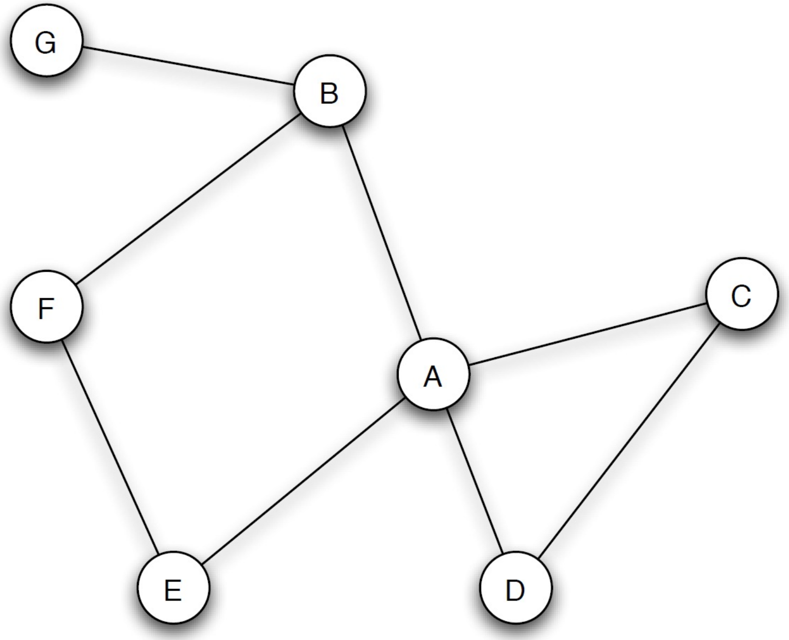
网络的演化：外部环境的力量

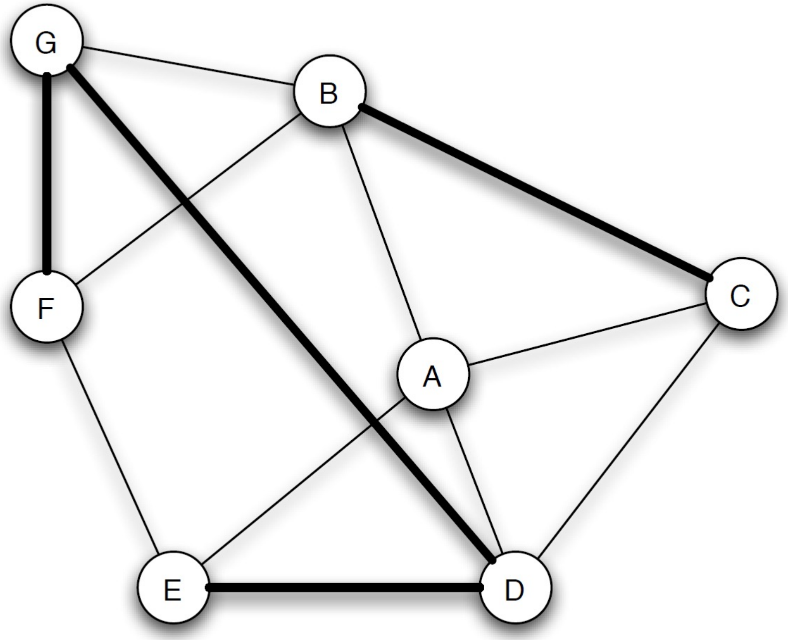
姜少峰

北京大学前沿计算研究中心
Center on Frontiers of Computing Studies, Peking University

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

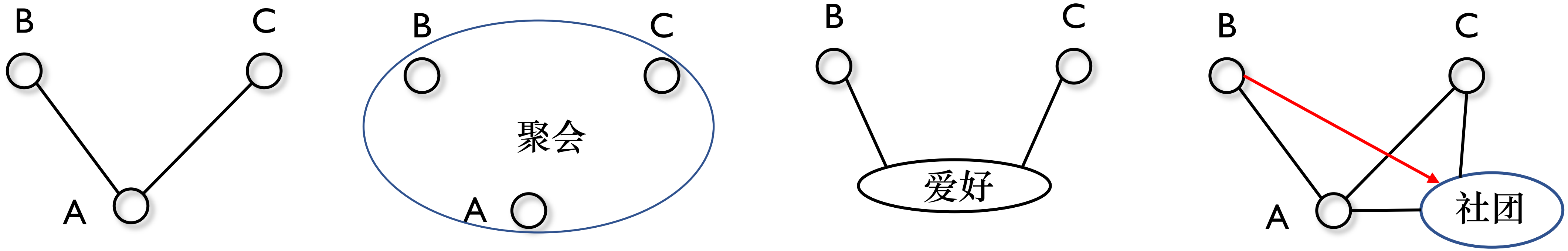




Opportunity, trusting, incentive

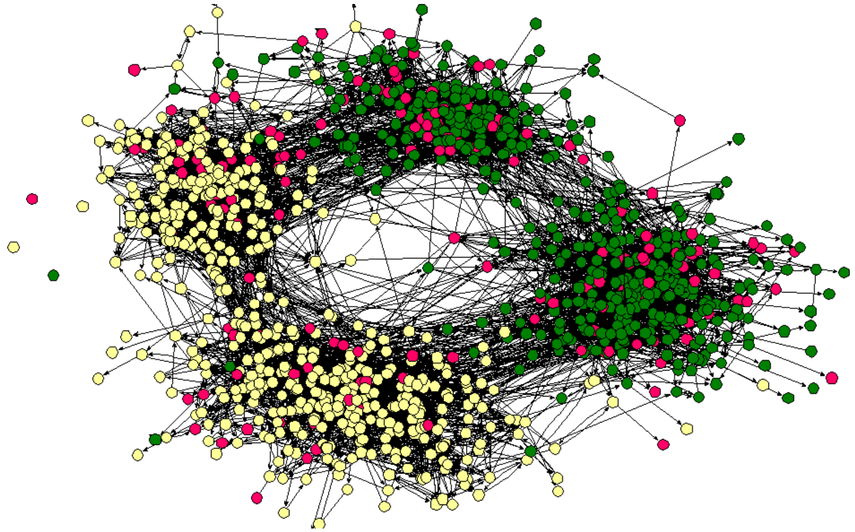
三元闭包的成因?

- 三元闭包中的"机会"，往往是在网络之外发生的
  - 例如A组织了一个聚会
  - B与C相遇的机会也可能是偶然的（与A并没有关系），相似性有助于成朋友
  - 有可能，A与B是同学关系上的朋友，A与C是同属某社团关系上的朋友，长此以往，B也对该社团产生了兴趣，加入该社团，进而和C成为了朋友



同质性 (Lazarsfeld & Merton, 1954)

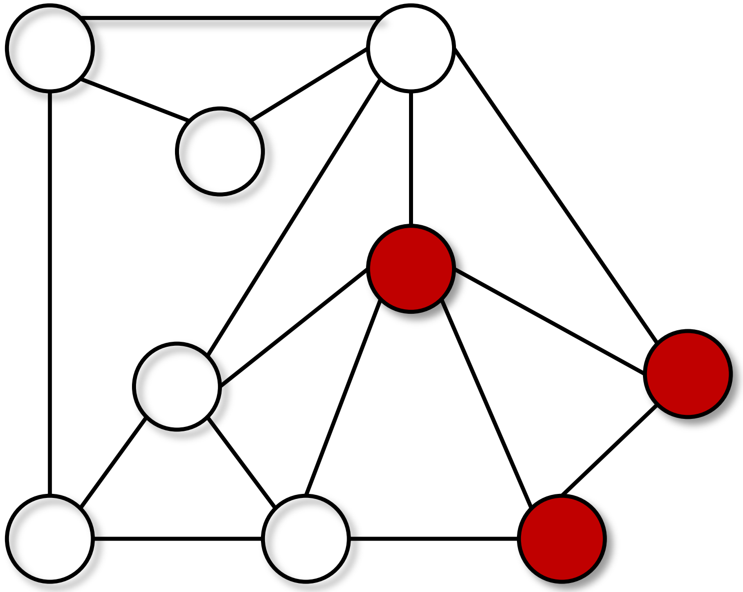
同质性homophily：朋友（相近的人们）之间具有某种特征相似性的社会现象



物以类聚，人以群分；近朱者赤，近墨者黑

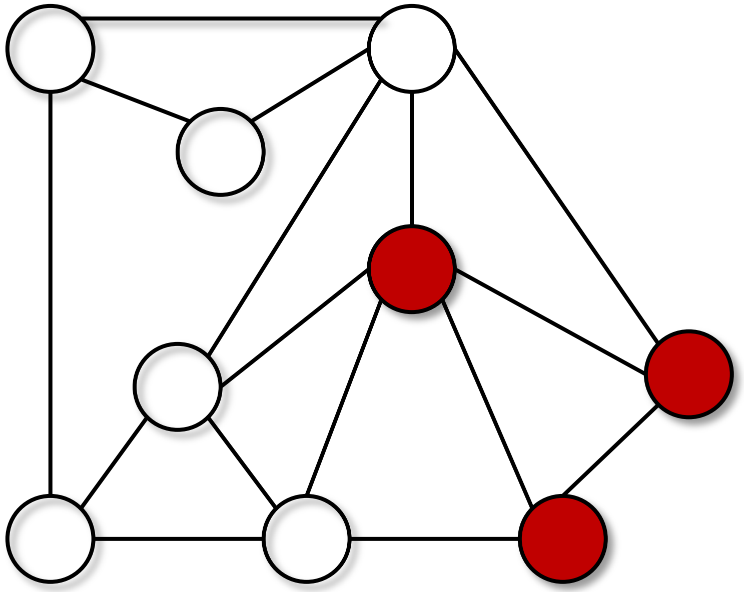
量化同质性

- 两个人之间究竟在哪些方面是“同质性”的?
  - 可能是两个人就职于同一个公司，拥有同一个爱好，或者更加隐秘的联系
- 重点：如何测试某个具体的“特征”是否表现出了“同质性”？
- 以性别为例进行讨论

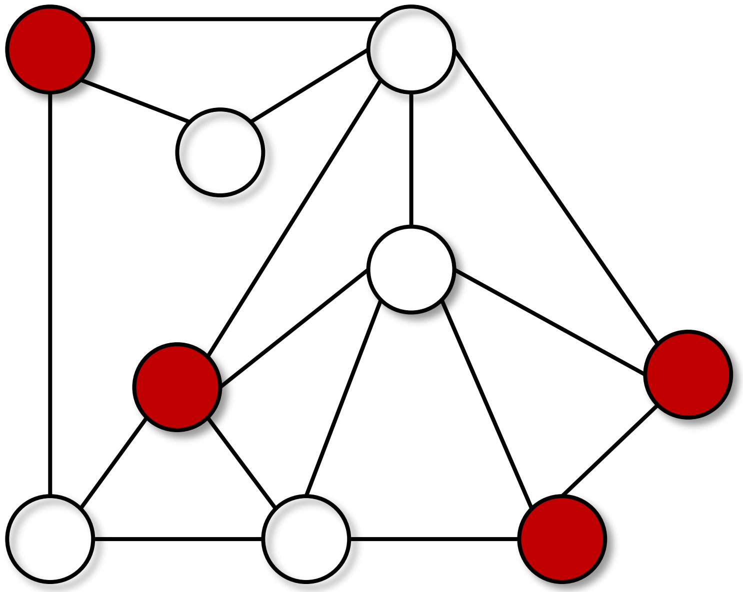


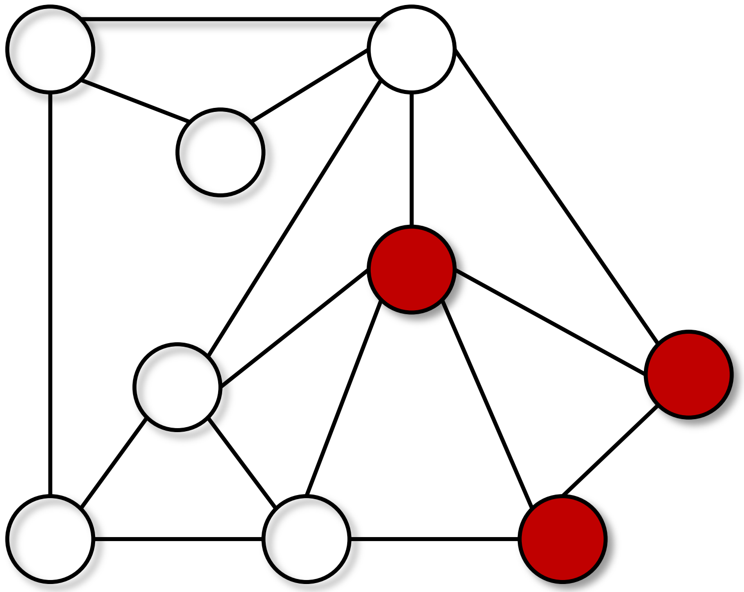
测试“同质性”

- 如果有“同质性”，我们预期看到什么?
  - 人们会依据性别来“聚类”
  - 比如女生和女生是朋友，而男生和男生是朋友



- 考虑反面：如果网络中没有关于性别的同质性？
  - 对于每个人，ta的邻居的男女比例应该与人群平均情况类似
  - 技术上，利用统计假设检验的思想

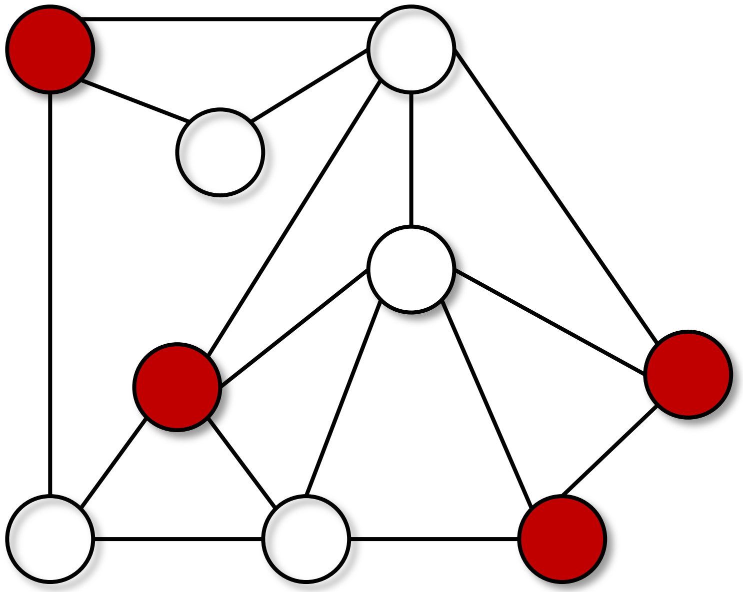


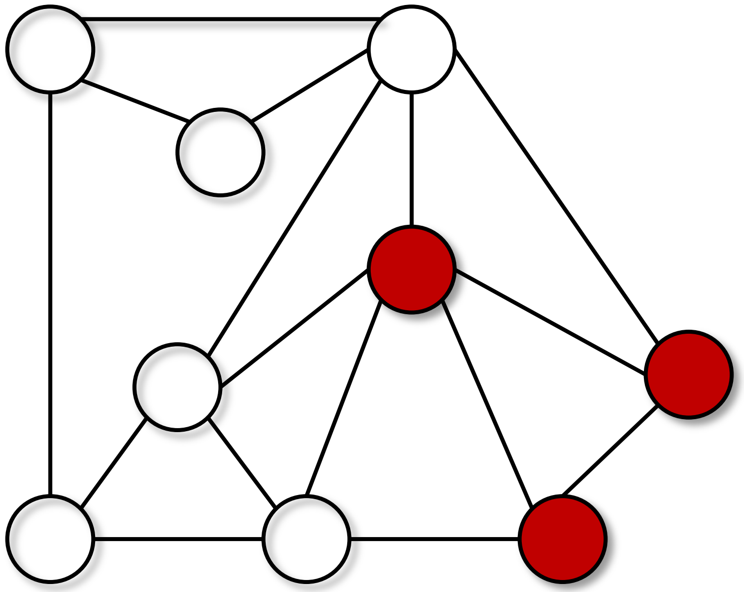


基于统计假设检验的测试

- 设在总体人群中有p的比例是女性，q = 1 – p比例是男性
- 若无同质性，则性别与边的连接应该“无关”
- 如何建模“无关”？
  - 考虑将每个点按照p概率随机分配女性，q分配男性
  - 如果“无关”，那么在我们的图中跨性别边的比例应该与上述随机情况类似

- 对于随机分配，有多少比例的边是跨性别的?
  - 考虑任意一条边，两个端点分配不同性别的概率是2pq
- 同质性测试：如果跨性别边占比显著低于2pq，则有性别同质性





几个问题

- 如何理解“显著”低于?
- 如果高于2pq呢?
- 这里只有两种“颜色”；多种呢?

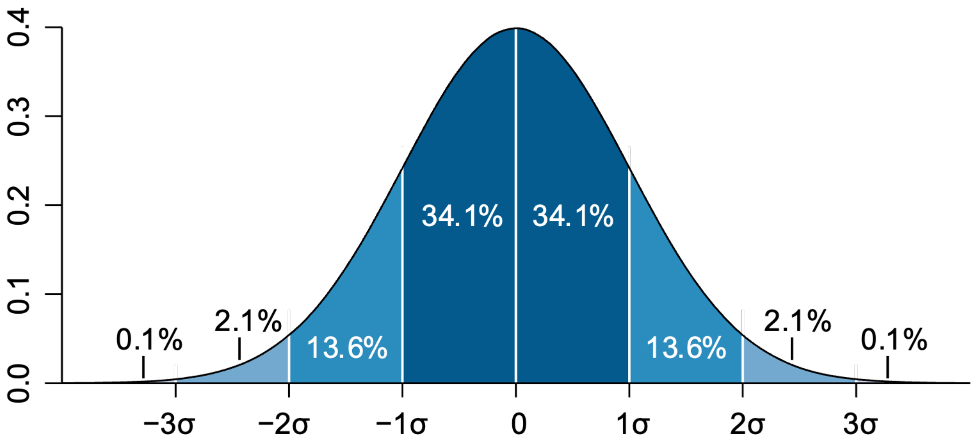
显著性的讨论

- 显著性：验证X是否与E[X]“足够”接近

找到t使得 $\Pr[|X - E[X]| > t] \leq 0.01$

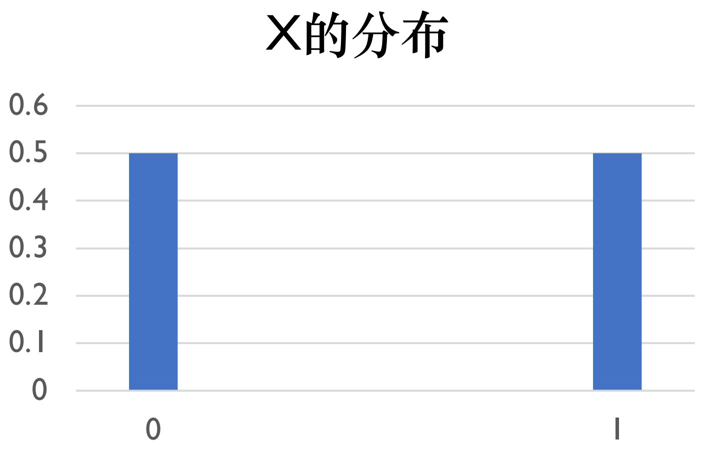
  - 如果实际量小于t，那么就“有信心”认为实际量并不服从X的分布
- 常见方法是利用中心极限定理，将X近似看作正态分布，采用3σ定则
- 在我们的应用中，X是随机安排下的异色边占所有边比例

我们这里可以利用中心极限定理吗?



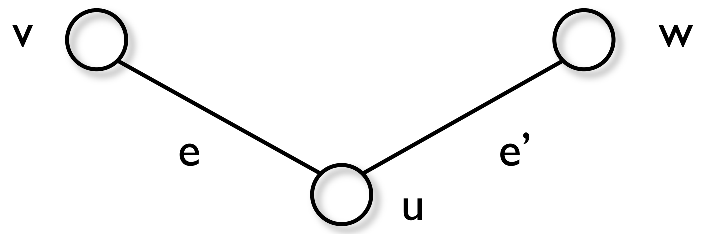
为何中心极限定理不适用?

- 中心极限定理：考虑n个独立同分布的{0, 1}随机变量，设他们的均值是X，那么当n趋向于无穷时，|X – E[X]|趋向于0的概率等于1
- 重点：“独立”
  - 如果不独立会怎样?
  - 考虑n个随机变量，他们永远相等，且每个的分布都是0.5概率从{0,1}取
  - 此时，E[X] = 0.5，那么|X – E[X]|永远等于0.5，不会趋向于0



为何中心极限定理不适用?

- 我们关心的是异色边的情况，而两条边是否异色并不是独立的
  - 具体：我们对每条边e有{0, 1}随机变量X_e表示e是否为异色边，关心均值X
  - 考虑两个边有重合端点的情况：e = u-v, e’ = u-w那么两个对应随机变量不独立



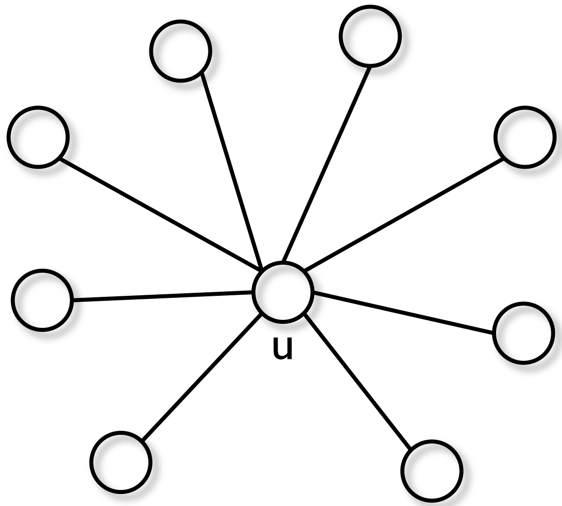
如果独立的话：

$$\Pr[X_e = 1 \mid X_{e'} = 0] = \Pr[X_e = 1]$$

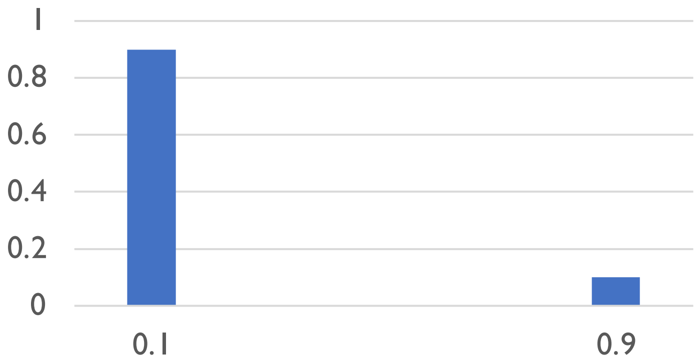
$$\Pr[X_e = 1 \mid X_{e'} = 0] = \frac{\Pr[X_e = 1 \cap X_{e'} = 0]}{\Pr[X_{e'} = 0]} = \frac{p^2 q + q^2 p}{p^2 + q^2} = \frac{pq}{p^2 + q^2} \neq \Pr[X_e = 1]$$

具体“反例”

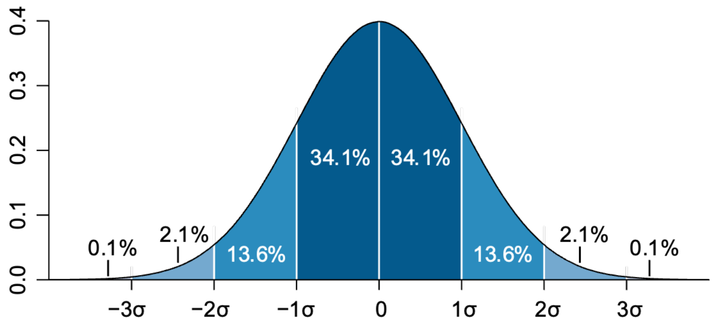
- 考虑星型图
  - u取红色时 (概率p)，每条边异色概率是q
  - u取蓝色时 (概率q)，每条边异色概率是p
  - 假定p > q (如p = 0.9, q = 0.1)，那么大概率 (p=0.9)看到接近q=0.1比例的异色边
  - 与典型值/期望值2pq = 0.18差大约2倍



异色边数概率分布**两极化**



较“好”的概率分布：**集中化**

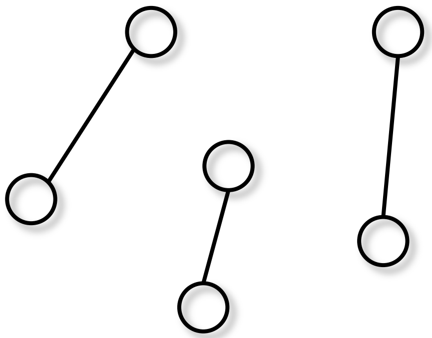


实际情况呢？

- 反例中u的度是O(n)的，这在实际中是不可能的
  - 一般来说，一个人只会有少数朋友（相对n来说，一般是o(n)甚至常数的）
  - 设每个节点的度不超过某个较小的d

若一个图中无孤立点且每个点的度不超过d，则存在一个大小为$\Omega(\frac{n}{d})$的匹配

- 匹配是一个任何点的度至多是1的图：边之间是独立的！



在匹配中的边没有交点，所以他们是独立的
可以用中心极限定理！
因此这些边中的异色边的比例大概率是2pq的

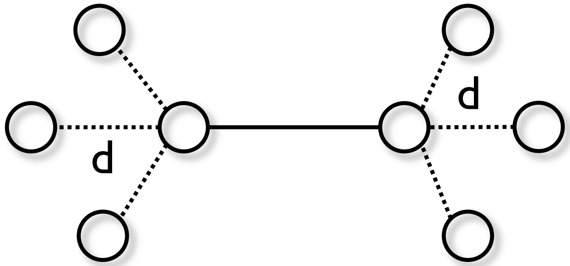
如何将边集剖分成“独立”的匹配

若一个图中无孤立点且每个点的度不超过d，则存在一个大小为$\Omega(\frac{n}{d})$的匹配

- 找到一个匹配后，将该匹配的边集记为E_1
- 将E_1删除，继续找匹配
- 因为删边只会让度更小，所以依然可以找到很大$\Omega(n)$的匹配
- 令此匹配的边集为E_2
- 重复上述过程，每次删除E_i然后找匹配E_{i+1}，直到不能继续进行下去
- 在每个E_i中都大概率有2pq比例的异色边，故总异色边比例大概率2pq

证明匹配定理

若一个图中无孤立点且每个点的度不超过d，则存在一个大小为$\Omega(\frac{n}{d})$的匹配



加1条边至多删除2d条边

- 将任意一条边u-v加入匹配，然后将与u和v相邻的边删除
- 继续加入任一剩余边，重复上述过程，直到不能找到任何新边
- 每加一条边，都删除至多2d条边；而总共有至少n / 2条边（无孤立点）
- 因此，加边可以进行$\Omega(\frac{n}{d})$次

“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)

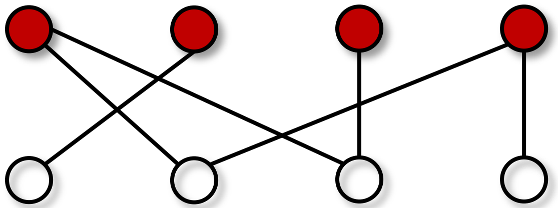
设$X_1, \dots, X_n$是独立的$[0,1]$上的随机变量。令$X = \frac{1}{n}\sum_{i=1}^{n} X_i$，$\mu = E[X]$。那么

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

- 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！
- 对于我们：匹配大小是log n就能让误差以1 – 1 /n概率<1%

回顾：几个问题

- 同质性测试：如果跨性别边占比显著低于2pq，则有性别同质性
  - 如何理解"显著"低于？
    - Concentration不等式
  - 如果高于2pq呢？
    - "反"同质性/异质性



    - 同样可以用concentration不等式来检测显著性
  - 这里只有2种"颜色"；多种颜色的同质性检测呢？
    - 在多种颜色下计算异色边比例期望，依然与此进行对比

将外部因素引入网络

- 在社会网络之上，增加一个新的网络来描述人与外部因素的关系
- 用“社团”依次代表一种抽象的同质性外部因素
  - 我们要将人与社团的关系进行建模

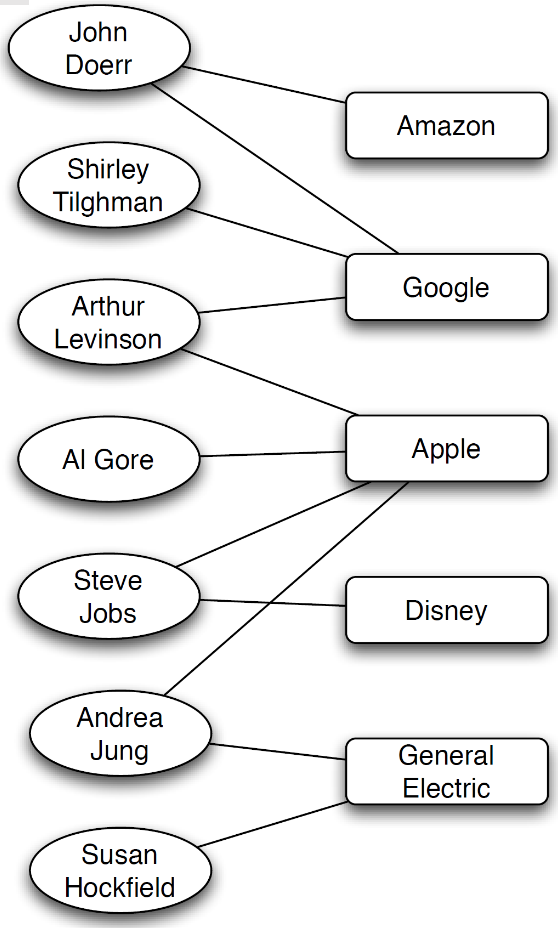
归属网络 (Breiger, 1974)

英文术语“foci”(Scott Feld, 1981)

- 归属网络：一个二分图，左边是人，右边是社团
- 在二分图中：
  - 节点分为左右两个集合L和R
  - 边仅出现在L和R之间

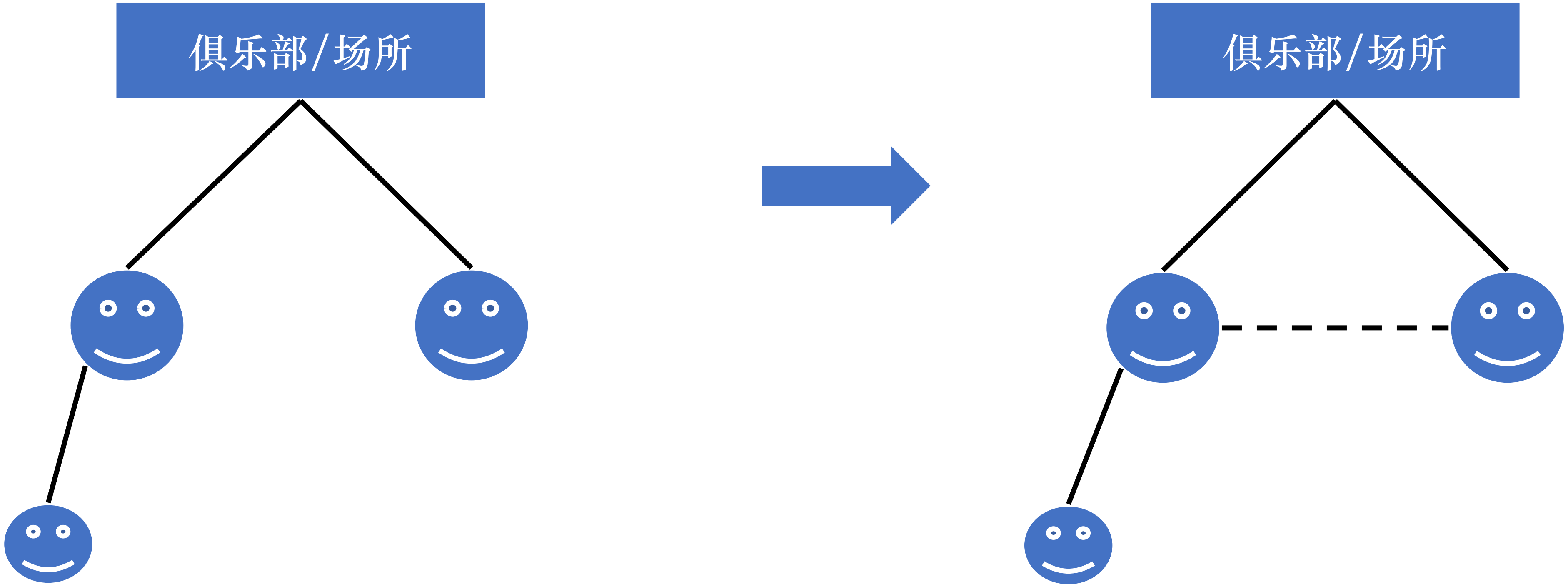
“社团”是一个抽象的“同质性”的代名词可以指代任何同质性特征

- 相应的，归属网络也有类似三元闭包的性质



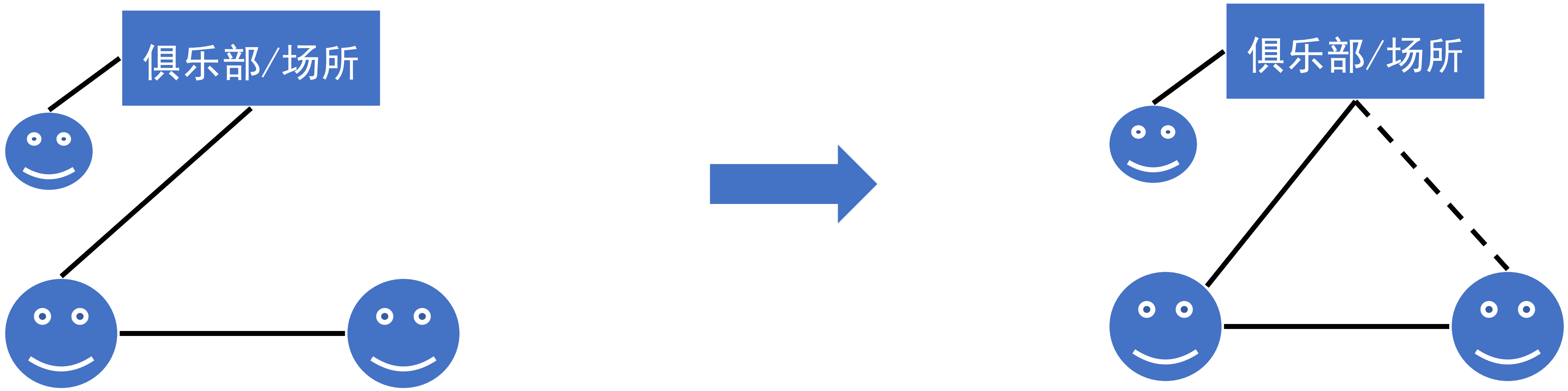
社团闭包（social focal clousure)

两个人不认识，但是通过共同兴趣认识，反映了“选择”



会员闭包（membership closure）

两个认识的人，经过一个介绍，另一个参加了俱乐部，体现了“影响”



对同质性内在机理的解释

古老智慧

物以类聚
人以群分

近朱者赤
近墨者黑

现代概念

选择
(selection)

社会影响
(influence)

抽象：
人、场合及其关系

社会归属
网

让我们能够讨论网络在一定环境下的演化过程

社团闭包

会员闭包

选择与影响的共同作用：实证研究

- “选择”和“社会影响”是同质性的内在产生动力
- 在Wikipedia数据上的实验
  - 每个编辑过wiki的用户是一个节点
  - 在某个时间点，用户A和B的相似度定义为

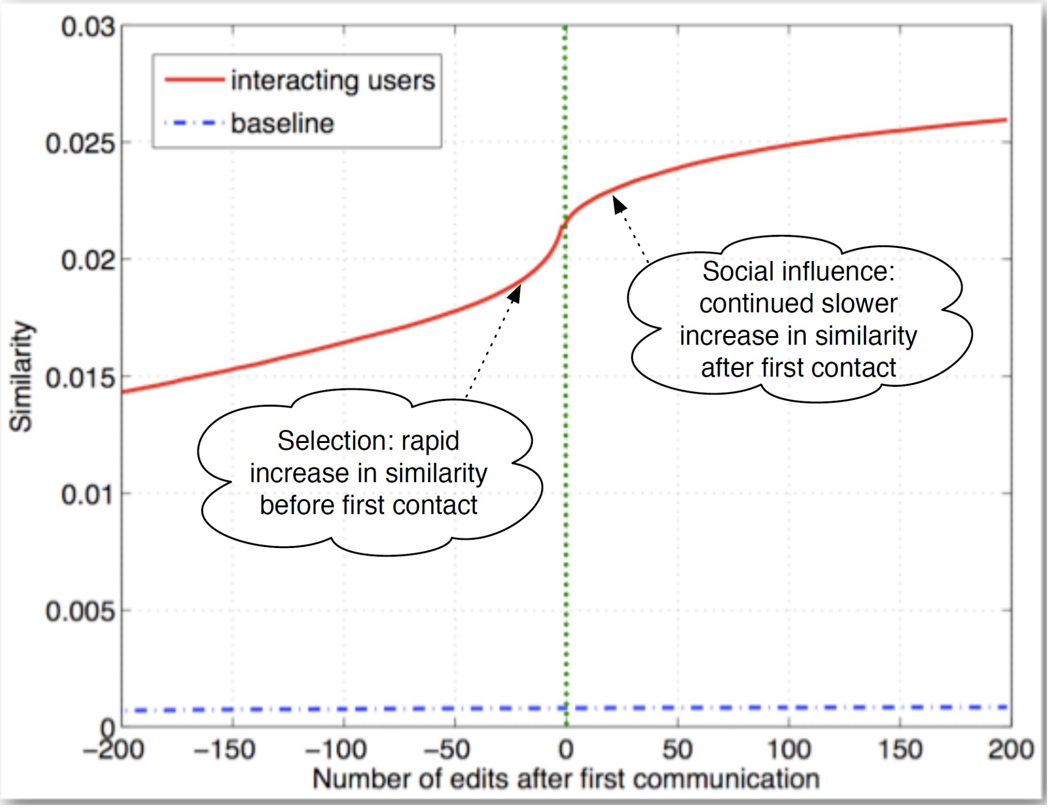
$$\frac{\text{用户A和B都编辑过的文章数}}{\text{至少用户A或B之一编辑过的文章数}}$$

  - Jaccard similarity coefficient/index (Jaccard, 1912) 是一种常用的衡量两个集合S和T相似度的度量:

$$J(S,T) = \frac{|S \cap T|}{|S \cup T|}$$

相似度变化曲线

- 对于每对用户u,v，定义一个曲线，横轴是时间，纵轴是他们的相似度
  - 这里，“时间”是一个离散的序列，u和v任何一人编辑过网页就算一个时间点
  - 将u和v第一次通信的时间定义为0时间（所以会有负时间）



MinHash计算Jaccard similarity

- 利用哈希的思想计算Jaccard similarity
  - 每个子集S定义一个哈希值
  - 希望仅看哈希值，而不需要求集合交并
- 设全集是U，h将U随机均匀映射到[0, 1]
- 对于U的子集S，令$h_{\min}(S)$为h在S元素上的最小值
- 则：对于任何$S, T \subseteq U$,

$$\Pr[h_{\min}(S) = h_{\min}(T)] = J(S, T) = \frac{|S \cap T|}{|S \cup T|}$$

基于MinHash的估计量

1. 构造m个独立的MinHash $h^{(1)}, \dots, h^{(m)}$
2. 对于给定的S, T，令$Z = \frac{\left|\left\{i : h_{\min}^{(i)}(S) = h_{\min}^{(i)}(T)\right\}\right|}{m}$ （也就是有多少比例的hash function上的min相等）

- $E[Z] = J(S, T)$
- 用Chernoff bound，取$\mathrm{m} = O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$，误差可达到$\epsilon$，失败概率$\delta$
- 计算上的好处：预处理每个集合的$h_{\min}$，以后查询$J(S, T)$都是$O(1)$

亚线性时间估算—对Jaccard similarity

- 回顾定义：$J(S,T)=\frac{|S\cap T|}{|S\cup T|}$
- 如果S和T都是m元集合，那么利用哈希表可以O(m)时间来计算
- 亚线性o(m)时间的估计?
- 注意到$|S\cup T|=|S|+|T|-|S\cap T|$，所以重点在于估计$|S\cap T|$

估计$|S \cap T|$

- Idea：用大小很小的S',T'代替原始S,T来进行求交计算
- 考虑如下随机过程:

1. 固定某个p，对每个S和T中的元素，独立以p的概率保留，得到S', T'
2. 返回$Z = |S' \cap T'|/p^2$作为估计

注意：这个算法在有随机访问下运行时间$O(|S'| + |T'|)$

- 数学期望分析：
  - 对每个$S \cap T$的元素i，定义X_i为i是否属于$S' \cap T'$的指示变量
  - 则$Z = \sum X_i / p^2$，且 $E[X_i] = \Pr[X_i = 1] = p^2$，故$E[Z] = |S \cap T|$

如何选取p?

Chernoff bound (Chernoff, 1952)

设$X_1, \dots, X_n$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

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

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

- 定义：对每个$S \cap T$的元素i，定义X_i为i是否属于$S' \cap T'$的指示变量
- $E[X_i]=p^2$，故$\mu=E[X]=p^2|S \cap T|$
- $\Pr[|X-\mu| \geq \epsilon\mu] \leq 2\exp(-\epsilon^2 p^2 |S \cap T|/3)$，令其$\leq \delta$
- $p^2=O\left(\frac{1}{\epsilon^2|S\cap T|} \cdot \ln\frac{1}{\delta}\right)$
- 算法复杂度$\mathrm{E}[|\mathrm{S}'|+|\mathrm{T}'|]=(|S|+|T|)p=O\left(\frac{|S|+|T|}{\sqrt{|S\cap T|}}\right)$
  - 若已知$J(S,T) \geq c$，那么上述复杂度$\approx O\left(\frac{\sqrt{|S\cap T|}}{c}\right)$

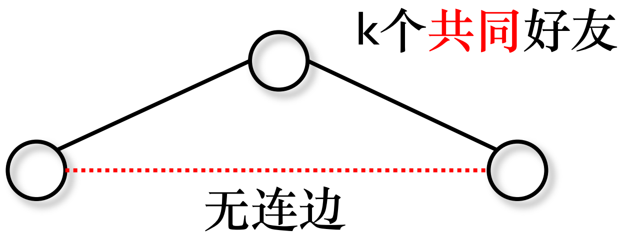
小论文：在实际数据中验证闭包性质

在实际数据中验证闭包性质

- 在较大规模真实数据上，验证三元闭包、社团闭包、会员闭包对网络演化的作用
- 数据集：SNAP http://snap.stanford.edu/data/index.html#temporal
- 以三元闭包为例
  - 问题：两个人成为朋友的几率，如何随着共同朋友的数量变化?

方法举例

- 方法举例：
  - 选取2次不同时间的网络快照
  - 对于每个k，在第一次快照中找出有多少个恰有k个公共朋友的未连边节点对



  - 定义T(k)为这些节点对第二次快照中形成边的比例 (亦即建立连接的几率)
  - 画出T(k)关于k的函数图像

我们可能看到什么？

- 简单模型：
  - 假定两人之间的一个共同朋友以某概率p在选定时间段内独立促成他们的连接
- 在选定时间段内，如果两人有k共同朋友，没有促成连接的概率至多是$(1-p)^k$
- 所以建立连接的概率至少是$1-(1-p)^k$

实际的T(k)图像能用这个模型拟合吗？如果不能，那么可能的原因是？

其他选项?

- Jaccard similarity coefficient/index (Jaccard, 1912) 是一种常用的衡量两个集合S和T相似度的度量:

$$J(S,T) = \frac{|S \cap T|}{|S \cup T|}$$

- 应该采用哪个?
- 考虑:
  - 计算可行性/复杂性? （采样？更多机器/核？近似算法？）
  - 辨析在所选应用、数据集上哪个更合适（有时需要做实验才能验证）
  - 有时需要两个都采用，尤其是各有利弊、各自说明了一些有趣事实的情况下

你的任务

- 从SNAP选择一个合适数据集，用上述方法验证三种闭包中的一种
- 详细描述
  - 实验目标（要验证什么？）、内容（所采用的实验设计？）
  - 数据集的选取依据，数据集主要特征，实验环境以及任何需要注意的假设
  - 用图表展示实验结果
  - 尝试解释实验结果（比如提出一个简化的理论模型），讨论你的发现
- 一些建议/工具
  - 用Latex；在线环境可以用Overleaf：www.overleaf.com
  - 按照科技文献写作的规范写一个自洽的小论文（注意引用等）
  - 画图可以使用Matplotlib（比excel的好处：自动化，与实验程序输出对接）