# 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

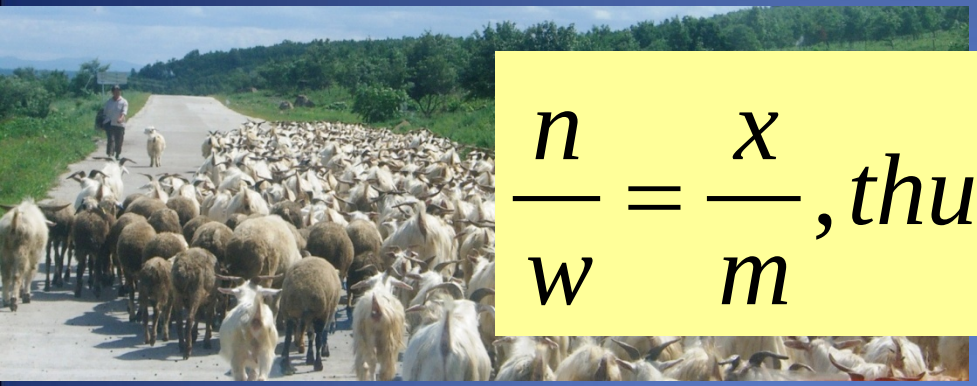
# “size of the web”→google

# 人们关心Web规模问题

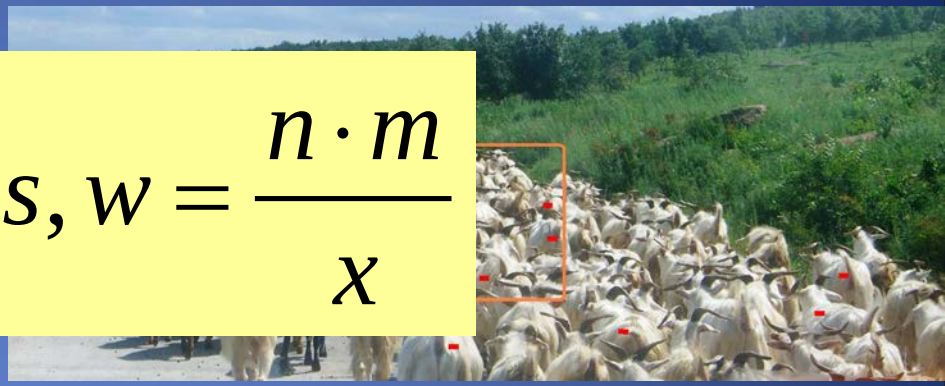
- People seriously report it from time to time
  - 1997, ~200 million, K. Bharat and A. Broder
  - 1998, ~800 million, S. Lawrence and C. Giles
  - 2000, ~2.1 billion, Shayna Keces
  - 2005, ~11.7 billion, A. Gulli and A. Signorini
- People maintain websites to talk about it
  - http://www.worldwidewebsize.com
  - 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

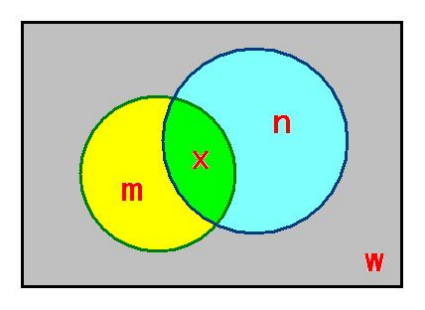


$$\frac{n}{w} = \frac{x}{m}, thus, w = \frac{n \cdot m}{x}$$



- 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




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

- 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.

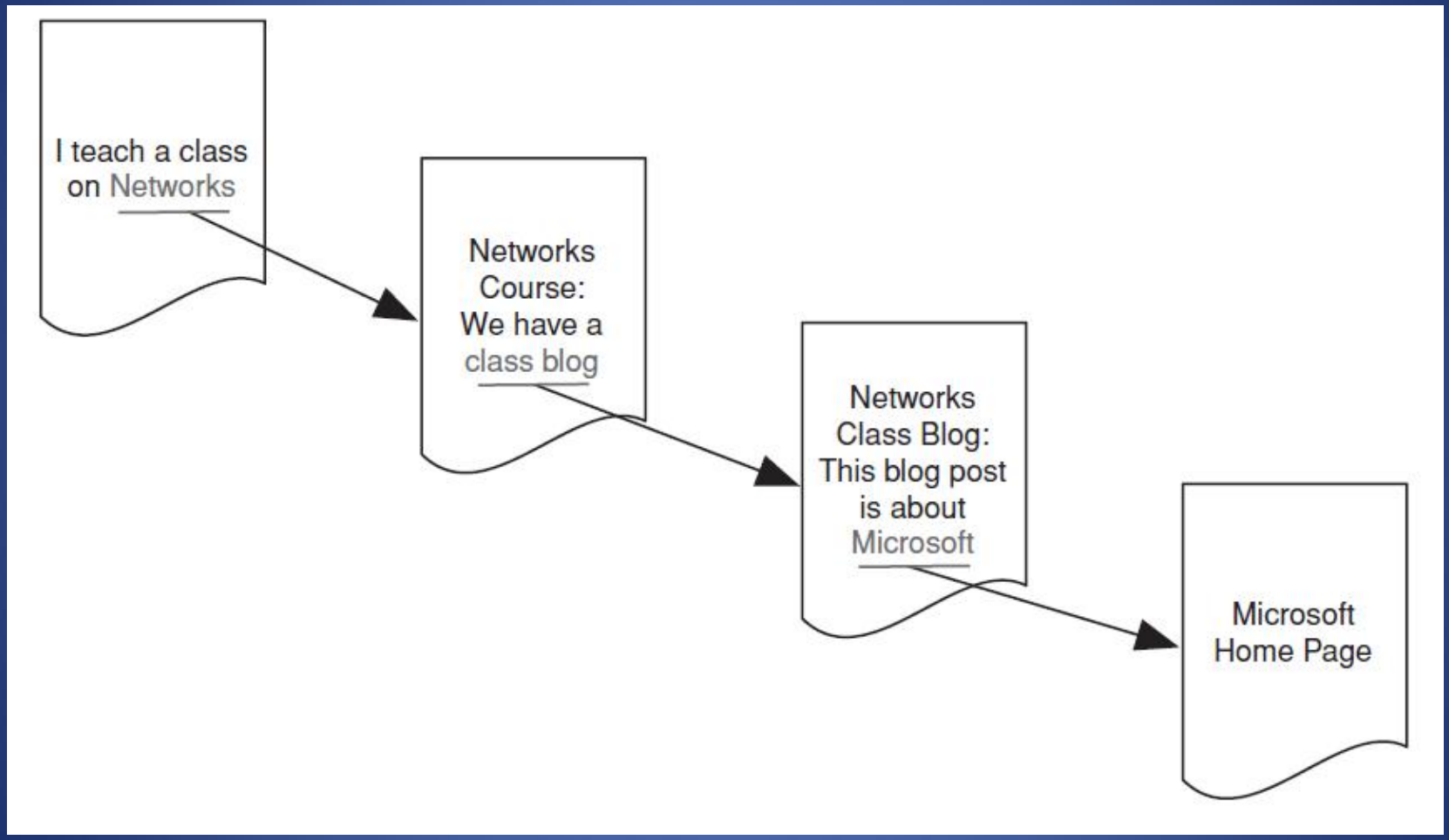
# 信息的组织（一般意义的）

- 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

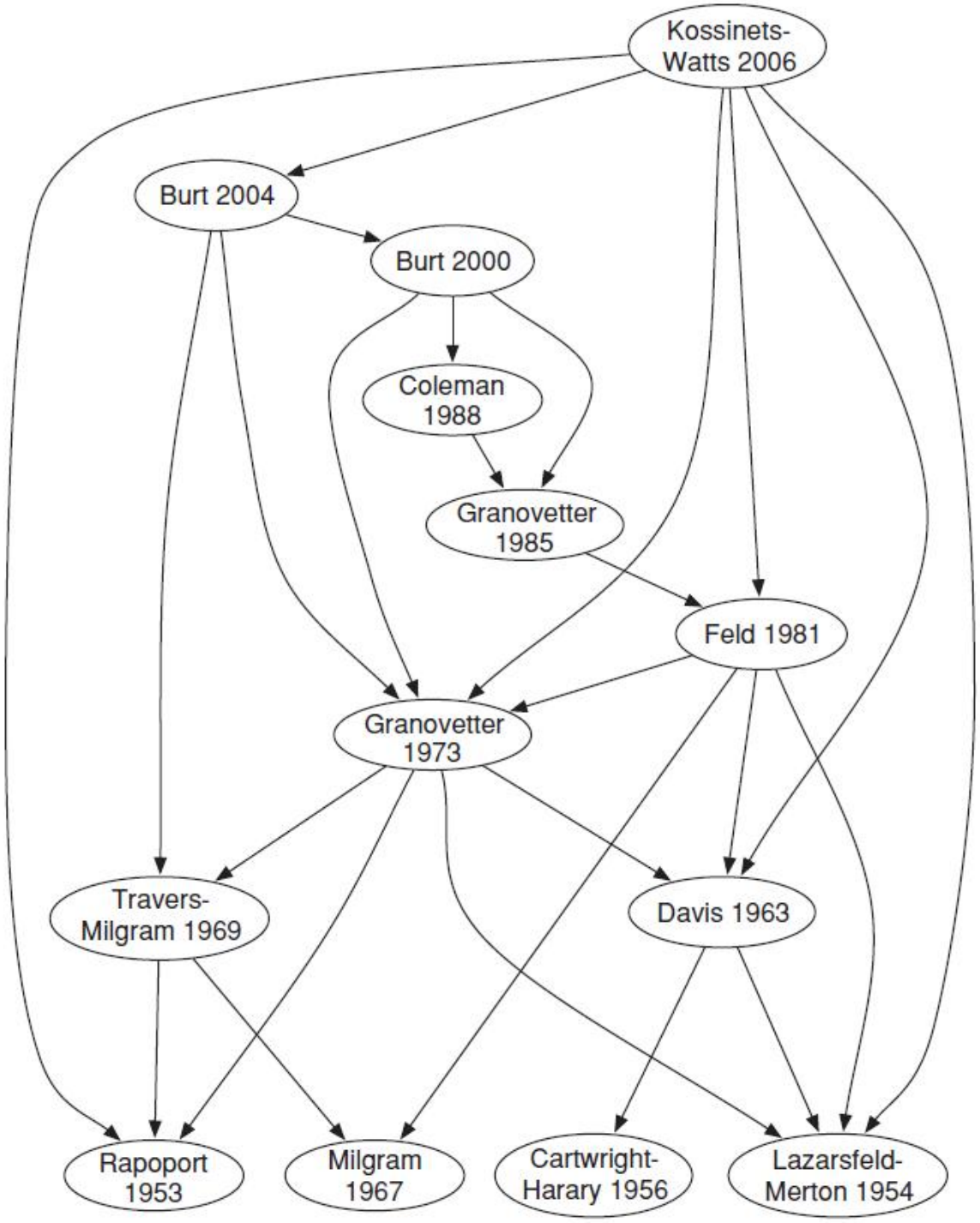
# Web信息的组织方式

- 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。

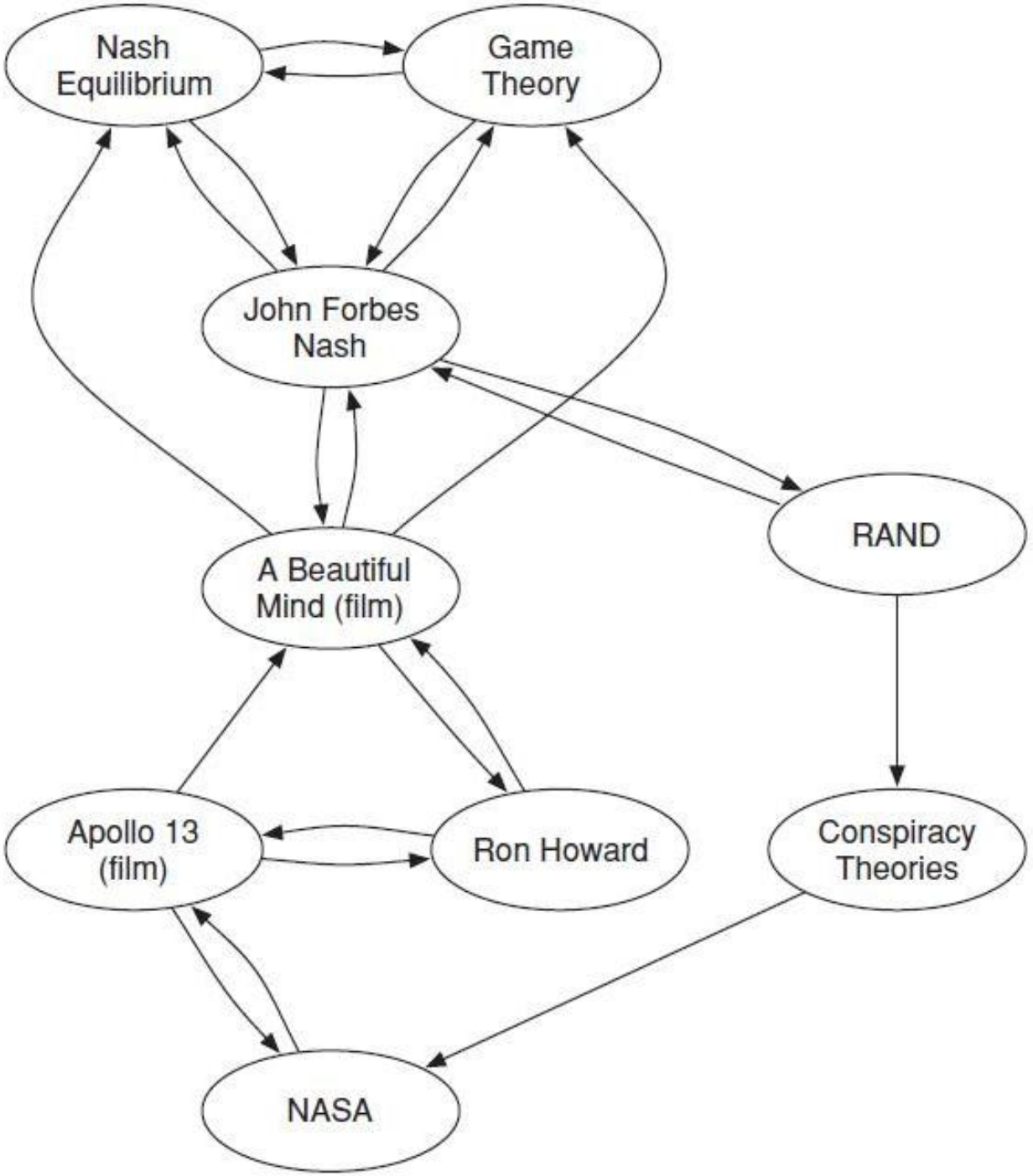


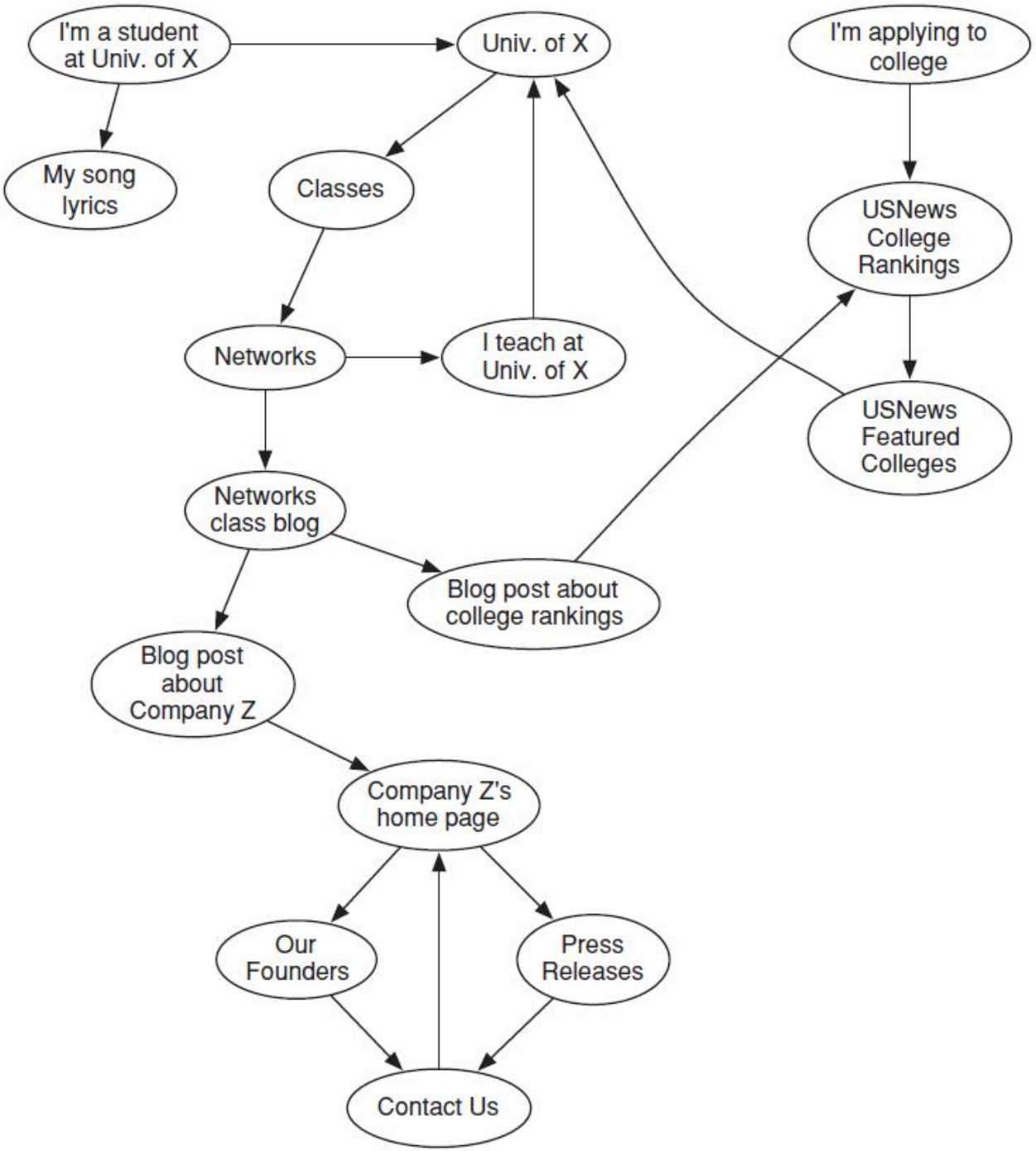

“链接”不仅用于表达网页之间的关系

- 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

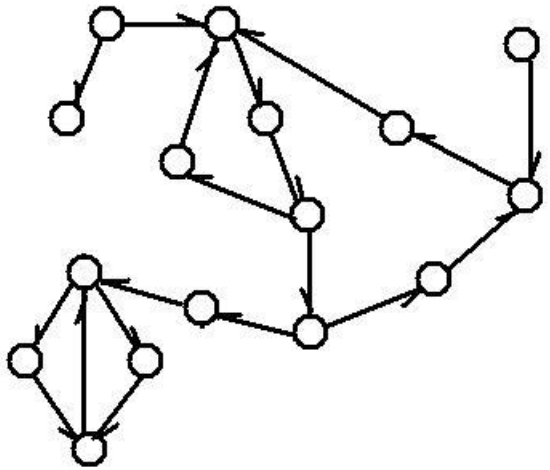
- 维基百科文章的交叉参考关系构成信息网络
- 大百科全书的词条之间的引用网络





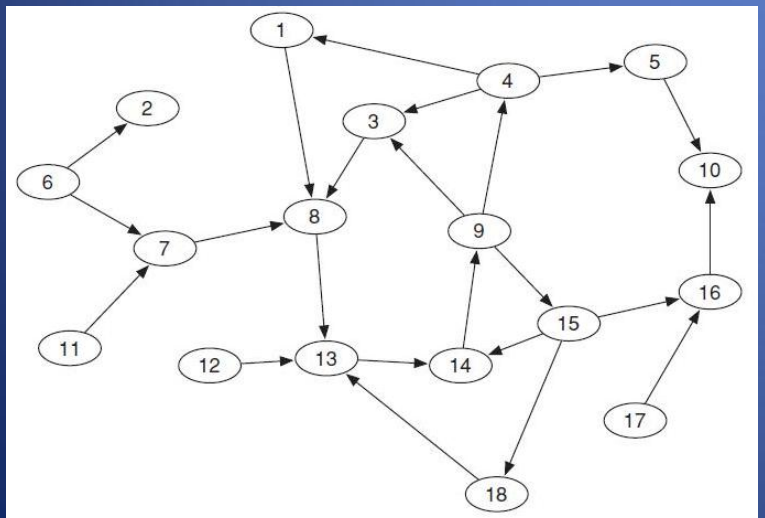

# 一组网页之间构成的一个有向图示例

* 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



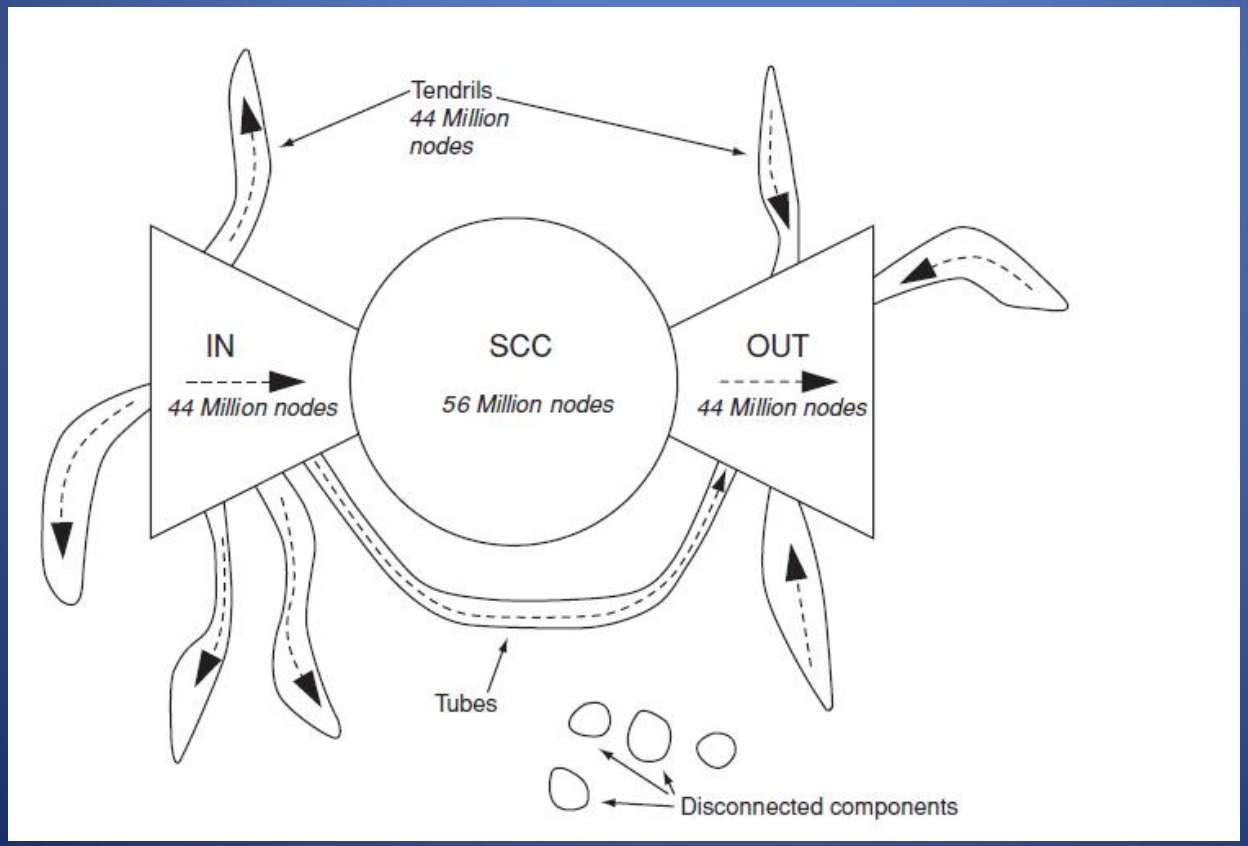

- 与图（无向图）对比
  - 路径
  - 连通
  - 分量
- 易见，强连通概念视觉上不如连通概念直观

# 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



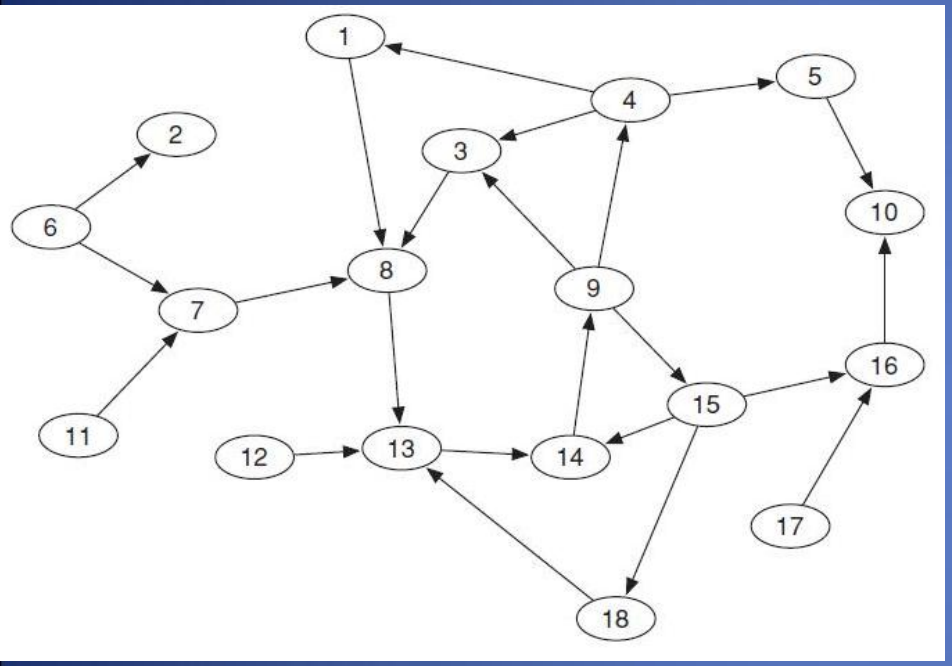

这是怎么知道的？

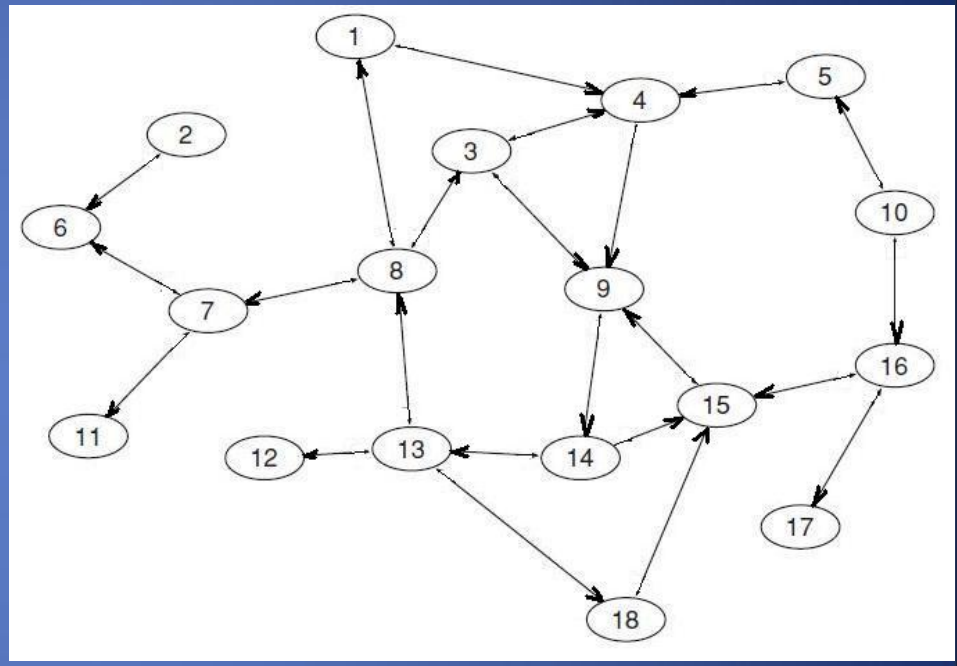
# 基本问题

- 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？

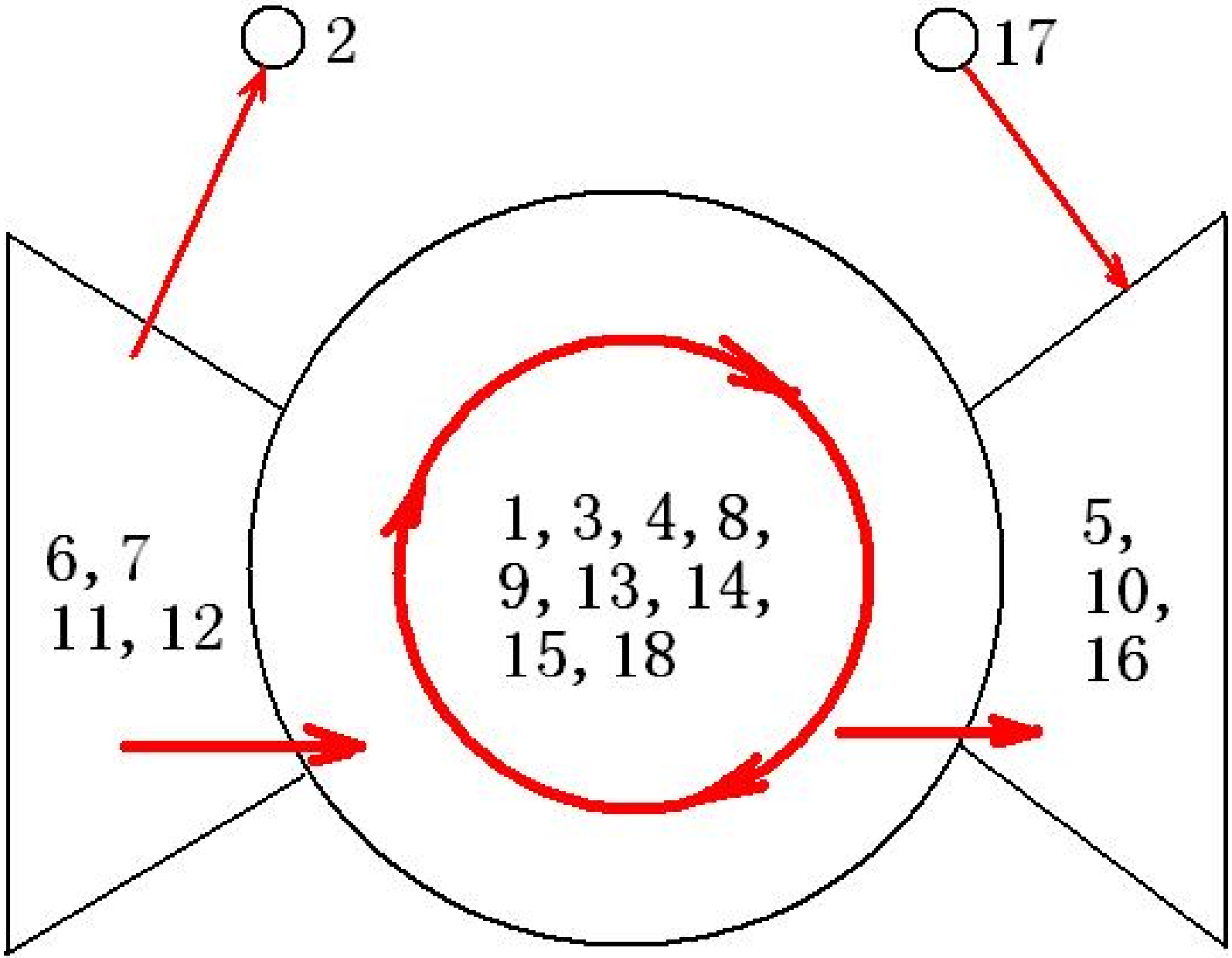
为了回答第一个问题，我们问一个更具体些的问题：给定一个节点，如何确定包含它的强连通分量？

# 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

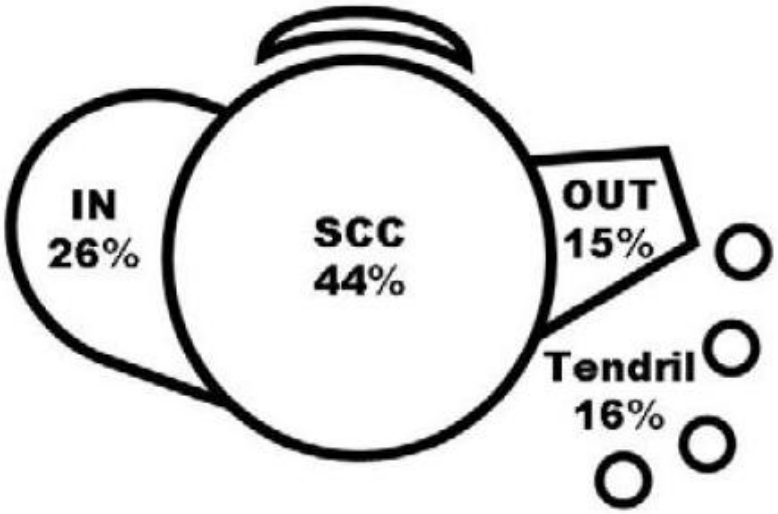
# 计算领结结构的方法（算法）

- 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

# 结果:
# 自相似、层次性

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 | 16.9M (hosts) | 0.79M (domains) |
| N of Links | 1.15G | 194.1M | 298.1M | 43.28G | 43.28G | 43.28G |

[1] Taken from [2].