社会网络的大致结构

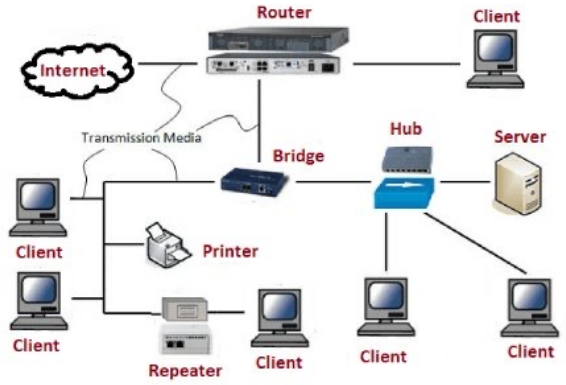
姜少峰

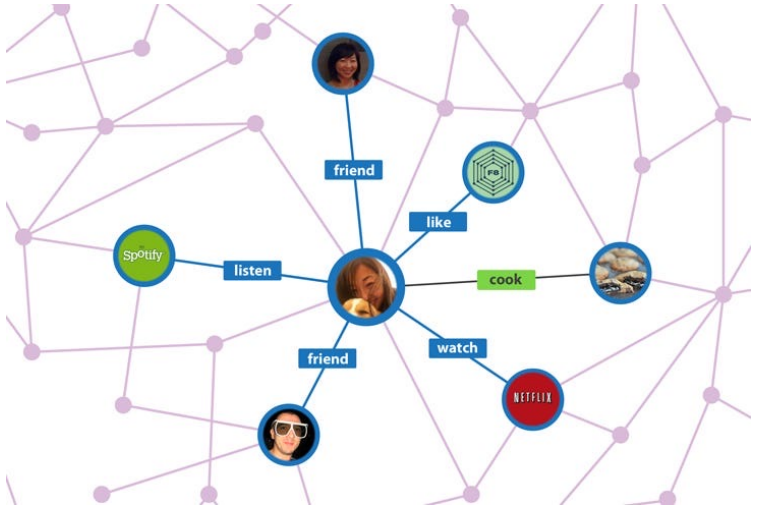
北京大学前沿计算研究中心

Center on Frontiers of Computing Studies, Peking University

图论基本知识

- 图是一组对象之间两两关系的抽象描述
  - 通常图表示为$G = (V, E)$
  - V称作节点集，可以代表任何一组抽象对象（如人，计算机）
  - $E \subseteq V \times V$是边集，如果$(u, v) \in E$那么代表u到v符合某种关系
  - 称两个节点为邻居如果他们有边相连

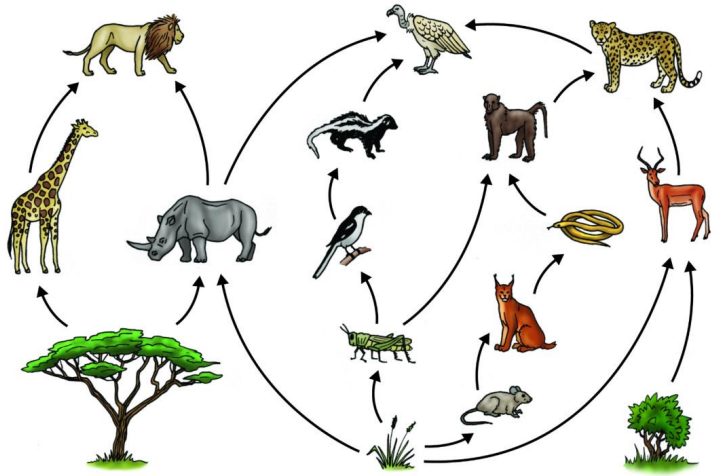




无向图与有向图

- 无向图：$(u,v) \in E \Rightarrow (v,u) \in E$，即关系是对称的；否则为有向图
- 我们默认考虑无向图

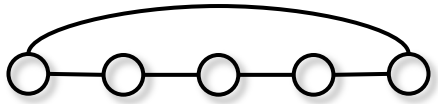
无向图



有向图

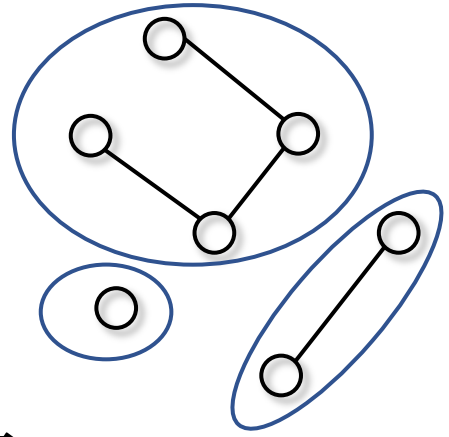
图上的基本结构

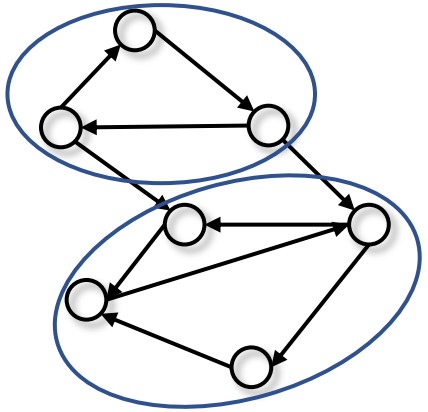
- 度：一个点的度是它连边的条数（有向图又分出度和入度）
- 路径（path）
  - 一个点集的序列，并且相邻点都有边相连；简单路要求没有重复点
  - 长度：定义为路径上的边数
- 回路（cycle）
  - 路径 + 一条边将首尾相连



连通分量与最短路

- 连通分量（connected component）
  - 最大的点集S，使得S中任何两点都有路径连接
  - “最大”：加入任何其他点都不能满足定义

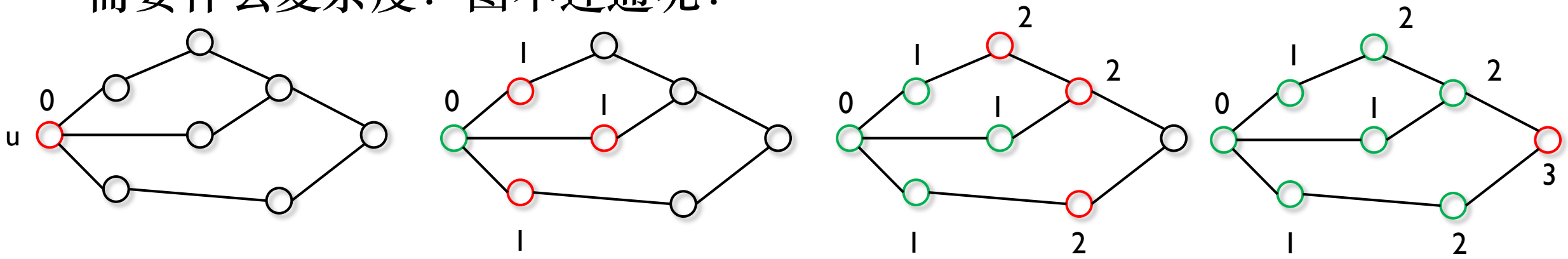




- 距离/最短路
  - 两点间长度最短的路径的长度；这条长度最短的路径叫做最短路

BFS求最短路

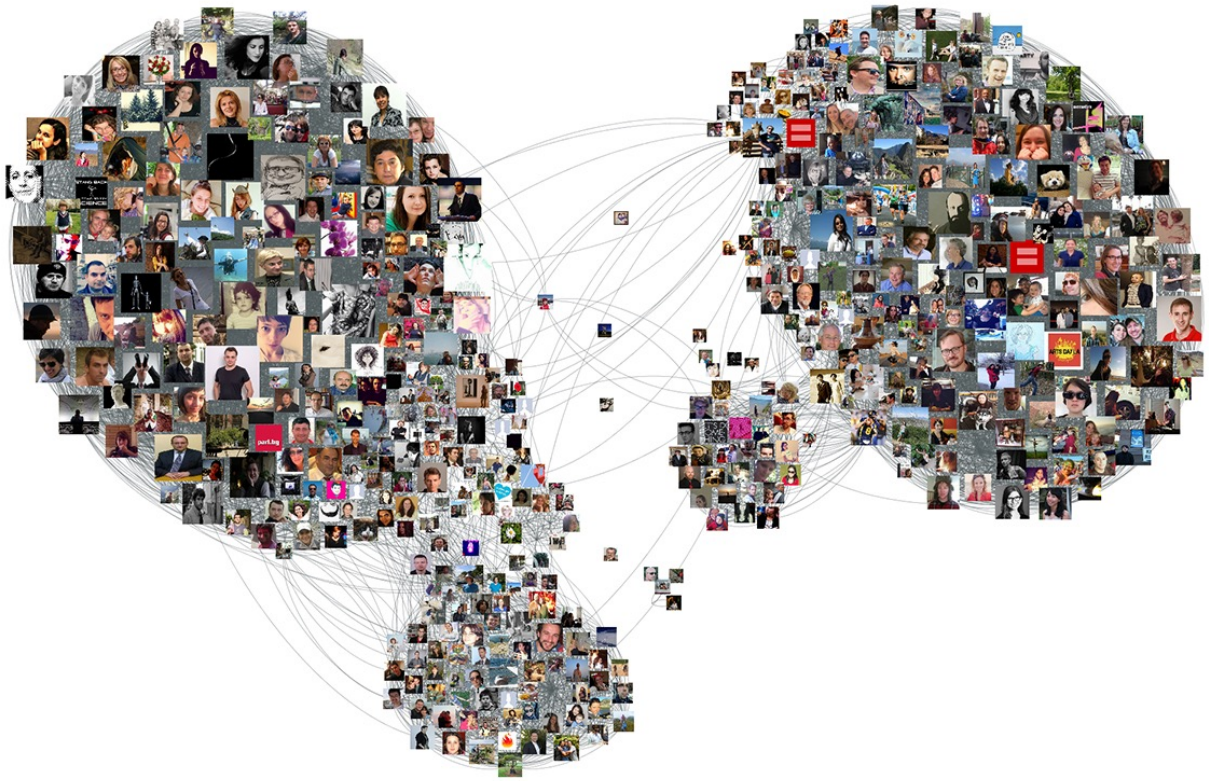
- BFS（Breath-first Search）广度优先搜索
  - 接受一个节点u，找出其他每个节点v到u的距离
- 方法：找到BFS树
  - 从u开始，u的邻居都是到u距离1的，将这些设置为第一层
  - 继续从每个第一层的点v开始，v的邻居都一定是到u距离2的
  - 以此类推，对图进行了分层，第i层的点就是距u恰是i的点的集合
- 需要什么复杂度？图不连通呢？



社交网络的基本结构

超大连通分量

- 现象：社会网络中，通常“超大”连通分量都只有一个



- 为何?
  - “超大”非确指，我们这里考虑有$\Omega(n)$个点（n是图的点数）
  - 设有两个“超大”连通分量A和B
  - A和B之间潜在的边有很多，$\Omega(n^2)$
  - 只要这些里面有一条A到B的边则A和B连通
- 在社会网络中的意义
  - 当两个超大连通分量合并时，往往伴随剧烈的网络变动
    - 《枪炮、病菌与钢铁》集中论述了欧洲人500年前抵达西半球对其文明造成的灾难。
    - 约5000年前，全球社交网络有美洲和欧亚大陆两个超大连通分量
    - 当500年前这两个连通分量合并时，其中一个分量的技术、疾病会迅速传播并压倒对方

小世界现象

- John Guare在1990年写的一个戏剧《Six Degrees of Separation》
  - I read somewhere that everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. The president of the United States. A gondolier in Venice. Fill in the names.
  - 这个“somewhere”是Stanley Milgram在1960年代的一个送信实验
- 绝大多数点不直接相连，但是任何两点都有短路径相连
- 其他例子如Coauthor network：Erdős number
  https://www.csauthors.net/distance/paul-erdos/shaofeng-h-c-jiang

随机图Erdős-Renyi模型

- 随机图Erdős-Renyi模型：
  - 给定n和p，G(n, p)的点集是{1, …, n}，每对点有p的概率独立产生一条边
  - 我们关心n趋于无穷的行为
- 事实上，随机图一点也不“随机”，其行为具有典型性
  - $p = \Omega(1 / n)$时，G(n, p)有一个超大连通分量
- Erdős-Renyi是一个“简化”模型，有很多不符合社会网络结构的地方（习题）

ER的小世界现象

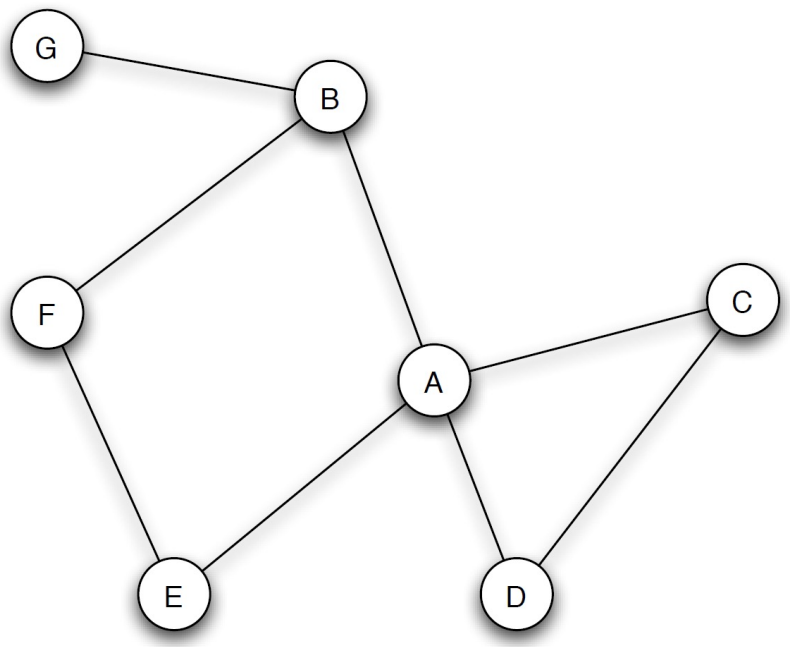
- 固定u和v，u和v可以通过第三边连接的概率是?
  - 即存在w，使得u-w, v-w都有边
  - 对于一个w，u-w、v-w都有的概率是p^2
  - 考虑反面：对w，u-w、v-w至少有一个没有的概率是$1-p^2$
  - 因为有n - 2个w，要求上述独立事件的“与”，所以不存在w使得u-w, v-w都有边的概率是

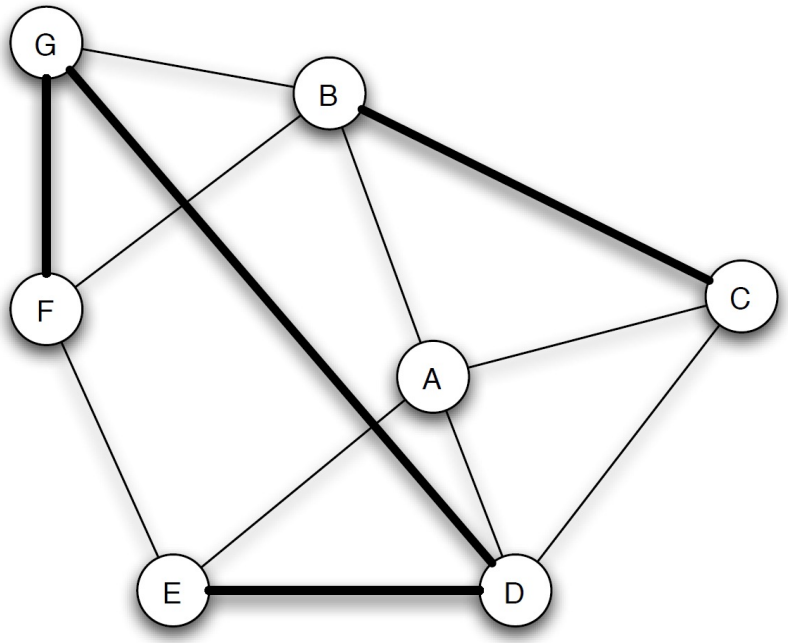
$$(1-p^2)^{n-2}$$

- 结论：当$p \geq 1/\sqrt{n}$时，上述概率$\leq O(1)$，所以说是大概率会有边

形成朋友的典型机制：三元闭包

- 三元闭包（Rapoport, 1953）：
  - 若两个人有一个共同朋友，则这两人以后成为朋友的可能性就会提高





Opportunity, trusting, incentive

聚集系数（clustering coefficient）

- 聚集系数衡量了三元闭包的密集程度
- 局部聚集系数
  - 对于一个点i，令N_i代表i邻居集合，并且令$k_i = |N_i|$

$$C(i) := \frac{|\{u, v \in N_i : (u, v) \in E\}|}{k_i(k_i - 1)}$$

  - **C(i)代表i的两个随机朋友彼此也是朋友的概率**
- 全局聚集系数
  - 图中三角形占所有三元组的比例
  - 三角形：节点u, v, w两两之间都有边

用随机方法
高效估计聚集系数/三角形个数

聚集系数（clustering coefficient）

- 聚集系数衡量了三元闭包的密集程度
- 局部聚集系数
  - 对于一个点i，令N_i代表i邻居集合，并且令$k_i = |N_i|$

$$C(i) := \frac{|\{u, v \in N_i : (u, v) \in E\}|}{k_i(k_i - 1)}$$

  - C(i)代表i的两个随机朋友彼此也是朋友的概率
- 全局聚集系数
  - 图中三角形占所有三元组的比例
  - 三角形：节点u, v, w两两之间都有边
  - 朴素/暴力算法：$O(n^3)$复杂度，大数据不适用！

随机工具：concentration不等式

“定量”版本的中心极限定理

- 中心极限定理描述的是n趋于无穷的情况
  - 收敛速度如何？定量描述？

Chernoff bound (Chernoff, 1952)

设$X_1, \dots, X_n$是独立的$[0,1]$上的随机变量。令$X = \frac{1}{n}\sum_{i=1}^{n} X_i$，$\mu = E[X]$。那么

$$\forall t \in (0,1) \qquad \Pr[|X - \mu| > t\,\mu] \leq 2\exp(-t^2\mu n/3)$$

- 抛掷均匀硬币n次，有多大概率看到0.45n到0.55n个正面？
  - $\mu = 0.5, t = 0.1$，概率至多2exp(-O(n))
  - 所以如果想要p概率看到0.45n到0.55n个正面，只需要抛掷n = O(log 1/p)次！

Chernoff bound (Chernoff, 1952)

设$X_1, \ldots, X_n$是独立的$[0,1]$上的随机变量。令$X = \sum_{i=1}^{n} X_i$，$\mu = E[X]$。那么

$$\forall t \in (0,1) \quad \Pr[|X - \mu| \geq t\,\mu] \leq 2\exp(-t^2\mu/3)$$

$$\forall t > 0 \quad \Pr[|X - \mu| \geq t\mu] \leq 2\exp(-t^2\mu/(2+t))$$

Hoeffding's inequality (Hoeffding, 1963)

设$X_1, \ldots, X_n$是独立的$[a,b]$上的随机变量。令$X = \sum_{i=1}^{n} X_i$，$\mu = E[X]$。那么

$$\forall t > 0 \quad \Pr[|X - \mu| > t\,] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- X_i不必同分布
- Hoeffding's是相加误差，可以转化成Chernoff（但在μ小时稍弱）
- 通常叫做measure concentration inequalities，tail bounds等

关键条件：独立性和有界性

Hoeffding's inequality (Hoeffding, 1963)

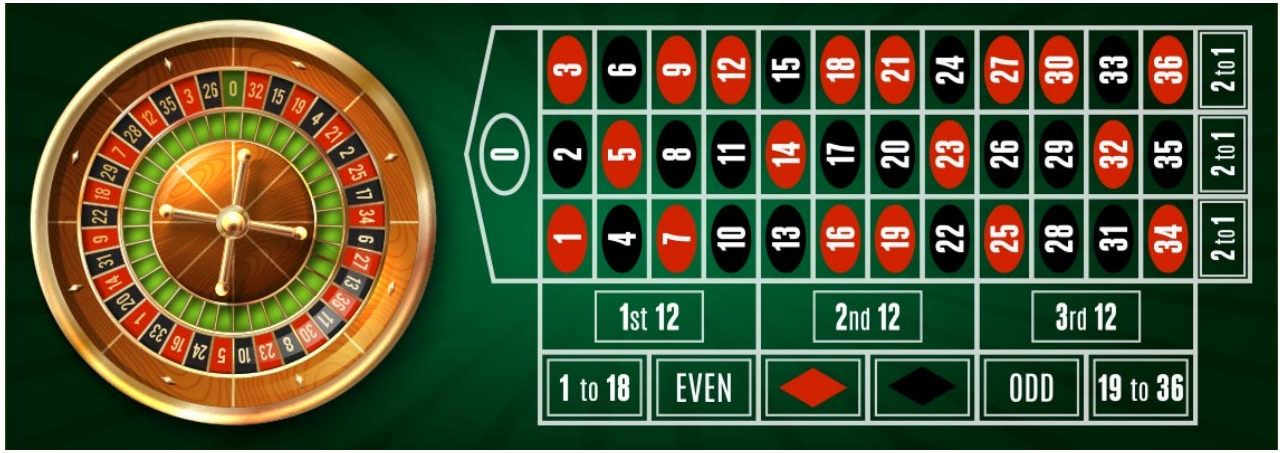
设$X_1, \ldots, X_n$是独立的$[a, b]$上的随机变量。令$X = \sum_{i=1}^{n} X_i$，$\mu = E[X]$。那么

$$\forall t > 0 \qquad \Pr[|X - \mu| > t] \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- 我们已经讨论了为何n个随机变量必须是独立的
- 那么有界性有多重要呢?

赌场的“限红”

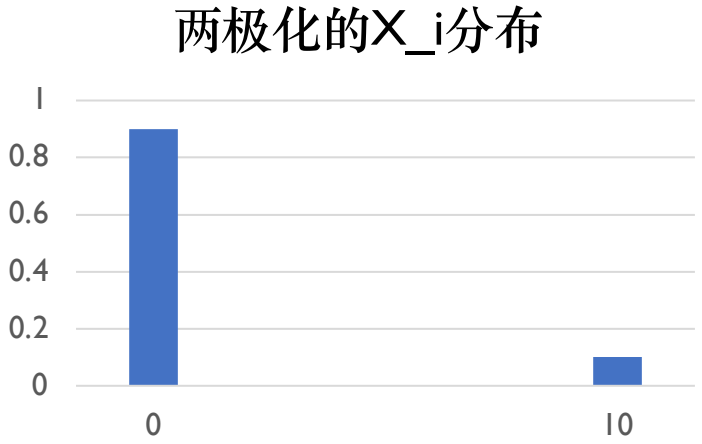
- 赌场想要盈利，必然需要使设计出的游戏的数学期望有利于赌场
- 这就足够了吗?

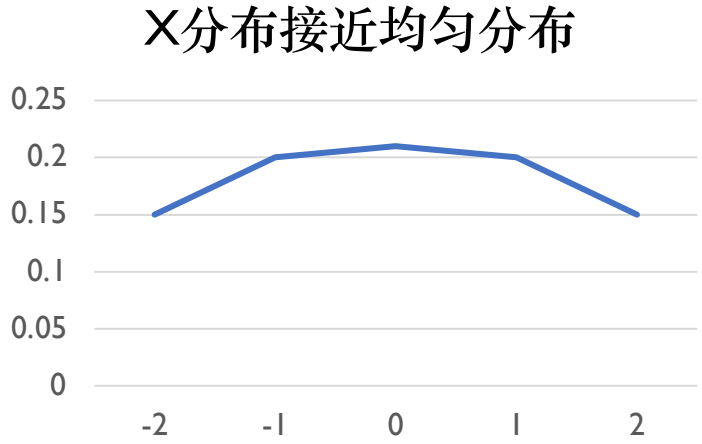


- 轮盘赌奇偶：赔率1：1（如果轮盘是非零且奇偶与所押相同则赢）
- 假设赌场资本T，而赌徒押注也是T，则赌场一局就破产的概率是0.5
- “限红”一箭双雕：控制赔款上界，让赌徒玩更多局

无界变量的反concentration行为

- 如果每个X_i都是{0,T} 随机变量，$\Pr[X_i = T] = 1/T$，故$E[X_i] = 1$
  - 考察$X = \frac{1}{n}\sum_{i=1}^{n} X_i$；我们有$\mu = E[X] = 1$
  - 存在某个$X_i = \mathrm{T}$（也就是$X \geq T/n$）的概率是：$n \cdot \frac{1}{T} \cdot \left(1 - \frac{1}{T}\right)^{n-1}$
  - 所以$\Pr\left[X \geq \frac{T}{n}\right] \geq n \cdot \frac{1}{T} \cdot \left(1 - \frac{1}{T}\right)^{n-1}$
  - 如果$n \leq 0.25\,T$，则有$\Pr[X \geq 4\mu] \geq 0.25 \cdot e^{-0.25} \approx 0.195$



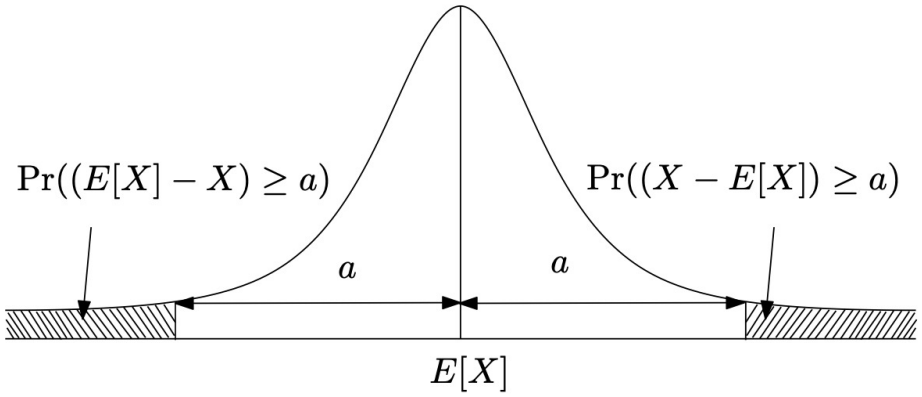


没有独立、有界，但方差小？

Chebyshev's inequality (Chebyshev, ~18XX)
设X的期望是μ方差是σ^2。那么

$$\forall t > 0 \qquad \Pr[|X - \mu| \geq k\sigma\,] \leq \frac{1}{k^2}$$

- 类似正态： $|X - \mu|$在3σ内的概率是1 – 1/9 = 8/9
- 仅当σ较小的时候有作用

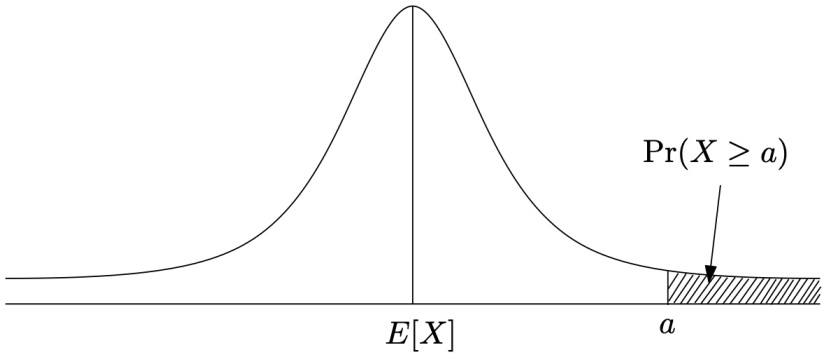


没有独立、有界，无其他条件?

Markov's inequality (Markov or Chebyshev, ~18XX)
设X是非负随机变量。那么

$$\forall t > 0 \qquad \Pr[X \geq t\,] \leq \frac{E[X]}{t}$$

- 如果令t = 2 E[X]，那么就是$\Pr\left[X \geq 2E[X]\right] \leq \frac{1}{2}$
- 另一侧不能保证
- “非负”十分重要；没有“非负”的反例?



随机方法估计三角形个数

- 设三角形个数是T，图有n点m边
- 计划：设计一个随机变量Z，使得E[Z] = T

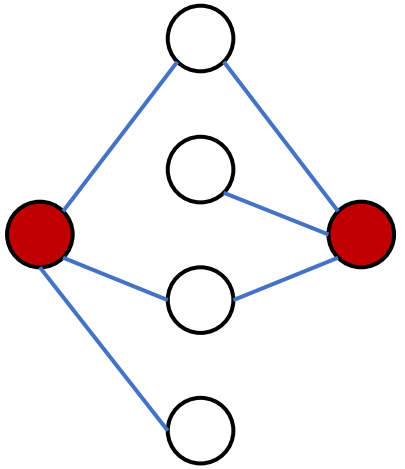
1. 均匀随机选取一条边e = {u, v}
2. 设c_e是e的两端点u和v的公共邻居的数量
   （也就是以e为边的三角形的个数）
3. $Z = m \cdot c_e$

$$E[Z] = \sum_e \frac{1}{m} \cdot m \cdot c_e = \sum_e c_e = 3T$$

- 因此Z/3是一个无偏估计
- 该方法用时：$O(n + m)$

多次试验取平均值

- 将上述随机试验独立重复L次，设估计量是$Z_1, \ldots, Z_L$
- 设最终的$Z = \frac{1}{L} \cdot \sum_{i=1}^{L} Z_i$
- 能用Chernoff/Hoeffding吗?
  - 回忆：每个Z_i都是$m \cdot c_e$
  - 独立性没问题，但c_e上界可以到n
  - 引入假设：所有点的degree都至多是d
  - 可以推出：$c_e \leq d$
  - 因此可以有$Z_i \leq md$



至多连出d条边，
形成d个三角形

套用concentration不等式

Hoeffding's inequality (Hoeffding, 1963)
设$X_1, \ldots, X_n$是独立的$[a, b]$上的随机变量。令$X = \sum_{i=1}^{n} X_i$，$\mu = E[X]$。那么

$$\forall t > 0 \qquad \Pr[|X - \mu| > t] \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- 如何套用：令$X_i = Z_i/3$，$\mu = L \cdot T$，$t = \epsilon\mu$
- $\Pr[|X - \mu| > \epsilon\mu] \leq 2 \exp\left(-\frac{2 \cdot \epsilon^2 \cdot L^2 T^2}{L \cdot m^2 d^2}\right) = 2 \exp\left(-2\epsilon^2 \cdot L \frac{T^2}{m^2 d^2}\right)$
- 令该概率$\leq \delta$，我们可以取$\mathrm{L} = O\left(\frac{m^2 d^2}{\epsilon^2 T^2} \ln \frac{1}{\delta}\right)$

总结

- 独立运行$\mathrm{L} = O\left(\frac{m^2 d^2}{\epsilon T^2} \ln \frac{1}{\delta}\right)$次随机试验，每次随机试验

1. 均匀随机选取一条边$\mathrm{e} = \{\mathrm{u}, \mathrm{v}\}$
2. 设c_e是e的两端点u和v的公共邻居的数量
   （也就是以e为边的三角形的个数）
3. $Z = m \cdot c_e$，这里m是边数

- 总共运行时间$O\left(\frac{m^2 d^2}{\epsilon^2 T^2} \ln \frac{1}{\delta} \cdot (n+m)\right)$，误差$\epsilon$，失败概率$\delta$

- 注意：
  - 仅当T（相对m）足够大时才有效；但通常不难达成：T是n^3量，m是n^2量
  - 看似需要某个对T的预估；实际可以重复试验直到收敛
  - 问题：如果多次试验都找不到三角形，能说明什么？

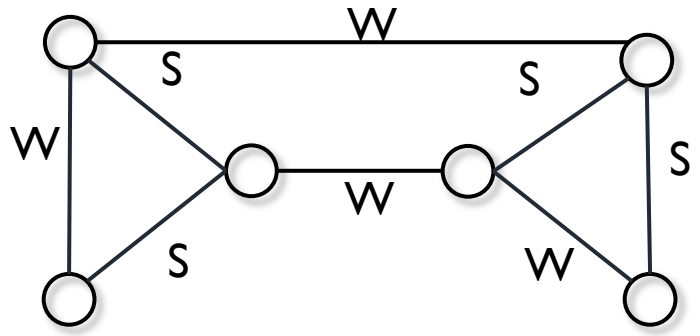
强联系与弱联系

引例

- 在对最近换过工作的人的采访中发现，他们心仪的工作机会多是从“熟人但又不是密友”的朋友那里得知/引荐的
- 有何道理？为何不是密友？为何不是毫不相关的人？

关系的强度

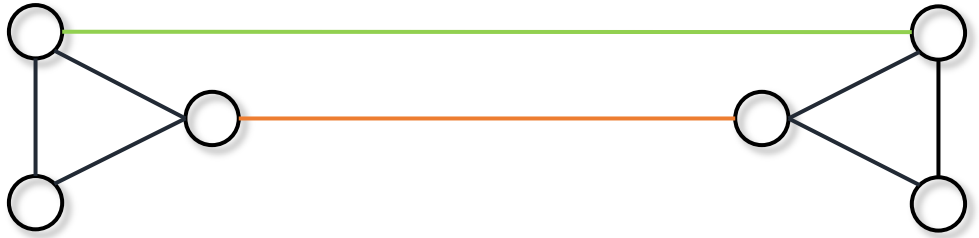
- 关系（也就是边）的“强度”：强度越大代表越亲密
  - 考虑最简单的二分情况：朋友（强联系） vs 熟人（弱联系）
- 让每个人（节点）标记自己到邻居连边的强弱



关键联系：桥与捷径

- 桥：
  - 一条删掉后就会增加连通分量的边
  - 社交网络通常不会出现桥——人与人之间常有一些隐秘的联系

- 捷径：
  - 一条边(u, v)的两个端点u, v没有共同的邻居
  - 捷径的跨度定义为去掉该边后的距离



  - 捷径沟通了两个社群，是信息流通的关键桥梁

捷径和弱联系

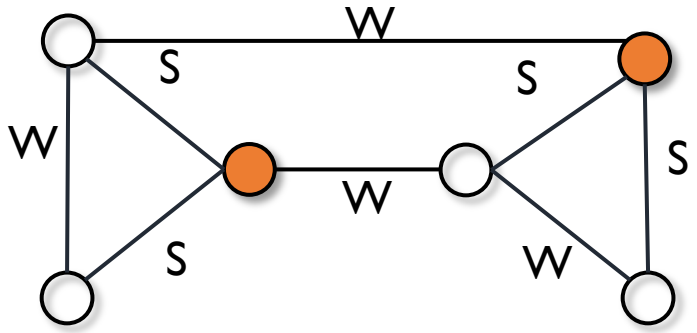
定理：社交网络中，若节点u满足强三元闭包性质，并有至少两个强联系邻居，则与其相连的任何捷径都是弱联系

大致说明：捷径都是弱联系带来的

强三元闭包

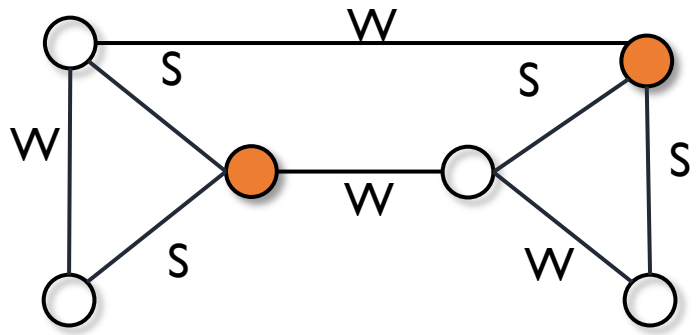
- 强度与三元闭包：设A-B,A-C有边且都是强联系，则B-C很可能形成

- 强三元闭包（一个理想化定义）：对节点A，若存在B, C与A为强联系且B, C无连接，则A违反强三元闭包性质，否则A满足强三元闭包性质



捷径和弱联系

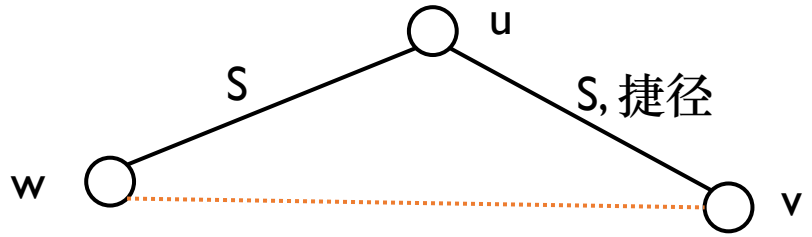
定理：社交网络中，若节点u满足强三元闭包性质，并有至少两个强联系邻居，则与其相连的任何捷径都是弱联系



证明

定理：社交网络中，若节点u满足强三元闭包性质，并有至少两个强联系邻居，则与其相连的任何捷径都是弱联系

- 反证法：假设满足条件的u存在一条捷径是强联系
- 令v, w为u的两个强联系邻居，并且u-v是一条捷径
- 关键问题：v和w之间是否有边？
  - 因u-v是捷径，那么u-w和v-w不能同时存在，故v-w无边
  - 另一方面，因u满足强三元闭包性质，并且u-v, u-w都是强邻居，那么v-w有边
  - 矛盾！



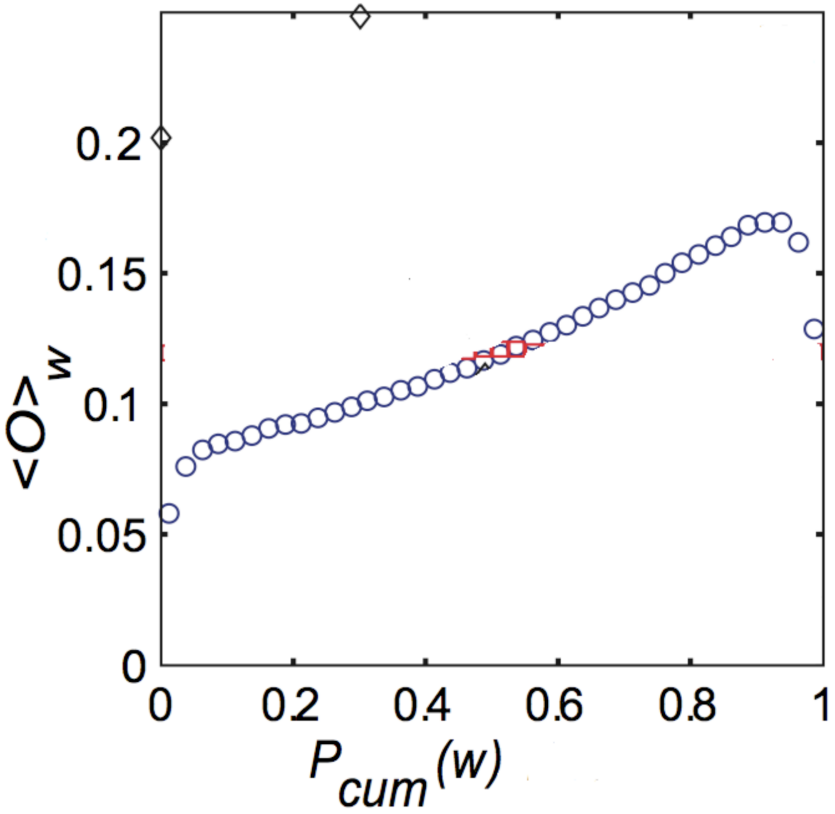
实际数据上的关系强度与网络结构

- “谁和谁讲话”网络
  - 某国20%人口、18周内手机通信数据
  - 手机用户是节点，如果两个用户“频繁”通话，则连边
  - 手机通信一般是私人通信，多在已认识人中进行 -> 网络结构研究的合理样本
- 弱联系概念的推广
  - 之前的定义太二元/极端，实际上应该是某个联系强度数值
  - 这里采用两端通话总分钟数作为“强度值”
- 捷径定义的放松：A-B邻里重叠度（分母不含AB）
  - 分母=0设比率=0，此时对应于捷径

$$\frac{\text{与A、B均为邻居的节点数}}{\text{与A、B至少一个为邻居的节点数}}$$

交集大小除以并集大小

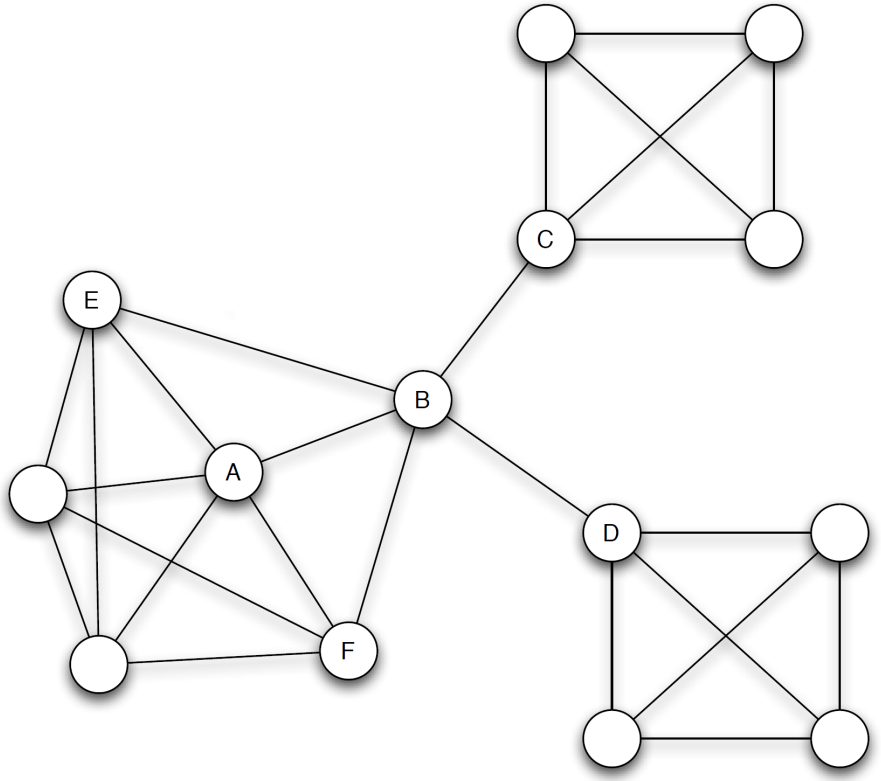
- 对于每条边A-B，计算邻里重叠度
- 将所有边按照强度排序，绘制对应的邻里重叠度的曲线



网络的中心性

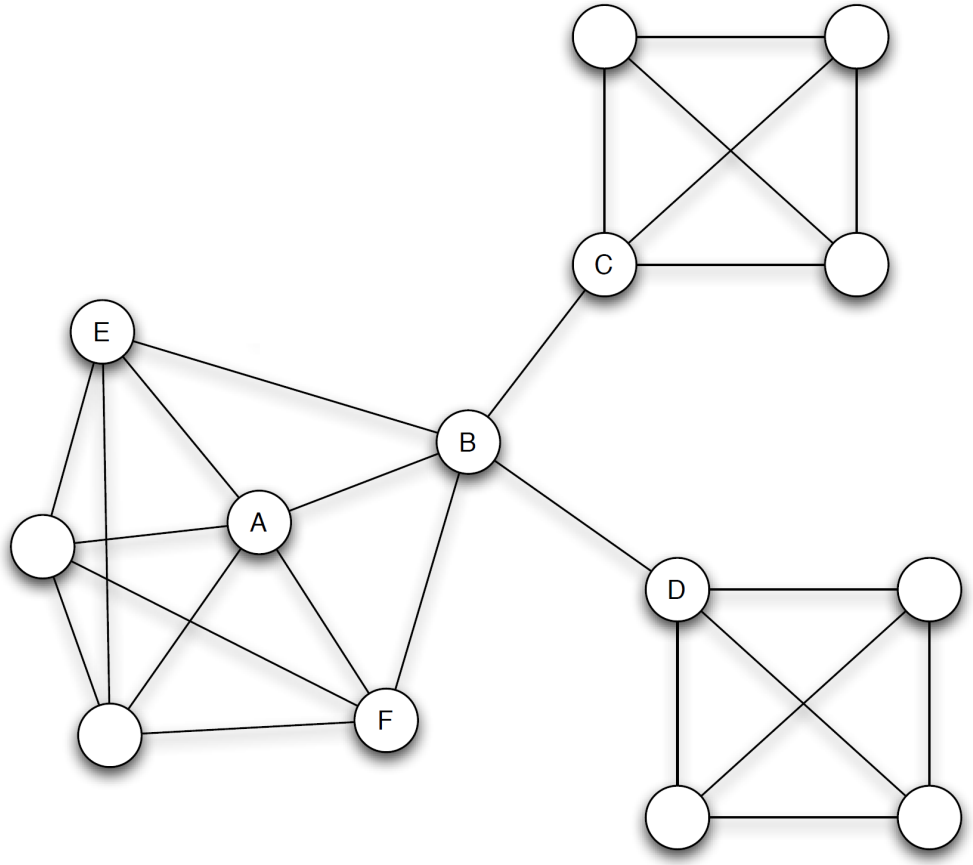
点所处的地位与社会资本

- 嵌入性embeddedness：一条边u-v的嵌入性为uv共同邻居的数量
  - 捷径是嵌入性=0的边
- 社会学研究：两个由高嵌入性边相连的点通常会更加信任彼此
- A被很多人信任



结构洞 structural hole

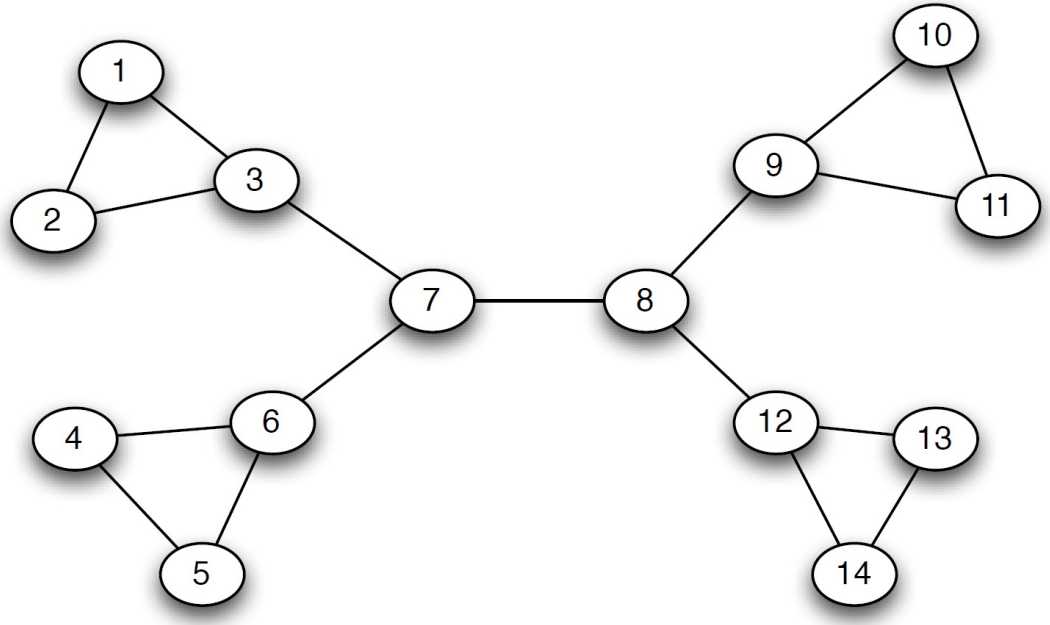
- 节点B的特点？有什么社交优势？
  - “结构洞”：多条捷径的交点
  - 更早获得、整合多个社群的信息
  - 社交“把关”；阻碍连接的形成
  - 同时较少获得信任



边的中心性：介数 (betweenness)

- 对于节点u, v，如果有k条最短路，则赋予每条最短路上的边1 / k流量
- 介数（Freeman 1977）：
  - 一条边的介数是其承载流量和，所有节点对的流量都计算在内

7-8的介数：7 x 7 = 49
3-7的介数：3 x 11 = 33
1-3的介数：1 x 12 = 12
1-2的介数：1



如何计算介数值：框架

- 对每个节点u:
  - 计算u到各个节点的最短路的条数
  - 利用该信息，计算每条边上u出发的流量
- 汇总所有节点u的流量

从定点出发的最短路条数

- 固定u，计算u到其他所有节点的最短路的条数

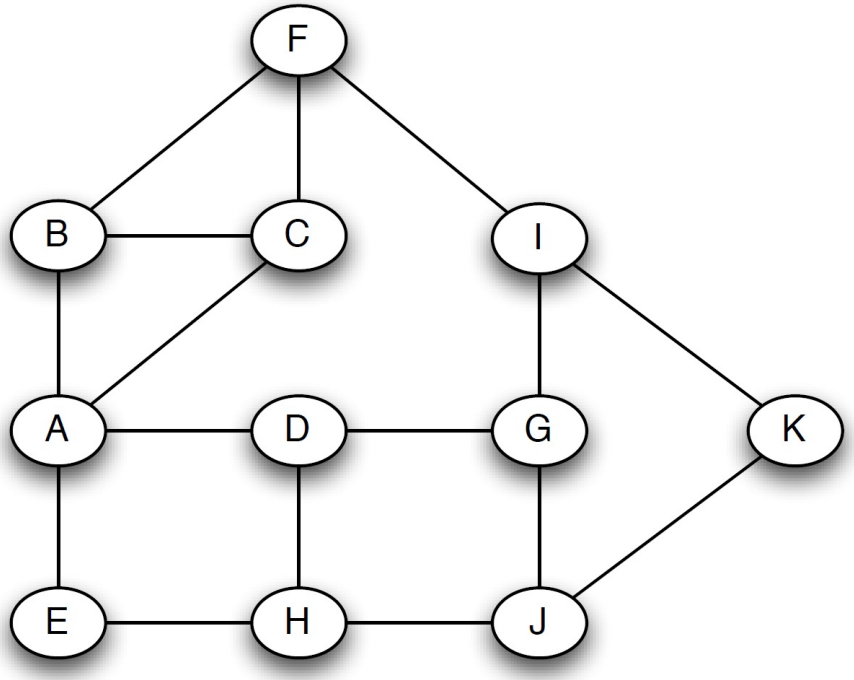
设f(v)是到v的最短路条数

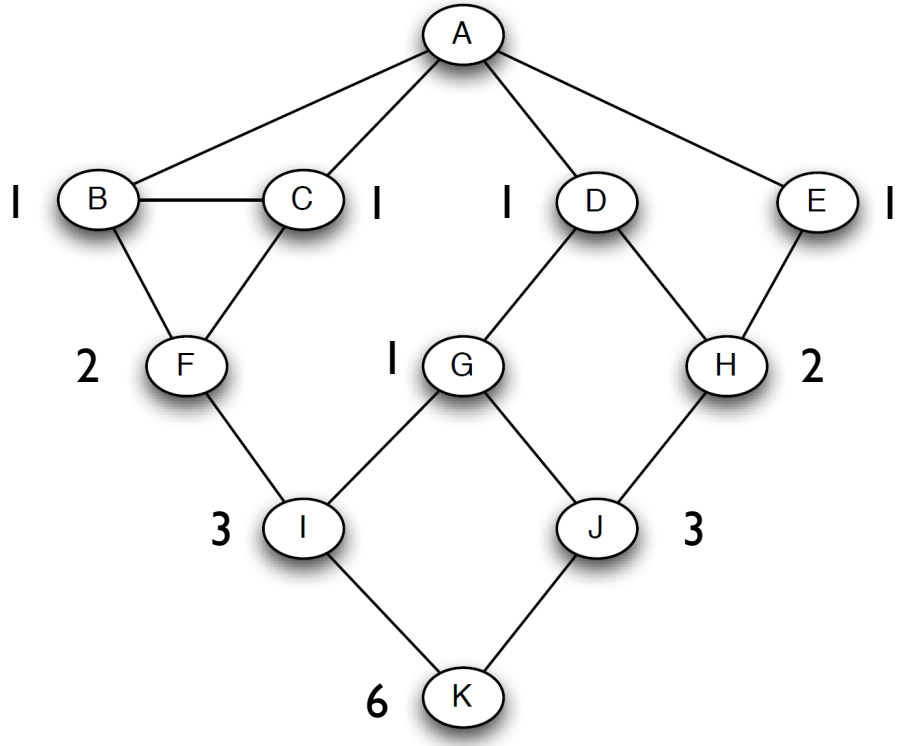
1. 计算u出发的BFS
2. 对于每个点v，设其在BFS上一层的邻居集合是N，则

$$f(v) = \sum_{v' \in N} f(v')$$

实现方法：自顶向下

BFS求最短路条数





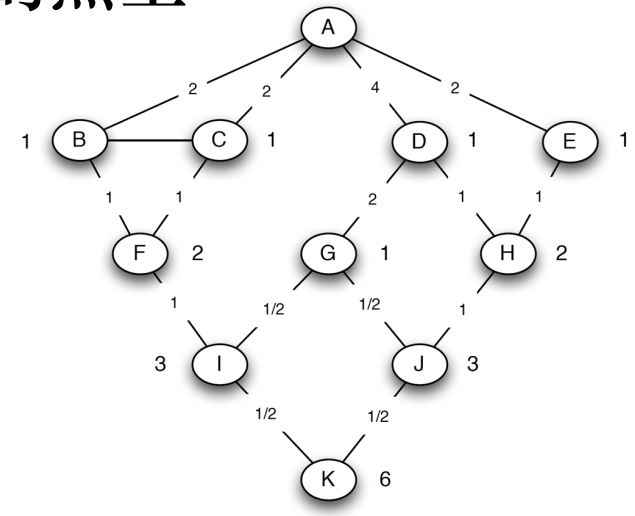
计算流量：固定起点u

- 自底向上：从BFS层数最大的点开始，每个点上有1的流量，往起点方向所有边按最段路条数均匀分配（一条边有k条最短路经过，则分 1/k）
- 对于一个点v，考虑其BFS上一层的邻居集合N
- 观察：对$v' \in N$，边(v', v)上的流量等于v流出的总流量按照$\frac{f(v')}{f(v)}$比例分配

  因为是按照最短路条数均匀分配，因此相对比例就是f的比例

- 继续下一层之前，需要把当前层边的流量累计到连接的点上

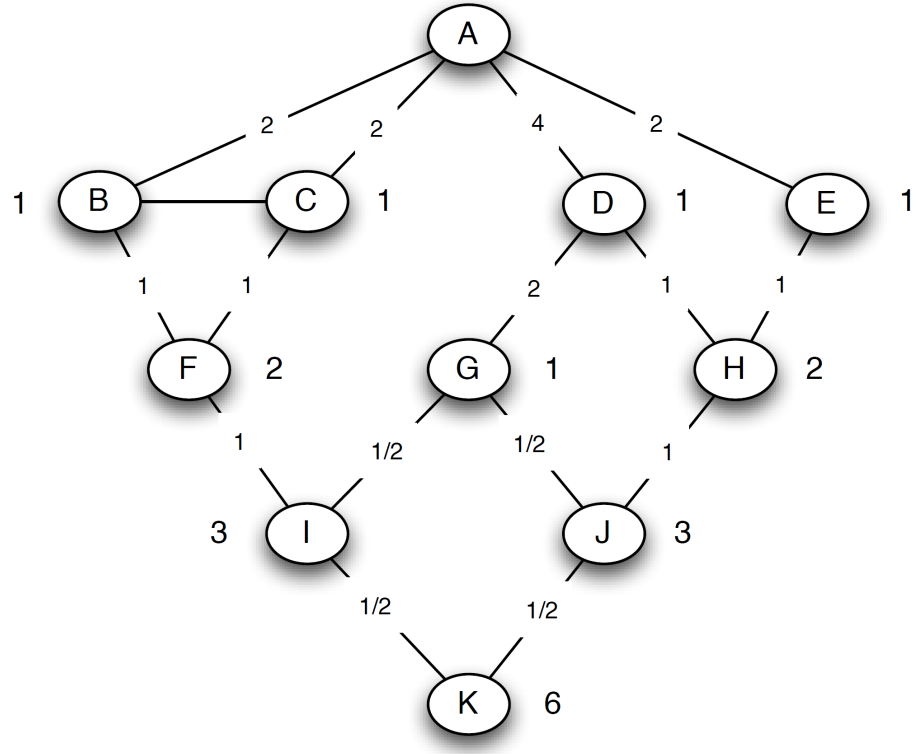
举例：
对于v = K：N = {I, J}，K流出1单位，故I-K和J-K各分1/2
对于v = I：N = {F, G}，I流出1 + 1 / 2单位，故F-I，G-I分1和1/2
一般地，我们继续从底向上进行，直到A为止



计算流量：固定起点u

设BFS有m层，设f[v]为起点u到v的最段路条数，初始化所有点v上的流量c[v] = 1

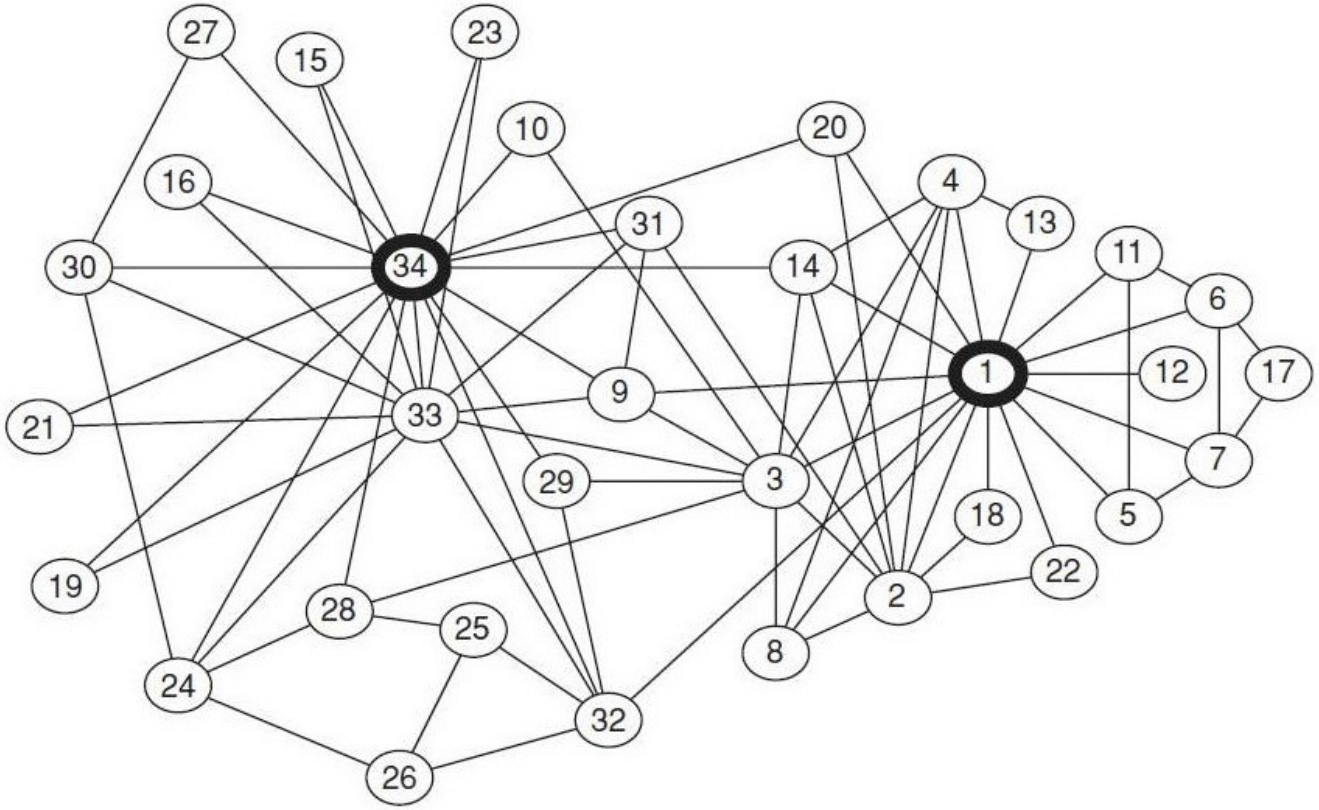
- for $i = m, \dots, 1$
  - for v in level i
    - 设N为$i - 1$层中v的前趋集合
    - 对$v' \in N$，给边(v', v)赋流量c[v]$\frac{f(v')}{f(v)}$
    - 对$v' \in N$更新$c[v'] := c[v'] + c[v]\frac{f(v')}{f(v)}$



- 最后，应该对所有u运行该流量算法，然后累计边上流量来得到介数

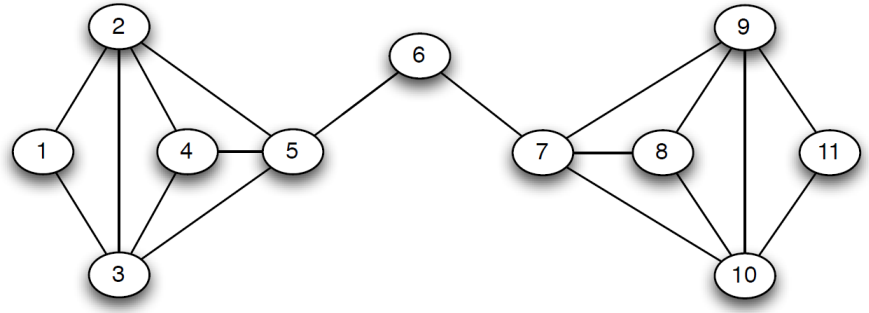
图划分

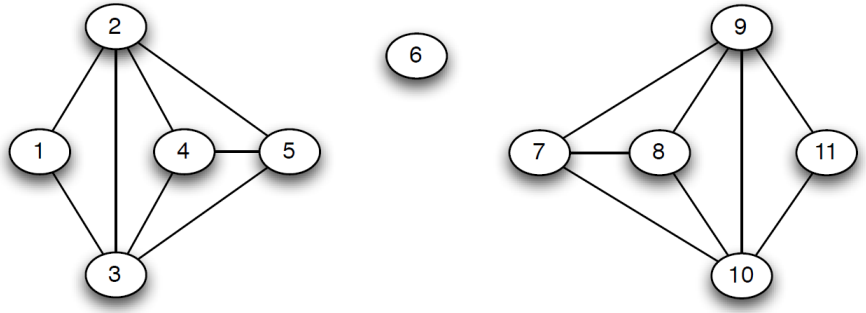
- 弱联系沟通了“关系紧密”的区域
- 图划分：将图分割成关系紧密的区域，区域间稀疏地互联

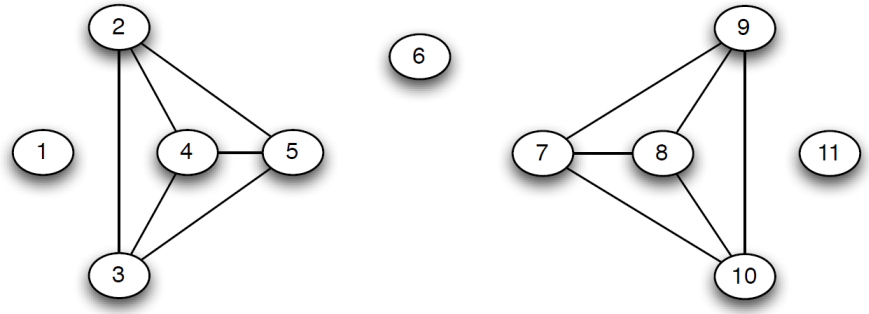


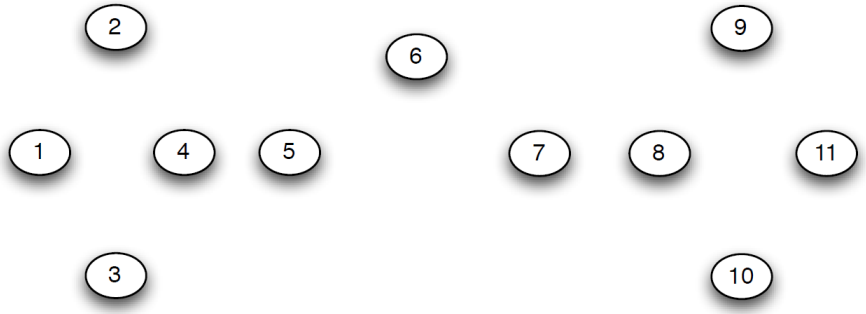
Girvan-Newman

- 每次找到并删除联系紧密区域地“跨接边”，然后在剩余区域递归
- 如何找到“跨接边”：“介数”（betweenness）最大的边



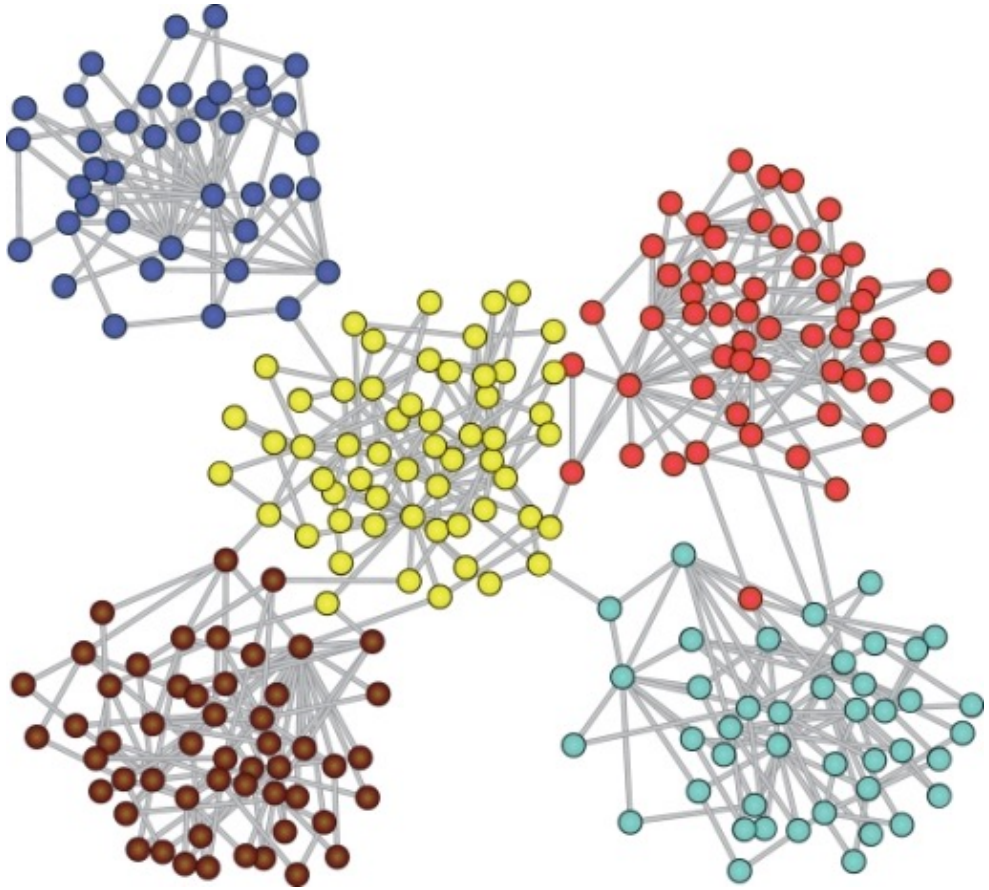






Community Detection

- Girvan-Newman方法是community detection的重要方法



Girvan-Newman方法的效率?

- 然而，运行速度无法扩展到大图上（试分析其时间复杂度?）
- Girvan-Newman方法的（近似）大数据算法是极有意义的科研问题

其他图/空间划分算法选讲

一些基本概念

- 算法的近似比
- 度量空间（metric space）

k-聚类

- 需要定义一个距离函数
  - 距离函数：是一个metric space；常见用欧氏空间$\mathbb{R}^d$
- (k, z)-Clustering：输入数据是一个$X \subseteq \mathbb{R}^d$以及一个参数k，目标是找到一个k个点的$C \subseteq \mathbb{R}^d$使得

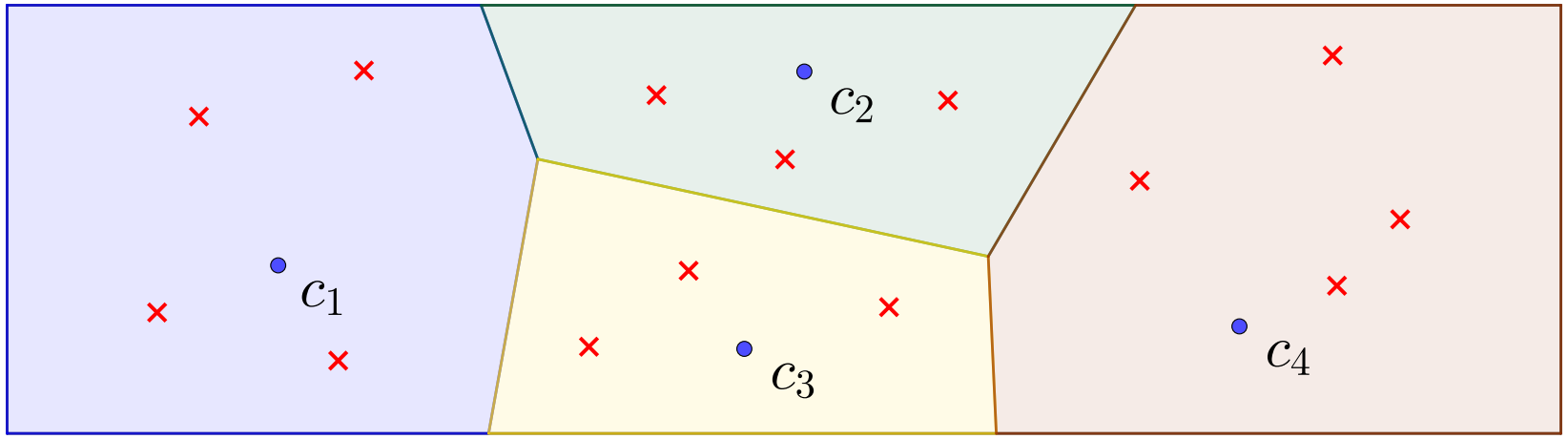
$$cost(X, C) := \sum_{x \in X} \mathrm{dist}^z(x, C)$$

这里：

$$\mathrm{dist}(x, C) := \min_{c \in C} dist(x, c)$$

$$\mathrm{dist}(x, y) := ||x - y||_2$$

$z = 1$是k-median，$z = 2$是k-means



求解算法

- k-means用的最广泛，算法也最多
- 最经典的：Lloyd heuristic (Lloyd, 1982)；历史上是用来定义k-means

选取初值C_0，令i = 0
Do
$i := i + 1$
根据最近邻规则C_{i-1}对应的聚类，对每个类找mean作为聚类中心，形成C_i
Until C_i与C_{i-1}足够接近

- 一个点集S的mean：$\frac{1}{|S|}\sum_{x \in S} x$，即几何中心；可证是1-mean的最优解
- 每轮循环可以在O(nk)时间实现，但轮数和精确度很依赖初值的选取

k-means++ (Arthur, Vassilvitskii 2007)

- k-means++ 是一种为k-means选初值的方法
- 事实上，k-means++本身就是一个近似算法

初始化：选取任意数据点进入候选集合C
For $i = 1, \ldots, k - 1$
　　采样$x \in X$，每个x的概率正比于$\mathrm{dist}^2(x, C)$
　　$C := C \cup \{x\}$
返回C

- 这个初值本身是$O(\log k)$近似的，后面接Lloyd可以继续改进这个解

k-center

- k-center是另一个标准k-聚类变种
- 目标函数：$cost(X, C) := \max_{x \in X} \mathrm{dist}(x, C)$
- 算法： Gonzalez (1985)

初始化：选取任意数据点进入候选集合C
For $i = 1, \ldots, k - 1$
    找到使$\mathrm{dist}(x, C)$最大的$x \in X$
    $C := C \cup \{x\}$
返回C

- 该算法是2-近似的

相似度/距离度量

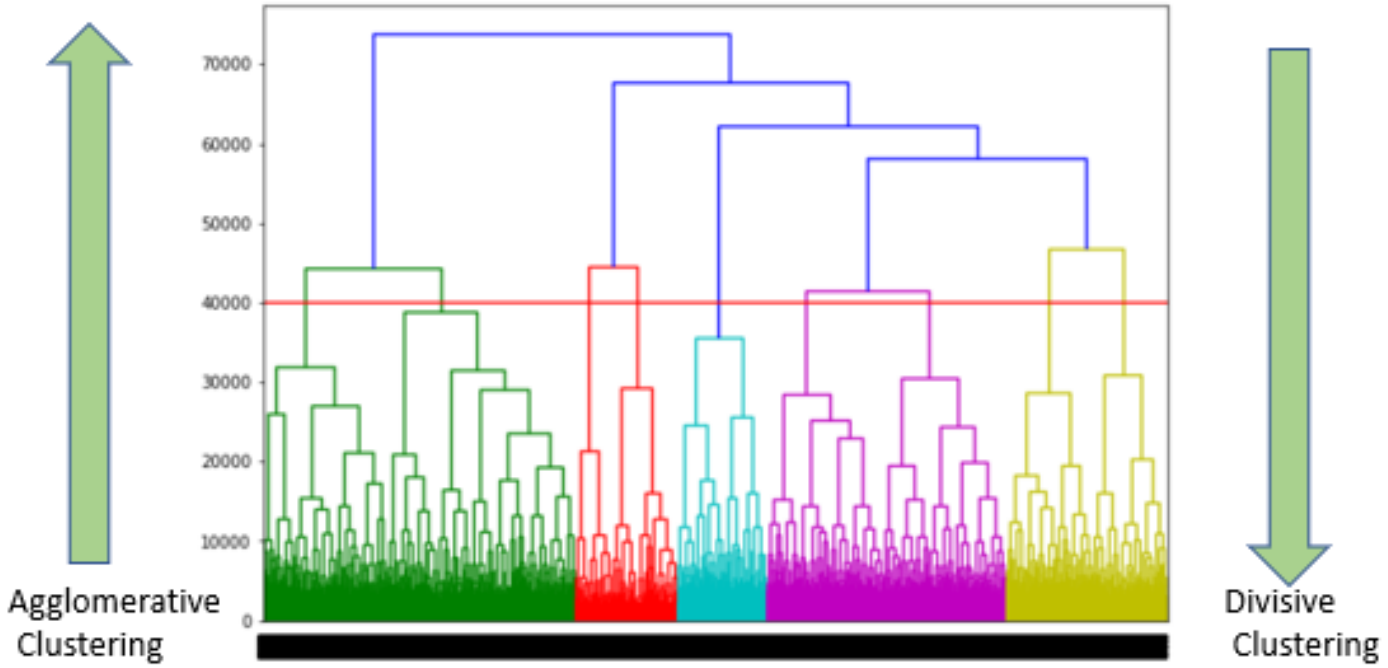
- 如果数据可以embed到$\mathbb{R}^d$，衡量两个点之间的距离
  - ℓ_p-norm，常见$\ell_2, \ell_1, \ell_\infty$ – 有三角形不等式
  - cosine similarity：$\arccos \frac{\langle a,b \rangle}{||a||_2 ||b||_2}$，衡量a和b向量之间的夹角 – 无三角形不等式!
- 如果是一般的图数据，衡量两个节点之间的相似度
  - 基于结构等价：两个点如果连到完全一样的邻居那么就是结构等价的
    - 设A是邻接矩阵，则$\text{dist}(i,j) = ||A_i - A_j||_2$，这里$A_i$是i这一行
  - 类似的：基于相交占比$\frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}$，其中N(i)代表i得邻居的集合
  - 另一个类似的：A_i和A_j的Pearson correlation
  - 这些度量一般不满足三角形不等式

k-聚类的优劣

- 优势：成熟且简单易行，快速算法、甚至大数据算法有很多可用
- 劣势：需要预先确定k，一般需要$\mathbb{R}^d$表示

层次聚类

- 层次聚类：不需要预先知道k
  - 图中每条边需要赋予一个权重，衡量的是两个端点的相似度
  - 分为top-down和bottom-up方法



层次聚类: Bottom-up

- Bottom-up方法是最常用的，效果一般不错
- 设给定图是G = (V, E)，并且变权函数是w（边权衡量相似度）
- Linkage方法：
  - 初始所有点都各自一类（即都是一个元素的类）
  - 每次从当前所有类中找“距离”最近的两个合并
  - 直到剩下全集
- 两种具体的“距离”定义：
  - Single linkage clustering：集合S和T的“距离”定义为 $\min_{s \in S, t \in T} w(s, t)$
  - Complete linkage clustering：类似，但是min改成max

层次聚类:Top-down

- 传统上没有很好的一般方法
- 对于能表示成$\mathbb{R}^d$的情况：一个经典方法是bisecting k-means，即递归运行一个k = 2的k-means来划分每个集合

对层次聚类的系统研究

- 层次聚类很长一段时间都是用算法来定义问题的，即没有目标函数
- (Dasgupta, STOC 2016)给出了一个目标函数的定义
  - 证明了实际表现良好的single-linkage确实在新的目标函数下也是好的
- 由此目标函数提出了一个新的top-down的算法
  - 每次不是找2-means，而是找sparsest cut
  - 对于一个集合V和子集S，S定义的cut的sparsity定义为$\frac{w(S,V\backslash S)}{|S|\cdot|V\backslash S|}$
  - 每次递归到一个集合V，找到（近似）最sparse的S进行二划分
  - 相对bisecting k-means，该方法对一般的图都适用，而不只是有$\mathbb{R}^d$表示的

- Dasgupta之后又诞生了若干其他对目标函数建模的文章
  - Cohen-Addad et al., JACM 2019
  - Moseley et al., NeurIPS 2017
- 对于层次聚类的系统性研究依然是科研热点
  - 新近似算法
  - 适用于大数据的算法
  - （特殊图上）更合适的目标函数