链接分析点

击编辑主字幕样式

大数据分析|何铁科
http: // hetieke.cn

南京大学
南京大学

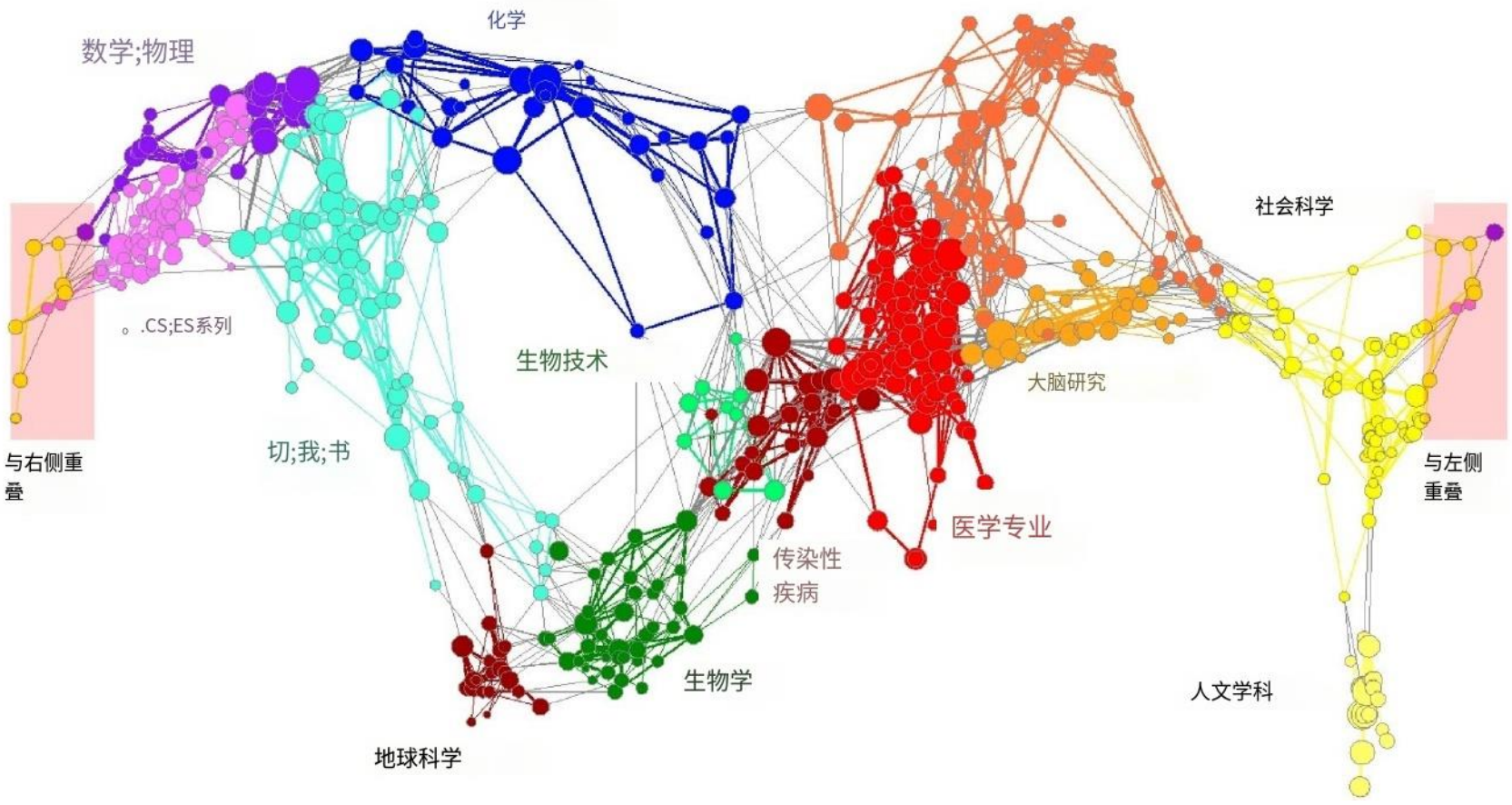
新型数据:图数据

高维数据	图数据	流数据	机器学习	应用
局部敏感哈希算法	**PageRank,** SimRank	数据流过滤	支持向量机	推荐系统
聚类	社区检测	网络广告	决策树	关联规则
降维	垃圾邮件检测	流查询	感知机 kNN	重复文档监测

图数据:社交网络

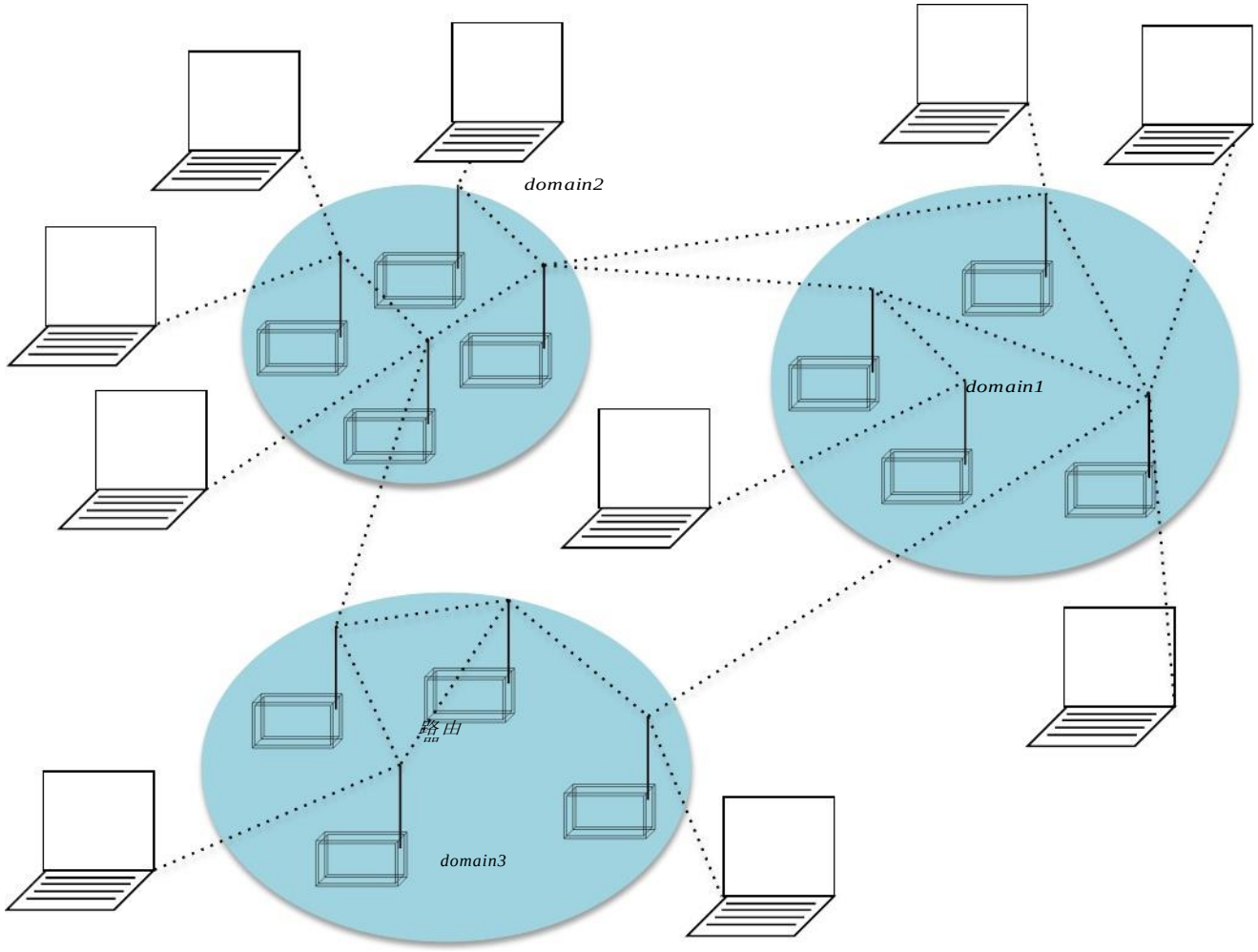
Facebook 社交图谱 4 度分离 **[Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]**

图数据:信息网络

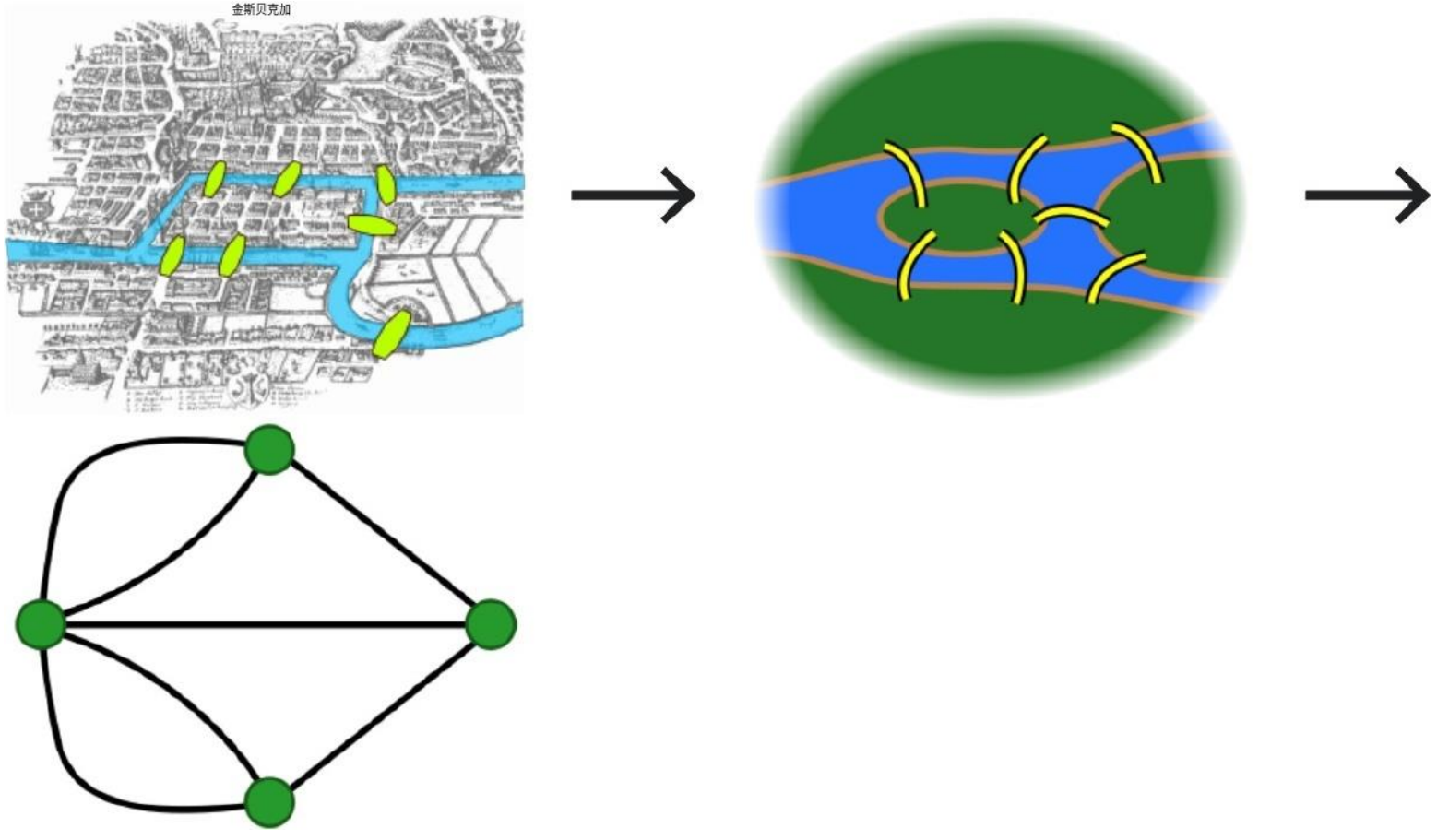


科学引文网络与地图[Börner 等人， 2012]

图数据:信息网络



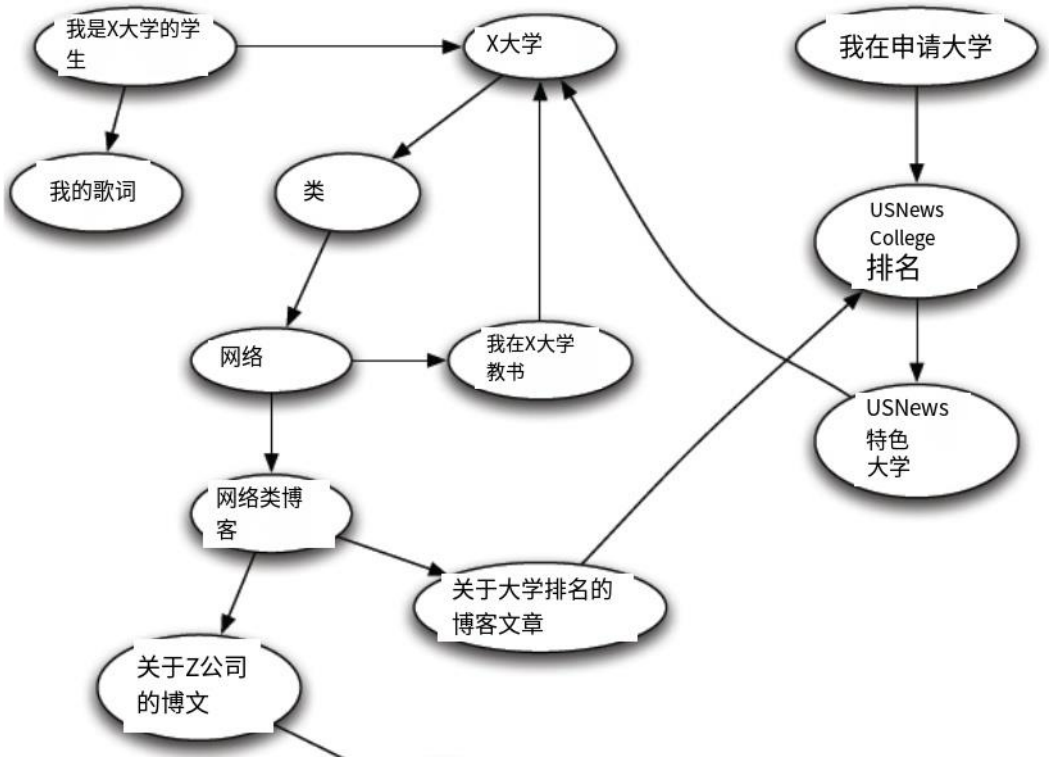
图数据:技术抽象



柯尼斯堡七桥问题

将Web表示为图

- 网站表示为有向图
  - 节点:网页
  - 边:超链接


我是X大学的学生
X大学
我在申请大学
我的歌词
类
USNews College 排名
网络
我在X大学教书
USNews 特色大学
网络类博客
关于大学排名的博客文章
关于Z公司的博文

如何组织网页？

- **方式一:网页索引(人工编辑)** § 雅虎、DMOZ、LookSmart

- **方式二:Web 搜索**
  - § **信息检索调查:在一个小而可信的集合中找到相关的文档。** § 报纸文章、专利等。
  - § **缺陷:** 网络是巨大的，充满了不可信，过时和随机的东西。

网页搜索中的挑战

网页搜索中的两项挑战:’(1)网络中存在多个来源的数据该“信任”谁?

- 诀窍:可信的页面彼此相互引用和链接

¡ (2) 查询“数据”的最佳回答是什么?

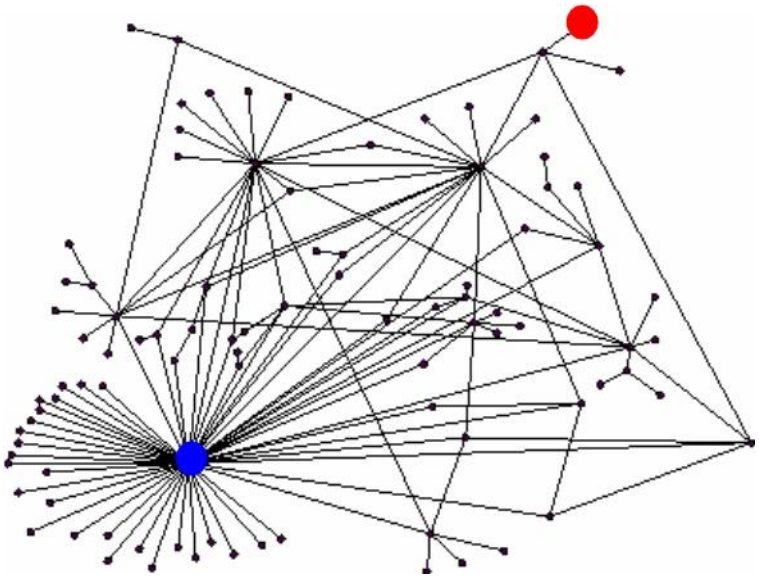
- 没有单个的最佳答案
- 诀窍:实际关于“数据”的页面往往指向许多“数据”

在图中作节点排序

- **所有网页的重要性都不是“平等”的**

在网络图节点的连接中有极高的多变性。

我们通过链接结构来对页面进行排序



链接分析算法

’我们将介绍以下**链接分析方法计算图中节点的重要性**：§页面排名

§特定主题(个性化)页面排名 §网页垃圾邮件检测算法

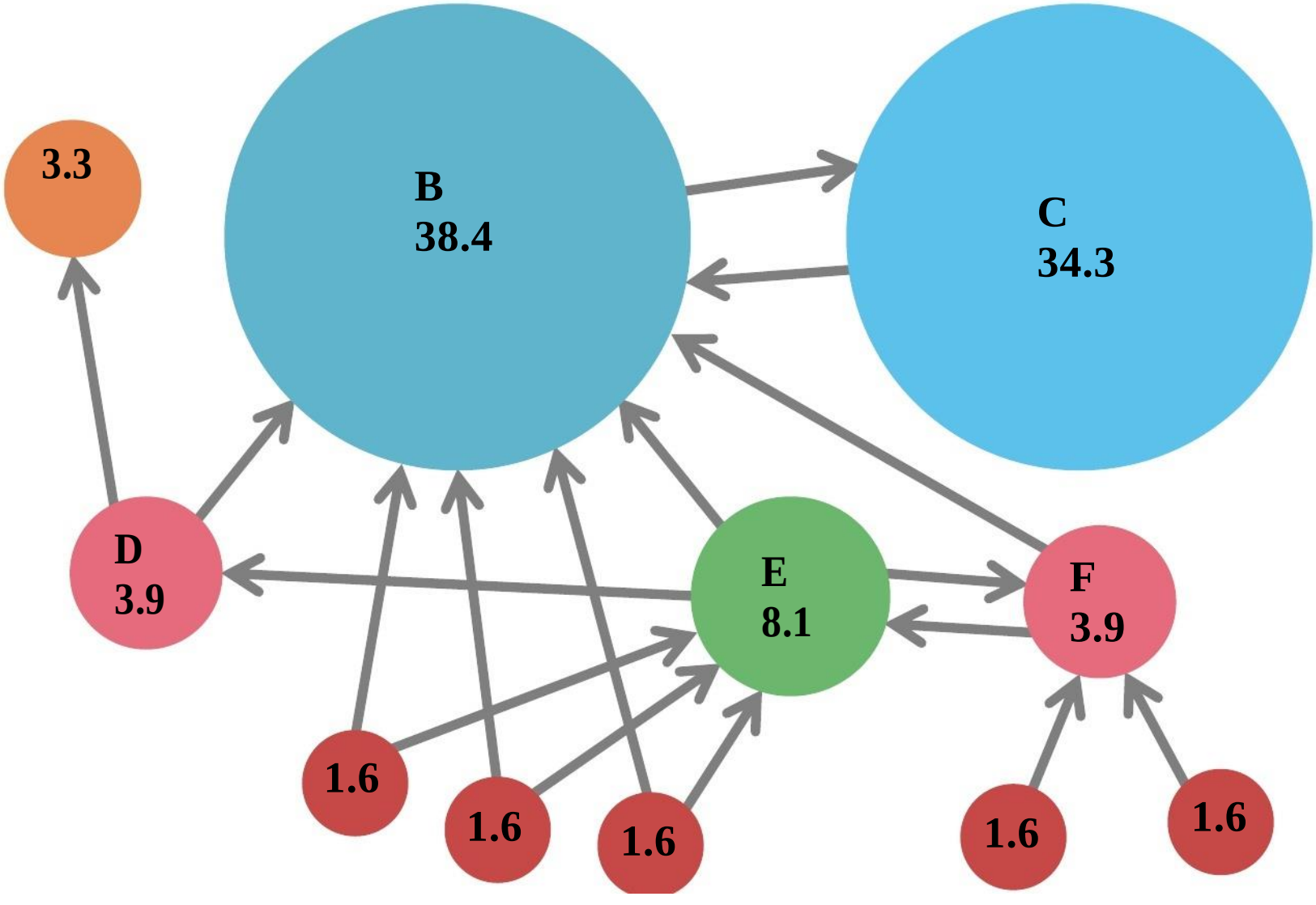
链接投票

- **Idea: 链接投票**
  - 页面拥有的链接越多越重要
  - 入链？ 还是出链？

- **考虑来自外部网站的链接**: § www.stanford.edu 有 23400 个 in-link § www.joe-schmoe.com 有 1 个 in-link

- **所有的入链都同等权重吗？** §
  来自重要的链接占更大权重 § 递推问题

PageRank网页排名



搜索引擎的难题

对搜索结果的重要性排序？

搜索引擎的核心框架


张洋 博客
找到约 2,630,000条结果（用时0.13秒
爱雨的蓝色心情 新浪博客
blog.sina.com.cn/mimacleyang
2011年12月16日-爱雨的蓝色心情_新浪博客，爱雨的蓝色心情，2011年12月16日，终于开了围脖….……换个地方听我说.迎接你的到来---我的奥边情人….张浑的BLOG..
张洋视觉 新浪博客
blog.sina.com.crvzhangyangmv
2011年11月26日-张洋视堂_新浪博客，张洋视觉，许飞我们终究会牵手旅行mv正式版，吹子《失态排行榜》mv花絮，欢子<我们回不去了》mv花絮，欢子Mv <可是你是他的…
张洋 新浪博客
blog.sins.com.cn/haonanerzhangyang-网页快照
2010年1月15日-张洋_新浪博客，张洋、红色细篮开播了，2009年08月23日，2009年08月04日.生活在继续.，上海我来了，在火焰山吃火锅，在北极吃冰棍，梦到XXX，共和国…
12啦菌体-博客园
www.cnblogs.com/leoo2sk/·两页快照
2012年6月16日一个人博客已迁移至codinglabs.org，博客园不再更新 个人博客已迁移至codinglabs.org，欢迎访问，发布一个查看PHP opcode的扩展模块及Web服务…
CodingLabs
www.codinglabs.org/-胃页快照
另外我也不想在虚拟机中写博客，于是一直在寻找Live Writer的替代品。……属觉、表演、放映、广播或通过信息网络传播本博客的文章，但期间必须保留作者姓名张洋及adinglabs.org02012

解决步骤

1.建立资料库——爬虫
2.建立一种数据结构——倒排索引

”张洋“：{1, 3, 6, 8 ,11, 15}”博客“：{1, 6, 11, 12, 17, 20, 22}

结果:{1,6,11}

核心难题

对查询结果排序

高质量的页面

早期的一些做法

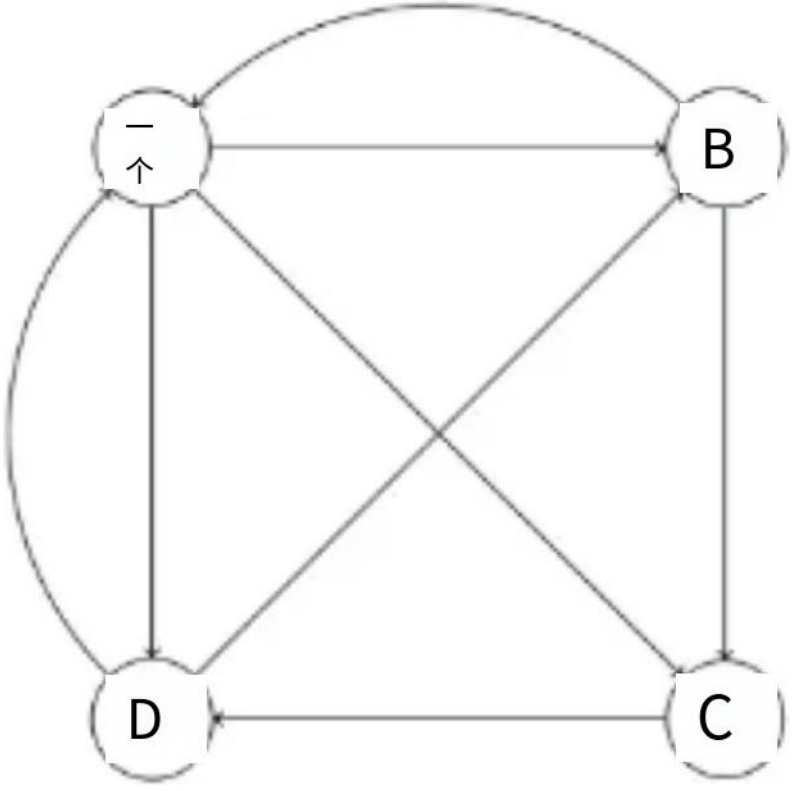
1.不评价（自然顺序）

2. 基于检索词的评价(例如:TF-IDF)

垃圾邮件

1.目标页面排名靠前
2.干扰其他关键词（“亚运会”）

Pagerank示例



20.

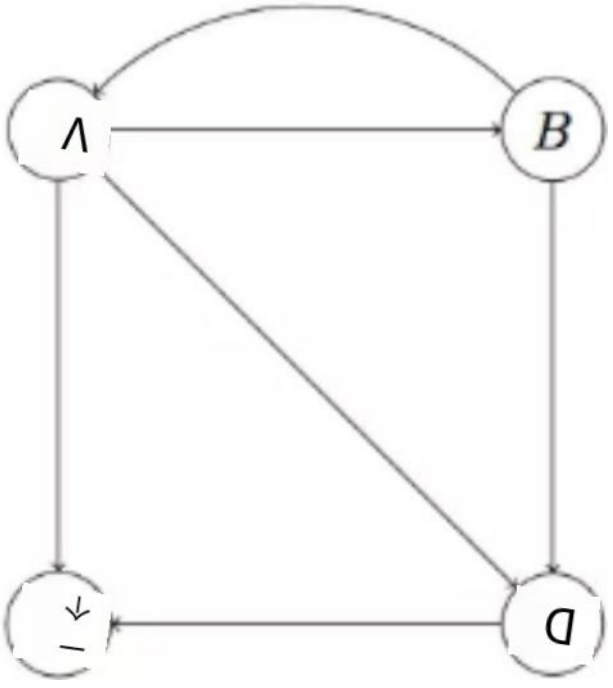
Pagerank示例

$$M = \begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \\ 1/3 & 0 & 1 & 0 \end{bmatrix} \qquad v = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

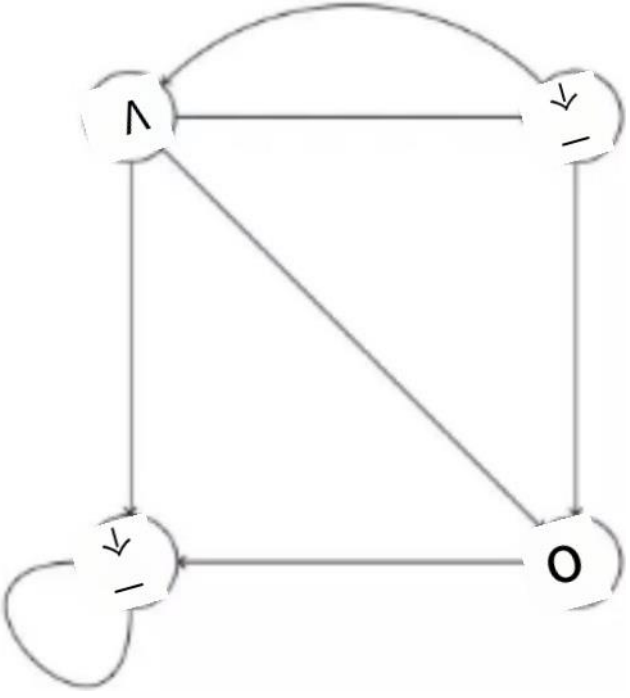
$$Mv = \begin{bmatrix} 1/4 \\ 5/24 \\ 5/24 \\ 1/3 \end{bmatrix}$$

收敛结果：(1/4, 1/4, 1/5, 1/4)

死胡同



蜘蛛陷阱

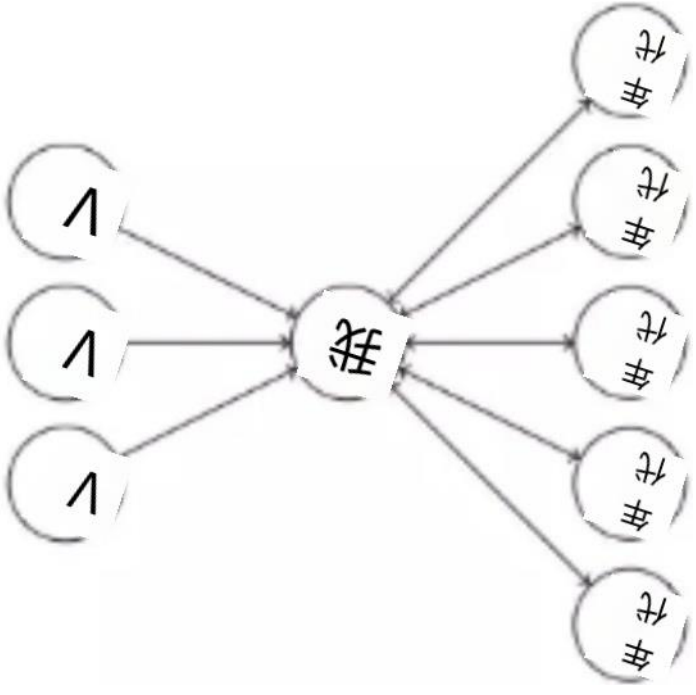


主题敏感网页排名

1.话题分类
2. 网页话题归属
3.分向量计算
4. 用户话题倾向

针对PageRank的攻击

1.目标页
2.支持页
3.可达页
4.不可达页



链接垃圾邮件

1.网络拓扑分析

2.TrustRank