互联网与计算智能

张彦如
yanruzhang@uestc.edu.cn

2020

Course Syllabus

Overview

Instructor Information

张彦如

计算机科学与工程学院（网络空间安全学院）

Office/Office Hour

创新中心B318

By appointment

Email

yanruzhang[at]uestc.edu.cn

Course website

https://faculty.uestc.edu.cn/yanruzhang/zh_CN/index.htm

TA Information

林璨

- 大三

胡瑞

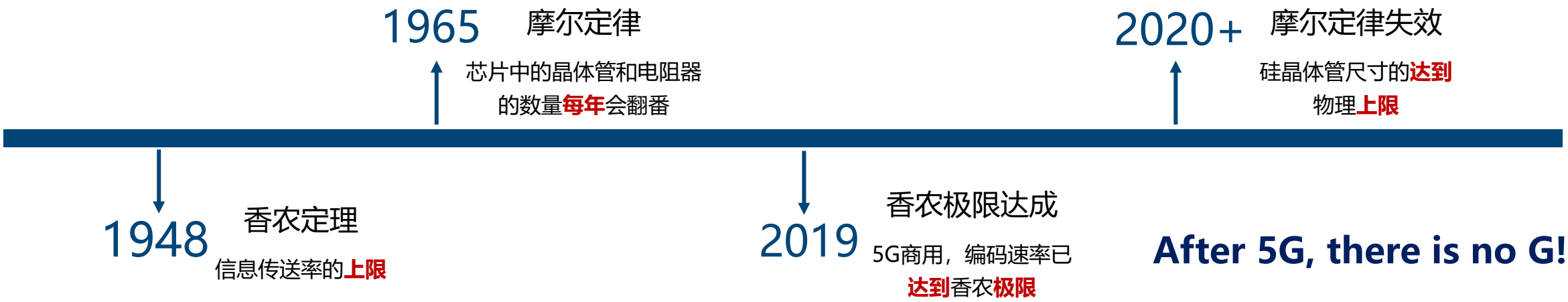
- 大三

- 职责：课程设计、考勤、作业、答疑

为什么要“互联网+”？

❑信息科学领域

- **香农定理**：描述通信行业的信息传播效率和带宽的关系
- **摩尔定律**：描述硅片计算能力发展的普遍性规律



5G之后，基于香农定理的无线通信已经发展完毕

这是一个最好的时代，也是一个最坏的时代

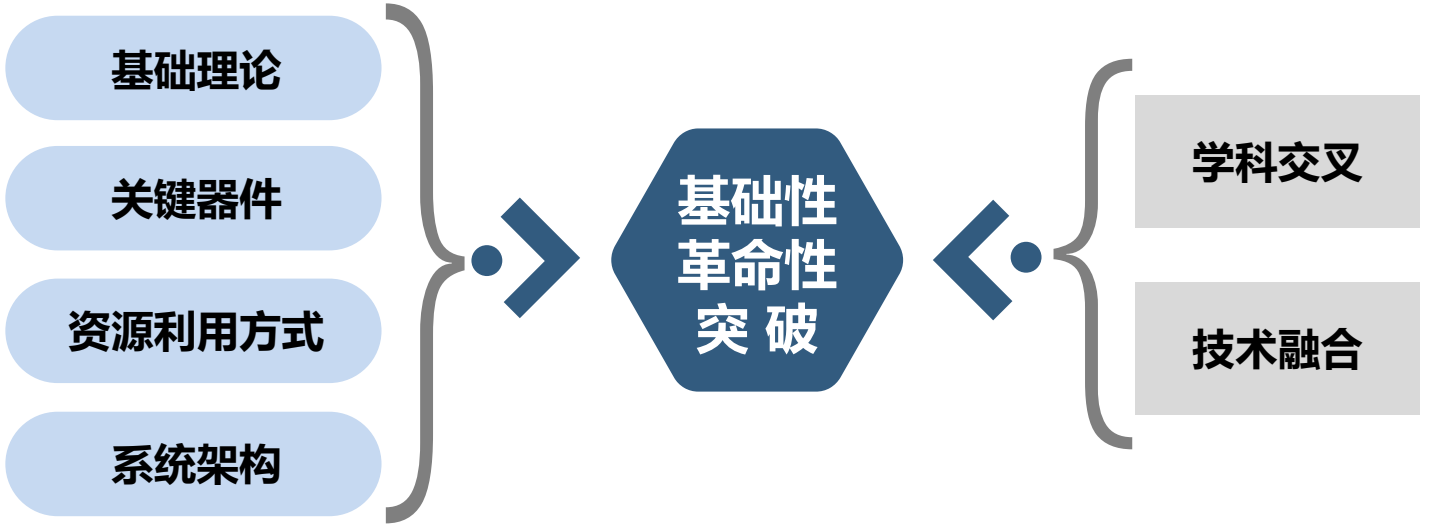
随着逐步**逼近香农定理、摩尔定律的极限**，而对大流量、低时延的理论还未创造出来，华为已感到前途茫茫，找不到方向。华为已前进在迷航中。

——任正非 全国科技创新大会

为什么要“互联网+”？

用户和业务的增长速度已经超过了系统扩容的速度，网络运行面临巨大挑战

- 关键问题：如何解决**有限的资源**与**迅速增长的业务需求**之间的矛盾？

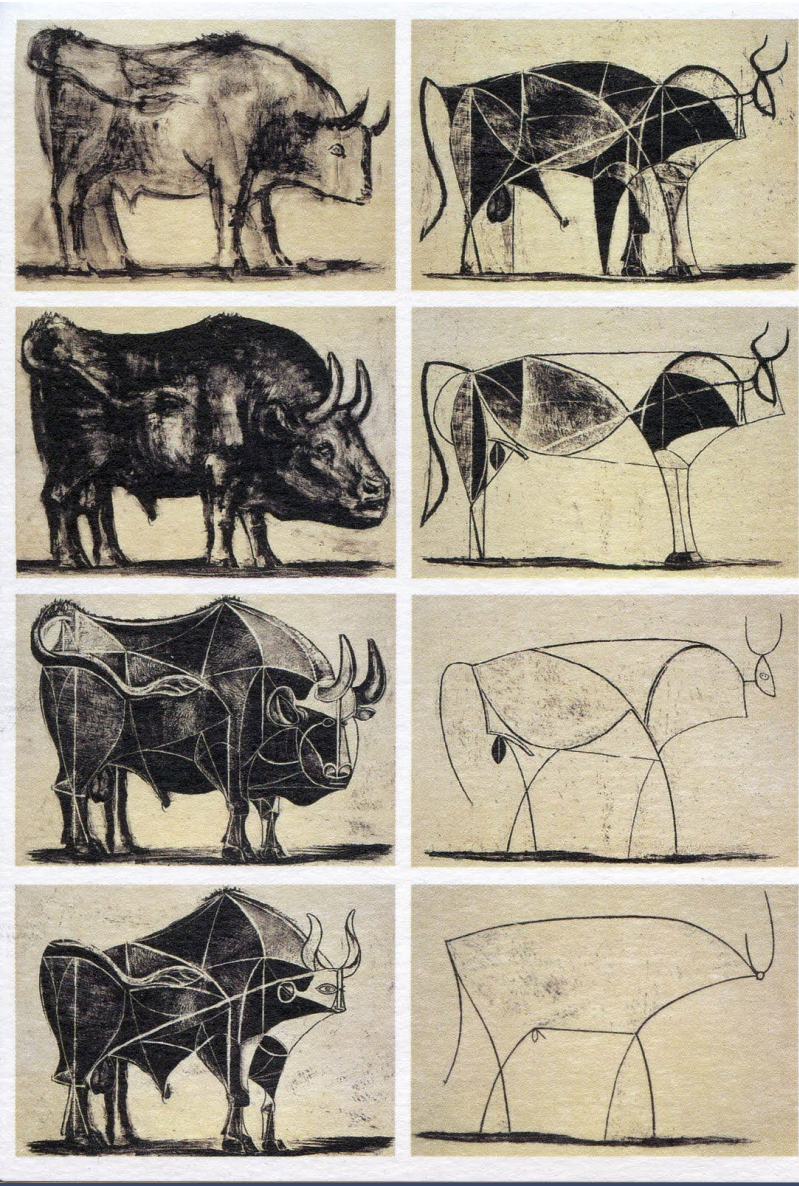


互联网+

- 多学科交叉，融合
- 计算机学科的起源，离不开多学科交叉

总体思路与教学方式

- 总体思路
  - ✓ 互联网应用和人工智能的相辅相成
  - ✓ 从基础知识（导论、数学、算法）进阶，进入人工智能的专业问题
  - ✓ 核心内容为人工智能的几大模块：预测、学习、规划、决策、交互
  - ✓ 预测、学习为基础；规划、决策为进阶；多智能体交互为高阶
  - ✓ 每一模块均从现实互联网应用出发，以数学模型为核心，引出具体理论
- 教学方式
  - ✓ 对象：面对学生为大二下半期，处于过渡期
  - ✓ 宗旨：抛弃难、偏问题，特别注重承前启后
  - ✓ 流程：应用案例→分解需求→抽出框架→引出理论
  - ✓ 操作：要经常让学生回顾数学在互联网与人工智能前沿问题中的意义



第一部分：互联网情境下的智能预测

- 背景调查
  - ✓ 介绍在线社交网络如QQ、微信、Facebook等
  - ✓ 讨论区别和相同的核心机制、各自的亮点和弱点
- 框架剖析
  - ✓ 巨大的图上的建模 （离散数学）
  - ✓ 大量的节点，统计意义 （概率论）
  - ✓ 不同场景有不同的约束条件，最好的预测？（线性代数、微积分）
  - ✓ 在线朋友关系预测的数理本质？ （图论+微积分+线性代数+概率论）
- 讲授核心
  - ✓ 图的定理、概率分布、矩阵的特征值在社交网络中的意义
  - ✓ 朋友关系的最优预测和偏微分的关系
  - ✓ 抛出谷歌PageRank、社团、聚类、分类等概念，引起兴趣

第二部分：互联网情境下的智能学习

- 背景调查
  - ✓ 介绍网上购物系统如京东\网易严选\唯品会（贸易）、天猫/淘宝/拼多多（平台）、
  - ✓ 是否有本质区别？为何表现如此不同？
- 框架剖析
  - ✓ 用户vs商品的建模与分析（二部图）
  - ✓ 用户画像、商品分类的数学本质是什么？ （特征值、特征向量）
  - ✓ 不同平台的数学模型的特点的不同？ （图结构的不同）
  - ✓ 在线商品推荐的数理依据是什么？（微分方程、梯度下降）
- 讲授核心
  - ✓ 二部图的建模和其上的学习训练
  - ✓ 矩阵分解、矩阵的秩的处理，及其在学习中的作用
  - ✓ 偏微分与梯度下降下的学习最优化
  - ✓ 引出统计学习、主成分分析等，为高年级学习引起兴趣

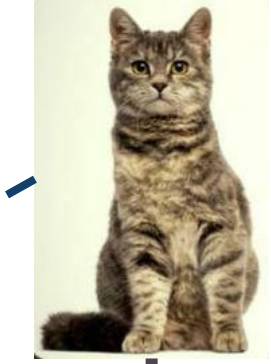
第三部分：互联网情境下的规划与决策

- 背景调查
  - ✓ 介绍在线博弈Google Master，AlphaGo，AlphaZero
  - ✓ DotA在线对战平台影魔机器人
  - ✓ 与传统游戏的区别：在线数据获取、在线交互、在线提升
- 框架剖析
  - ✓ 智能体vs环境的建模：人工智能的核心和最终目标
  - ✓ 强化学习与动态决策
  - ✓ 梯度下降、线性方程组、数学期望、动态规划算法在智能体交互的地位
- 讲授核心
  - ✓ 智能体所处环境的建模（线性方程组）
  - ✓ 智能体长期收益（数学期望）
  - ✓ 智能体动态决策（递归算法、动态规划）
  - ✓ 最优规划（偏微分方程）
  - ✓ 引出所需的高年级的自动机理论、随机过程

第四部分：互联网情境下的智能交互

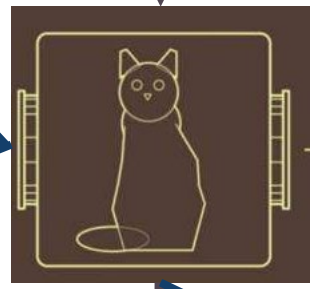
- 应用背景
  - ✓ 介绍谷歌、百度的广告位拍卖，5G网络的频谱共享和竞争，交通网络拥塞
  - ✓ 讨论背后的问题：多用户交互/竞争下的资源分配
- 框架剖析
  - ✓ 系统的各个参与者都是智能的个体
  - ✓ 多个个体需求不一样，智能体需求如何刻画？（线性方程组求解）
  - ✓ 从个体角度出发，“智能”体现在最优化自身的利益（偏导数）
  - ✓ 从系统的角度出发，“智能”体现在系统的整体最优
  - ✓ 但整体最优和个体最优往往冲突（博弈的基本思维）
- 讲授核心
  - ✓ 多智能体系统的建模、资源分配问题
  - ✓ 线性方程组、概率分布、偏微分在以上问题中的意义
  - ✓ 引出博弈问题、运筹学、凸优化，以供高年级时学习参考

近几年计算机热点技术研究路径



数据收集、处理

大数据，数据挖掘与分析



识别、分类

监督学习、非监督学习、深度学习

猫咪

老虎

决策、控制

强化学习、博弈

撸

跑

Game Theory VS Operation Research

运筹学方法 | 最优化 | 规划论 | 图论 | ……

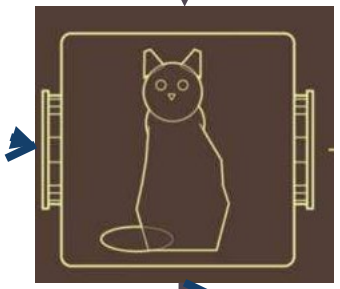
同样都能做决策和规划，博弈论和运筹学有什么区别?

资源本身 → **资源管理对象的演化** → **掌握或使用**资源**人或组织**

计算机与经济学的强关联性

数据管理

大数据采集、储存、清洗、打标



识别、分类、预测

数据挖掘与分析

（非）监督学习、深度学习

宏观经济学

猫咪

老虎

决策、控制

博弈论

（深度）强化学习

Agent

微观经济学

智能体

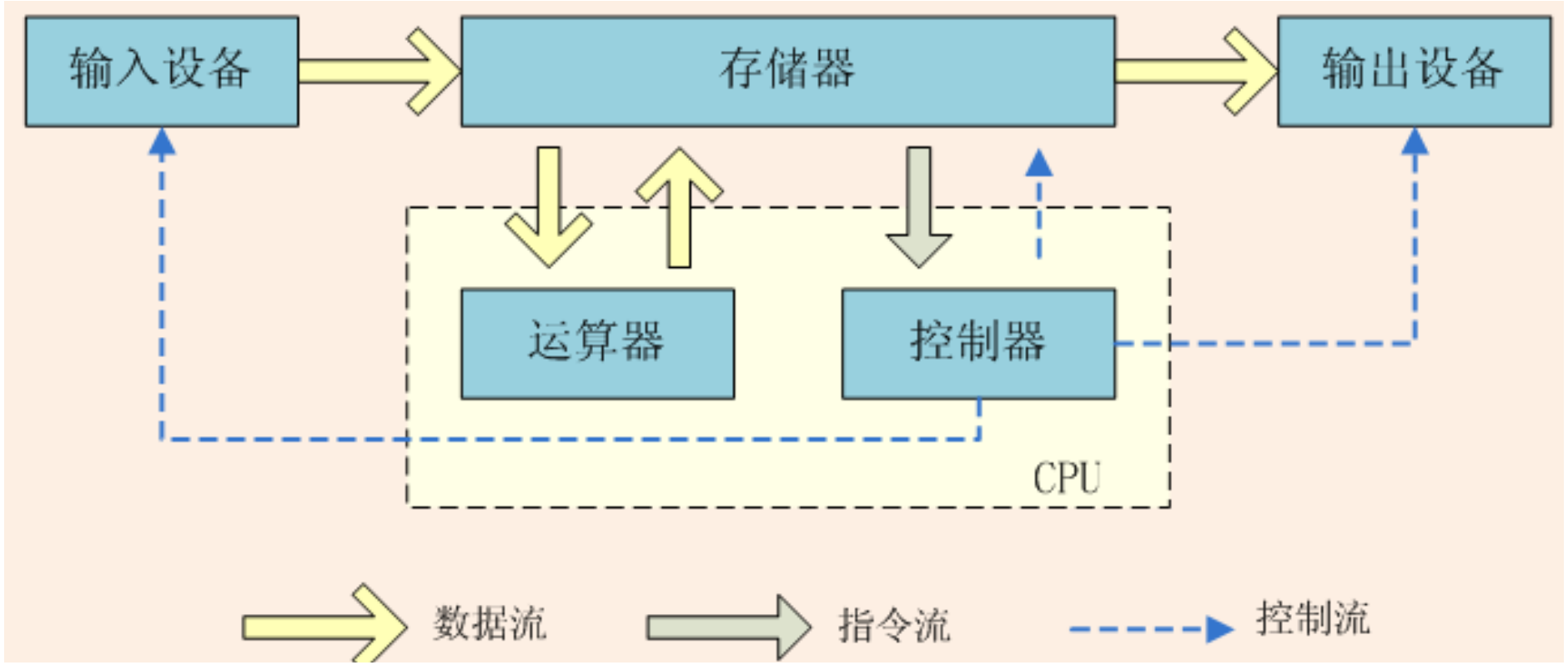
撸

跑

博弈论只跟经济学有关吗?

冯·诺依曼（John Von Neumann）

“计算机之父”：冯诺依曼体系结构（1946年）



“博弈论之父”：《博弈论与经济行为》（1944年出版）

博弈论的发展历史

起源

1928年，冯·诺依曼证明了博弈论基本原理，宣告博弈论的正式诞生

发展

1944年，冯·诺依曼和摩根斯坦共著的划时代巨著《博弈论与经济行为》将二人博弈推广到n人博弈结构并将博弈论系统地应用于经济领域，从而奠定了这一学科的基础和理论体系

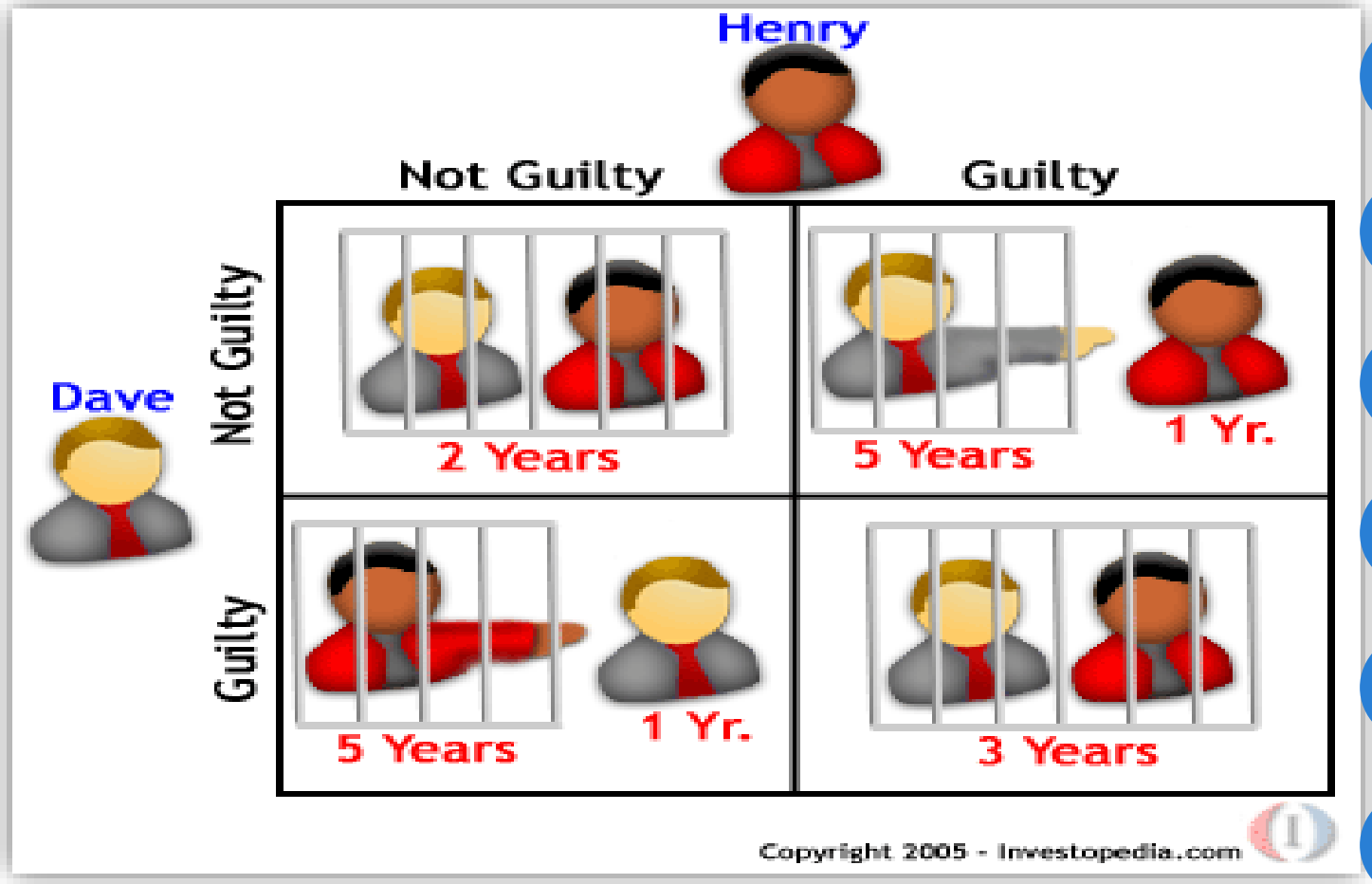
成熟

1950～1951年，约翰·福布斯·纳什（John Forbes Nash Jr）利用不动点定理证明了均衡点的存在，为博弈论的一般化奠定了坚实的基础。纳什的开创性论文《n人博弈的均衡点》（1950），《非合作博弈》（1951）等等，给出了纳什均衡的概念和均衡存在定理

成名

1994年，授予加利福尼亚大学伯克利分校的约翰·海萨尼（J.Harsanyi）、普林斯顿大学约翰·纳什（J.Nash）和德国波恩大学的赖因哈德·泽尔滕（Reinhard Selten）诺贝尔经济学奖

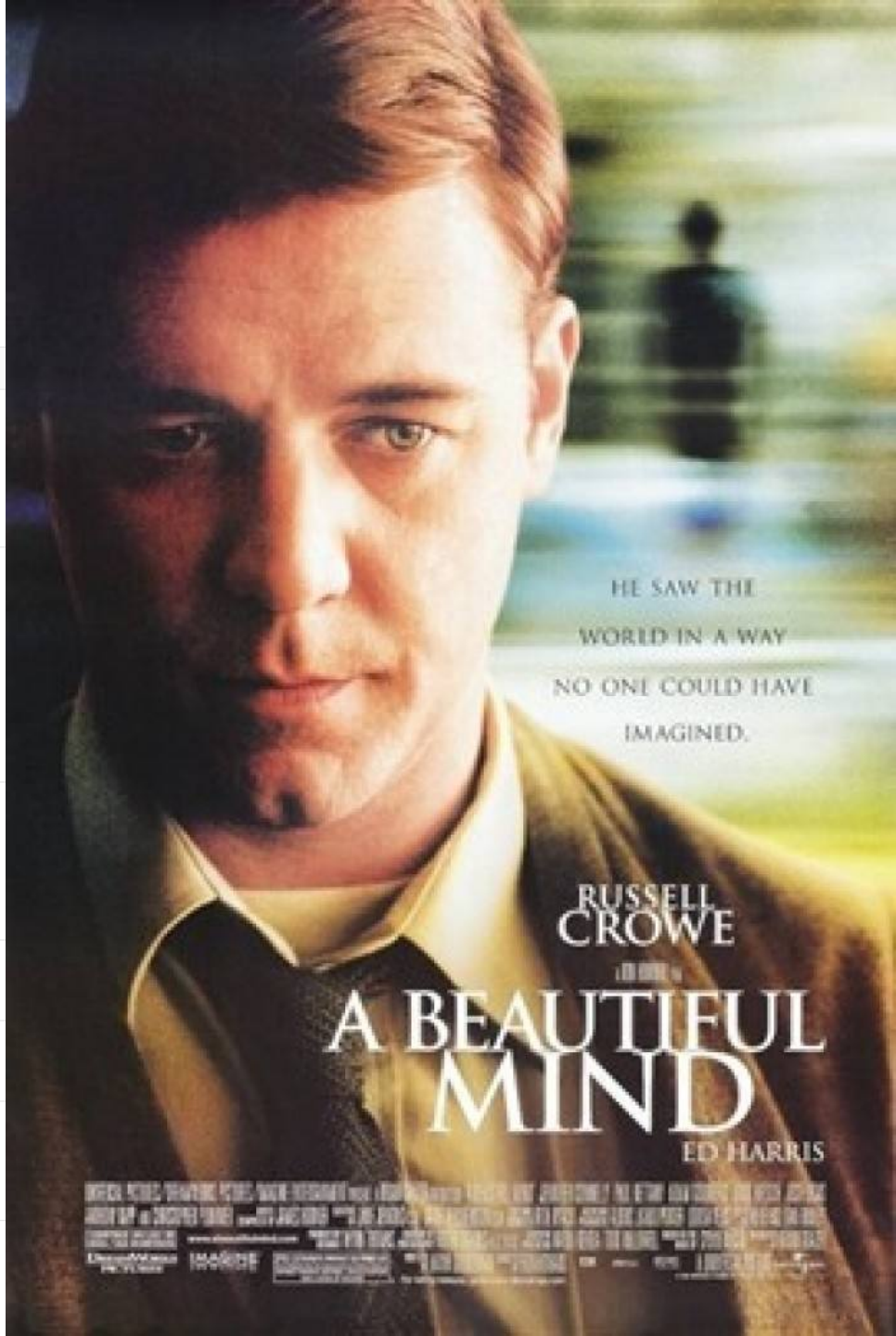
Nobel Prize in Economics



- **Game Theory**
  Nobel prize 1994
- **Mechanism Design**
  Nobel prize 1996
- **Information Asymmetric Market**
  Nobel prize 2001
- **Repeated game**
  Nobel prize 2005
- **Myerson's Lemma**
  Nobel prize 2007
- **Matching Theory**
  Nobel prize 2012
- **Contract Theory**
  Nobel prize 2014/2016
- **Behavior Economics**
  Nobel prize 2002/2017

A Beautiful Mind

年份	颁奖方	奖项	获奖方	结果
2002	第74届奥斯卡金像奖	最佳影片 最佳女配角 最佳导演 最佳改编剧本	《美丽心灵》 詹妮弗·康纳利 朗·霍华德 阿齐瓦·高斯曼	获奖
		最佳男主角 最佳电影剪辑 最佳化妆 最佳配乐	罗素·克劳 Mike Hill、Daniel P. Hanley、Greg Cannom Colleen Callaghan 詹姆斯·霍纳	提名
2002	第59届美国金球奖	剧情类最佳男主角 最佳女配角 剧情类最佳影片 最佳编剧	罗素·克劳 詹妮弗·康纳利 《美丽心灵》 阿齐瓦·高斯曼	获奖
		最佳导演 最佳电影配乐	朗·霍华德 詹姆斯·霍纳	提名
2002	第55届英国电影学院奖	最佳男主角 最佳女配角	罗素·克劳 詹妮弗·康纳利	获奖
		最佳影片 最佳剧本改编 大卫·林恩导演奖	《美丽心灵》 阿齐瓦·高斯曼 朗·霍华德	提名

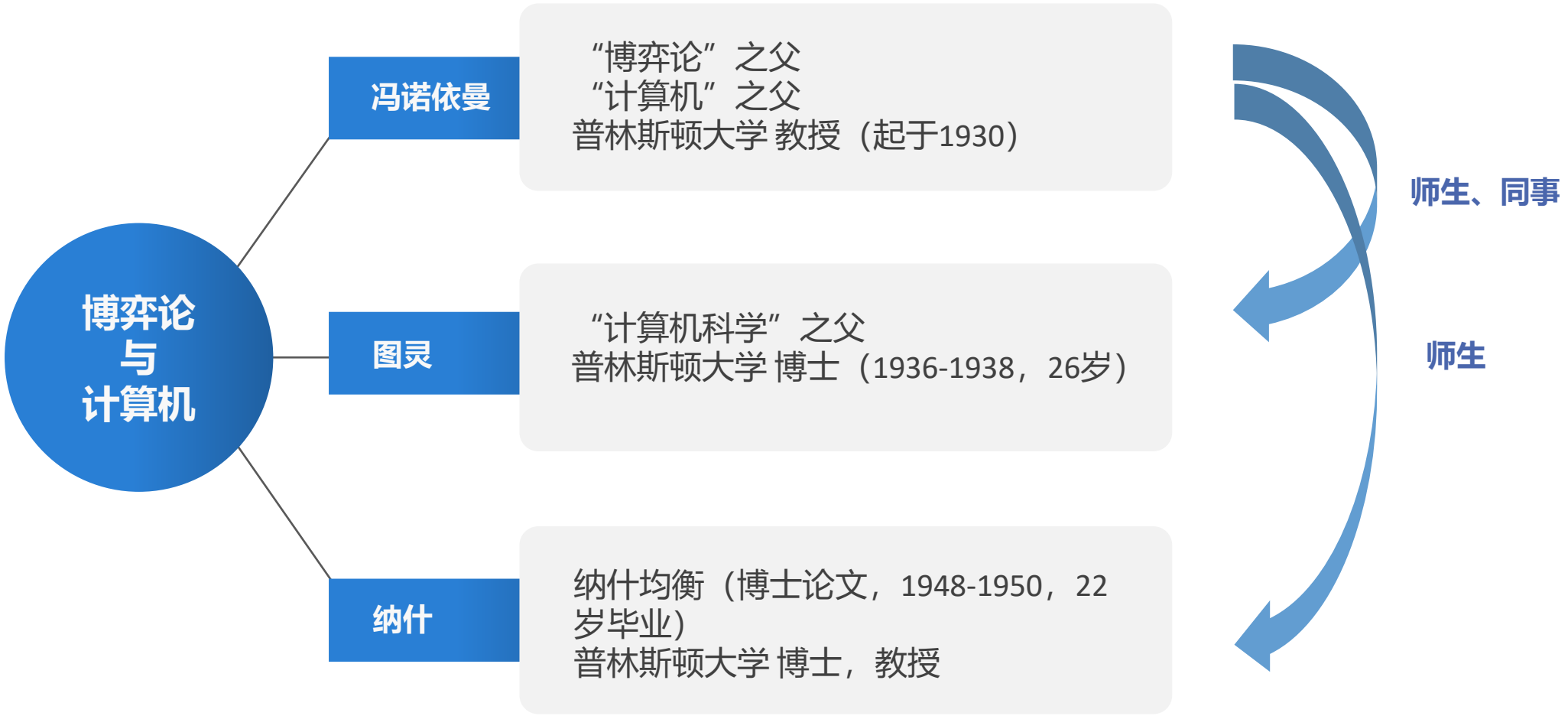


The Imitation Game

英国电影学院奖				
2015	第68届	最佳服装设计	Sammy Sheldon	（提名）
2015	第68届	最佳改编剧本	格拉汉姆·摩尔	（提名）
2015	第68届	最佳剪辑	William Goldenberg	（提名）
2015	第68届	最佳男主角	本尼迪克特·康伯巴奇	（提名）
2015	第68届	最佳影片		（提名）
2015	第68届	最佳女配角	凯拉·奈特莉	（提名）
2015	第68届	最佳艺术指导	Maria Djurkovic，Tatiana Lund	（提名）
2015	第68届	杰出英国电影		（提名）
2015	第68届	最佳音效		（提名）

奥斯卡金像奖				
2015	第87届	最佳艺术指导	Maria Djurkovic，Tatiana Lund	（提名）
2015	第87届	最佳改编剧本 [3]	格拉汉姆·摩尔	（获奖）
2015	第87届	最佳导演	莫腾·泰杜姆	（提名）
2015	第87届	最佳影片		（提名）
2015	第87届	最佳男主角	本尼迪克特·康伯巴奇	（提名）
2015	第87届	最佳剪辑	William Goldenberg	（提名）
2015	第87届	最佳女配角	凯拉·奈特莉	（提名）
2015	第87届	最佳原创配乐	亚历山大·迪斯普拉特	（提名）

The Relationships Between Gamers and Computers



教材

- 教材
  - ✓《人工智能：一种现代方法》，Stuart Russell and Peter Norvig，清华大学出版社，北京，ISBN: 7302128294 ，2006.
- 辅助教材
  - ✓《网络、群体与市场：揭示高度互联世界的行为原理与效应机制》，David Esley ， Jon Kleinberg，清华大学出版社，北京，ISBN: 9787302264170 ，2011.
  - ✓Reinforcement Learning: An Introduction, Richard S.Sutton and Andrew G. Barto, MIT Press, Second Edition, 2018.

第1部分 人工智能
第1章 绪论
1.1 什么是人工智能
1.2 人工智能的基础
1.3 人工智能的历史
1.4 最新发展水平
1.5 本章小结
参考文献与历史注释
习题
第2章 智能Agent
2.1 Agent和环境
2.2 好的行为：理性的概念
2.3 环境的性质
2.4 Agent的结构
2.5 本章小结
参考文献与历史注释
习题

第II部分 问题求解
第3章 通过搜索进行问题求解
3.1 问题求解Agent
3.2 问题实例
3.3 通过搜索求解
3.4 无信息搜索策略
3.5 有信息（启发式）的搜索策
3.6 启发式函数
3.7 本章小结
参考文献与历史注释
习题
第4章 超越经典搜索
4.1 局部搜索算法和最优化问题
4.2 连续空间中的局部搜索
4.3 使用不确定动作的搜索
4.4 使用部分可观察信息的搜索
4.5 联机搜索Agent和未知环境
4.6 本章小结
参考文献与历史注释
第5章 对抗搜索
5.1 博弈
5.2 博弈中的优化决策
5.3 a-p剪枝
5.4 不完美的实时决策
5.5 随机博弈
5.6 部分可观察的博弈
5.7 博弈程序发展现状
5.8 其他途径
5.9 本章小结
参考文献与历史注释
习题
第6章 约束满足问题
6.1 定义约束满足问题
6.2 约束传播：CSP中的推理
6.3 CSP的回溯搜索
6.4 CSP局部搜索
6.5 问题的结构
6.6 本章小结
参考文献与历史注释
习题

第III部分 知识、推理与规划
第7章 逻辑Agent
7.1 基于知识的Agent
7.2 Wumpus世界
7.3 逻辑
7.4 命题逻辑：一种简单逻辑
7.5 命题逻辑定理证明
7.6 有效的命题逻辑模型检验
7.7 基于命题逻辑的Agent
7.8 本章小结
.....

第IV部分 不确定知识与推理
第V部分 学习
第VI部分 通讯、感知与行动
第VII部分 结论
参考文献

第1章 概述
第一部分 图论与社会网络
第2章 图论
第3章 强联系和弱联系
第4章 网络及其存在的环境
第5章 正关系与负关系
第二部分 博弈论
第6章 博弈
第7章 进化博弈论
第8章 网络流量的博弈论模型
第9章 拍卖
第三部分 网络中的市场与策略性互动
第10章 匹配市场
第11章 具有中介的市场网络模型
第12章 网络中的议价与权力
第四部分 信息网络与万维网
第13章 万维网结构
第14章 链接分析和网络搜索
第15章 商业支持的搜索市场
第五部分 网络动力学：总体模型
第16章 信息级联
第17章 网络效应
第18章 幂律与富者更富现象
第六部分 网络动力学：结构模型
第19章 网络中的级联行为
第20章 小世界现象
第21章 流行病学
第七部分 机构及其聚合行为
第22章 市场与信息
第23章 表决
第24章 产权

考核方式及成绩构成

- 最终成绩
  - 平时成绩50%，期末成绩50%
- 平时成绩50%
  - 第一次课程设计占15%
  - 第二次课程设计占15%
  - 最终提交课程论文占20%
- 课程设计内容：强化学习
  - 知识筹备
    - Python编程基础
    - 强化学习的基础知识
    - 简单了解深度学习

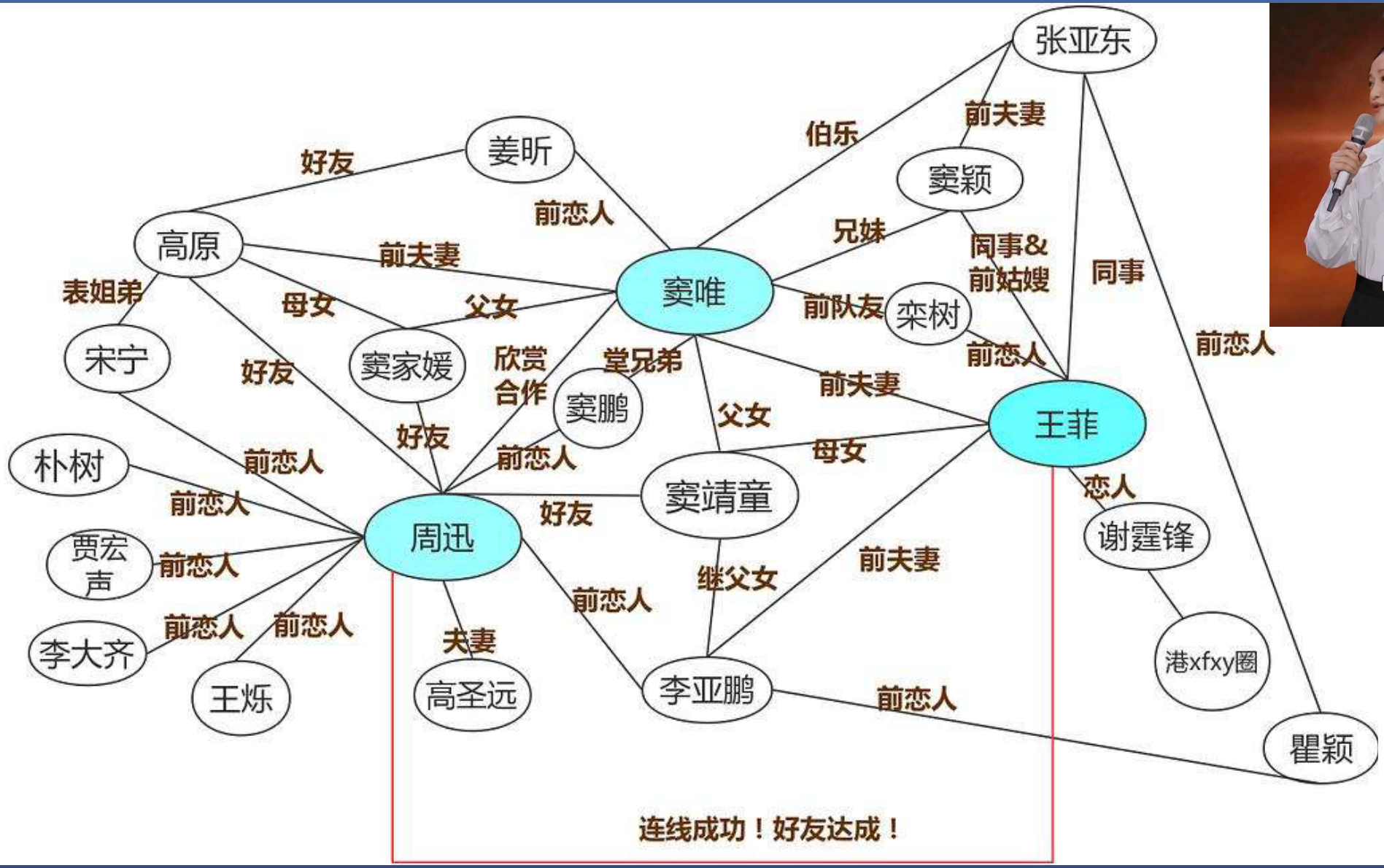
互联网与计算智能

第一部分

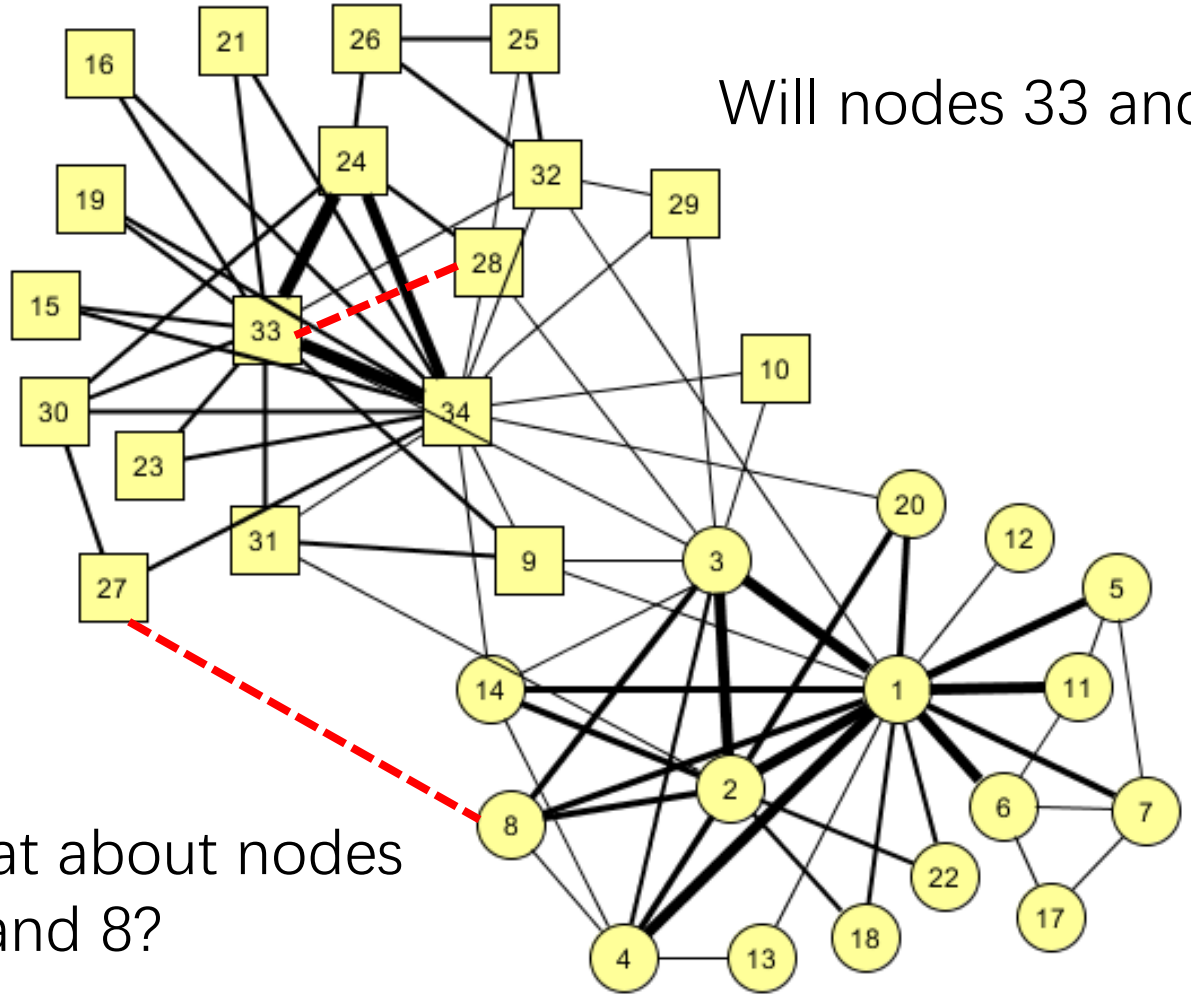
互联网情境下的智能预测

张彦如
yanruzhang@uestc.edu.cn

第一部分：互联网情境下的智能预测



第一部分：互联网情境下的智能预测



Will nodes 33 and 28 become friends in the future?

Does network structure contain enough information to predict what new links will form in the future?

What about nodes 27 and 8?

强连接与弱连接

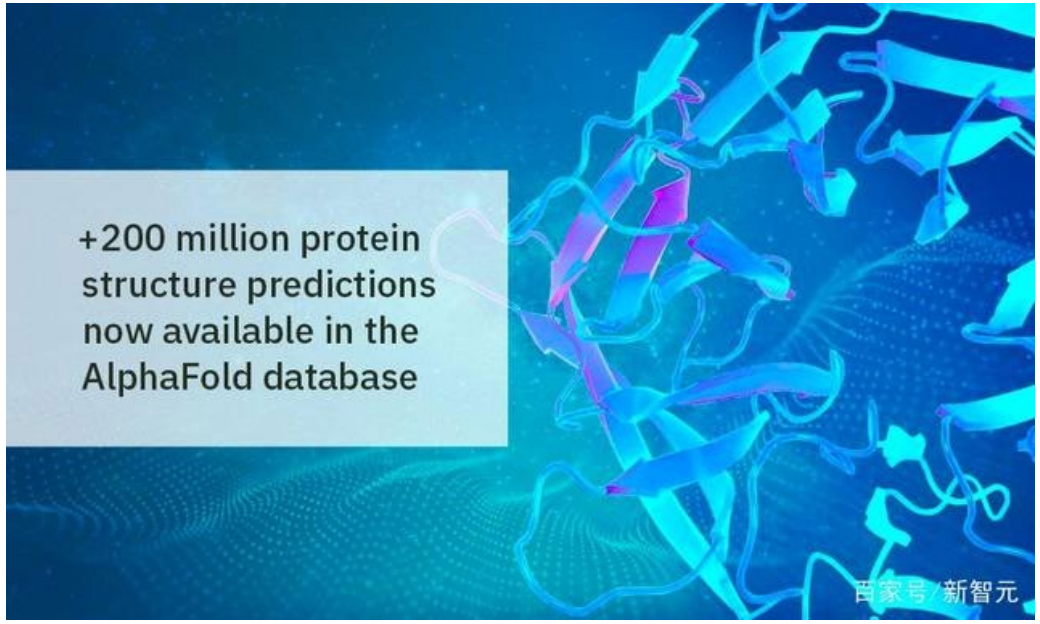
- Strong ties 强连接
  - Surrounded by *many* mutual friends
  - Characterized by lots of shared time together
  - 亲人、同学、朋友、同事……这是十分稳定的，然而传播范围有限的社会认知，这是“强连接”
- Weak ties 弱连接
  - Have few mutual friends
  - Serve as bridges to diverse parts of the network
  - Provide access to novel information
  - 例如无意间提到或者打开收音机偶然听到的一个人

社会网络中的智能预测

- The Link-Prediction Problem for Social Networks

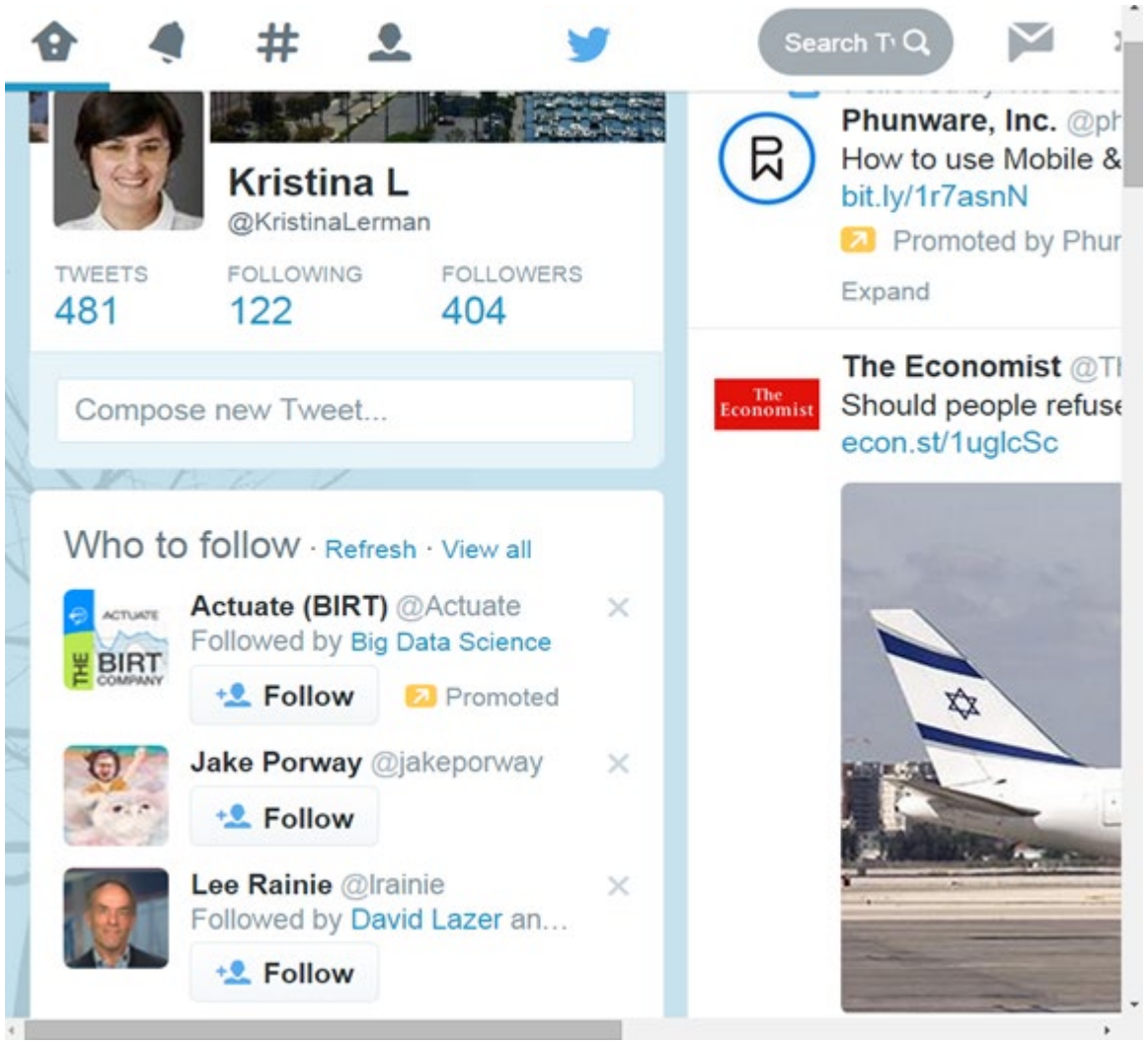
  (Liben-Nowell & Kleinberg)

  - Estimate the likelihood of the existence of a link between two nodes, based on *observed links* and the *attributes of nodes*
    - **未知链接（exist yet unknown links）**
    - **未来链接（future links）**



- Applications
  - Biological networks: costly to identify links between nodes through field/laboratorial experiments
  - **DeepMind祭出预测新冠病毒“AlphaFold”重磅武器**
  - Online social networks: predicting friendship and recommending new friends

关注与好友推荐


Search T
Kristina L
@KristinaLerman
TWEETS
481
FOLLOWING
122
FOLLOWERS
404
Compose new Tweet...
Who to follow · Refresh · View all
Actuate (BIRT) @Actuate
Followed by Big Data Science
Follow
Promoted
Jake Porway @jakeporway
Follow
Lee Rainie @lrainie
Followed by David Lazer an...
Follow
Phunware, Inc. @ph
How to use Mobile &
bit.ly/1r7asnN
Promoted by Phur
Expand
The Economist @T
Should people refuse
econ.st/1uglcSc

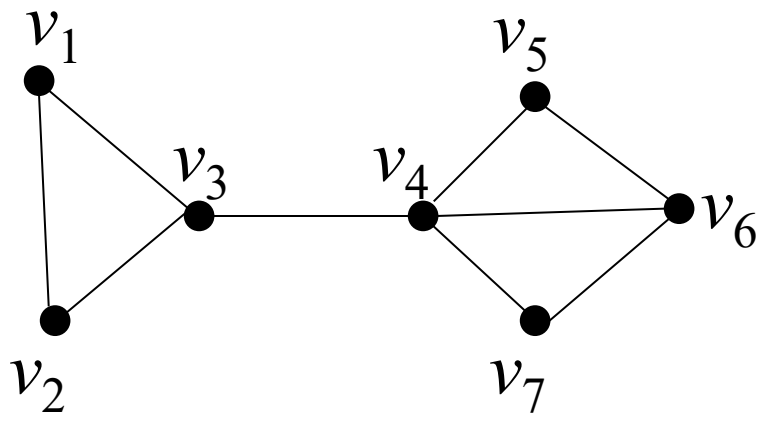

QQ好友推荐
为你找到了开通微博的QQ好友，快来收听他们吧！
(已选择33)
选择你要收听的QQ好友
马鑫
qq好友-马鑫
马震天
qq好友-马震天
博玉蓉
qq好友-
许士崇病圣子
qq好友-病圣子
黄海娟
qq好友-遗心锁爱
伤心城市
qq好友-笑云听
Crusader_yu
qq好友-於立
夏敏
qq好友-小夏
谭星怡
qq好友-
全选
收听已选


1 关注他们
2 邀请其他人
你的MSN联系人中已有32人在搜狐微博中了，立即关注他们吧。
全选
魔人
@hotmail.com
陽光不銹
@hotmail.com
水深火热
@hotmail.com

Review of Graph-UnDirected

Definition: A graph $G = (V, E)$ consists of V, a nonempty set of vertices (or nodes), and E, a set of edges.

Each edge has either one or two vertices associated with it, called its endpoints. An edge is said to connect its endpoints.

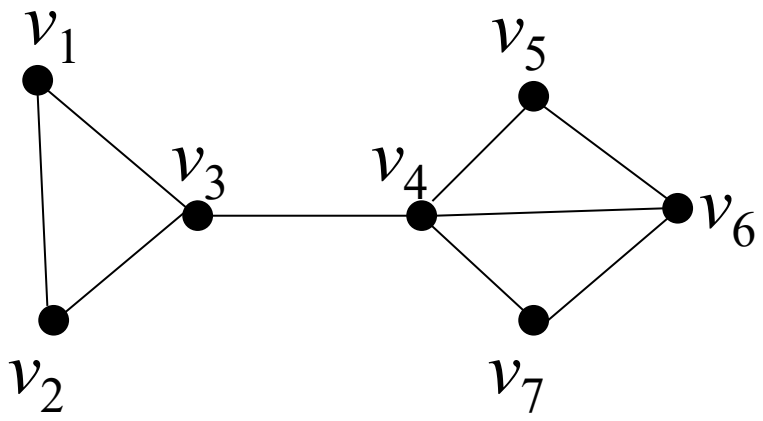


$G=(V, E)$, where

$V=\{v_1, v_2, \ldots, v_7\}$

$E=\{\{v_1,v_2\}, \{v_1,v_3\}, \{v_2,v_3\}$
$\{v_3,v_4\}, \{v_4,v_5\}, \{v_4,v_6\}$
$\{v_4,v_7\}, \{v_5,v_6\}, \{v_6,v_7\}\}$

Adjacency Matrix 邻接矩阵



$G=(V, E)$, where

$V=\{v_1, v_2, \ldots, v_7\}$

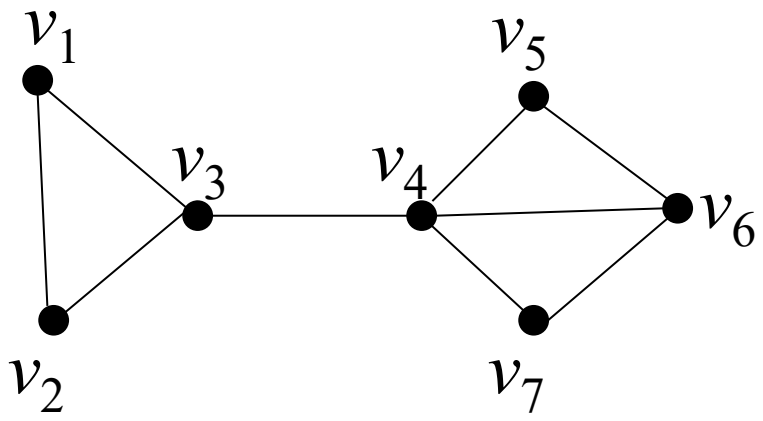
$E=\{\{v_1,v_2\}, \{v_1,v_3\}, \{v_2,v_3\}$
$\{v_3,v_4\}, \{v_4,v_5\}, \{v_4,v_6\}$
$\{v_4,v_7\}, \{v_5,v_6\}, \{v_6,v_7\}\}$

	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_1							
v_2							
v_3							
v_4							
v_5							
v_6							
v_7							

邻接关系

连边存在，值为1，Otherwise，值为0

Adjacency Matrix 邻接矩阵



$G=(V, E)$, where

$V=\{v_1, v_2, \ldots, v_7\}$

$E=\{\{v_1,v_2\}, \{v_1,v_3\}, \{v_2,v_3\}$
$\{v_3,v_4\}, \{v_4,v_5\}, \{v_4,v_6\}$
$\{v_4,v_7\}, \{v_5,v_6\}, \{v_6,v_7\}\}$

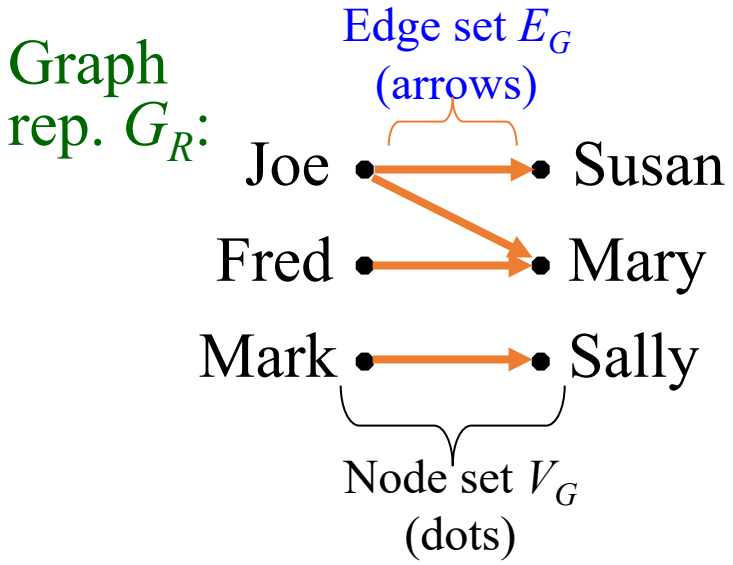
	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_1	0	1	1	0	0	0	0
v_2	1	0	1	0	0	0	0
v_3	1	1	0	1	0	0	0
v_4	0	0	1	0	1	1	1
v_5	0	0	0	1	0	1	0
v_6	0	0	0	1	1	0	1
v_7	0	0	0	1	0	1	0

矩阵特征

- **对角线上的元素都为 0：没有自己到自己的边**
- **对称矩阵：无向图**
- **不完全图：完全图**
- **顶点 v_i 的度：第 i 行/列中元素之和**

Review of Graph-Directed

- A ***directed graph*** or ***digraph*** $G=(V_G,E_G)$ is a set V_G of *vertices (nodes)* with a set $E_G \subseteq V_G \times V_G$ of *edges (links)*
- Visually represented using dots for nodes, and arrows for edges. Notice that a relation $R:A \times B$ can be represented as a graph $G_R=(V_G=A \cup B,\ E_G=R)$



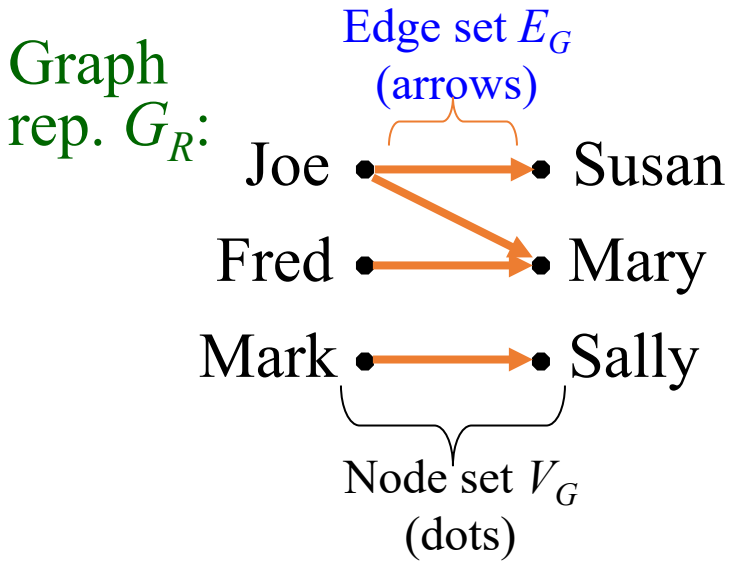
有向图邻接关系

- **第 i 行含义：从顶点 v_i 出发的边，出度边，箭头的尾**
- **第 i 列含义：到顶点 v_i 结束的边，入度边，箭头的头**

	Susan	Mary	Sally
Joe			
Fred			
Mark			

Review of Graph-Directed

	Susan	Mary	Sally
Joe	1	1	0
Fred	0	1	0
Mark	0	0	1



矩阵特征

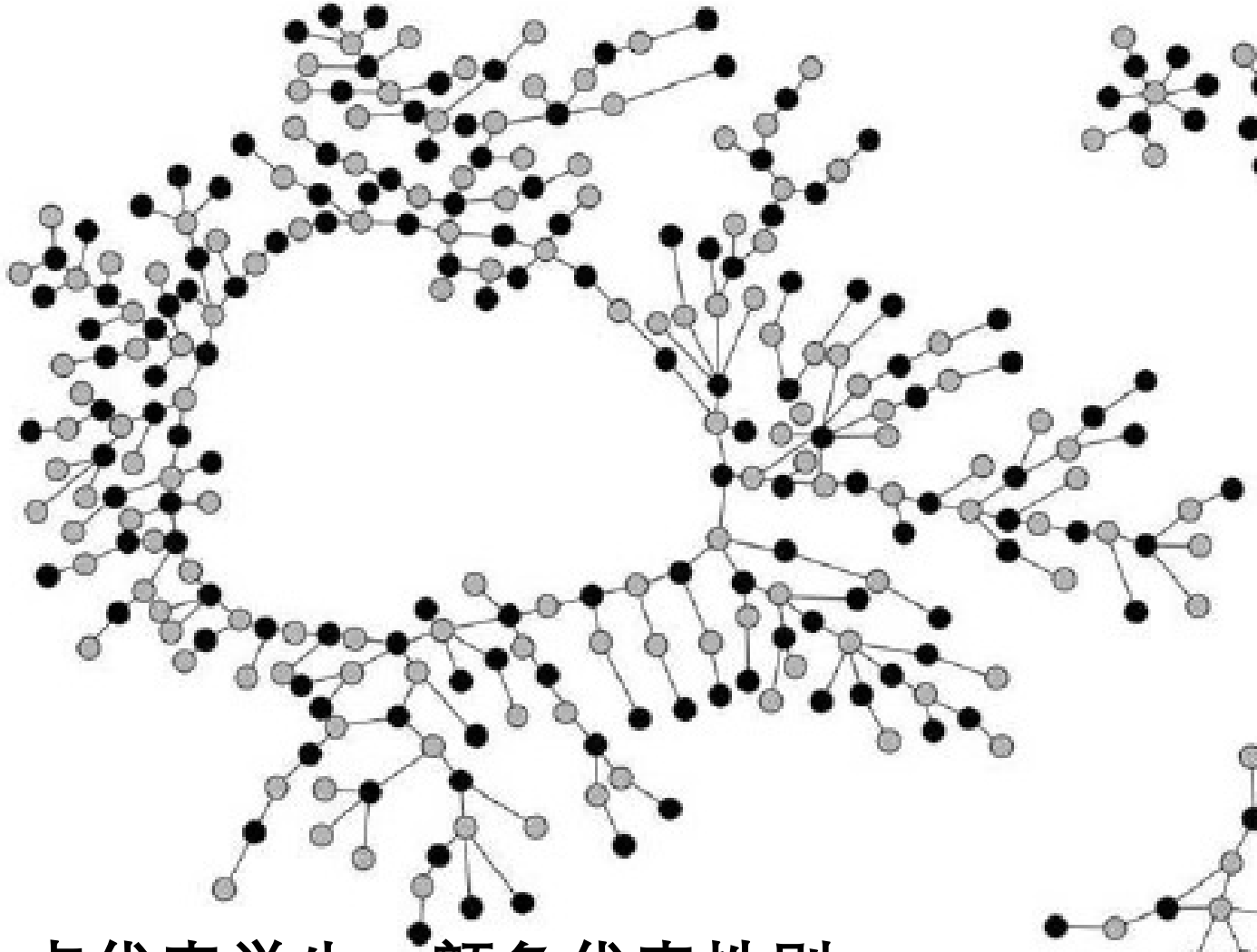
- **不对称矩阵：有向图**
- **有向图有可能是对称的吗?**

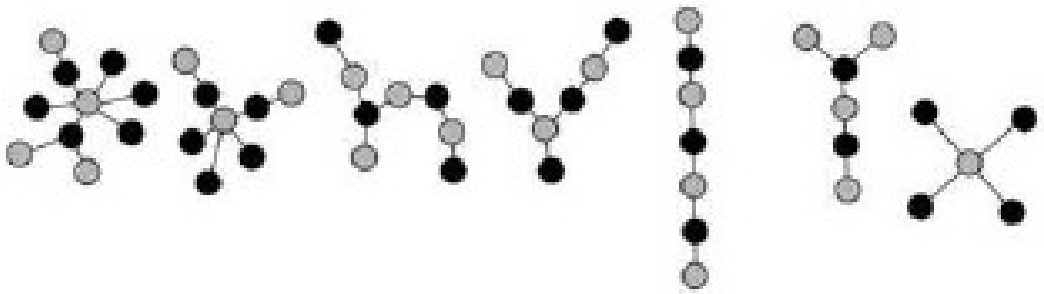
完全图

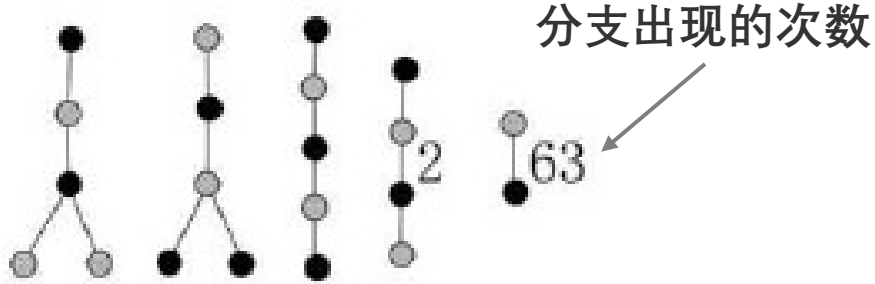
顶点 v_i 的度 =出度+入度

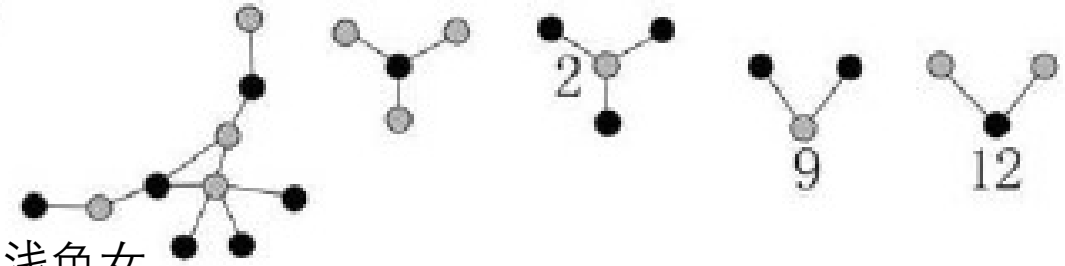
- **出度=第 i 行中元素之和**
- **入度=第 i 列中元素之和**

美国高中人际网络









点代表学生，颜色代表性别 深色男，浅色女

邻接矩阵的优缺点

	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_1	0	1	1	0	0	0	0
v_2	1	0	1	0	0	0	0
v_3	1	1	0	1	0	0	0
v_4	0	0	1	0	1	1	1
v_5	0	0	0	1	0	1	0
v_6	0	0	0	1	1	0	1
v_7	0	0	0	1	0	1	0

优点

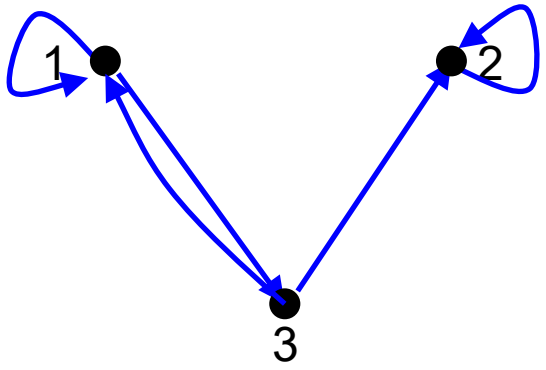
- 直观、简单
- 方便检查是否有连边存在
- 方便寻找任意顶点的邻接点
- 方便计算任意顶点的度

缺点

- 不便于增加和删除顶点
- 浪费空间：稀疏图（点多，边少），有大量元素
- 浪费时间：稀疏图统计需要大量时间

邻接矩阵的n次方

R contains edges between all the nodes reachable via 1 edge
$R \circ R = R^2$ contains edges between nodes that are reachable via 2 edges in R
R^n = contains edges between nodes that are reachable via n edges in R
R^* contains edges between nodes that are reachable via any number of edges (i.e. via any path) in R



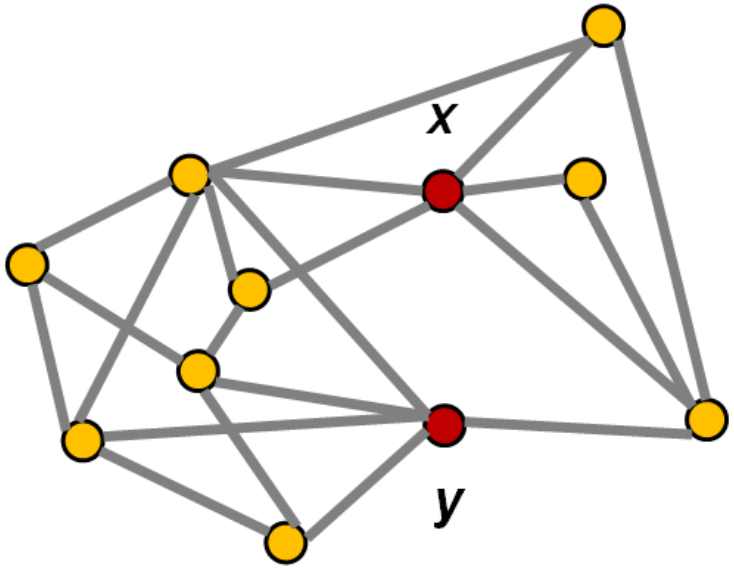
$$\mathbf{M}_R = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

Six Degrees of Separation

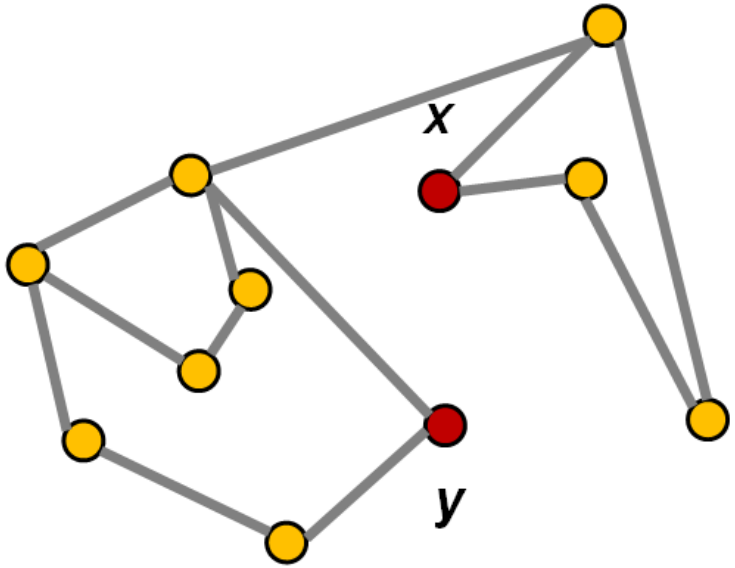
$$\mathbf{M}_R^{[2]} = \mathbf{M}_R \times \mathbf{M}_R = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Problem Description

- In many networks, people (objects) who are “close” belong to the same social circles and will inevitably encounter one another and become linked themselves.
- Link prediction algorithms measure how “close” people are



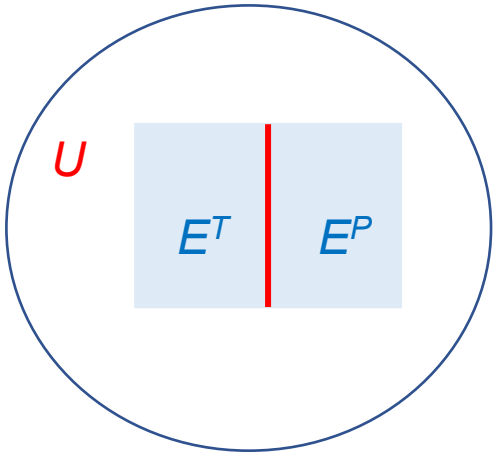
Red nodes are close to each other



Red nodes are more distant

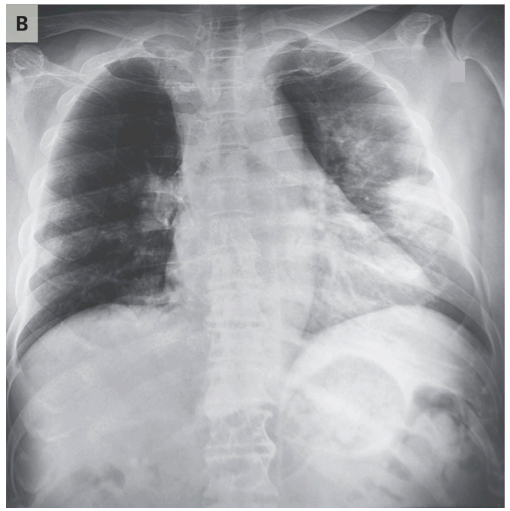
链路预测 Problem Description

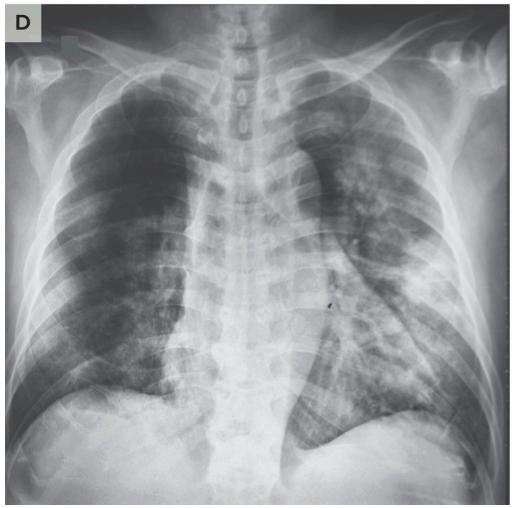
- Un-Directed network $G = (V, E)$
- Universal set U containing $|V|(|V|-1)/2$ possible links
- Objective: Find out missing links in $U - E$
- Prediction: randomly split E into two sets: training set E^T, probe/validation set E^P
  - E^P is a subset of $U - E^T$



新冠肺炎辅助确诊

在临床特征上，138 个病例中，常见症状包括发烧（136 人），疲劳（96 人）和干咳（82 人）。97 例患者出现了淋巴细胞减少症，97 例患者出现凝血酶原时间延长，55 位患者出现乳酸脱氢酶升高。同时，所有病人的肺部都有斑片状阴影或磨玻璃样阴影。





核酸检测确诊

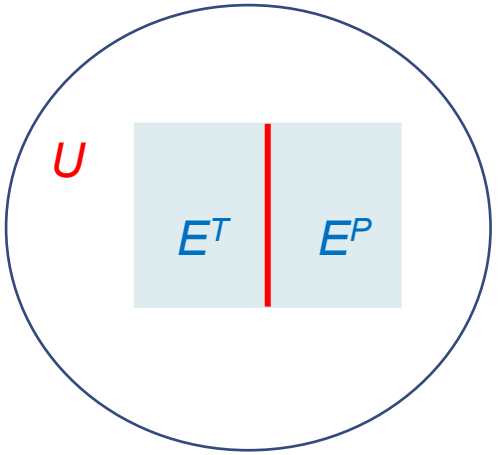
- 时间、人力成本
- 假阴性

特殊时期，临床确诊手段

体温、流行病史、血液常规、CT

机器学习辅助确诊

- 全集U
- 数据集E
  - 训练集E^T，测试集E^P
- 小样本问题



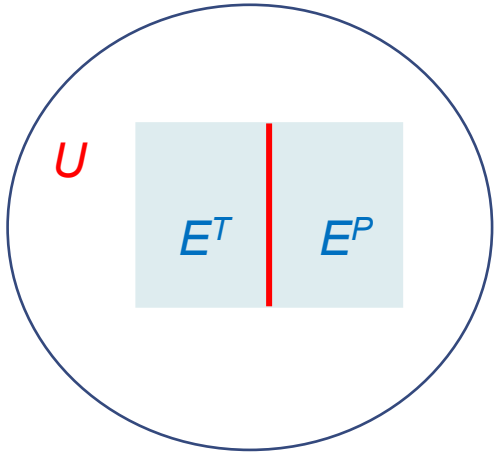
预测

We want to use the variables (Chest Pain, Good Blood Circulation, etc.)…

已知
是否患病

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
…	…	…	…	…

训练



…and predict if they have heart disease or not.

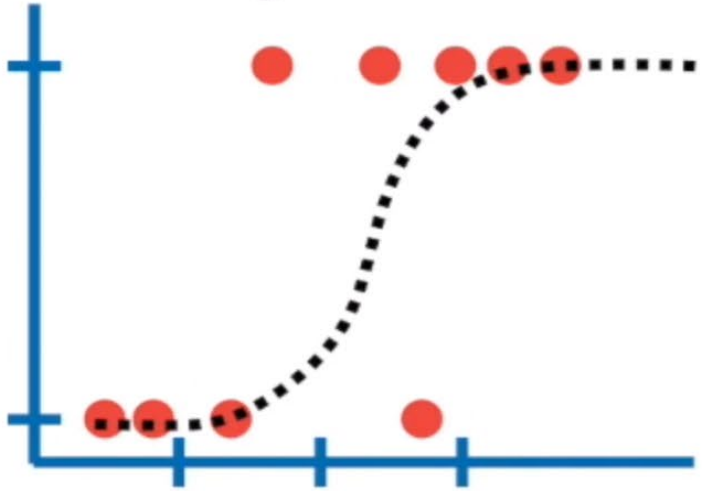
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	???

未知
是否患病

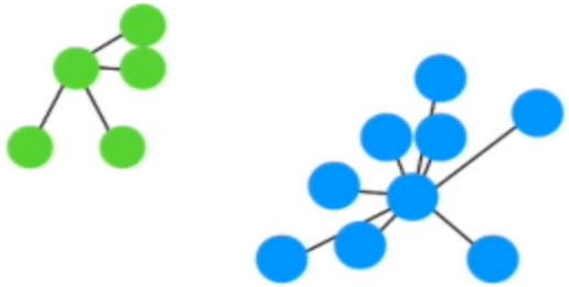
测试/验证

方法选择

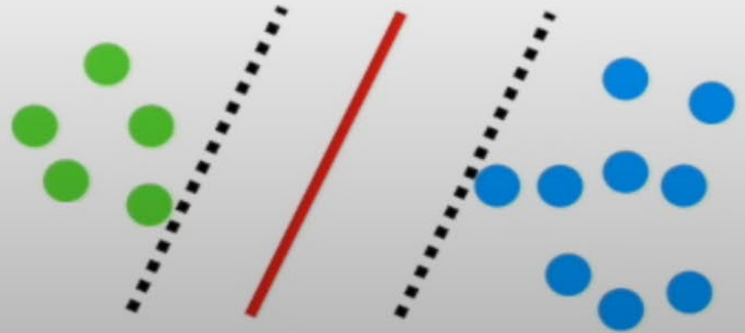
We could use Logistic Regression…



…or K-nearest neighbors…



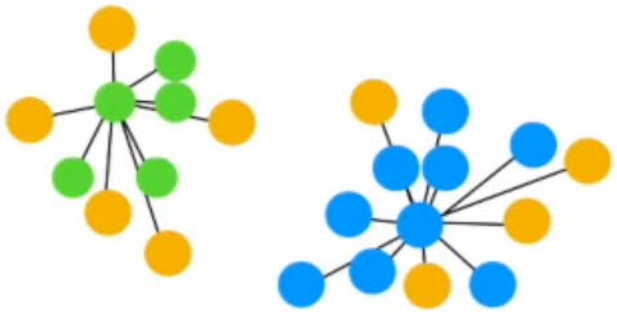
…or support vector machines (SVM)…

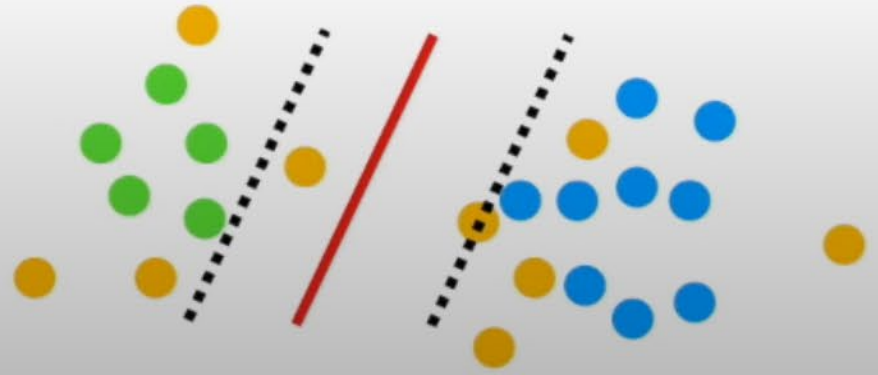


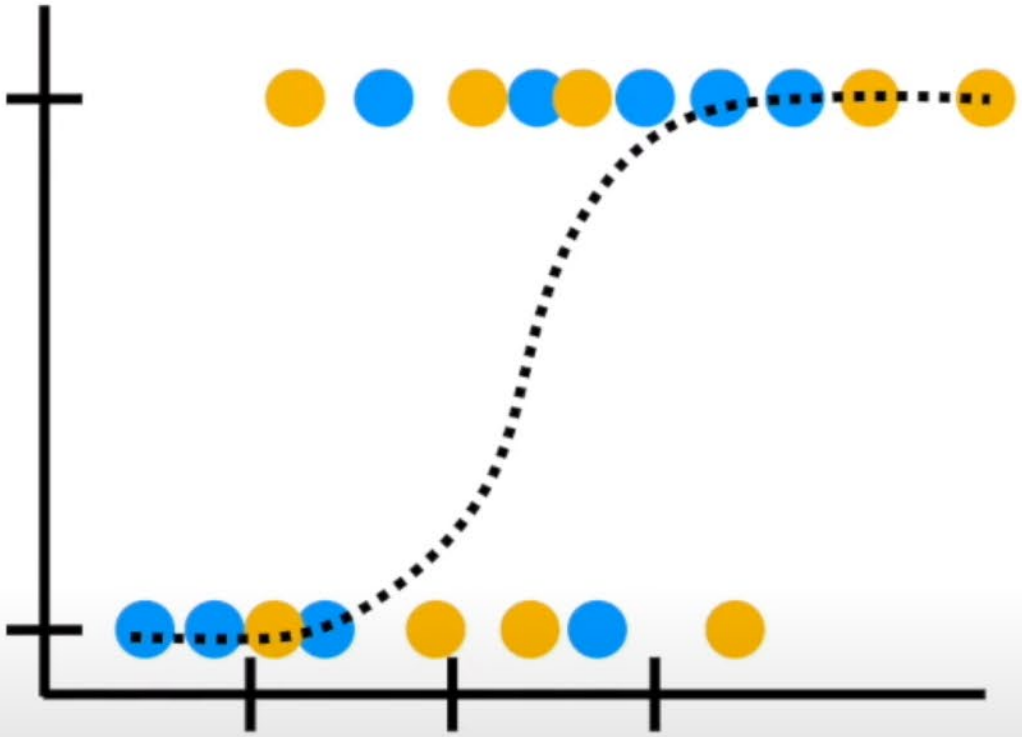
…and many more machine learning methods…

方法评价与验证

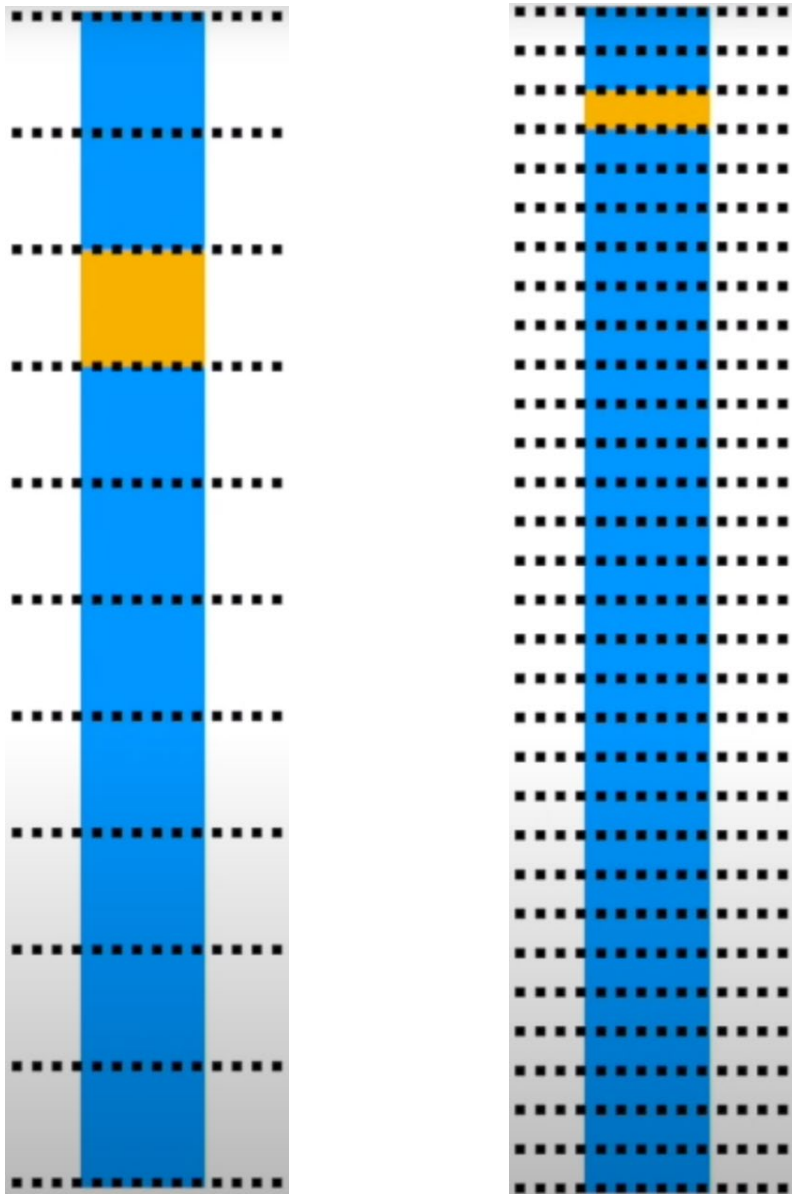
We could then compare methods by seeing how well each one categorized the test data.







Cross Validation 交叉验证



定义

Cross Validation也叫Rotation Estimation，是一种统计学上将数据样本切割成较小子集的实用方法

目的

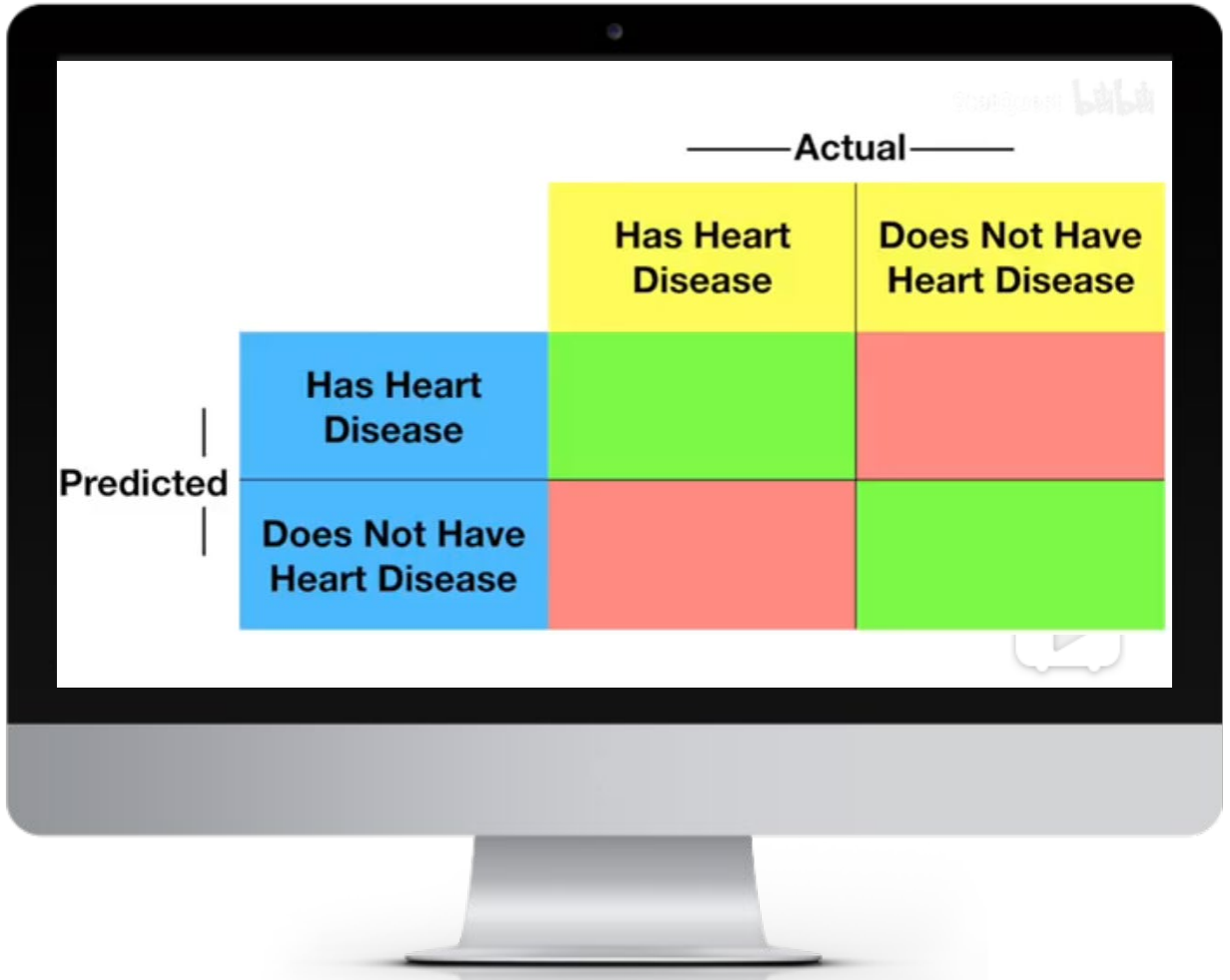
用交叉验证的目的是为了得到可靠稳定的模型

常见形式 K-Fold Cross-Validation

- Randomly partition into *k* subsets
- Each time one subset is selected as probe set, the others as training set
- Repeat *k* times, each with a different probe set

- Leave one out cross validation

混淆矩阵(Confusion Matrix)



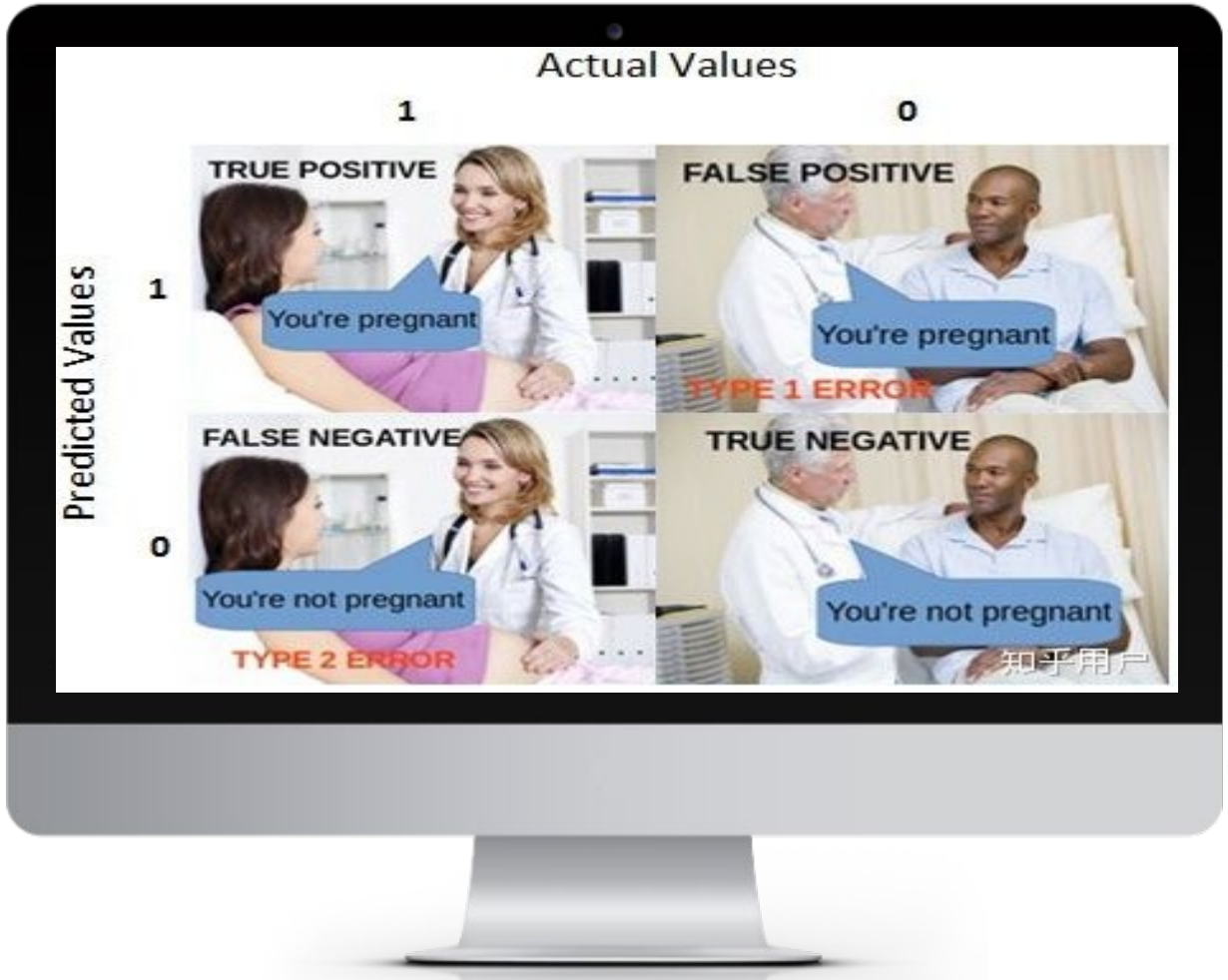
定义

也称误差矩阵，是表示精度评价的一种标准格式，用n行n列的矩阵形式来表示

二元分类-预测类别

- TP = True Positive = 真阳性
- FP = False Positive = 假阳性
- FN = False Negative = 假阴性
- TN = True Negative = 真阴性

混淆矩阵(Confusion Matrix)



二元分类-预测类别

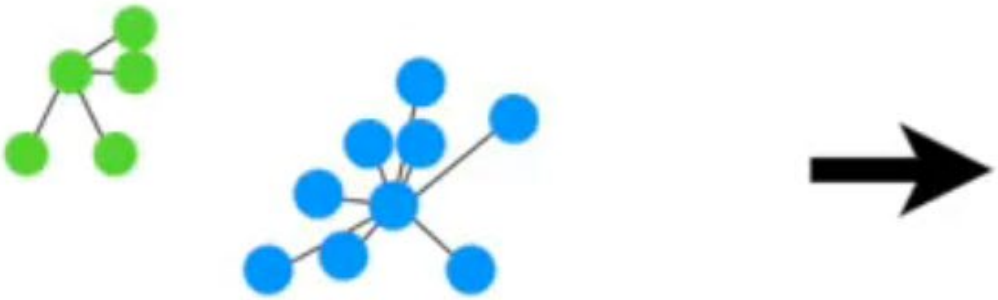
- TP = True Positive = 真阳性
- FP = False Positive = 假阳性
- FN = False Negative = 假阴性
- TN = True Negative = 真阴性

	实际的情况	
预测的情况	真阳性（TP） 预测为真，实际也为真。	假阳性（FP） 预测为真，实际为假。
	假阴性（FN） 预测为假，实际上为真。	真阴性（TN） 预测为假，实际上为假。

直接用混淆矩阵进行评价

	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	142	22
Does Not Have Heart Disease	29	110

哪一个方法更好?

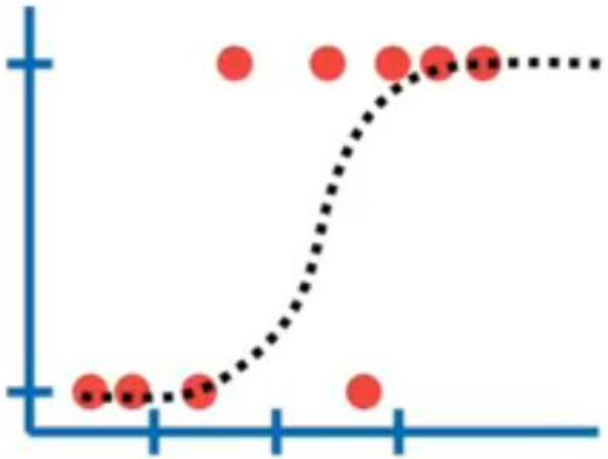


	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	107	53
Does Not Have Heart Disease	64	79

直接用混淆矩阵进行评价

	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	142	22
Does Not Have Heart Disease	29	110

哪一个方法更好?



	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	139	20
Does Not Have Heart Disease	32	112

机器学习性能评估指标

	实际的情况	
预测的情况	真阳性（TP） 预测为真，实际也为真。	假阳性（FP） 预测为真，实际为假。
	假阴性（FN） 预测为假，实际上为真。	真阴性（TN） 预测为假，实际上为假。

灵敏度 Sensitivity=召回 Recall

真阳性率，实际阳性案例中，检测出阳性的概率：检测出确实有病的能力

特异性 Specificity

真阴性率：实际阴性案例中，检测出阴性的几率：检测出确实没病的能力

Precision

- 被正确判为阳性的案例占所有检测为阳性的比例
- 被判为阳性的案例中实际为阳性的比例

Accuracy

正确分类的所有样本数与总样本数之比，该值越高，分类效果越好

链路预测/机器学习性能评估指标

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition positive	Condition negative	
Fecal occult blood screen test outcome	Test outcome positive	True positive (TP) = 20	False positive (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test outcome negative	False negative (FN) = 10	True negative (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

- **假阴性率**=假阴/（假阴+真阳） **=1-真阳性率**
- **假阳性率**=假阳/（真阴+假阳） **=1-真阴性率**

灵敏度 Sensitivity=召回 Recall

- 真阳性率，实际阳性案例中，检测出阳性的概率：检测出确实有病的能力
- 真阳性率=真阳/（真阳+假阴）

特异性 Specificity

- 真阴性率：实际阴性案例中，检测出阴性的几率：检测出确实没病的能力
- 真阴性率=真阴/（真阴+假阳）

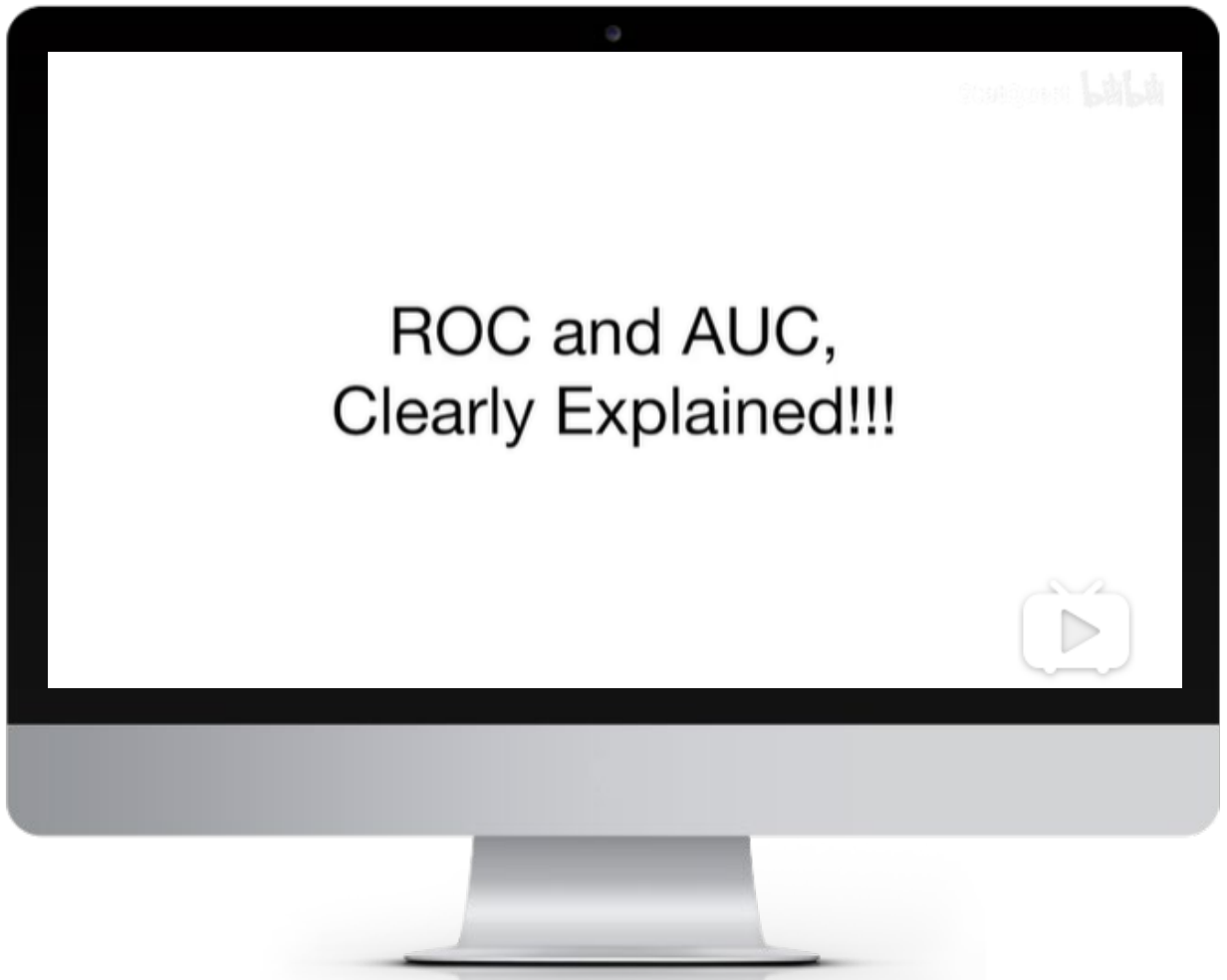
Precision

- 被正确判为阳性的案例占所有检测为阳性的比例
- 被判为阳性的案例中实际为阳性的比例
- Precision=真阳/（真阳+假阳）

Accuracy

- 正确分类的所有样本数与总样本数之比，该值越高，分类效果越好
- Accuracy= （真阳+真阴） /（真阳+假阳+真阴+假阴）

ROC & AUC



Receiver Operating Characteristic Curve (ROC)

是以假正率（FP_rate）和假负率（TP_rate）为轴的曲线

Area Under ROC Curve (AUC)

ROC曲线下与坐标轴围成的面积，显然这个面积的数值不会大于1

逻辑回归（Logistic Regression）

定义

- 是一种广义线性回归，其因变量通常是二分类的，也可以是多分类的
- 需要一条线，但不是去拟合每个数据点，而是把不同类别的样本区分开来

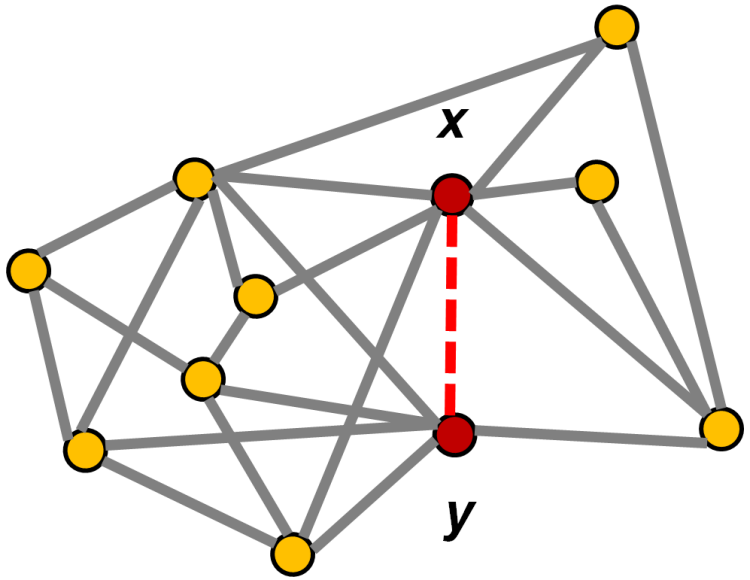
主要应用

- 点击率预估（CTR）
- 计算广告（CA）
- 推荐系统（RS）

链路预测算法 Similarity based Method

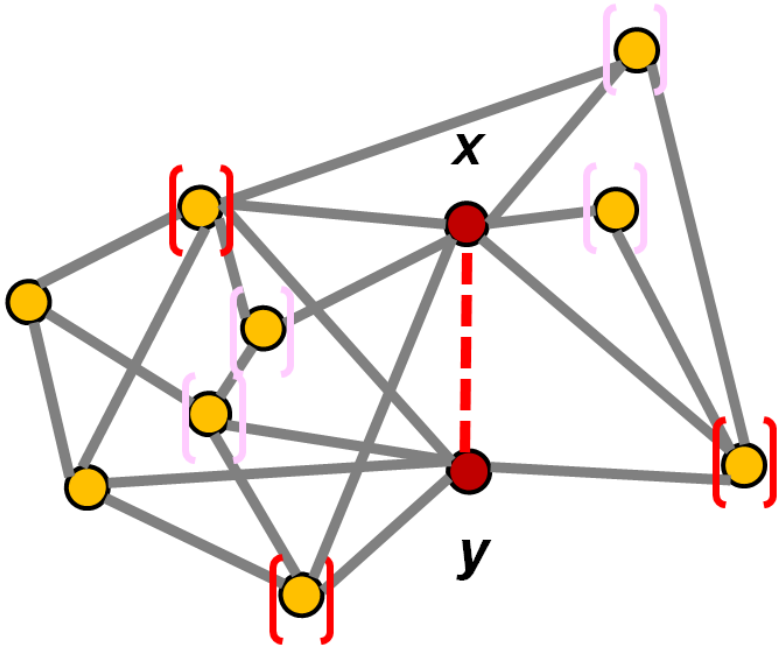
- Assign a score s_{xy} to each pair of nodes x and y
- The attributes of nodes are generally hidden
- Thus focus on *structural similarity*: two nodes are linked if they have similar network structure
- Similarity indices according to network structure
  - **Local similarity Indices**: only use local information (information around the nodes), not very accurate but fast
  - **Global similarity indices**: use global information (information among the whole network), more accurate but costly
  - **Quasi-local indices**: a tradeoff between local and global

链路预测算法分类



- Local Prediction
  - Common neighbors (CN)
  - Jaccard (JC)
  - Resource Allocation (RA)
  - Adamic-Adar (AA)
  - Preferential attachment (PA) …
- Global Prediction
  - Katz score
  - Hitting time
  - PageRank …
- Quasi-Local Prediction

Local information based link prediction

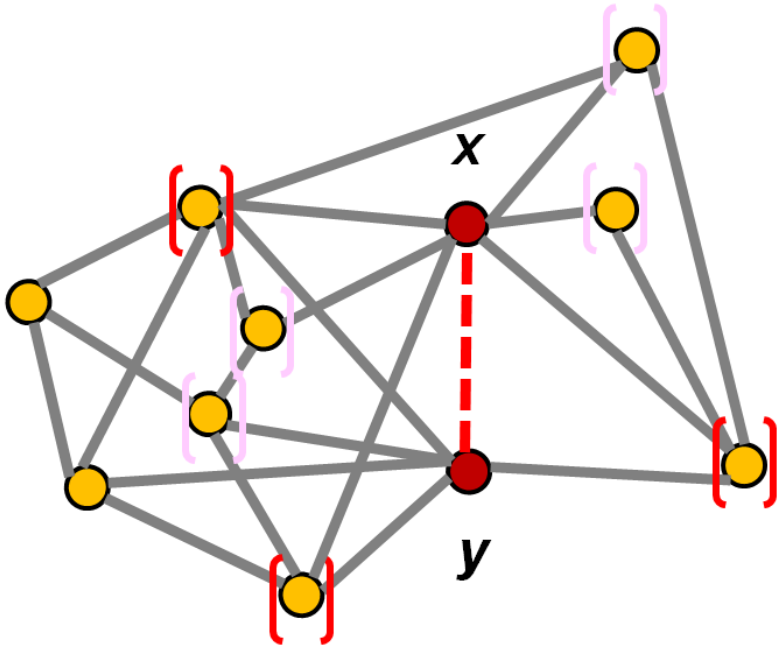


CN=?

- Link prediction algorithms
  - Common neighbors (CN)
    - Neighborhood overlap
  - Jaccard (JC)
  - Resource Allocation (RA)
  - Adamic-Adar (AA)
  - Preferential attachment (PA)

$$s_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)| \qquad \Gamma(x) \text{ set of neighbors}$$

Local information based link prediction



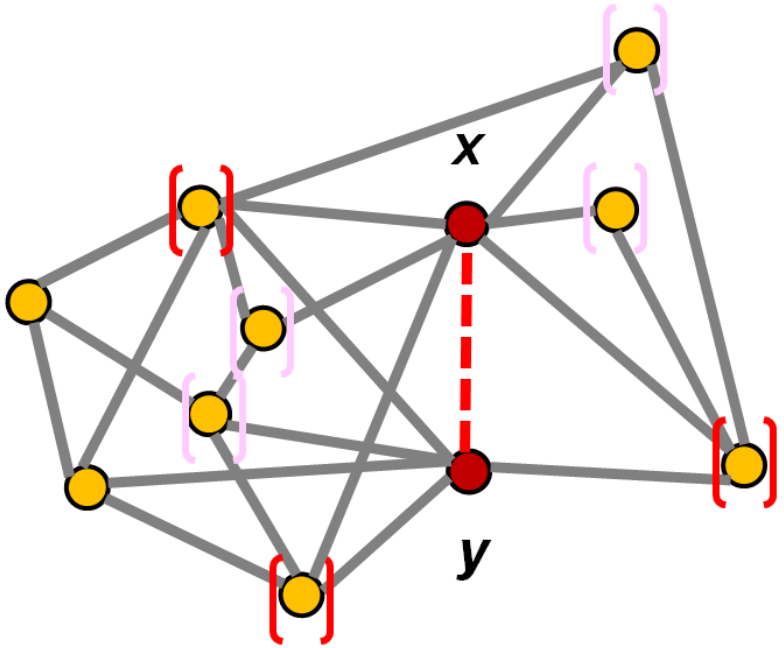
$$CN = 3$$

- Link prediction heuristics
  - Common neighbors (CN)
  - Jaccard (JC)
    - Fraction of common neighbors
  - Resource Allocation (RA)
  - Adamic-Adar (AA)
  - Preferential attachment (PA)

$$s_{xy}^{JC} = \frac{CN}{k_x + k_y - CN} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

k: degree of a node

Local information based link prediction



- Link prediction algorithms
  - Common neighbors (CN)
  - Jaccard (JC)
  - Resource Allocation (RA)
  - Adamic-Adar (AA)
  - Preferential attachment (PA)

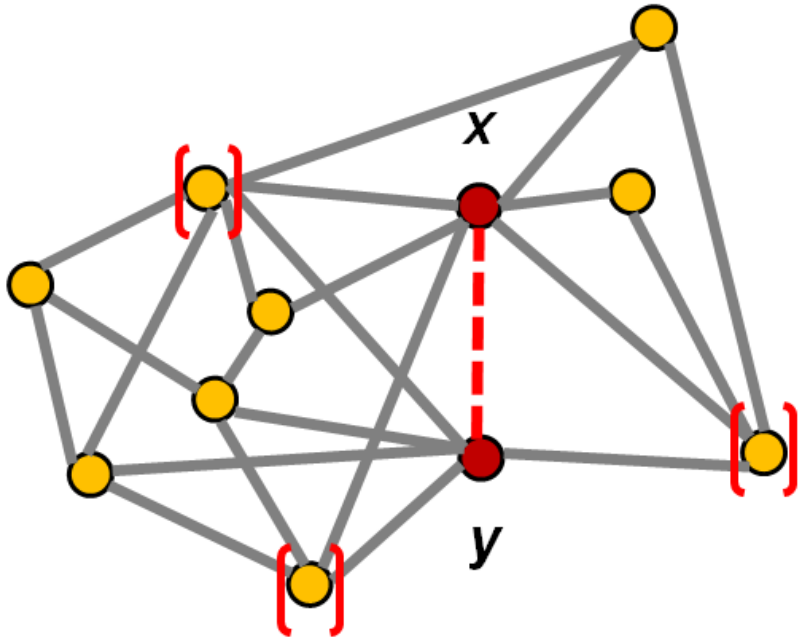
$$s_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}$$

k_z **degree of note z**

Intuition: A common neighbor would also be other nodes' neighbor and contributes for other nodes

- Similarity(*x, y*) = the amount of resource *y* received from *x* through their common neighbors
- *x* sends some resource to *y*, with their common neighbors *z* as transmitters
- Each transmitter *z* has a unit of resource and will equally distribute it to all its neighbors

Local information based link prediction

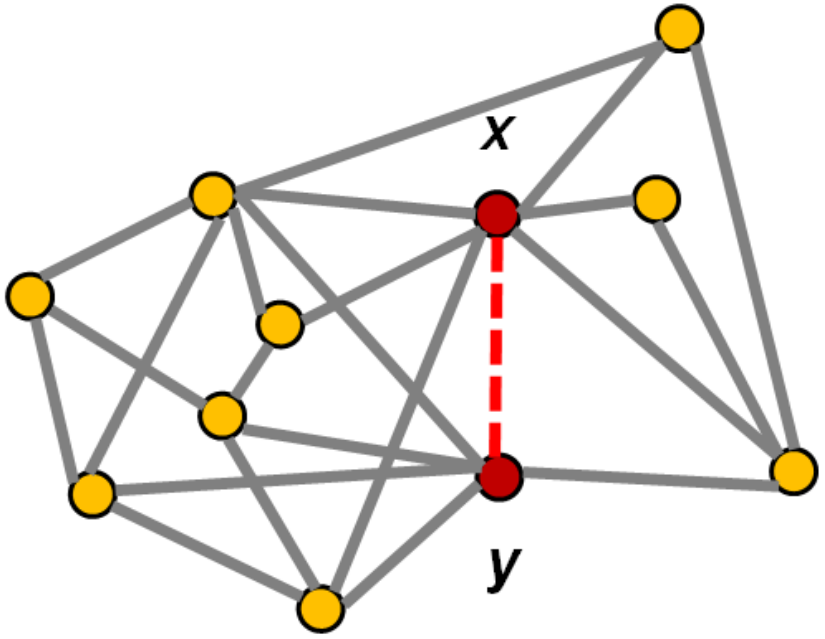


- Link prediction heuristics
  - Common neighbors (CN)
  - Jaccard (JC)
  - Resource Allocation (RA)
  - Adamic-Adar (AA)
    - Number common neighbors, with each neighbor z attenuated by log of its degree
  - Preferential attachment (PA)

$$s_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z}$$

对数变换能够很好地将**随着自变量的增加，因变量的方差也增大的模型**转化为我们熟知的模型。

Local information based link prediction



- Link prediction heuristics
  - Common neighbors (CN)
  - Jaccard (JC)
  - Resource Allocation (RA)
  - Adamic-Adar (AA)
  - Preferential attachment (PA)
    - Better connected nodes are more likely to form more links

$$s_{xy}^{PA} = k_x k_y$$

基于路径相似的指标

- **基于局部信息相似性的指标：**仅考虑全部节点二阶路径上的数目
- Pro：计算简单
- Con：预测精度低

局部、全局路径指标

- **局部路径指标 (Local path)**

$$S = A^2 + \alpha A^3$$

- $\left(A^n\right)_{xy}$ 为x和y之间长度为m的路径的数目
- 当 α 为0的时候，LP退化到Common Neighbor

基于路径相似的算法

- **局部路径指标 (Local path)**

$$S = A^2 + \alpha A^3$$

- $\left(A^n\right)_{xy}$ 为x和y之间长度为m的路径的数目。
- 当 α 为0的时候，LP退化到Common Neighbor

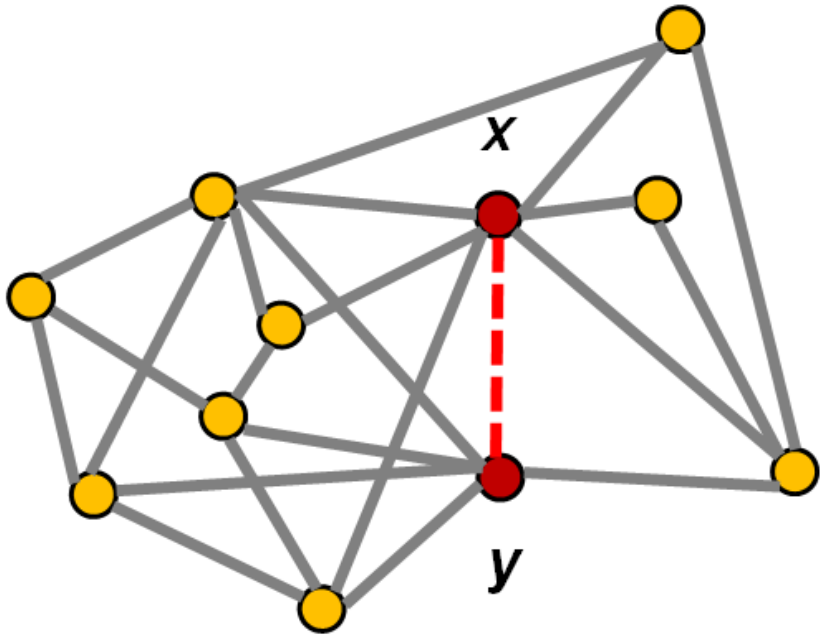
- **全局Katz指标**:

$$S = \beta A + \beta^2 A^2 + \beta^3 A^3 \cdots = \left(I - \beta A\right)^{-1} - I$$

- 各种长度路径的数目带权和。其中短的路径更重要?
- 为了保证以上数列的收敛性，需要使得 $\beta\lambda < 1$，其中 λ 是矩阵的最大特征值

Neumann Series $(\mathrm{Id} - T)^{-1} = \sum_{k=0}^{\infty} T^k$ https://en.wikipedia.org/wiki/Neumann_series

Global information based link prediction



- Link prediction heuristics
  - Katz score
    - Measures number of paths between two nodes, attenuated by their length
  - PageRank

Pros: more accurate than local indices

Cons: 1) time-consuming; 2) global topological information may not be available

- Quasi-Local Prediction
  - Local Path

PageRank

网页排序

- 竞价排名-课程第四部分
- **非竞价排名**

PageRank的核心思想

- **数量假设**：在Web图模型中，如果一个页面节点接收到的其他网页指向的入链数量越多，那么这个页面越重要。
- **质量假设**：指向网页的入链质量不同，质量高的页面会通过链接向其他页面传递更多的权重。所以越是质量高的页面指向网页，则这个页面越重要。

PageRank

- PageRank

外部网站j的重要程度

本网站的重要程度

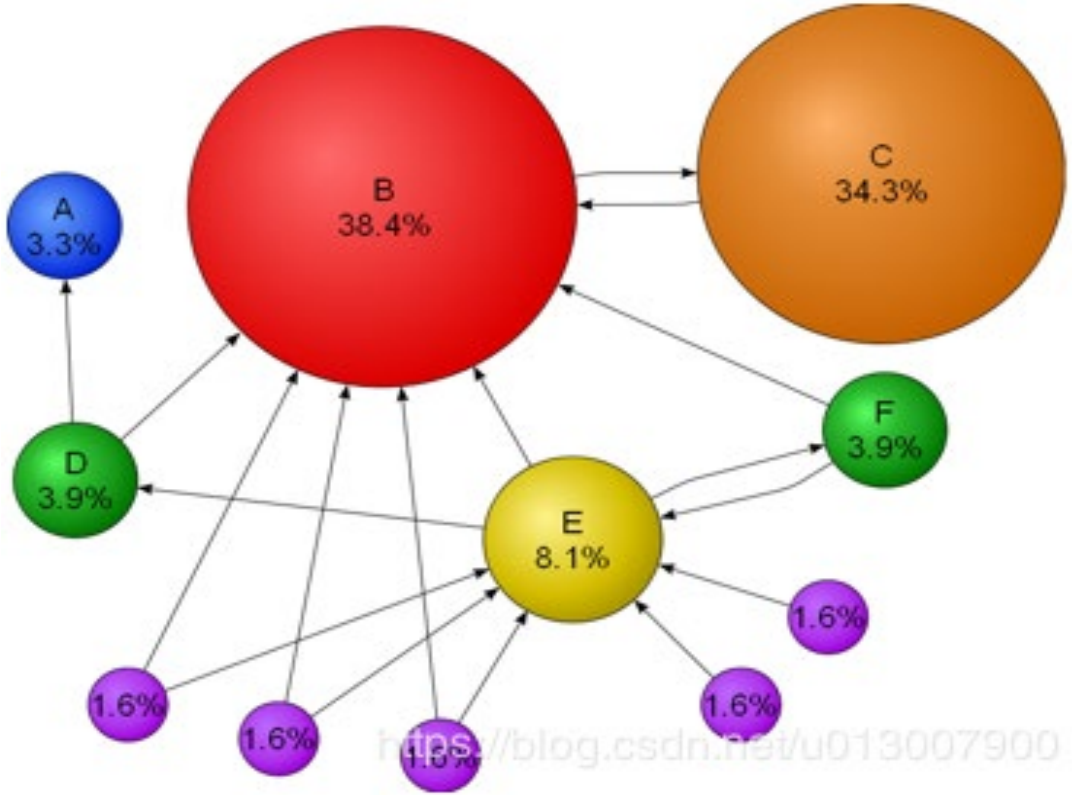
$$PR(i) = \sum_{j \in B_i} \frac{PR(j)}{O_j},$$

外部网站j的出度

指向本网站的外部网站集合

Oj – Number of outgoing links from page j 出度

Bi – *Set of web pages pointing to web page* i



在此定义下可能会出现哪些问题?

PageRank

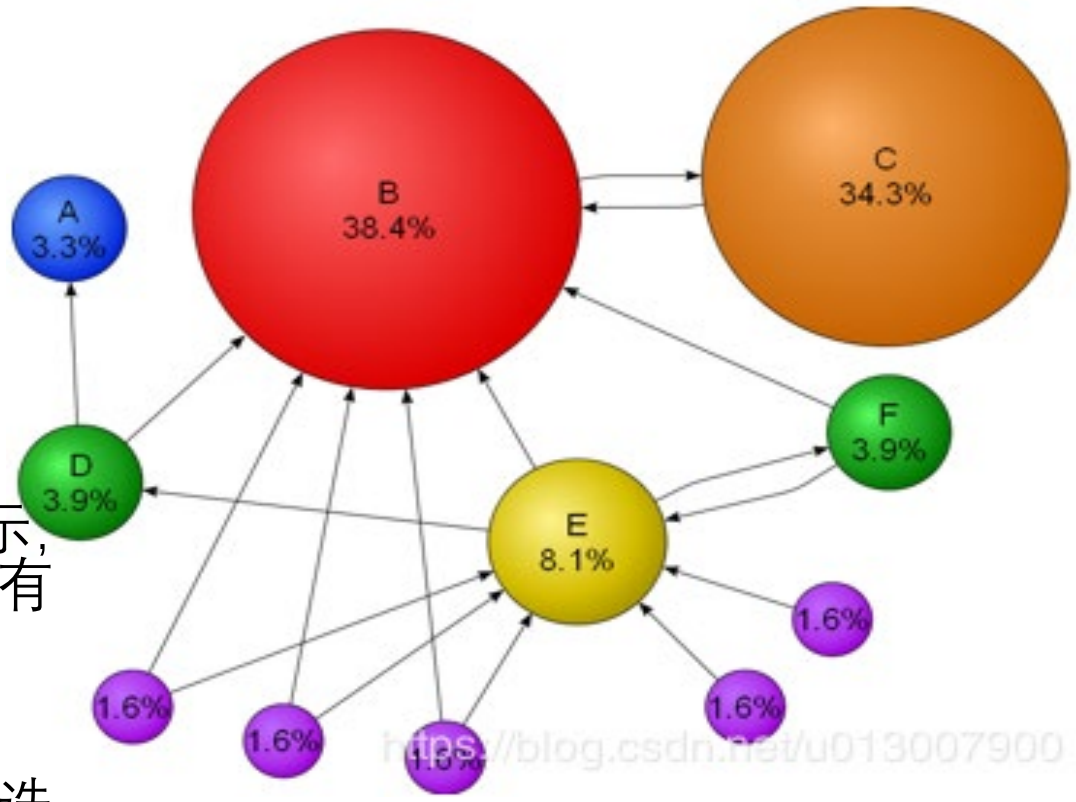
- PageRank

$$PR(i) = \sum_{j \in B_i} \frac{PR(j)}{O_j},$$

Oj – Number of outgoing links from page j （出度）

Bi – *Set of web pages pointing to web page* i

- **排名泄露**
  - 如果存在网页**没有出度**链接，如蓝色A节点所示，则会产生排名泄露问题，经过多次迭代后，所有网页的PR只都趋向于0
- **排名下沉**
  - 若网页**没有入度**链接，如紫色节点，经过多次迭代后，紫色节点的PR值会趋向于0
- **排名上升（站内导航链接）**
  - 互联网中一个网页只有对自己的出链，或者几个网页的出链形成一个循环圈。那么在不断地迭代过程中，这一个或几个网页的PR值将只增不减



- **对新网页不友好**
  - 一个新网页的入链相对较少，即使它的内容的质量很高，要成为一个高PR值的页面仍需要很长时间的推广

PageRank

- PageRank used by Google

$$PR(i) = \frac{(1-d)}{N} + d \cdot \sum_{j \in B_i} \frac{PR(j)}{O_j},$$

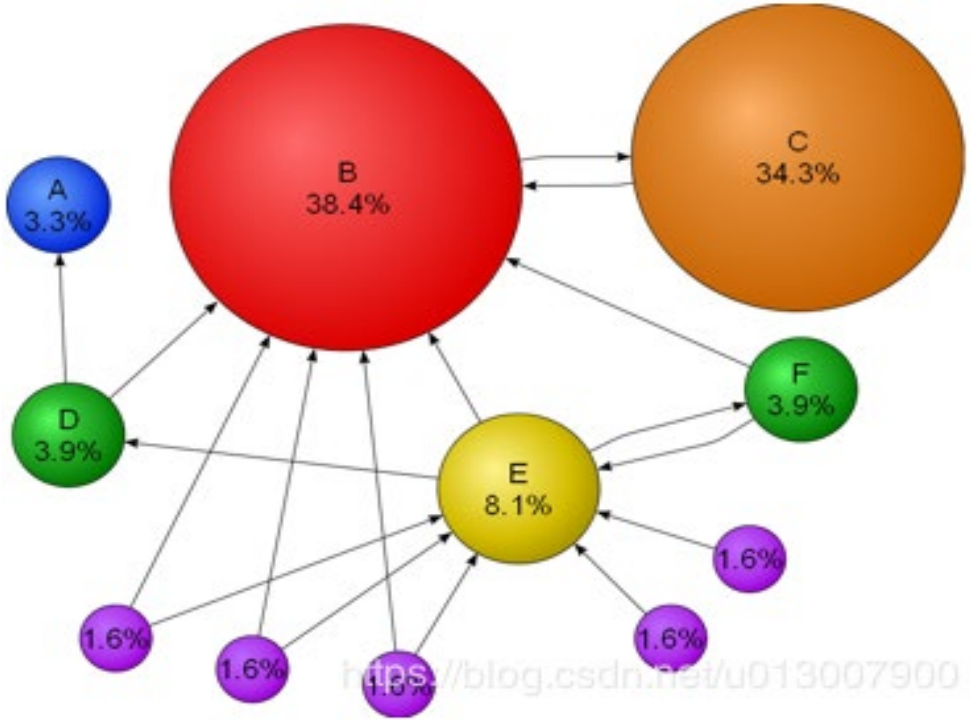
靠自己挣来的　　自己熟悉的圈子介绍的

N – Total number of webpages

Oj – Number of outgoing links from page j （出度）

Bi – *Set of web pages pointing to web page* i

d – *dampening factor* (usually set to 0.85)



Experiment: Prediction in social networks

- Collaboration networks of physicists
  - Core nodes: authors who published at least 3 papers during the training period and at least 3 papers during test period
- Training data: graph G(t_0, t_0') of collaborations during time period [t_0, t_0'] with V core nodes and E_{old} edges
- Test data: graph G(t_1, t_1') of collaborations during a later time period [t_1, t_1'] with V core nodes and E_{new} edges

	Training Period			Core		
	Authors	Articles	Collaborations[a]	Authors	E_{old}	E_{new}
astro-ph	5,343	5,816	41,852	1,561	6,178	5,751
cond-mat	5,469	6,700	19,881	1,253	1,899	1,150
gr-qc	2,122	3,287	5,724	486	519	400
hep-ph	5,414	10,254	47,806	1,790	6,654	3,294
hep-th	5,241	9,498	15,842	1,438	2,311	1,576

Experiment in social networks (cont'd)

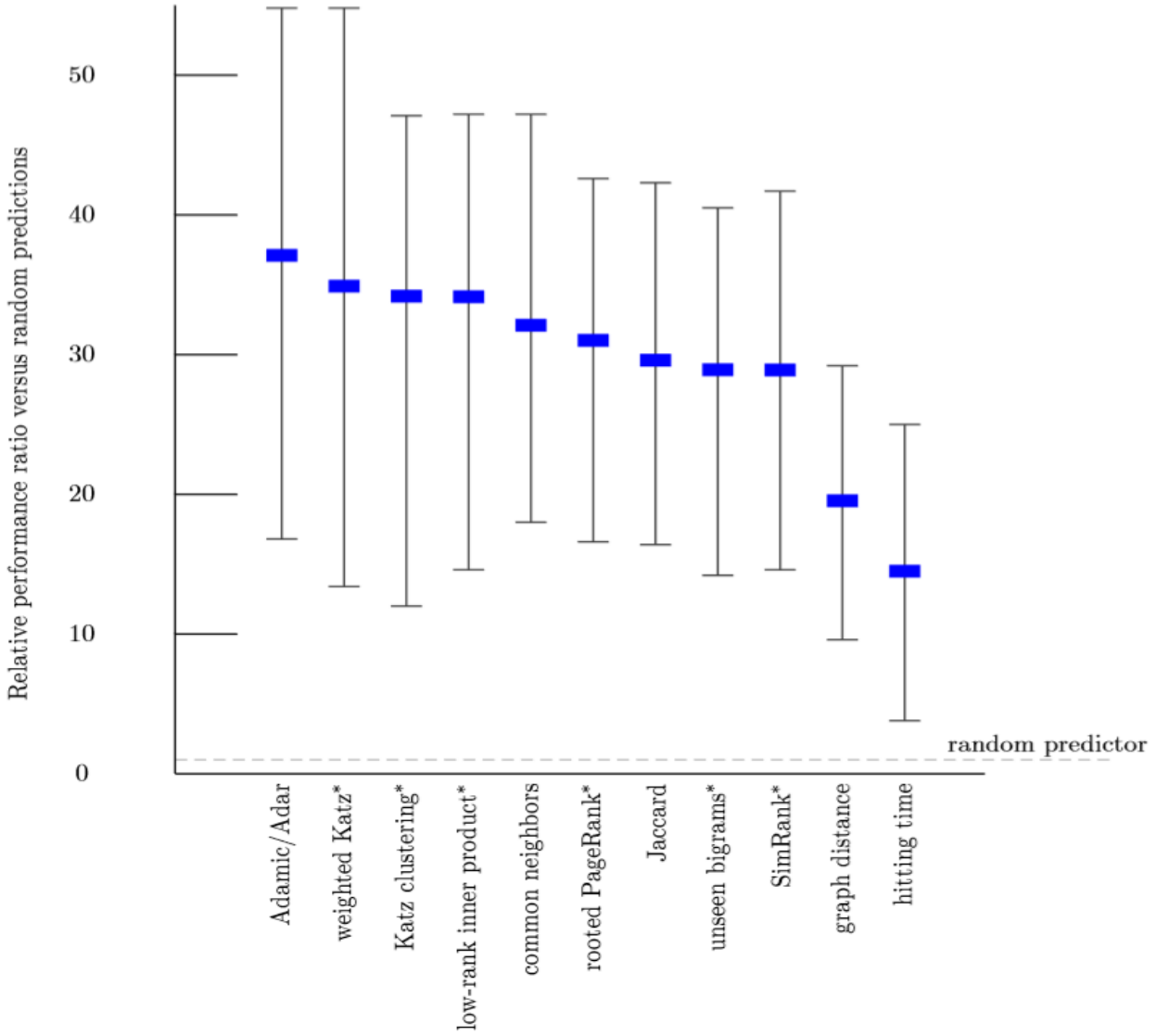
- Link prediction algorithm
  - Score all the node pairs using an algorithm ρ
  - New links more likely among high scoring pairs
- Each link prediction heuristic ρ outputs a ranked list L of new collaborations E_{new}^{*} : pairs in $V \times V - E_{old}$.
- Focus evaluation on new links E_{new}^{*} between core nodes
- Performance metric: How many of the top n pairs in ranked list L are the actual new nodes in E_{new}^{*} ?

Data Sets for Link Prediction Research

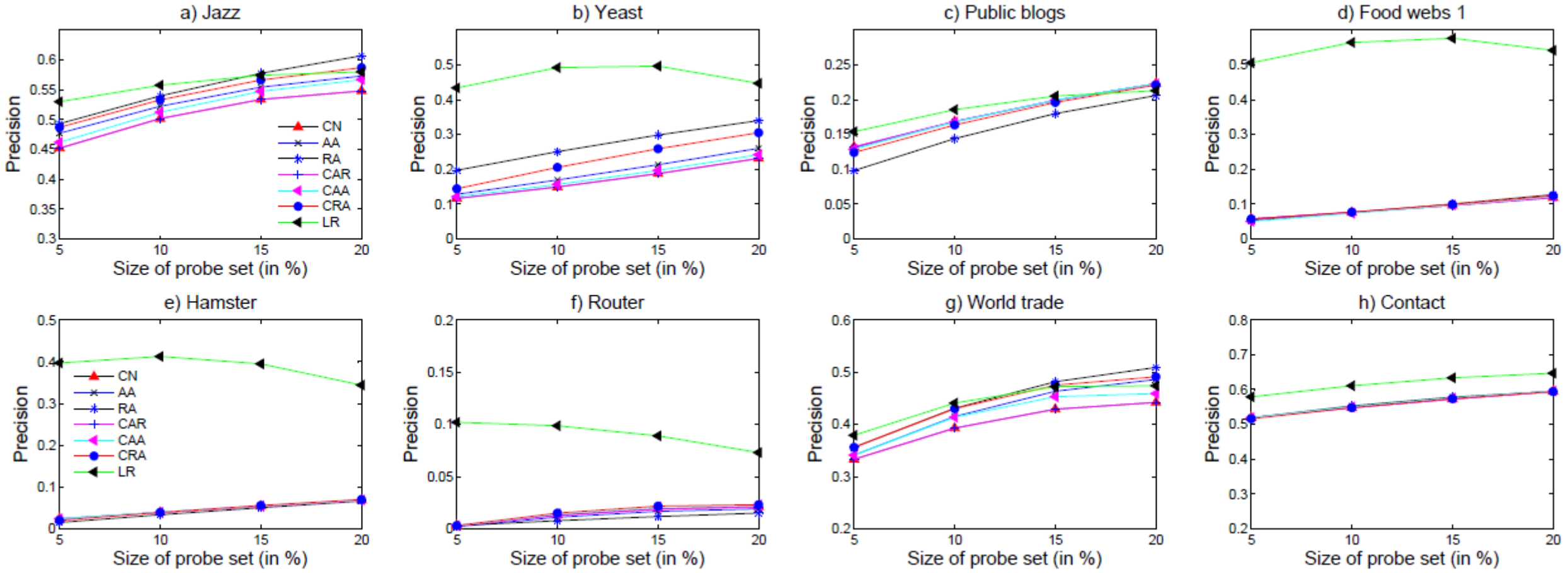
- [1] **爵士乐手关系网络 JAZZ**: Pablo M Gleiser and Leon Danon. Community structure in jazz. Advances in complex systems, 6(04):565{573, 2003.
- [2] **蛋白质关系网络 Yeast**: Christian Von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G Oliver, Stanley Fields, and Peer Bork. Comparative assessment of largescale data sets of protein{protein interactions. Nature, 417(6887):399{403, 2002.
- [3] **美国的政治门户网站之间的链接网络 Political Blogs**: Robert Ackland et al. Mapping the us political blogosphere: Are conservative bloggers more prominent? Mapping the US political blogosphere: Are conservative bloggers more prominent, in: Presentation to BlogTalk Downunder, Sydney, Available at:http://inscub.org/blogtalk/images/robertackland.pdj.2005.
- [4] **一个在线社交网络 Hamster** : Hamster friendships network dataset - konect, may. Hamsterster friendships network dataset - KONECT, May, 2015.
- [5] **因特网的路由器节点网络 Router**: Neil Spring, Ratul Mahajan, David Wetherall, and Thomas Anderson. Measuring isp topologies with rocketfuel. Networking, IEEE/ACM Transactions on, 12(1):2 16, 2004.
- [6] **佛罗里达的一个捕食者网络 FOOD Web**: Robert E Ulanowicz and Donald L DeAngelis. Network analysis of trophic dynamics in south orida ecosystems. FY97: The Florida Bay Ecosystem, pages 20688{20038, 1998.
- [7**] 94年全世界范围80个国家的金属交易网络 World Trade**: Wouter De Nooy, Andrej Mrvar, and Vladimir Batagelj. Exploratory social network analysis with Pajek, volume 27. Cambridge University Press, 2011.
- [8] **一个电话接触网络 Contact**: Jerome Kunegis. Konect: the koblenz network collection. In Proceedings of the 22nd international conference on World Wide Web companion, pages 1343{1350. International World Wide Web Conferences Steering Committee, 2013.
- [9] **美国的空中运输网络 USAir**: A. Batageli, V. Mrvar. Available at http://vlado.fmf.unilj.si/pub/networks/data/mix/USAir97.net.
- [10] **线虫神经网络 C.Elegant**: Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. nature, 393(6684):440 442, 1998.
- [11]**佛罗里达的另一个捕食者网络** FOOD Web 2 RE Ulanowicz, C Bondavalli, and MS Egnotovich. Network analysis of trophic dynamics in south orida ecosystem, fy 97: The orida bay ecosystem. Annual Report to the United States Geological Service Biological Resources Division Ref. No.[UMCES] CBL, pages 98{123, 1998.
-

Some Basic Simulation Result

Networks	CN	AA	RA
Jazz	0.502	0.521	0.533
Yeast	0.139	0.159	0.256
Political blogs	0.178	0.175	0.155
Hamster	0.037	0.038	0.033
Router	0.018	0.016	0.008
Food web 1	0.070	0.072	0.068
World trade	0.402	0.420	0.430
Contact	0.556	0.559	0.558



Simulation under Different Prob Set Size



Experiment in scientific/social networks

Metric: AUC (area under the receiver operating characteristic curve)

Each number averaged by 10 implementations (cross validation)

Real-world networks
PPI: protein-protein interaction
NS: co-authorship
Grid: electrical power-grid
PB: US political blogs
INT: router-level Internet
USAir: US air transportation

CN and AA have second best performance

RA performs the best

Indices	PPI	NS	Grid	PB	INT	USAir
CN	0.889	**0.933**	**0.590**	0.925	**0.559**	0.937
Salton	0.869	0.911	0.585	0.874	0.552	0.898
Jaccard	0.888	**0.933**	**0.590**	0.882	**0.559**	0.901
Sørensen	0.888	**0.933**	**0.590**	0.881	**0.559**	0.902
HPI	0.868	0.911	0.585	0.852	0.552	0.857
HDI	0.888	**0.933**	**0.590**	0.877	**0.559**	0.895
LHN1	0.866	0.911	0.585	0.772	0.552	0.758
PA	0.828	0.623	0.446	0.907	0.464	0.886
AA	0.888	0.932	**0.590**	0.922	**0.559**	0.925
RA	**0.890**	**0.933**	**0.590**	**0.931**	**0.559**	**0.955**