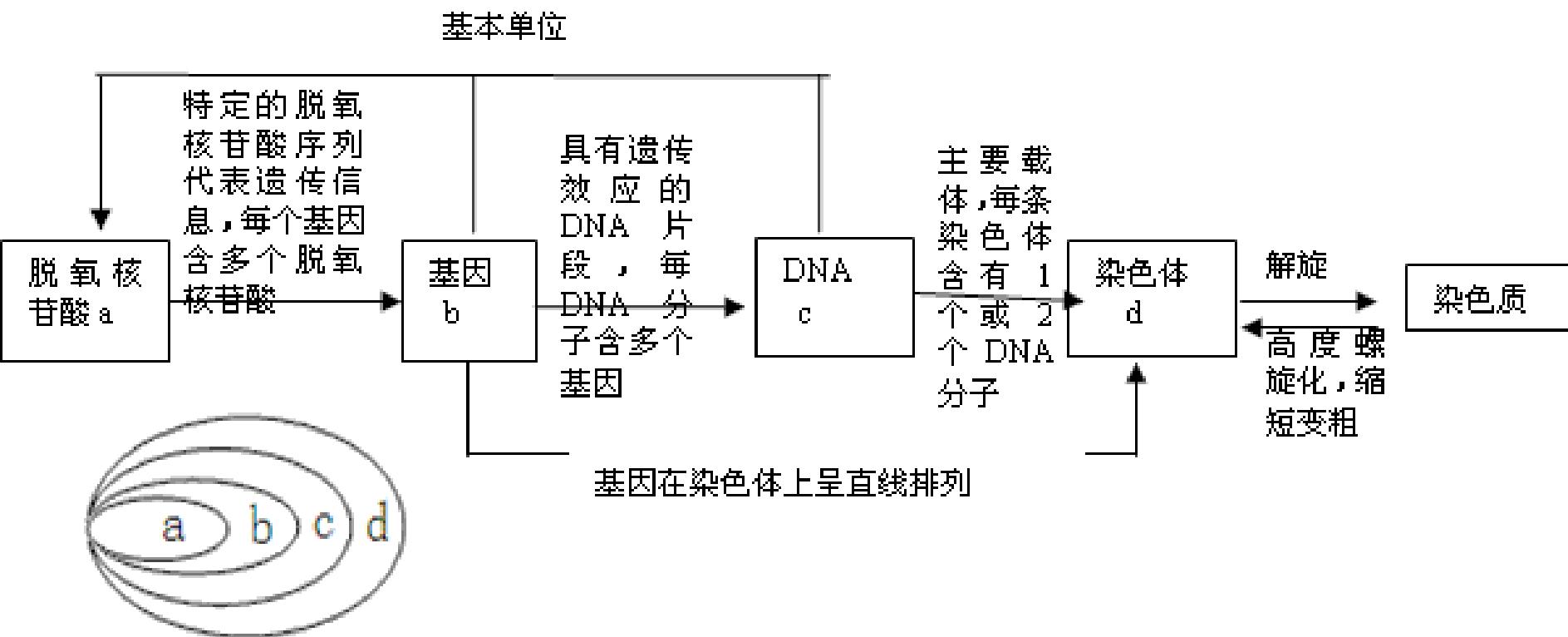
- 感染期不同时段传染性可以不同
- 不同个体传染性的不同
- 患病者可能在传染期间不同的阶段恢复
- 接触网络的边具有时间特性
-

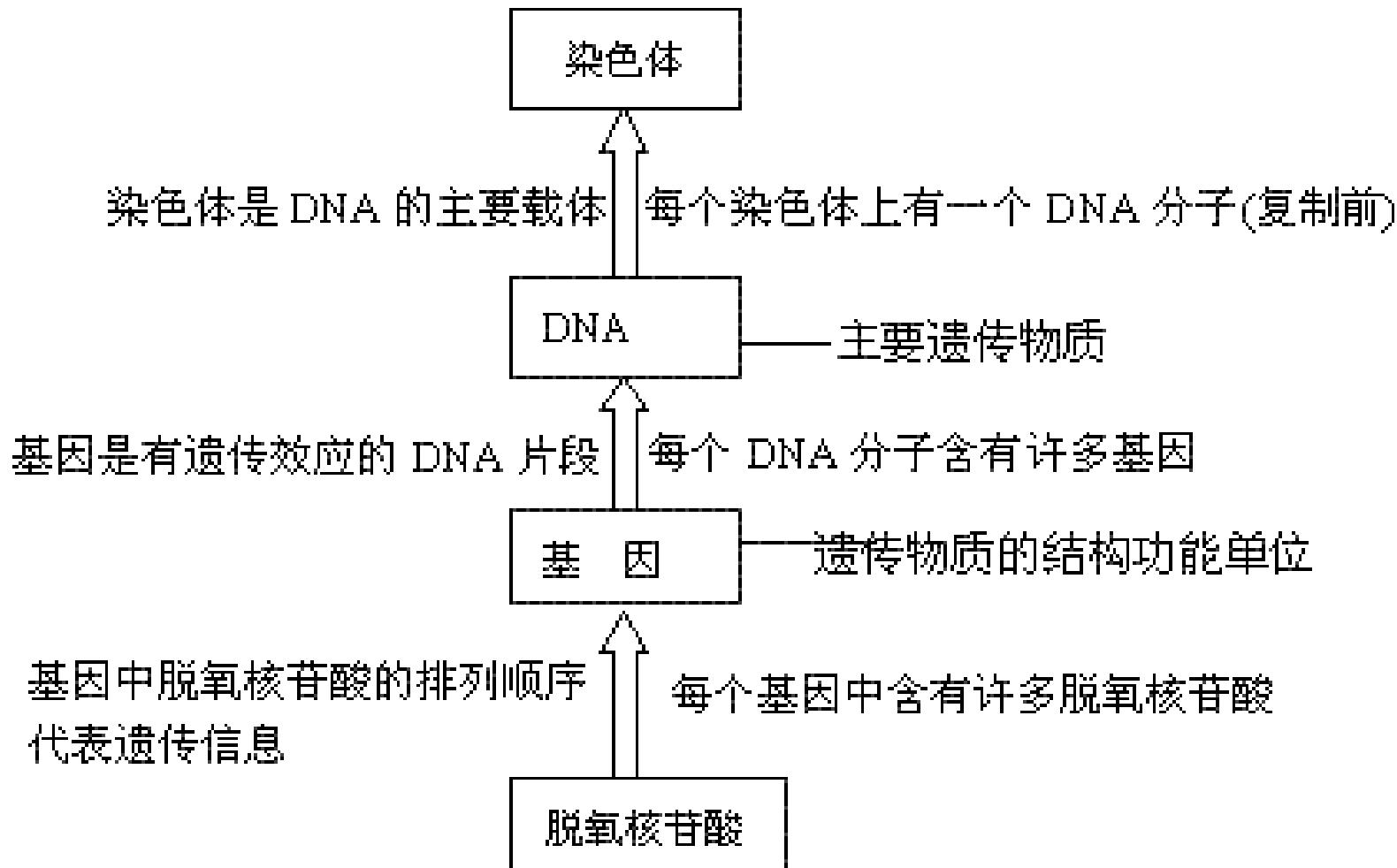


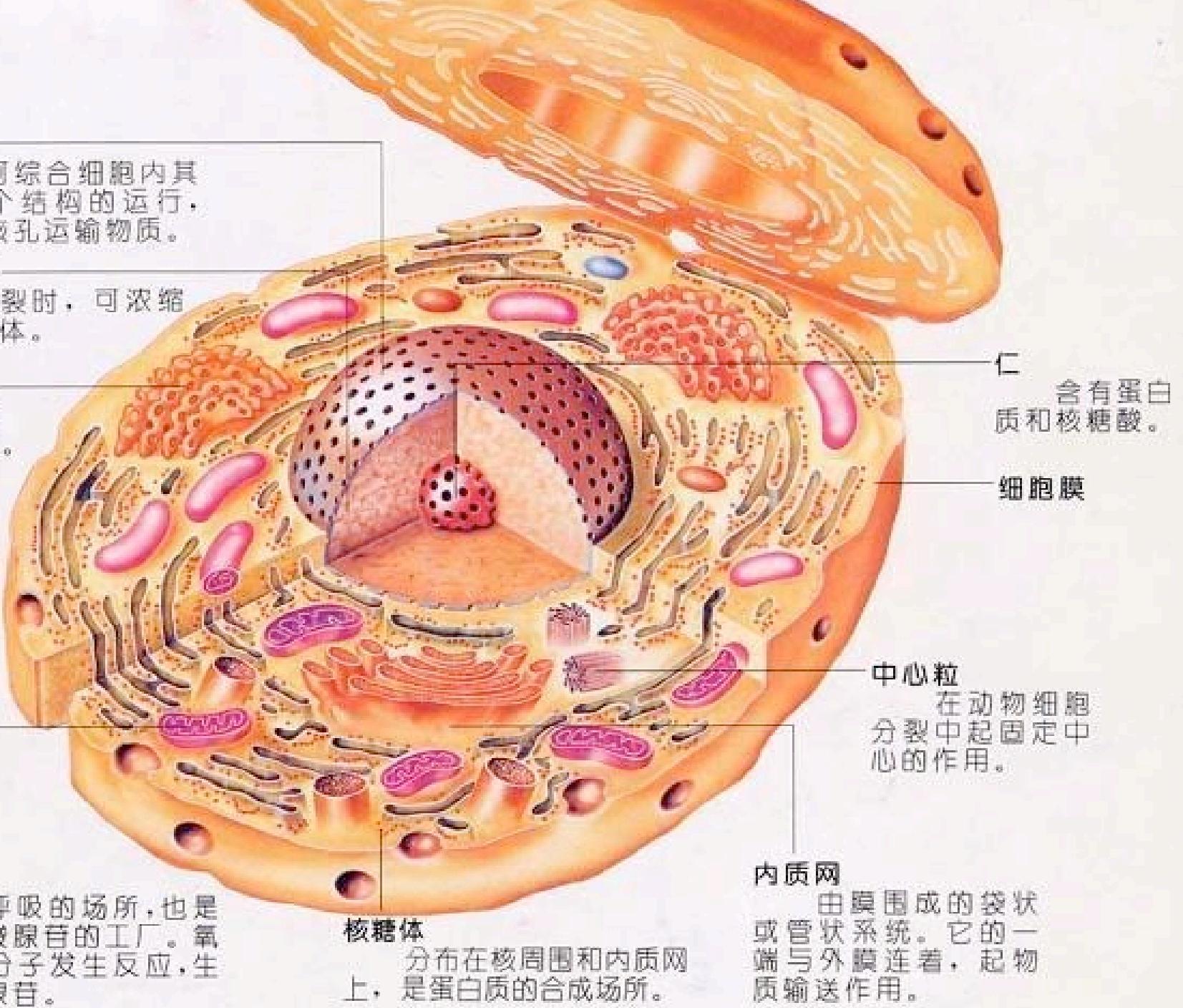
一个生物学基因遗传的研究

- 线粒体夏娃理论：1987年，Rebecca Cann、Mark Stoneking和Allan Wilson 在自然杂志上发表的一篇论文。
- 每个人都有一条母系祖先的线索，称为母亲的血统。所有这些血统源自10万至20万年前（14万年）的一个单一女子，可能是在非洲，她是我们所有人的母系祖先。

一点生物遗传知识



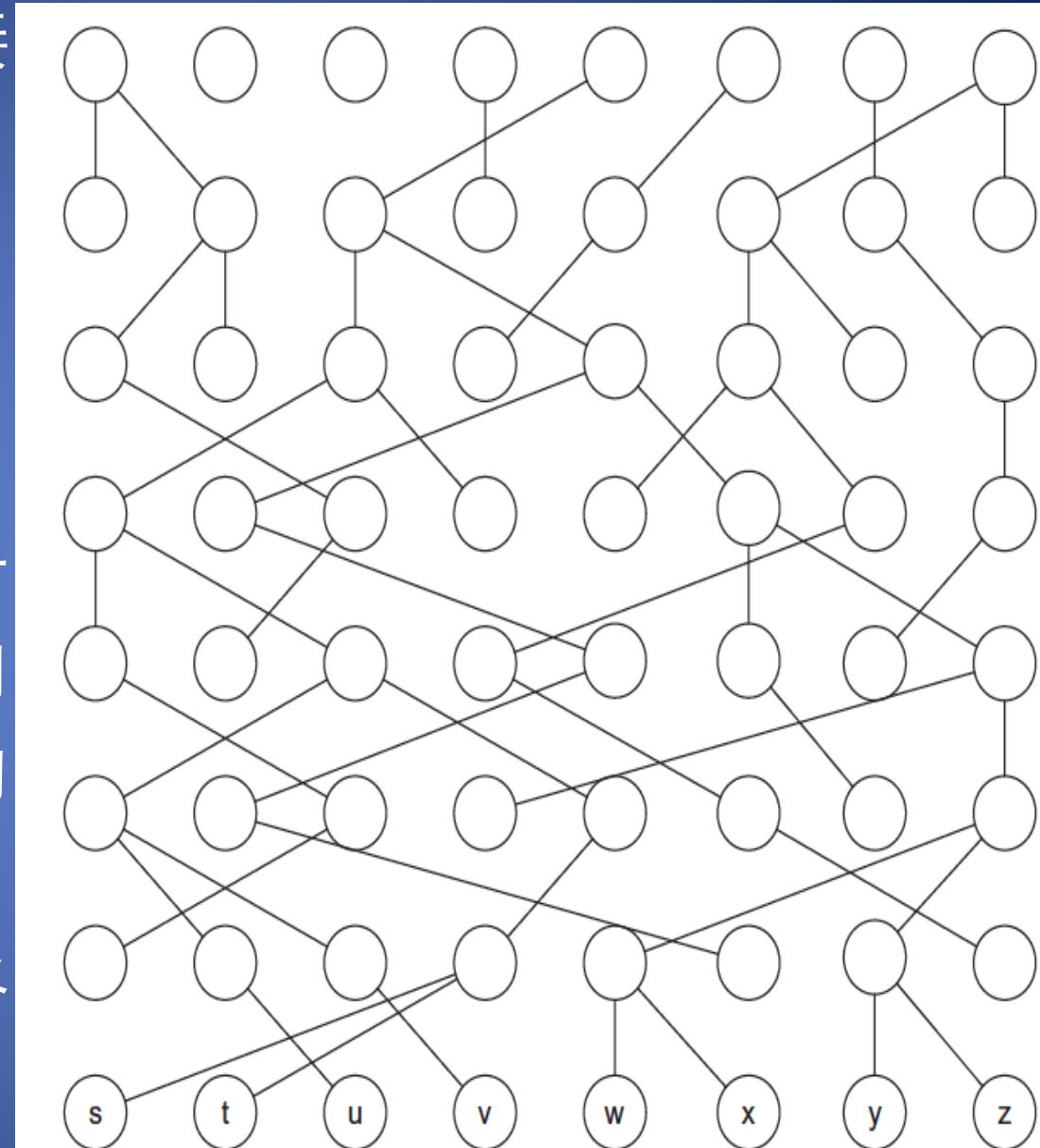




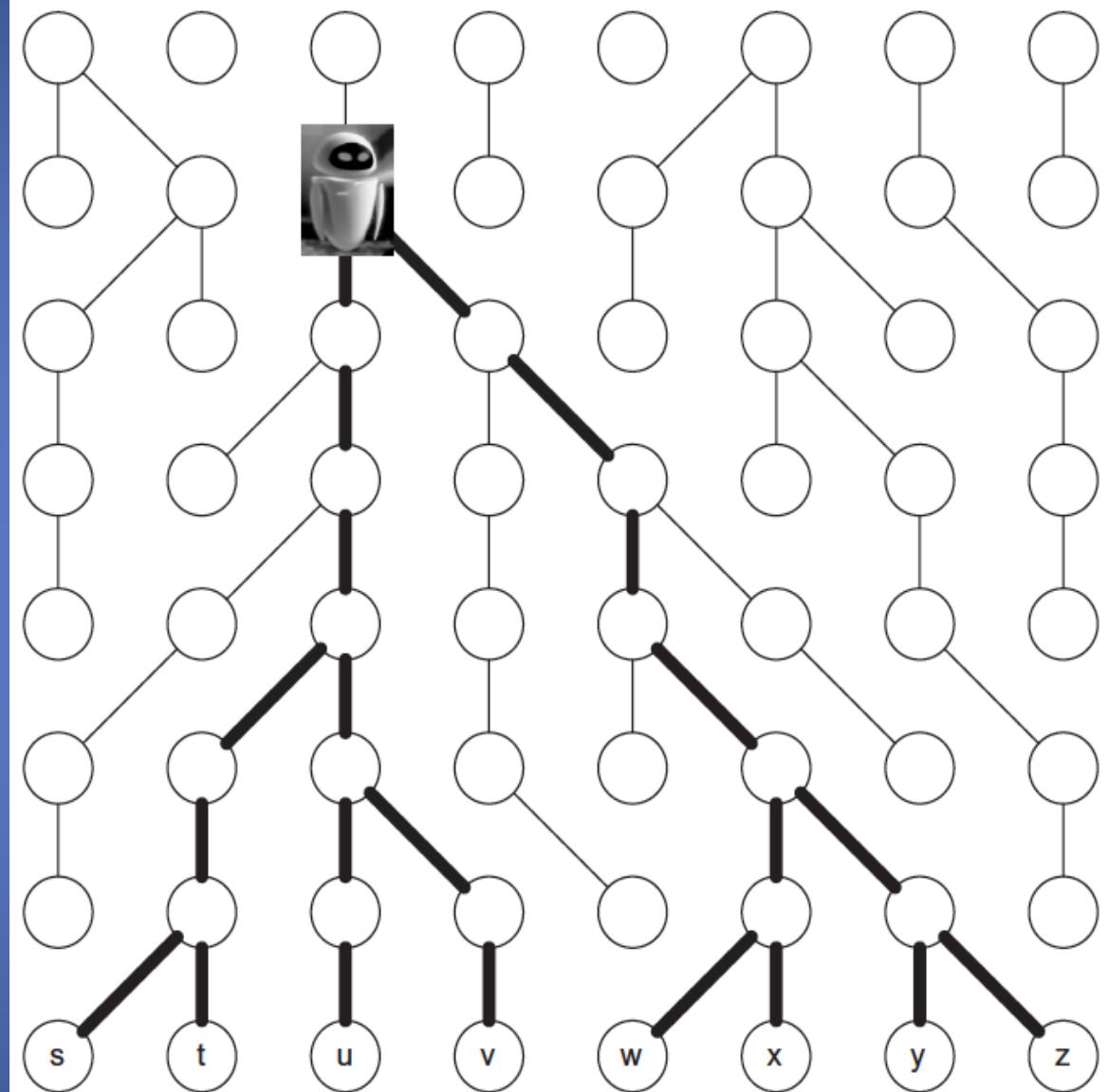
线粒体遗传的网络结构动力学解释

- Wright-Fisher单亲祖先模型：
 - 每一代的人口数量维持在固定的N
 - 新的一代是由当前一代N个个体产生总共N个后代
 - 每个后代由一个单亲产生，由当前一代中独立地均匀随机选择，当前一代特定的个体可能会有多个孩子

- 每个个体都向上连接到自己的母亲；
- 时间自上向下运行，最底层表示当前有N个个体
- 从底部开始，任何一个个体都可以沿着向上的边追溯他／她的单亲血统
- 能否说明这些边最终将汇聚到一点？



- 简单的概率分析，可以发现当前一定数量的个体在向上追溯单亲祖先的过程中会不断发生汇聚（发现同一个母亲），于是独立的线索越来越少，最终汇聚到一点
- 这个距当代最近的汇聚点就是线粒体夏娃



基因遗传学的解释

- 当时肯定同时生活着许多女人和许多男人，只不过她们的线粒体基因没有遗传到现在。
- 不仅是线粒体基因存在这种情形，任何一种基因，只要把它们的突变过程追溯得足够远，总能找到一个共同祖先，这些祖先生活的时间可能各不相同，甚至未必是人，有的可能要追溯到人进化出来之前的某个动物

Wright-Fisher单亲祖先模型的意义

- 可以分析单亲物种的遗传特点
- 可以分析双亲物种中那些只由单亲遗传的现象，如线粒体DNA
- 分析社会继承行为（包括学术继承等）
-

小结

- 疾病传染的简单网络模型：
 - 分支结构、SIR、SIS、SIRS
 - 不同模型下的不同现象与问题
- 生物遗传过程也可以自然看成是一个网络，从中可能发现出人意料的现象

作业

- 第21章，练习第1题。

匹配市场

(第10章)

背景问题

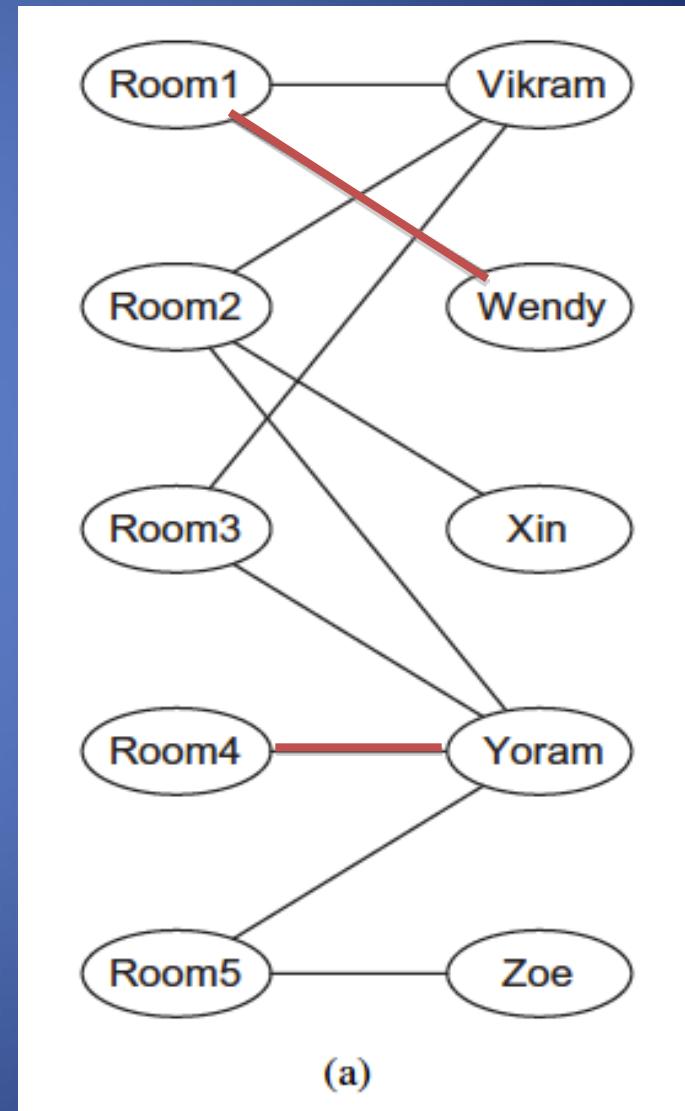
- 市场的匹配：买卖双方，数量相等，直接交易，买方有偏好，达成完全交易的条件？在不同价值期望下，大家能否都满意？

学习要点

- 匹配定理—达成完全交易的结构性条件
- 清仓价格—社会最优会“自动”实现！

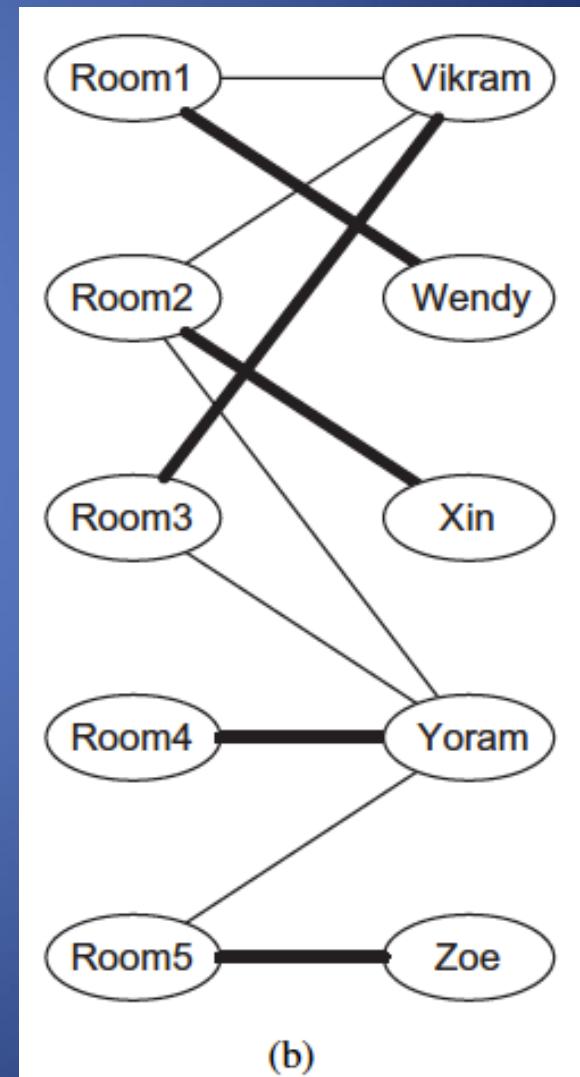
买卖双方，数量相等，直接交易，买家偏好

- 最方便的方法，就是将买家对卖家偏好关系用**二部图**刻画，讨论其性质
- 例子：宿舍供给与学生对宿舍需求的关系图
- 匹配：图中的一组边，其中所涉及节点没有重复



匹配的覆盖程度

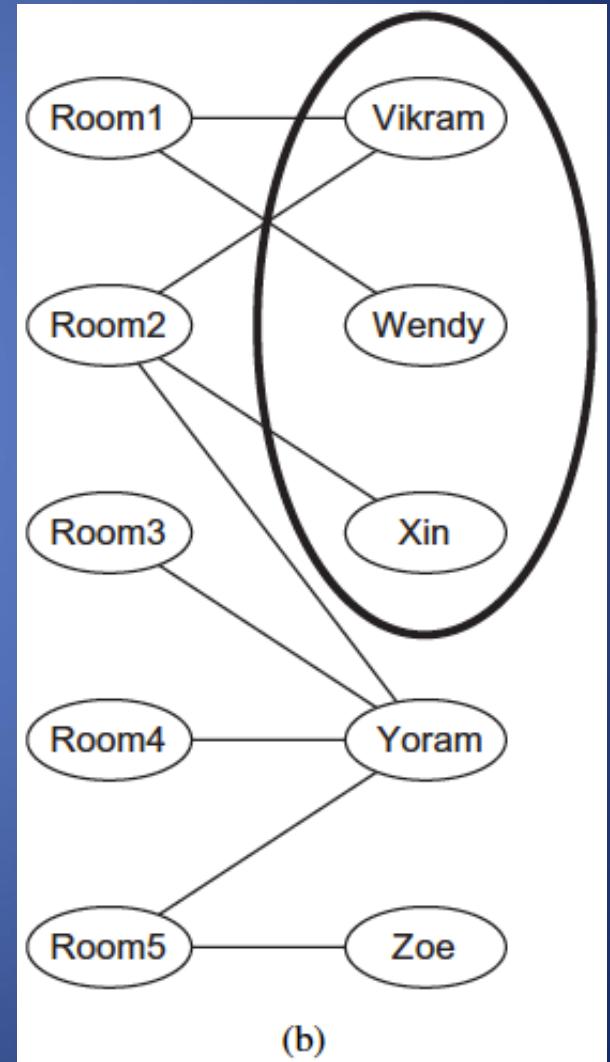
- 前面定义的匹配，仅仅表现了匹配的特征要求，即节点不能重复，并没有说涉及节点的个数
- 如果在两边相等的二部图中供需两组节点均完全出现在一个匹配中，则称该匹配为完美匹配（Perfect Matching）——见右图例



什么情况下，存在一个完美匹配

受限组

- 并不是所有二部图都含有完美匹配
- 受限组 (**Constricted Sets**) : 图中一边的一组节点, 其个数多于它们邻居节点的个数之和 (不算重复)
 - 相当于说供需之间在数量上不等



在市场双方数量相等情况下

- 若存在一组买方多于相应的卖方（即买方受限组），则不可能实现完美匹配
 - 亦即不能达成完全交易
- 若存在一组卖方多于相应的买方（即卖方受限组），则不可能实现完美匹配
 - 但可否达成完全交易？

若存在一组卖方受限组，则存在一组买方受限组，反之依然。

匹配定理

- 一个两边节点相等的二部图存在完美匹配，当且仅当不存在受限组。
- 若存在受限组，则不存在完美匹配——这个很容易看到；
- 若不存在完美匹配，则存在受限组——这不是显然的，需要详细证明（深度学习材料）。

引入买方价值估计

- 对一个物品的需求程度，与需方对其价值的估计密切相关，引入**价值估计**，有助于更加一般地探讨供需之间的匹配关系。

Valuations
12, 2, 4
8, 7, 6
7, 5, 2

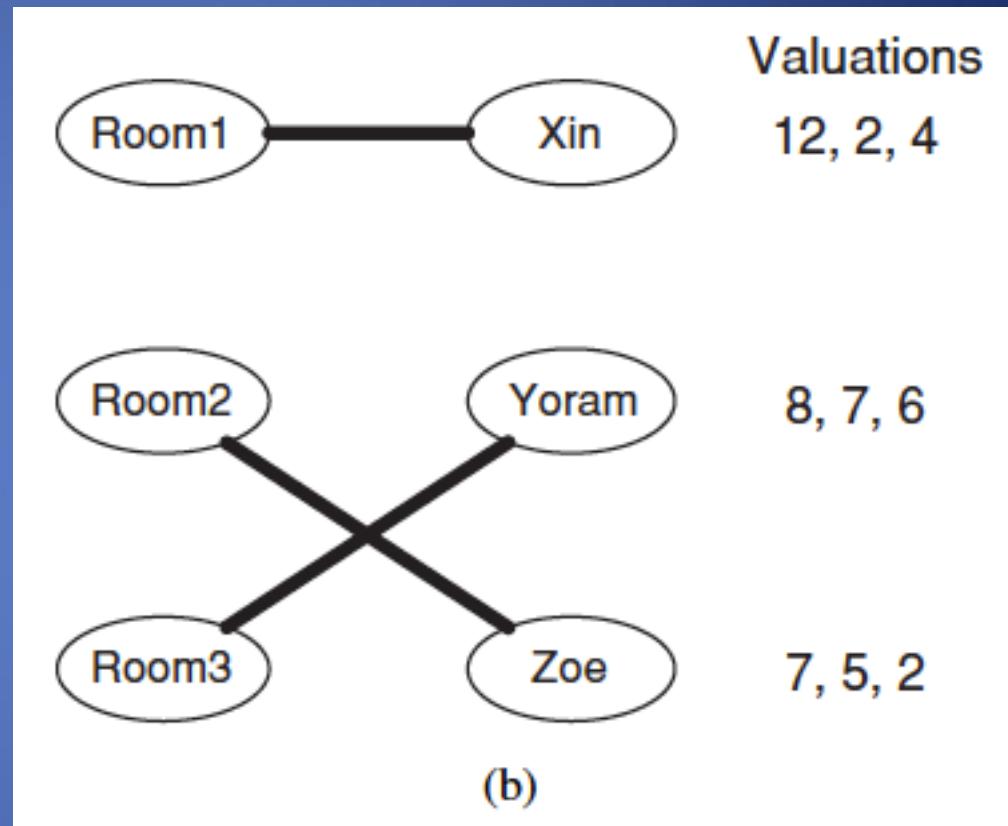
(a)

The diagram illustrates a matching problem between three items (Room1, Room2, Room3) and three buyers (Xin, Yoram, Zoe). Each item and buyer is represented by an oval. To the right of the ovals, their corresponding valuation lists are shown.

- Room1 is paired with Xin, with valuations 12, 2, 4.
- Room2 is paired with Yoram, with valuations 8, 7, 6.
- Room3 is paired with Zoe, with valuations 7, 5, 2.

最优分配

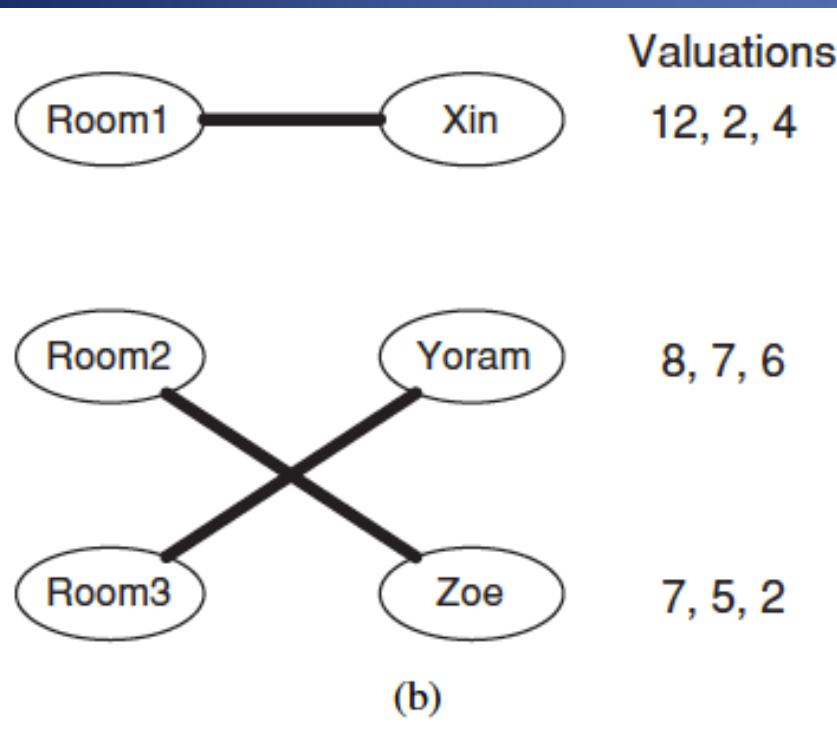
- 市场是一种分配机制，只要有供需，理论上就可以形成市场。
- 最优分配 (optimal assignments)，估值之和最大的完全分配方案



$$V_{\text{Max}} = 12 + 6 + 5 = 23$$

二部图匹配问题，是最优分配问题的一个特例：让看好的房间对应估值1，不看好的为0，希望 $V=N$ 。

最优分配问题

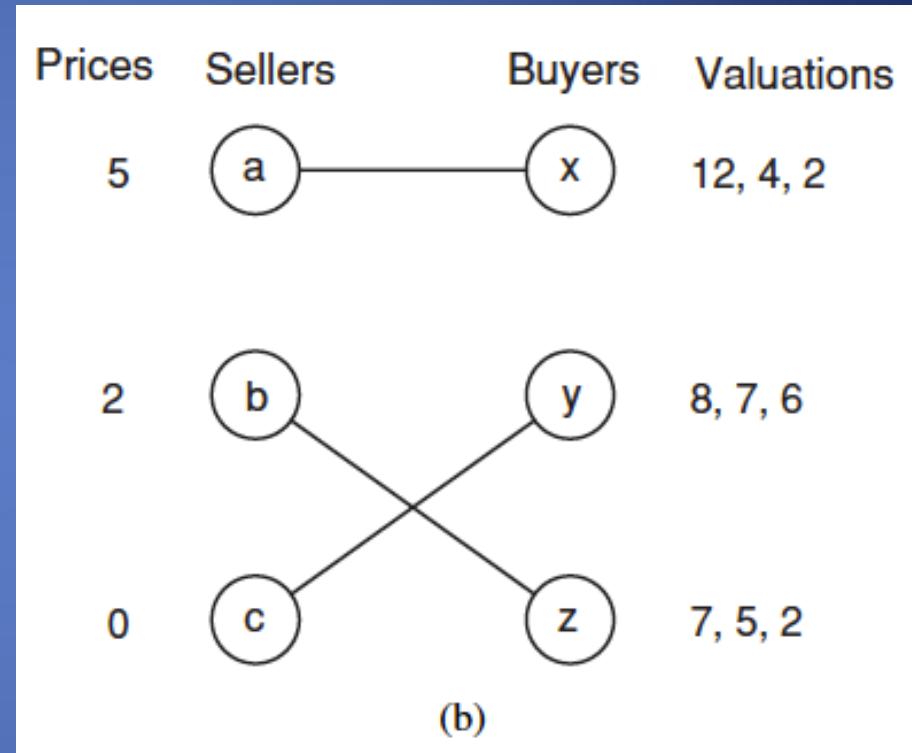


$$\begin{matrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{matrix}$$

一般性问题：给定一个 $N \times N$ 矩阵 (A)，从中选择 N 个不同行不同列元素， a_{ij} (即 i, j 分别在 $\{1, 2, \dots, N\}$ 中遍历)，使得和最大。

引入卖方价格

- 买方对房子有估值，卖方可能有**价格**
- 如果卖方的要价组合使每一个买方分别得到一个**回报相对最大的房子**，则称形成**市场清仓**
- 清仓时的出售价格，就是**市场清仓价格**（**Market-Cleaning Price**）

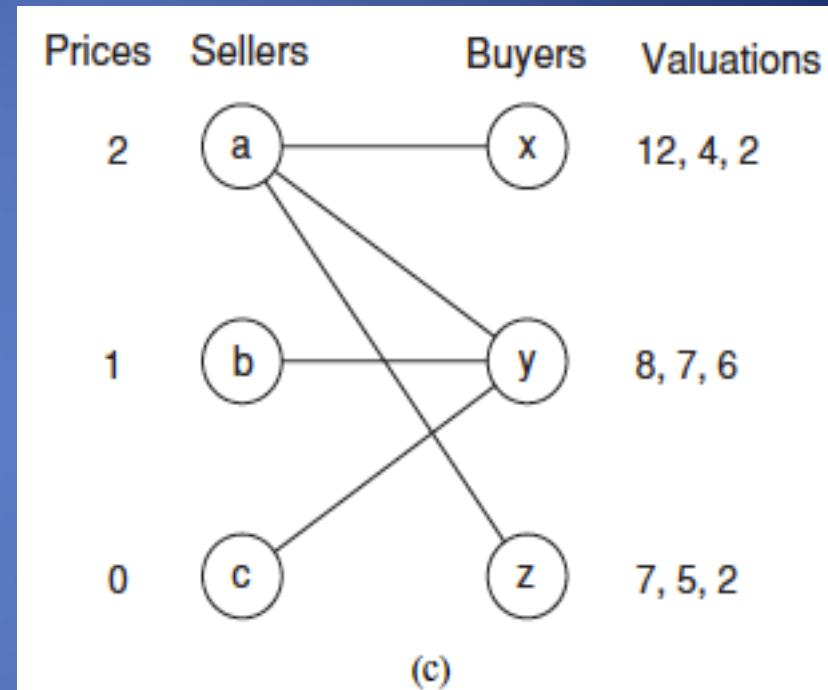


市场清仓价格

“**回报相对最大**”，分别针对每一买方而言。

偏好卖家图，描述买家喜好

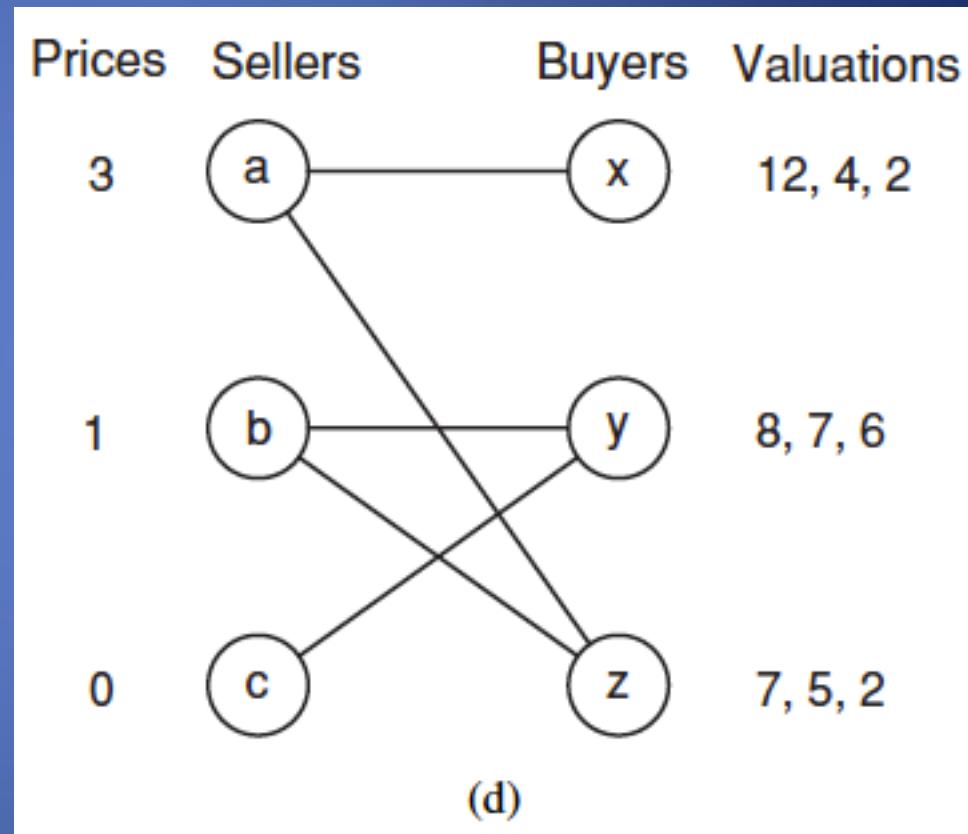
- 如果只有买方估值，匹配问题就是依据买方估值的最优分配问题
- 如果引入卖方价格，**偏好卖家图**（买方偏好图）的概念有助于我们的分析
- 偏好卖家图表达：“它们对我具有同样最划算”。
- 其上的完美匹配表明大家都等得到最满意结果
- 但偏好卖家图不一定有完美匹配，即可能存在**受限组**



例：没有完美匹配的偏好卖家图，
存在完美匹配的偏好卖家图中
卖家的价格称为“清仓价格”

市场：无形的手

- 市场协调，即价格调整，能够自动协调供需之间的关系，实现市场清仓
- 右图是在前页图的基础上，卖方a提高价格后的新卖家偏好图
 - 存在完美匹配，于是(3,1,0)为市场清仓价格
 - “无差异”（无所谓）：平手消解



价格调整后的偏好卖家图

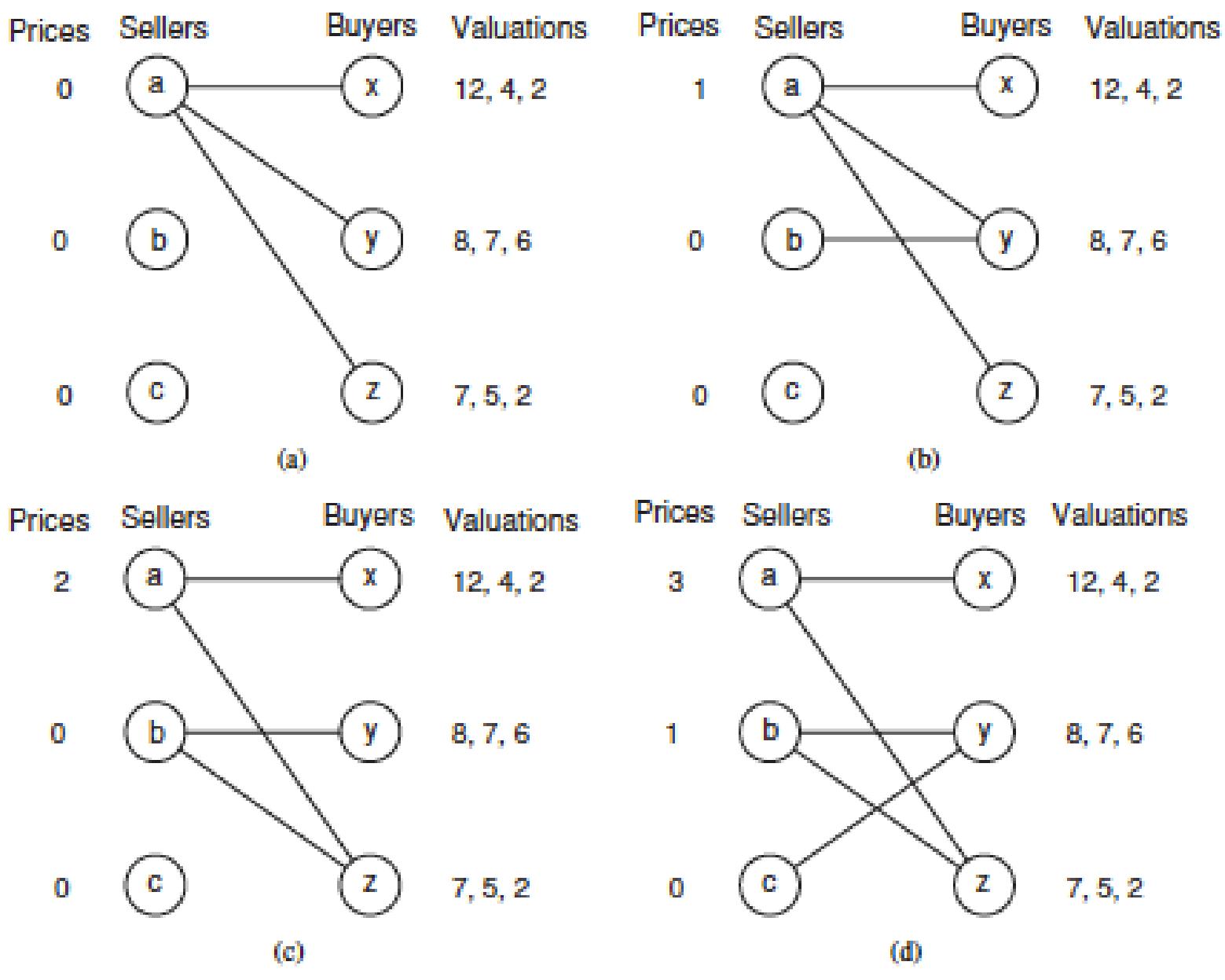
市场清仓价格的最优性

- “最优”的含义→社会最优。在任何可能的买卖配对安排下：
 - 买家群体可实现的估值之和最大。等价于：
 - 所有卖家收益之和 + 所有买家收益之和 最大
 - 这是因为：一个卖家的收益 = 价格
 - 一个买家的收益 = 购得物品的估值 - 所支付的价格
- 断言：在市场协调机制下
 - 对于任何买家的估值组合，都存在一组市场清仓价格，亦即对应的偏好卖家图存在一个完美匹配

市场清仓价格的形成：算法

- 卖方从 $(0,0,\dots,0)$ 开始，按照轮次进行下述：
 - 构造偏好卖家图
 - 识别一个买方受限组 (s) ，若没有，则存在完美匹配，结束。
 - 将受限组对应的卖方集合 $N(s)$ 中的价格都 +1 （也就是根据需求调整价格，“物以稀为贵”）
 - 若因此使所有价格都 >0 ，则统一约减最低价至 0。
 - 开始下一轮
- （市场清仓价格形成的过程类似于拍卖过程）

类拍卖过程促成市场清仓



在这个过程中，偏好卖家图不断被调整，最终结束在一个含有完美匹配的图上

这个过程为什么一定能结束？

- 定义市场的势能：所有参与者潜在回报之和
 - 卖方：价格， a_1, a_2, \dots, a_n ；
 - 买方(i)：最大的“估值减去价格”， $\max(v_{ij} - a_j)$
- 势能初值 ($a=0$)：

$$P_0 = \sum_{i=1 \dots k} \max_{j=1 \dots k} (v_{ij} - a_j) + \sum_{j=1 \dots k} a_j = \sum_{i=1 \dots k} \max_{j=1 \dots k} (v_{ij}) > 0$$

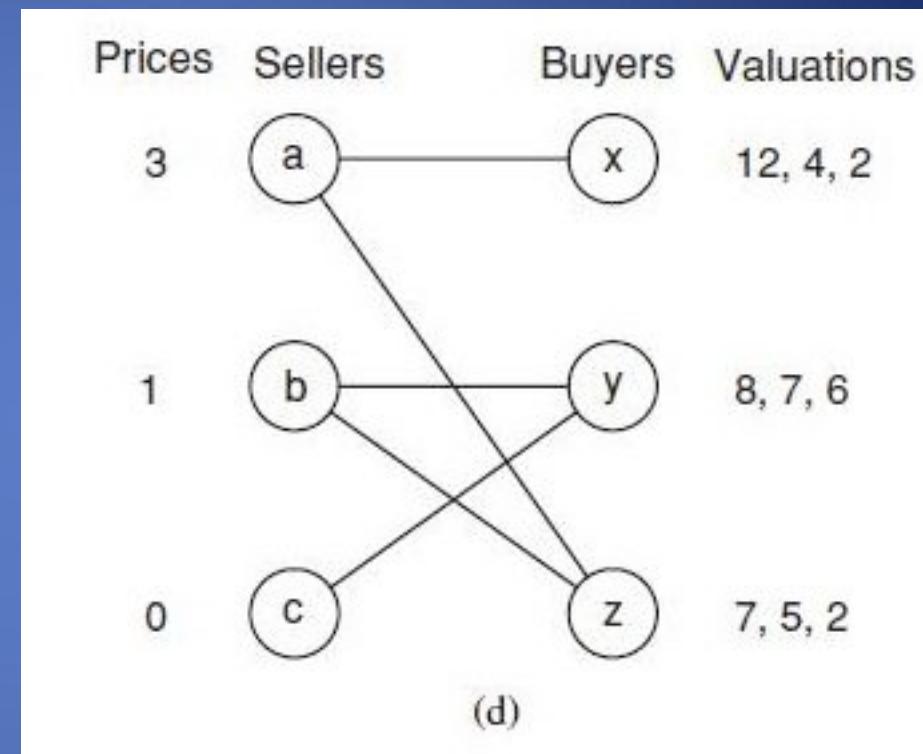
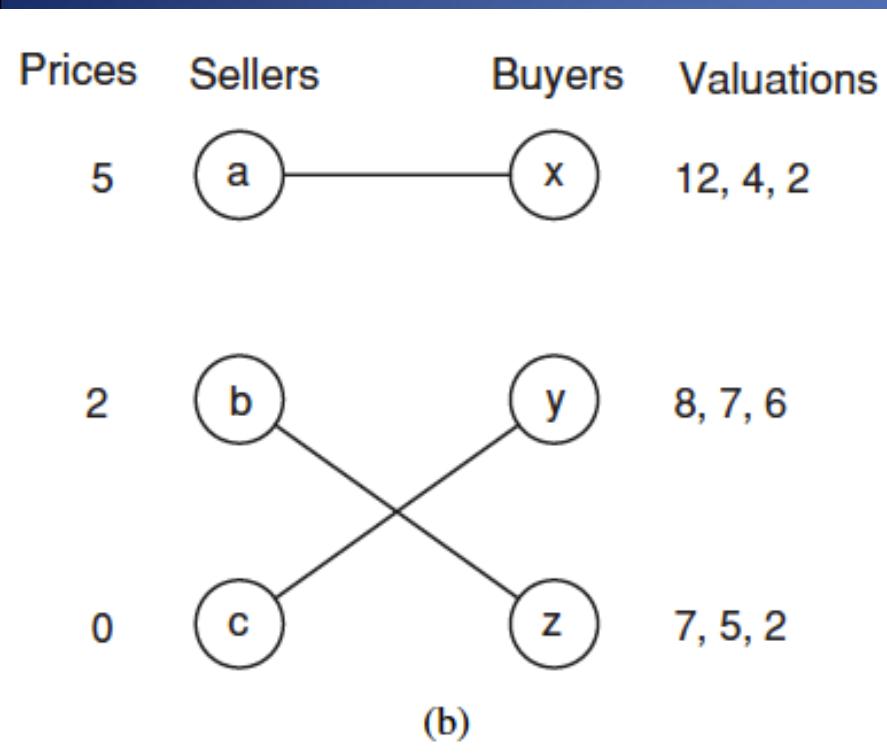
我们如果能说明在上述过程中，（1）势能每一轮单调减，（2）但总不会小于0；则就说明了过程一定结束。“结束” = “无受限集”。

这个过程为什么一定结束？（续）

$$P = \sum_{i=1 \dots k} \max_{j=1 \dots k} (v_{ij} - a_j) + \sum_{j=1 \dots k} a_j$$

- 观察势能在每一轮的变化，设买卖双方各有 K 人。可见只有价格 a 的变化会引起势能的变化。
- 在操作过程中有两处会引起 a 的变化
 - (1) 因受限集 S 造成的 $N(S)$ 中元素价格+1
 - (2) 统一约减 a 至最小价格为0
- 可见
 - 卖方势能之和，(1) 增加 $N(S)$ ，(2) 减少 K
 - 但总保持是 ≥ 0
 - 买方势能之和，(1) 减少 $S > N(S)$ ，(2) 增加 K
 - 结果也总是 ≥ 0 （因为 $v \geq 0$ ，且算法过程保证了总存在一个 $a=0$ ）
 - 于是市场势能在每轮都单调递减，且下界为0。

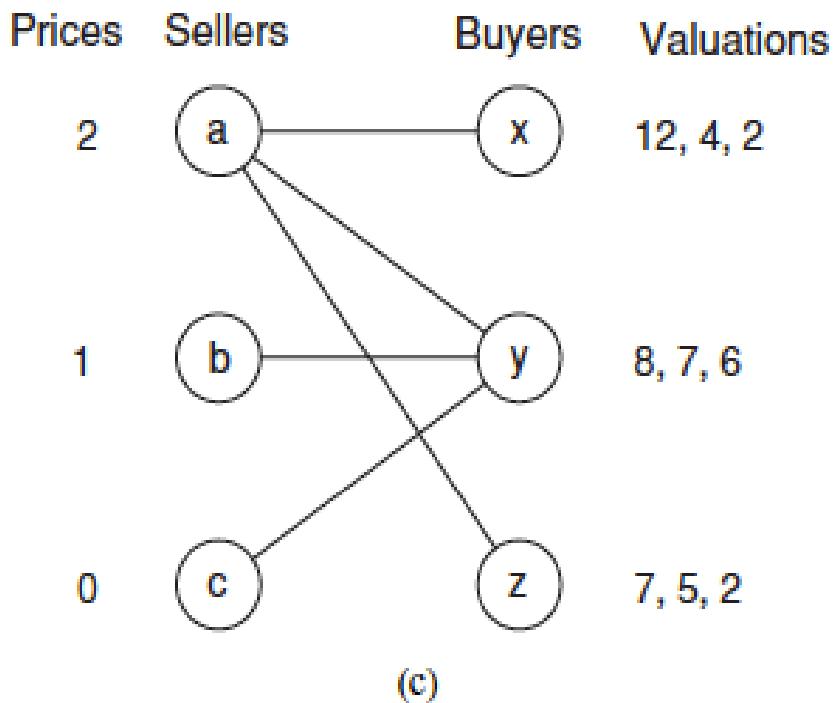
清仓价格的不唯一性



- 但都是“社会最优” ($12+6+5=23$)

再来看求矩阵不同行列 元素最大和的问题

$$P = \sum_{i=1 \dots k} \max_{j=1 \dots k} (v_{ij} - a_j) + \sum_{j=1 \dots k} a_j$$



$$\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{bmatrix}$$

- 给定一个 $N \times N$ 矩阵 (A)，从中选择 N 个不同行不同列元素， a_{ij} （即 i, j 分别在 $\{1, 2, \dots, N\}$ 中遍历），使得和最大。
- 可不可以利用前述算法来得到结果？复杂性如何？

作业

- 第10章，第3，5题

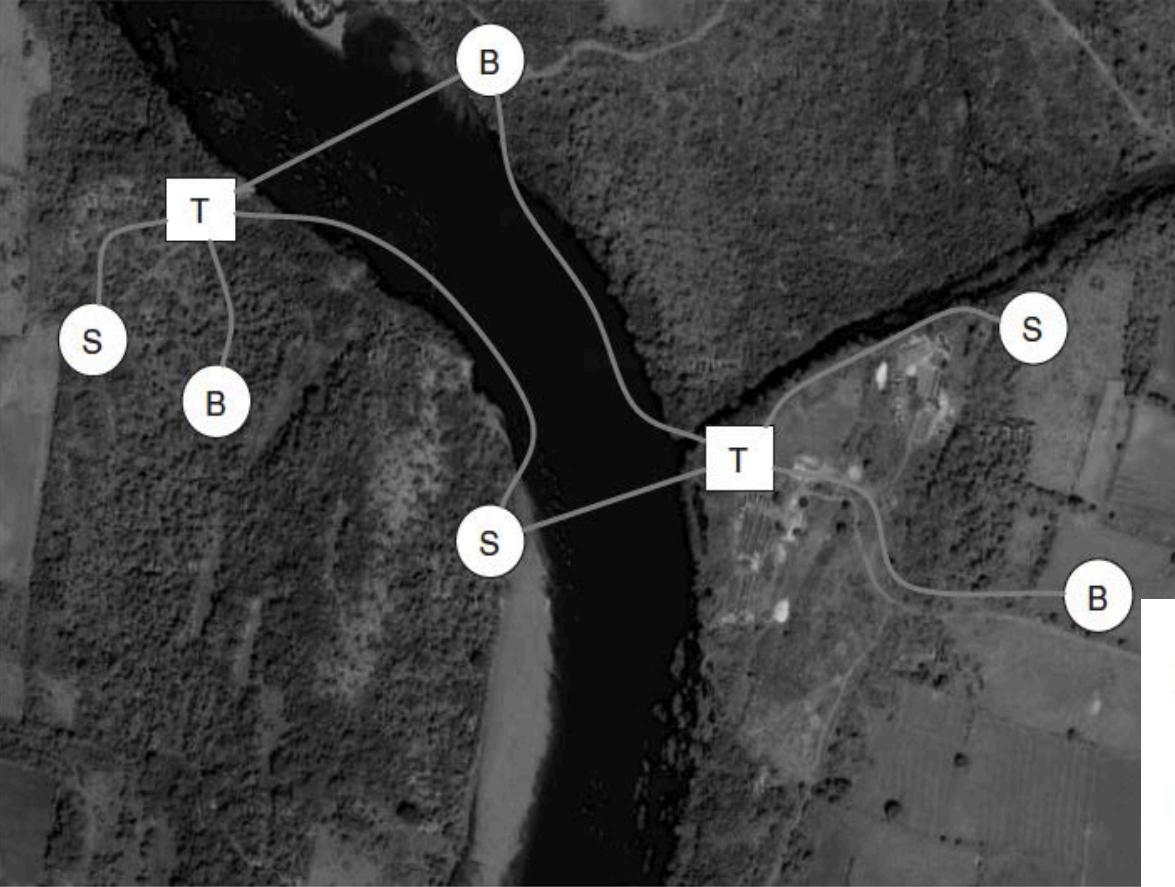
中间商市场网络模型

(第11章)

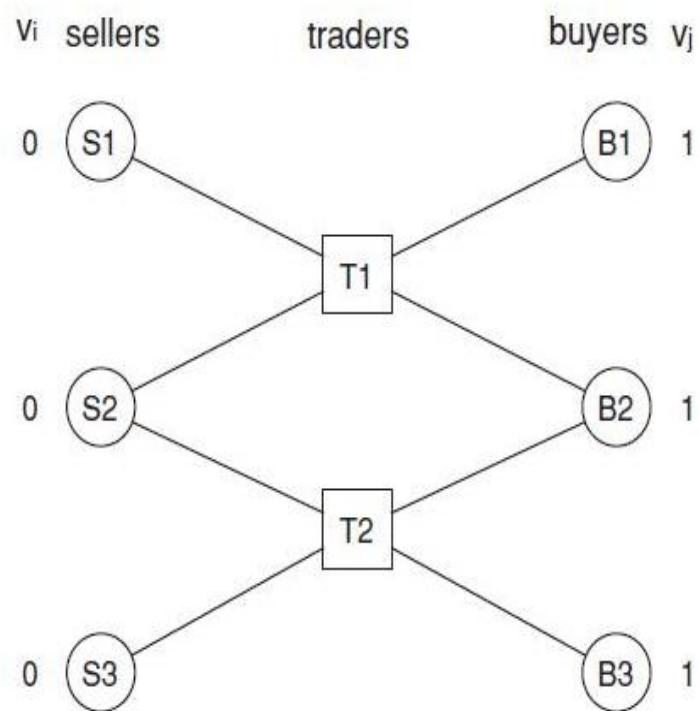
中间商作用的市场

- 股票市场、农产品市场、...（非直销市场）
- 地理位置的原因、运输条件的原因、掌握信息的原因、...
- 中间商（中间人、二道贩子），处于买卖双方之间，通过较低价收购，较高价出售，从差价中获利
 - 市场上，另外有一类中间人，其作用主要是中介（例如房屋中介），他们获利主要是佣金，不是差价。不在这里的讨论之列。
 - 佣金通常是公开的，差价通常是商业机密

地理位置与商业机会



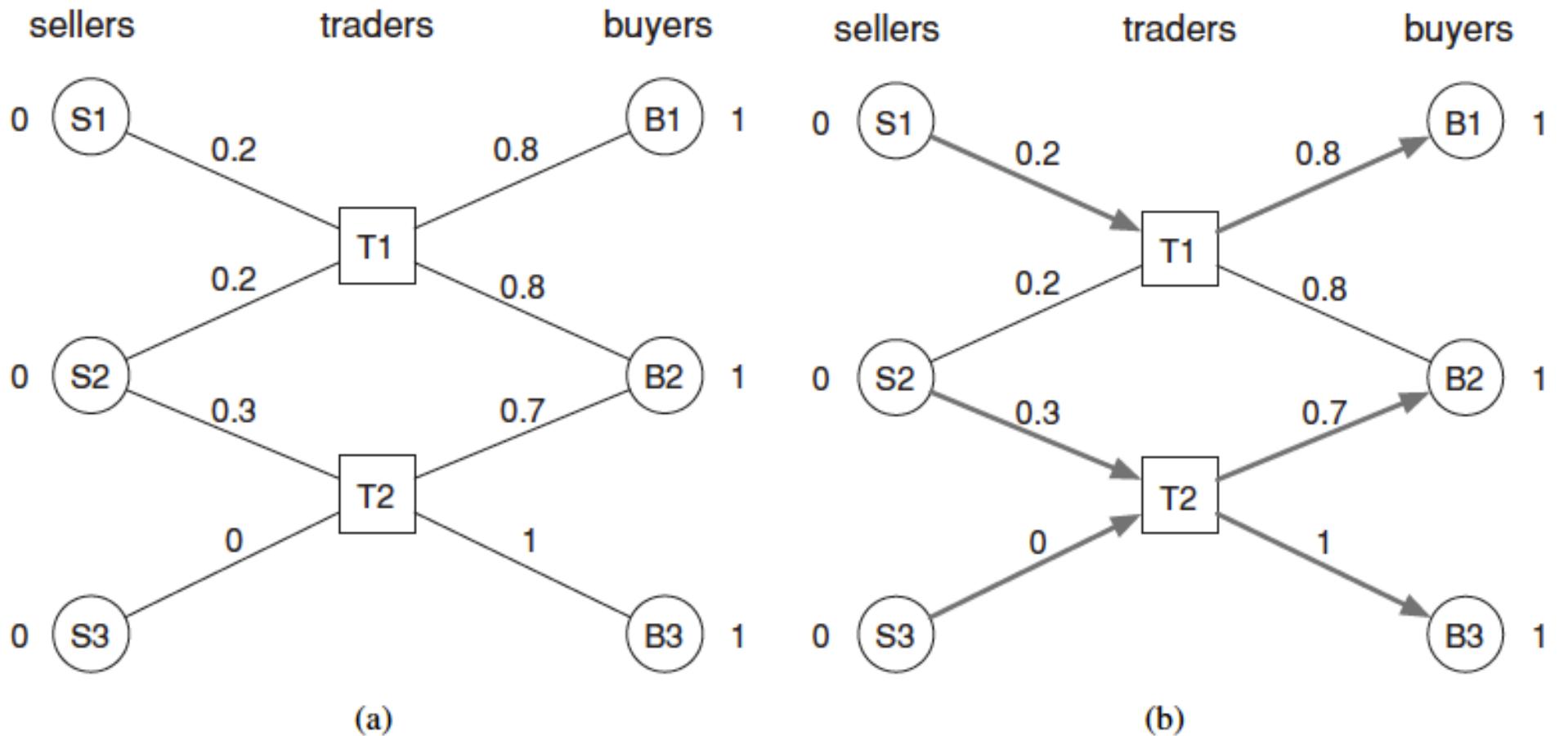
也可以用一种图来表达有中间商市场的潜在交易关系
 (S, T, B, V_i, V_j)



交易如何进行？

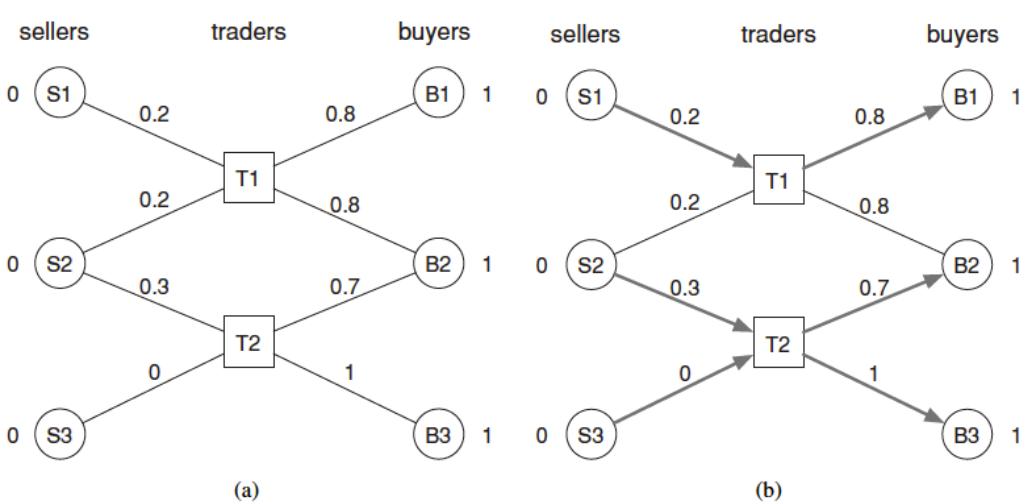
- 交易进行的规则
 - 第一步：中间商分别给出“出价”和“要价”
 - 第二步：买卖方给出反馈（选择一个中间商）
- 条件
 - 一旦价格公布，买方与卖方双方只能选一个中间商进行交易，或者不交易；但不能议价。
 - 卖方只有一件物品，买方也只要一件物品，如果交易达成，则可看成一件物品从卖方经中间商“流”到买方。
 - 中间商的定价不能造成“入不敷出”或“囤积居奇”现象
 - 在卖方或买方既有利益、也有权利的合理选择下，中间商一定要交易，否则可以想像一种“重罚”。

一个例子



中间商公布价格

卖方和买方的反馈
(形成物品流)



例子中值得讨论的几个问题

- 各参与人的决定是否正确（还有更好）？
 - 例如，下一次T1是否会通过调整自己的报价，将生意从T2争过来？（{下一次}* \rightarrow 均衡）
- 无差异（无所谓）现象的认识与处理
 - 例如，S3和B3在这个例子中都是“无利可图”，但这里的处理是认为他们都采取了参与交易的态度。（根据分析的需要决定，等价于可以认为S2-T2 价格0.0001，T2-B3价格0.9999之类）

将中介交易问题描述为一个博弈 (给定网络结构和 v_i, v_j)

- 参与人
 - 卖方, 中间商, 买方; 都可能有多个。
- 策略 (两阶段)

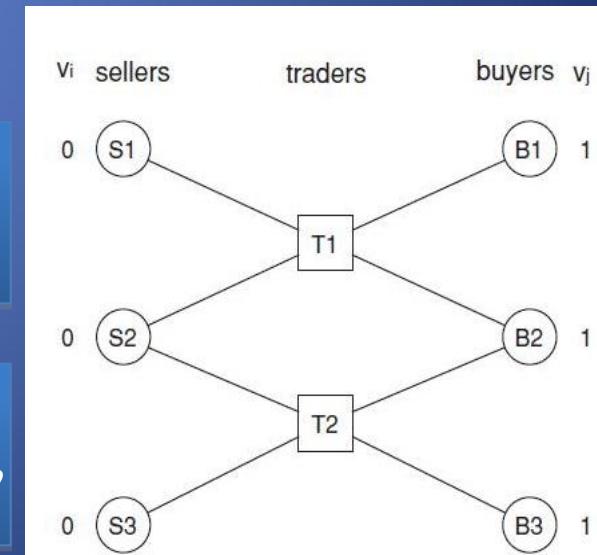
- 第一阶段: 中间商给出价格 b, a (即策略)
- 第二阶段: 卖方, 买方在**有接触的**中间商中选择对自己最有利的 (给出价格最合适)

- 回报
 - 卖方: 0或者 $b_{ti} - v_i$
 - 买方: 0或者 $v_j - a_{tj}$
 - 中介: 0或者 $a_{tj} - b_{ti}$

为简化讨论,
设 $v_i = 0$

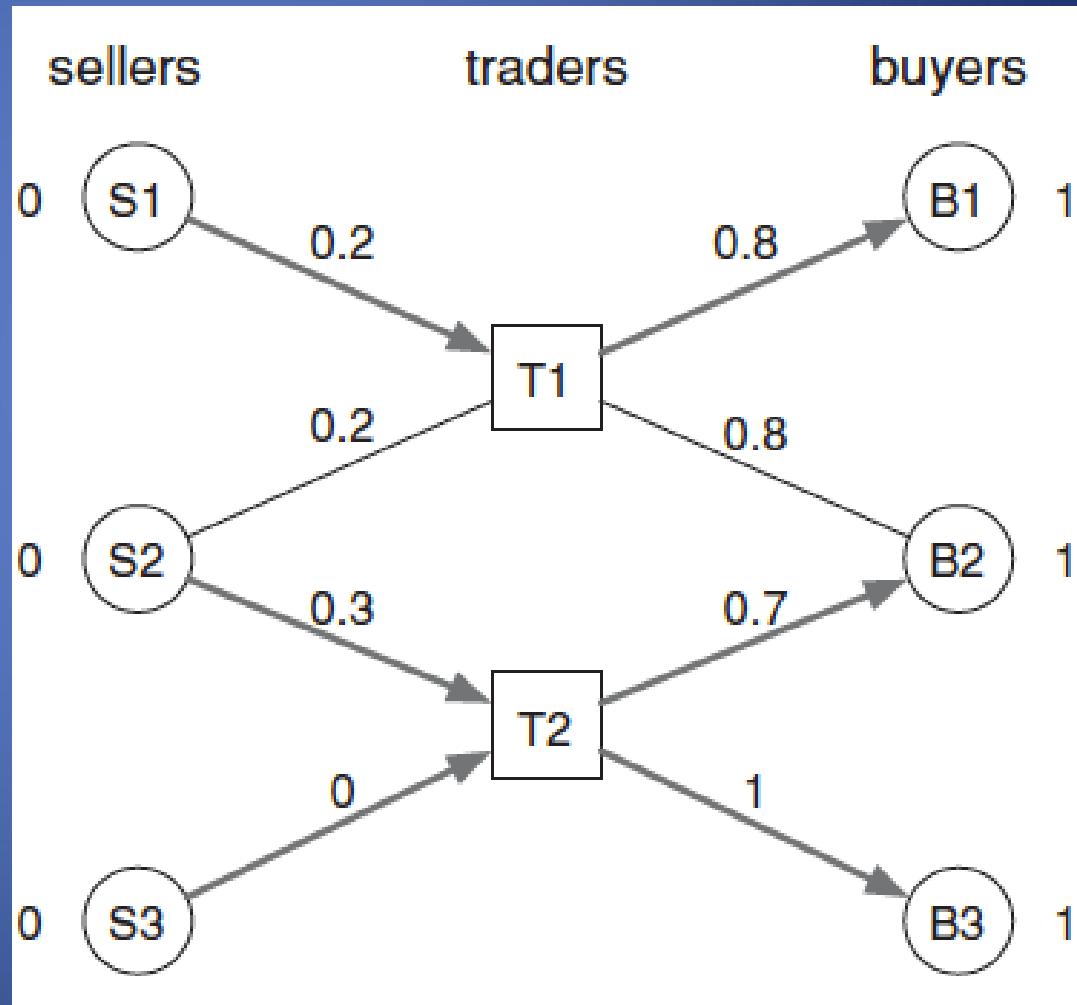
中间商不可
“恶意出价”

在无差异的情
形根据推理的
需要决定。

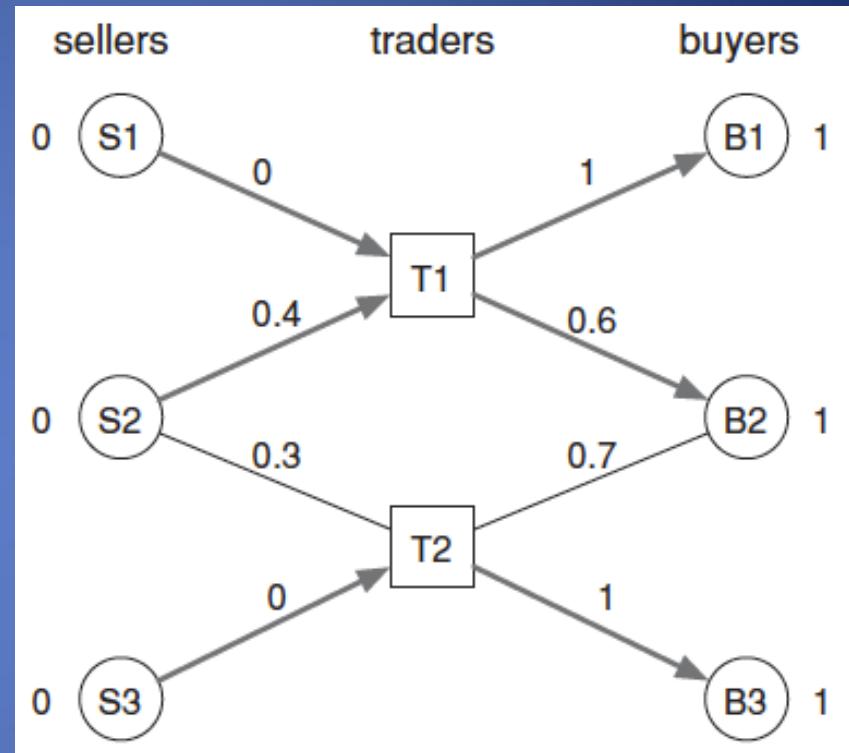
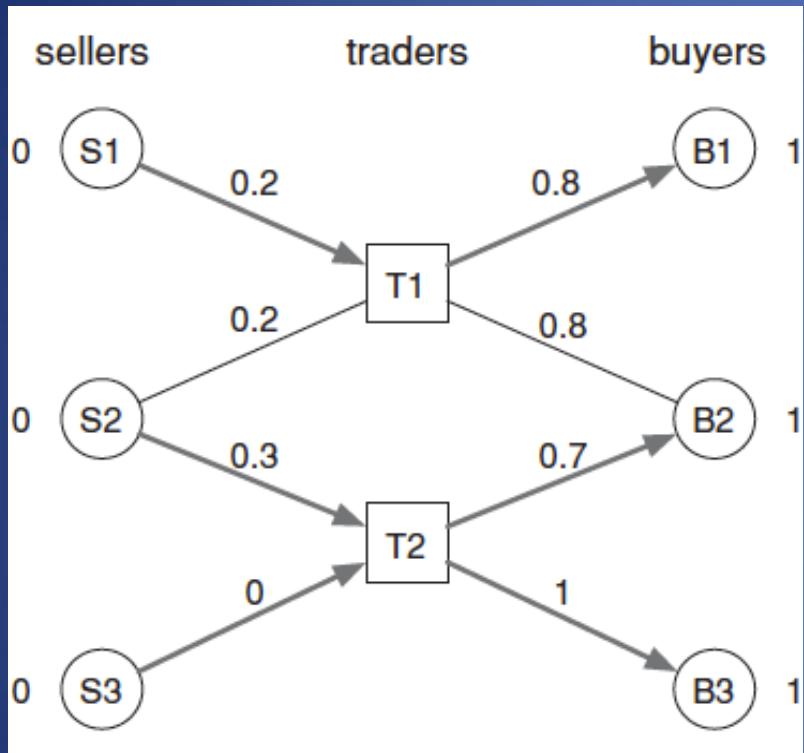


前例的计算

- 第一个中间商的回报: $0.8 - 0.2 = 0.6$
- 第二个中间商的回报: $1 + 0.7 - 0.3 - 0 = 1.4$
- 三个卖方的回报: $0.2, 0.3, 0$
- 三个买方的回报: $0.2, 0.3, 0$



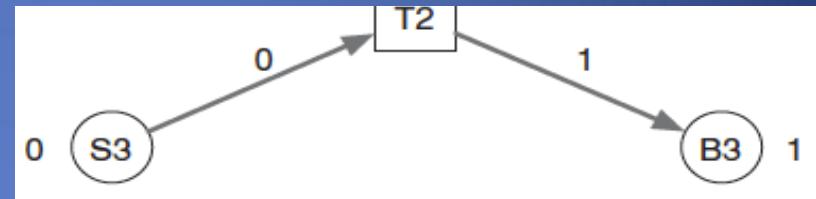
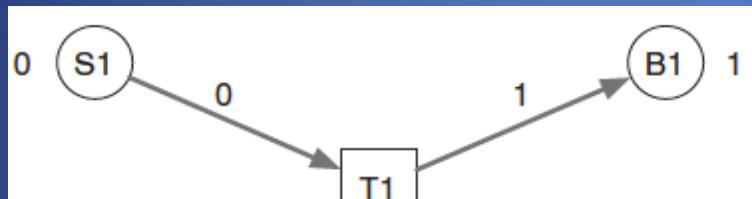
市场均衡 = 互为最佳应对（无改变动机）



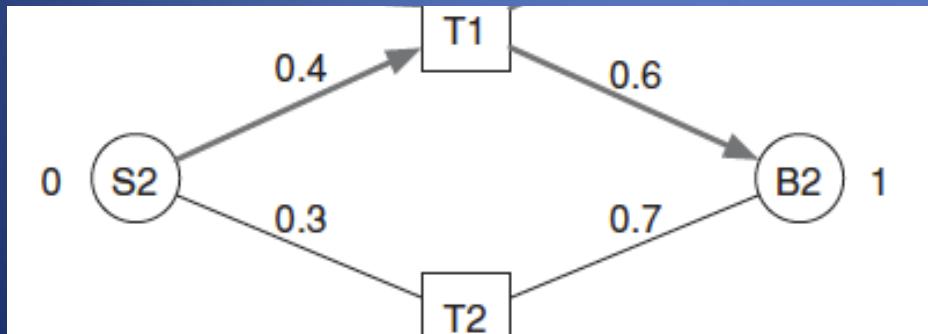
- 竞争的存在，会促使中间商调整价格，使其在不亏损的前提下赢得生意
- 垄断的存在，也会使中间商调整价格，争取获得最大回报
(可以想像均衡是长期市场互动的结果)

图中的现象：垄断和竞争

- 垄断：S1-B1被T1垄断； S3-B3被T2垄断

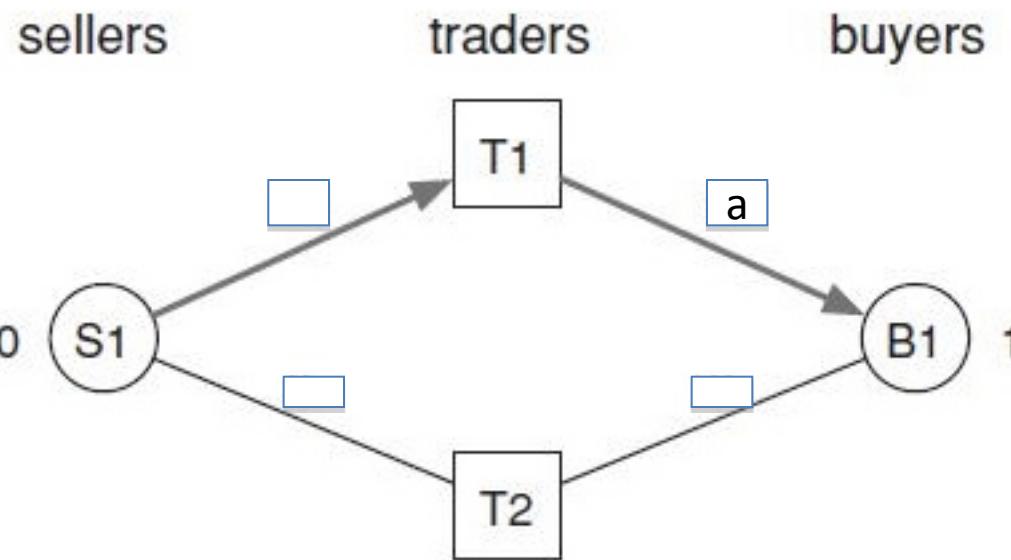


- 理想竞争（perfect competition）
 - T1 和 T2 在 S2-B2 之间呈理想竞争结构

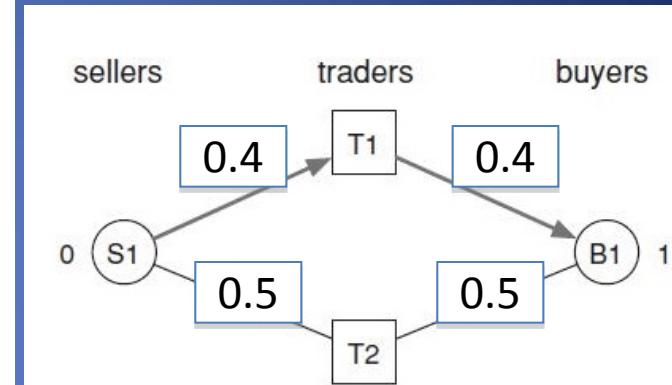


垄断结构的均衡？
理想竞争结构的
均衡？

关于理想竞争结构中的均衡



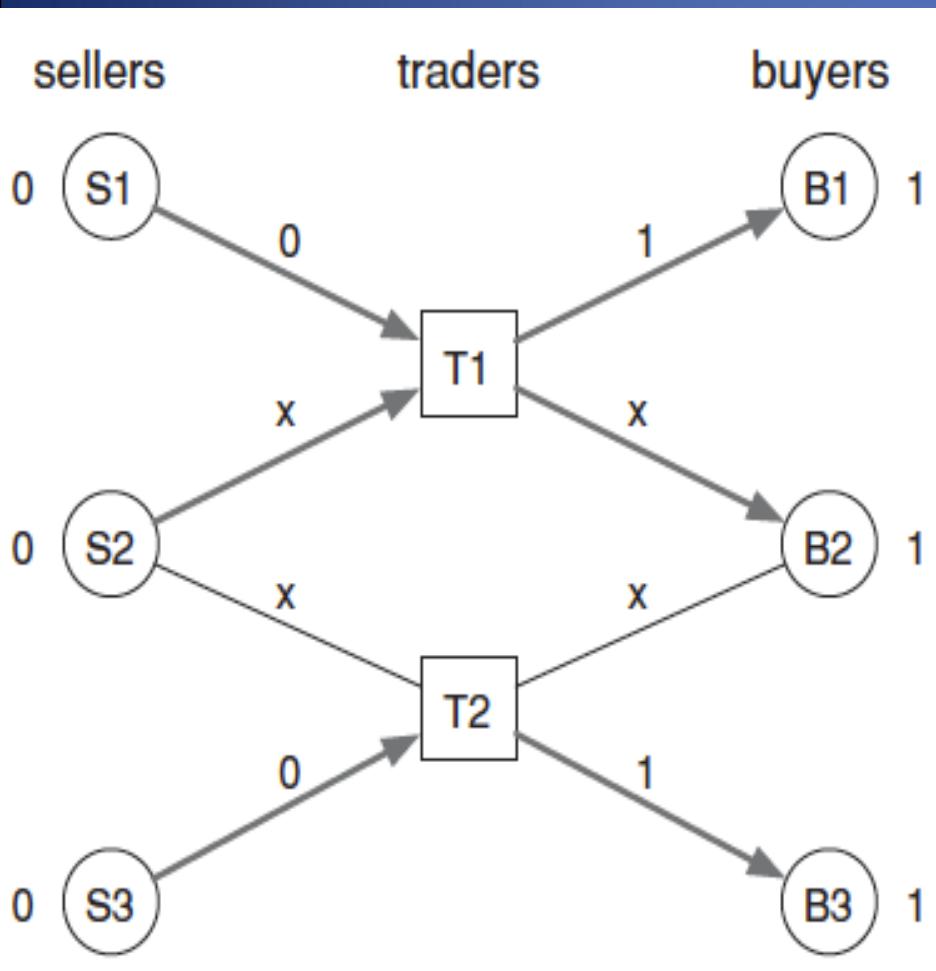
此时，S₁的选择需要依靠外部因素决定了



此时形不成商品流

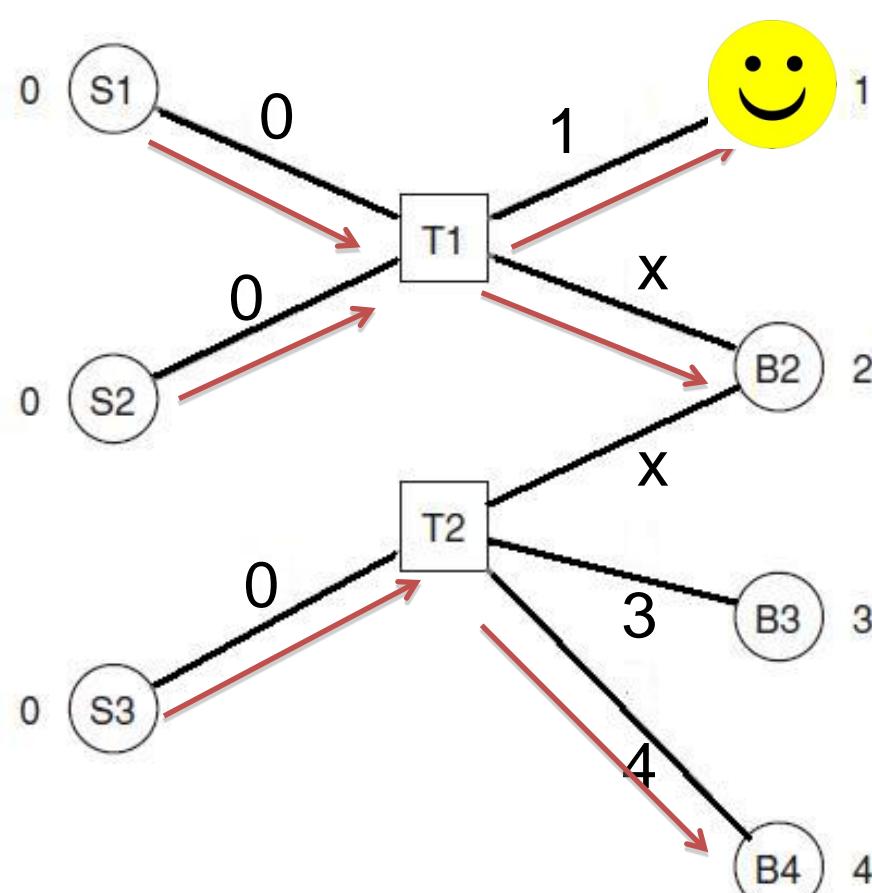
- 如果T₁的出价和要价不相同 ($a > b$)， T₂就有机会切入生意，改进自己回报为0的状态
 - “ $a=b$ ”（回报也是0，但“生意做成了”）是均衡的必要条件。
- 若T₂的两个报价不等，则T₁有机会调整他的价格，改进他的回报。如果T₂的报价不同于T₁，则形不成商品流。

例子市场网络中的均衡

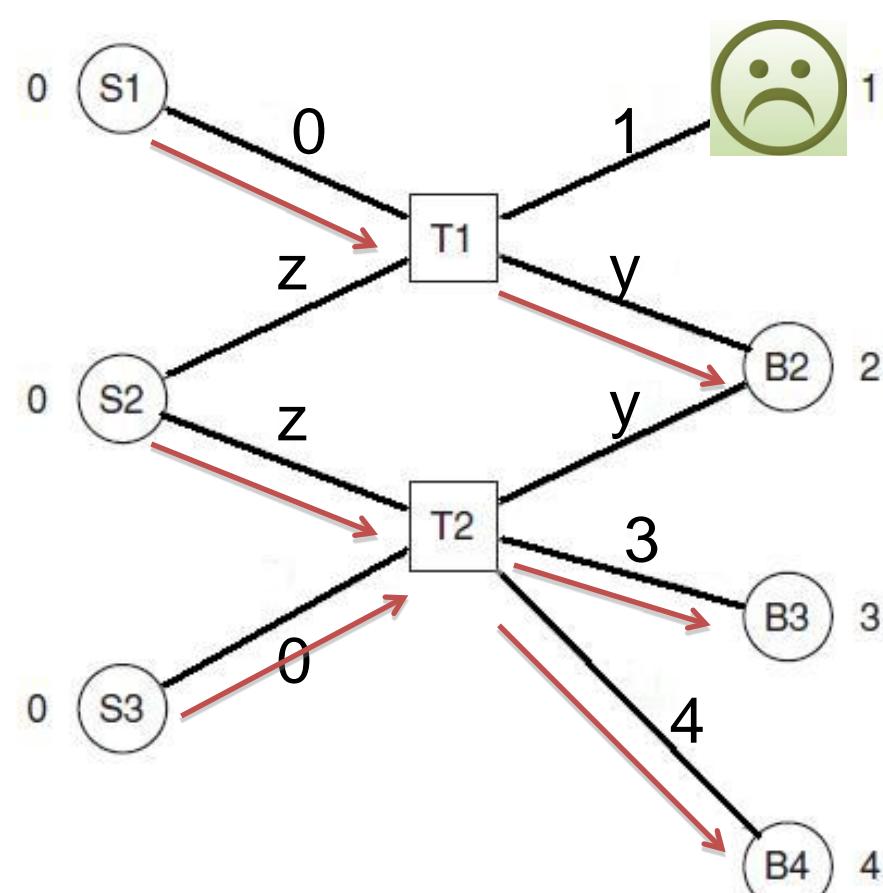


- x 在0, 1之间都是均衡
 - 由其他因素决定S2到B2的商品流是经过T1还是T2
 - 假定若S2选择了T1，则B2亦然。
- 在S2和B2之间，T1和T2的回报都是0
- x 的值对S2和B2的回报有影响

波及效应：均衡在不同网络结构下的变化

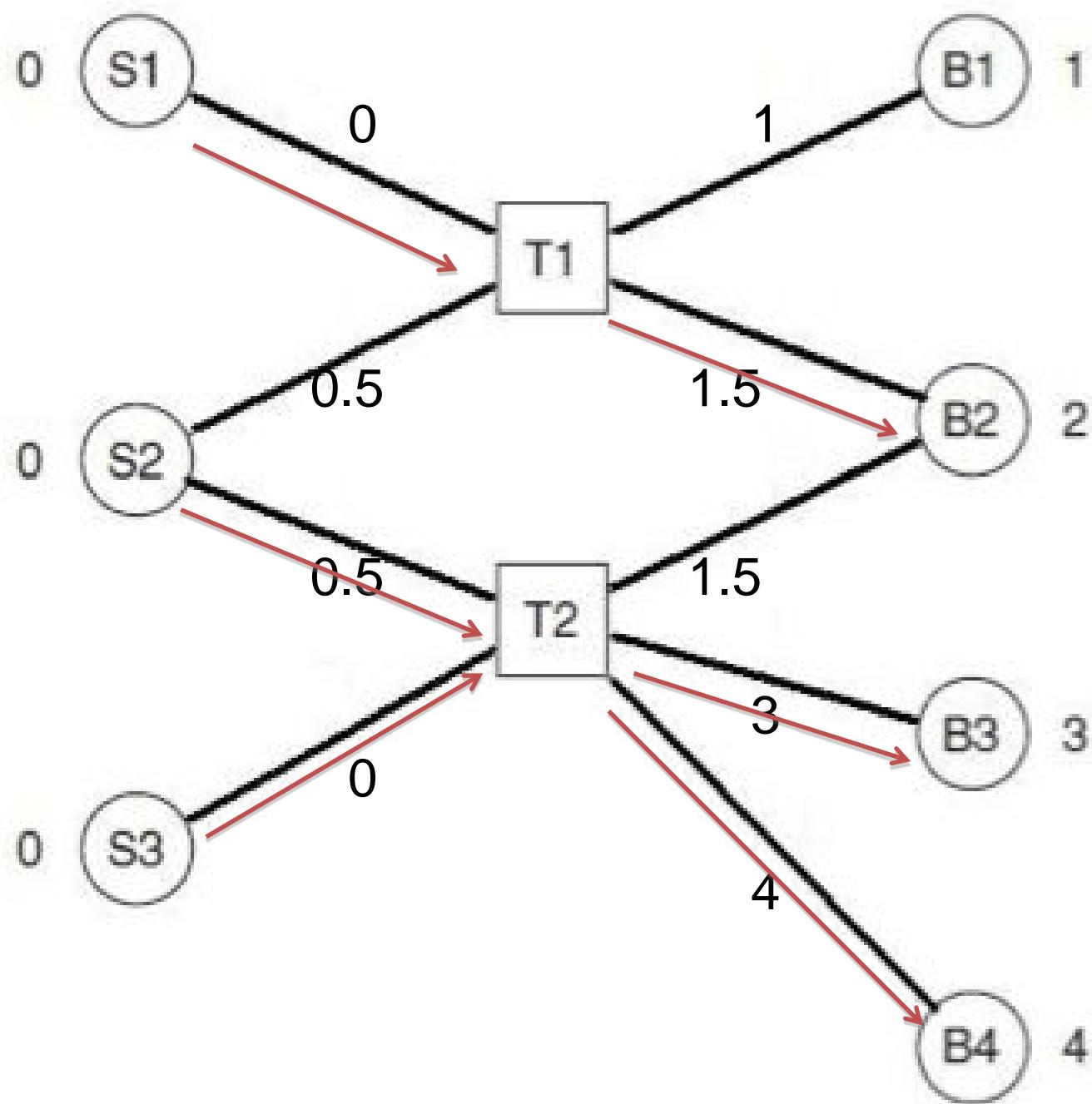


(a)



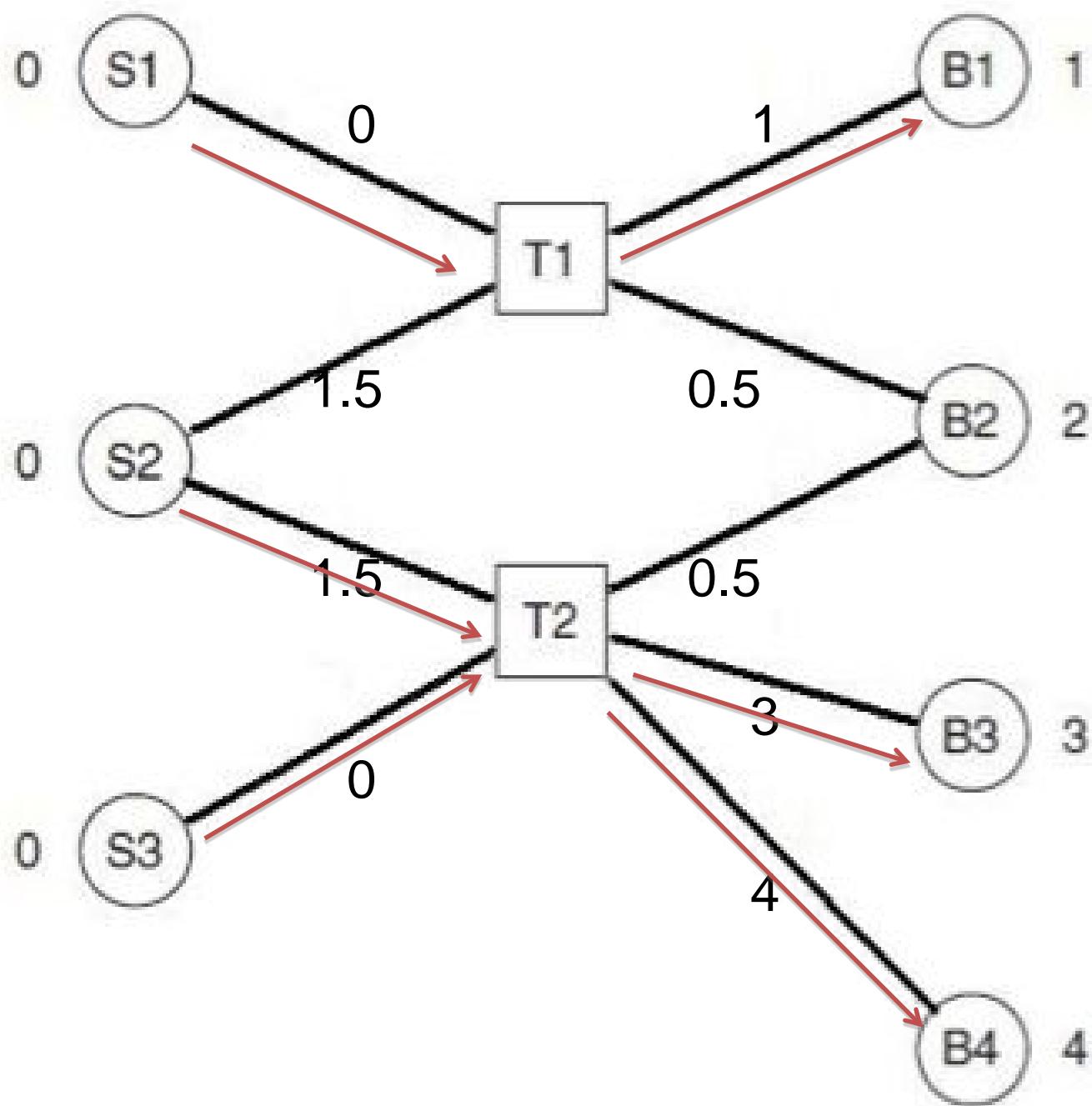
(b)

这是个均衡吗？



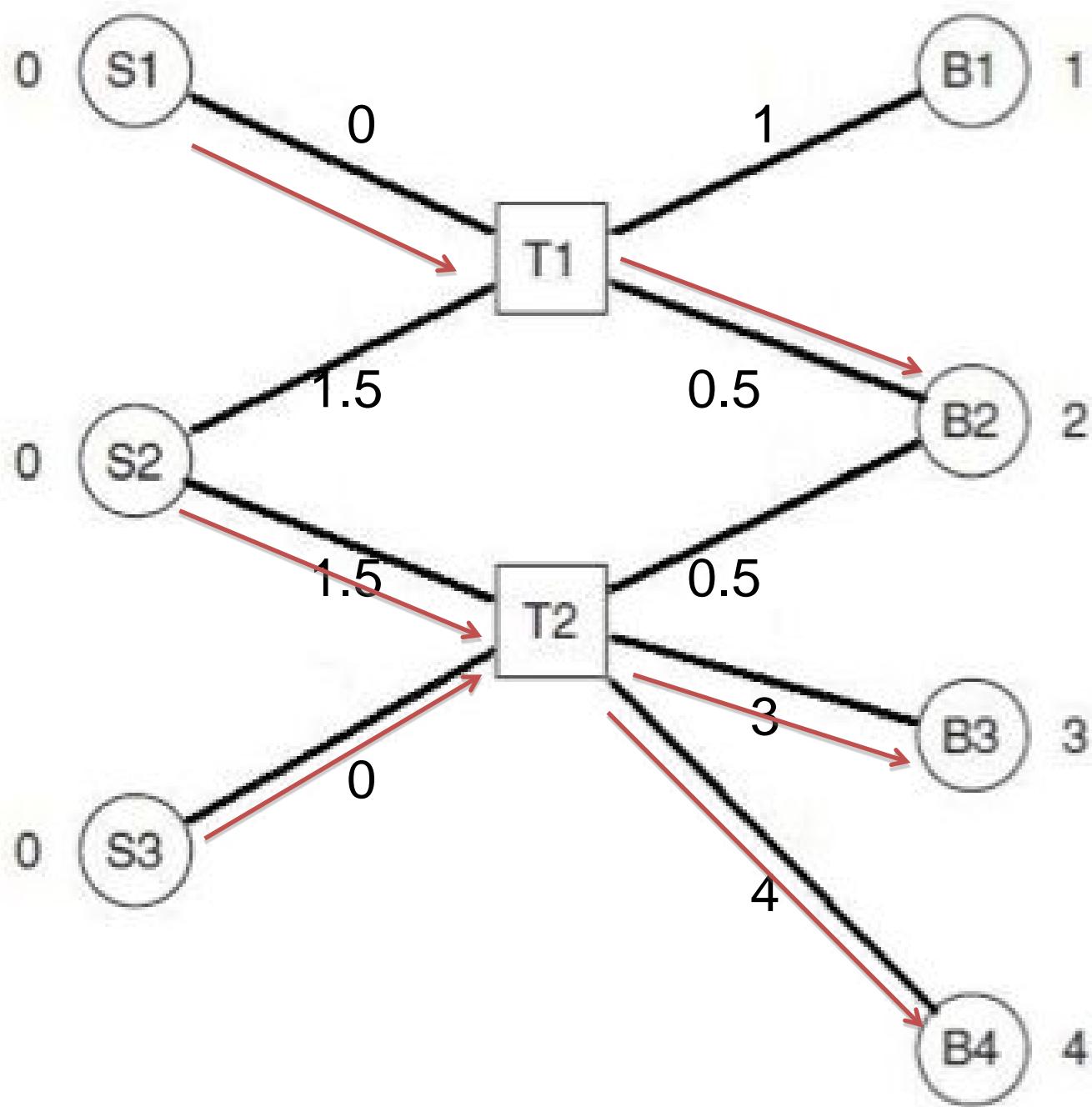
例如， T_1 的回报是1.5，若他将对 S_2 的出价提高到0.6，则可以将 S_2 赢过来，回报提高0.4

这是个均衡吗？



例中， B_2 有权利且有动机向 T 要商品，但他们没有给的，因此这种情形不符合规则

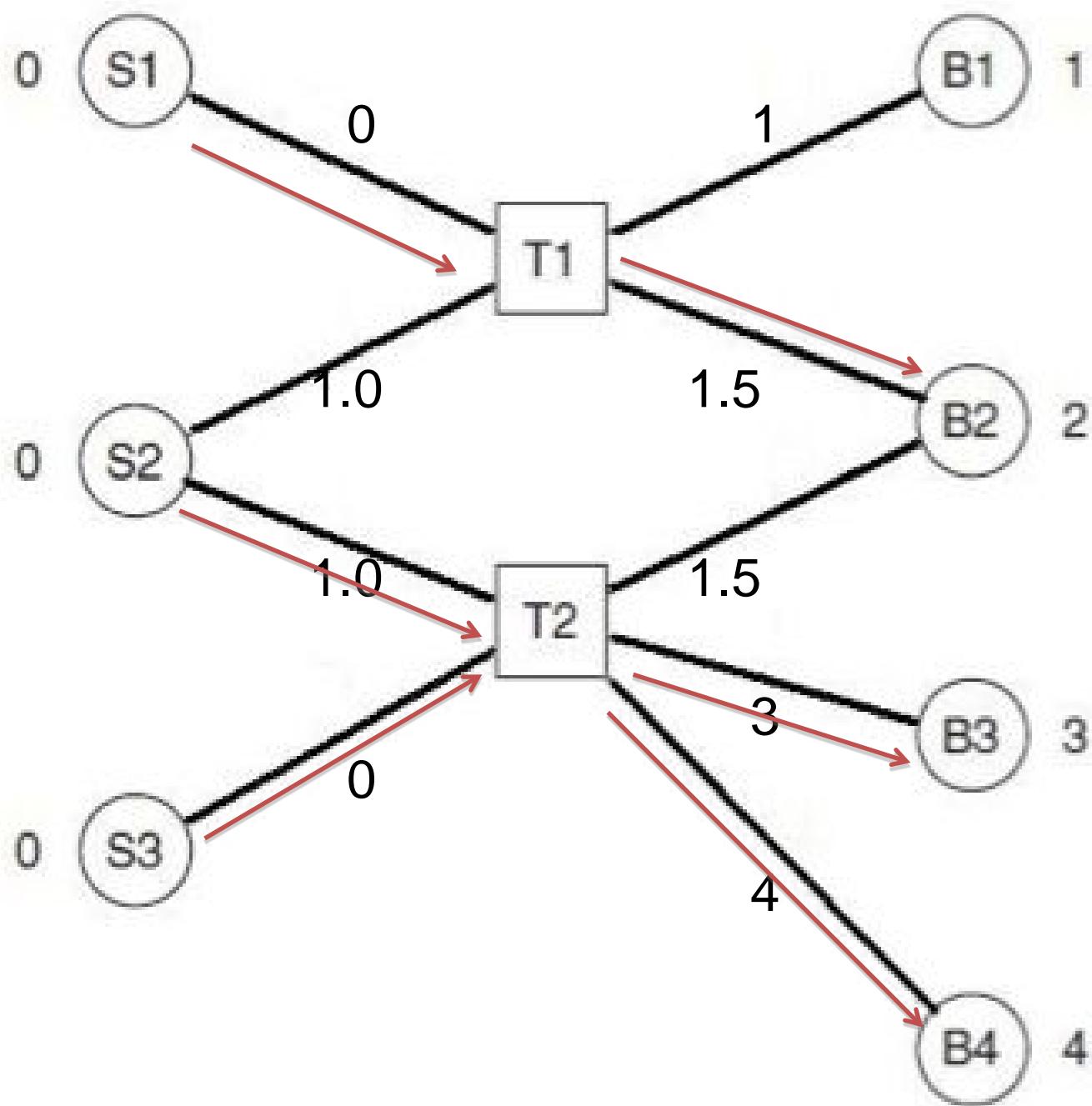
这是个均衡吗？



T1对B2的出价可以提高至少到1; 尽管看起来B2会选择T2, 但T2已经没有商品了, 因此他必须跟着提价, 向T1看齐。

这是个均衡吗？

例中，**B1**有权利但没动机向**T1**要商品，即“无所谓”，因此我们可按需裁决



T1有愿望
将S2争过来吗？

社会福利及其最优问题

- 如果一件商品实现了销售，设买家估值 v_j ，卖家价格 v_i ，中间商给的出价和要价分别为 b_{ti} 和 a_{tj} ，则三方的收益之和为

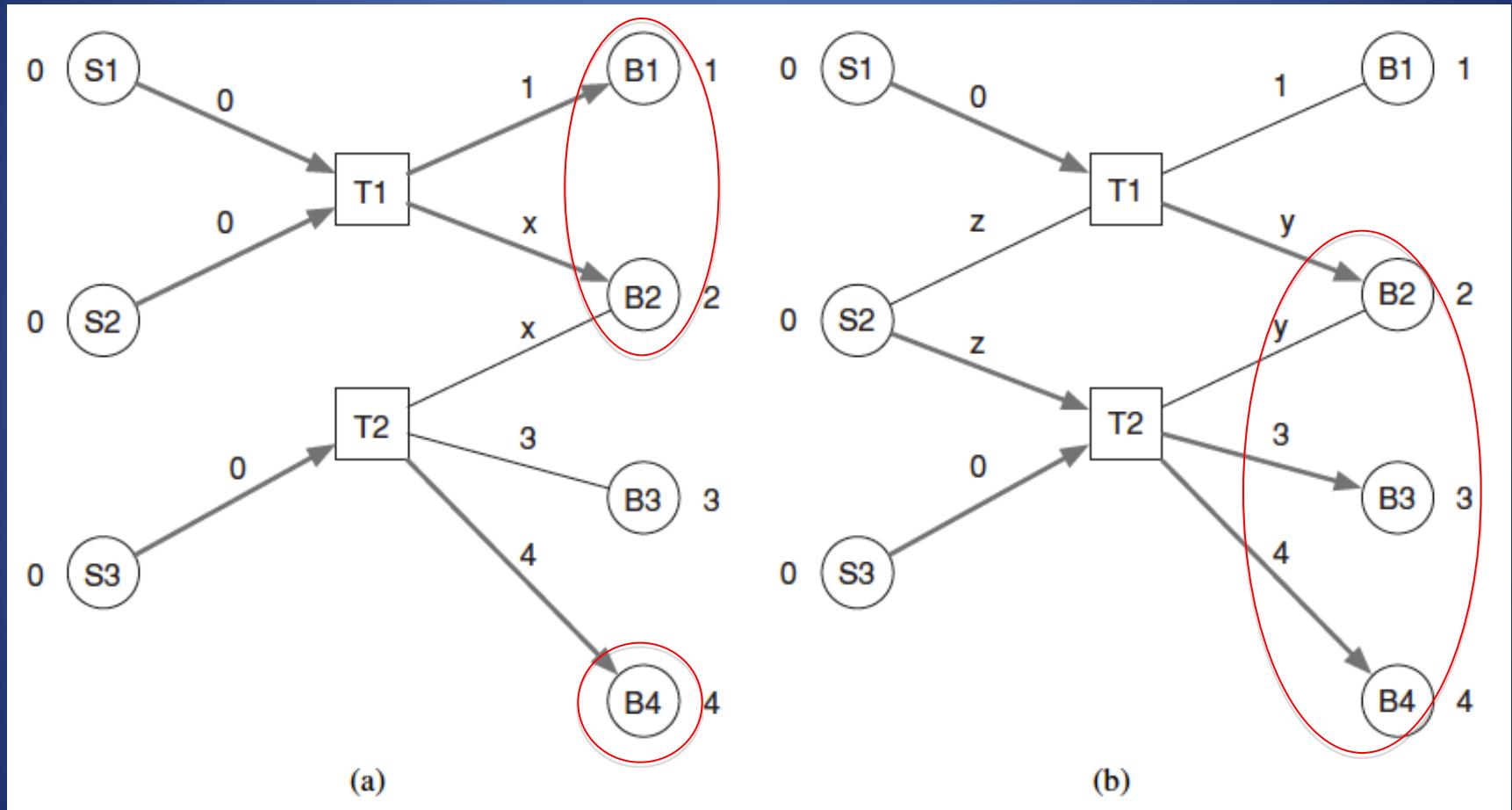
$$(b_{ti} - v_i) + (a_{tj} - b_{ti}) + (v_j - a_{tj}) = v_j - v_i$$

- 所有商品流对应的收益之和 $\sum(v_j - v_i)$ 称为社会福利

可以证明：每一个中间商交易网络都至少存在一个均衡，且每一个均衡所产生的商品流都达到社会最优。

- “Trading networks with price-setting agents,” ACM conference on electronic commerce, 2007

网络结构对社会福利的影响

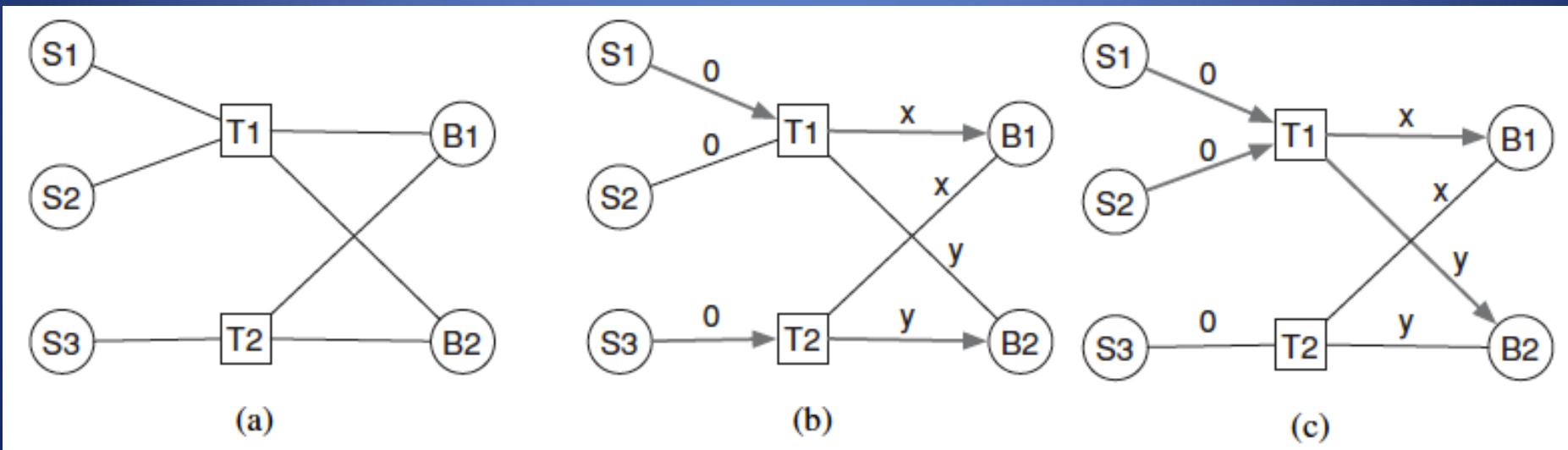


$$\text{社会福利} = 1+2+4 = 7$$

$$\text{社会福利} = 2+3+4 = 9$$

关于中间商在交易中的获利问题

- 前面的例子已证明，在理想竞争条件下，中间商获利的可能性为 0。
- 前面的例子也说明，在垄断条件下，中间商有利可图，但下例说明，不一定总如此！

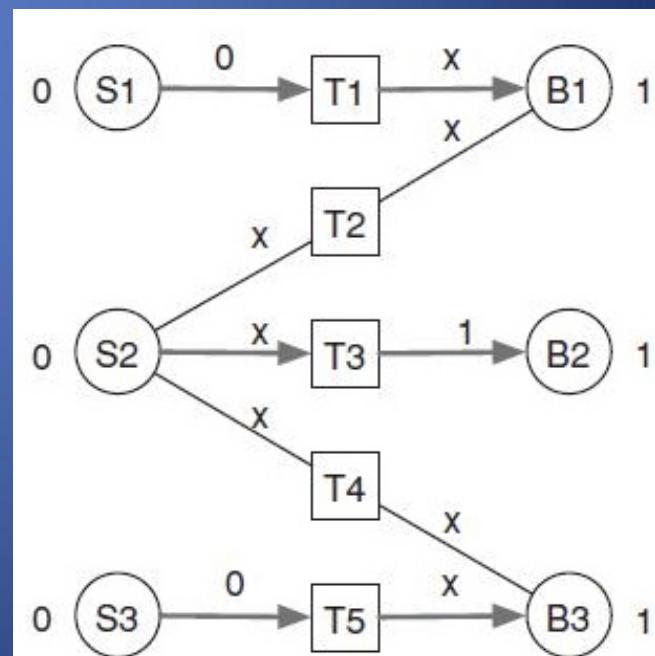


无论是否何种交易结果，均衡状态下中间商的获利均为 0

中间商获利的（结构性）条件

- 若， T 有一条连接买方或者卖方的边 e ，删除该边会改变网络的社会福利值，
 - 边 e 被称之为 T 的一条关键边
- 则存在一个均衡， T 可从中获利。
- 注：
 - 只是“存在一个均衡”，不一定是“每一个均衡”
 - 对于不同的 T ，结果可能不一样

随着 x 在0, 1之间改变，社会福利的分配在 S , T , B 之间变动



本章小结

- 依赖中间商的市场：卖家—中间商—买家
 - 中间商决定价格（策略性价格）
 - 中间商的利润取决于网络结构
- 市场（看不见的手）→ “社会最优”
 - 预期微观行为促成理想宏观效果关系的例子
- 结构与行为分析的互动
 - 策略行为的选择依赖于结构
 - 收益情况有可能驱使人们改变结构（例如在适当的地方增加一条边）

作业

- 第11章 第3, 4题