社交网络中的计算传播

阳德青

复旦大学大数据学院

Email：yangdeqing@fudan.edu.cn

引　言

- 社交网络信息传播的过程定义为：
  - 通过社交网络中的个体间的交互行为，将一条信息（知识、创新、产品等）传播并到达多数个体的过程

    大众传播 → 人际传播

- 传播过程的三要素
  - 传播者
    - 整个网络中通常是一个或一小部分人发起信息传播的过程
  - 接收者
    - 接收者数量通常远大于发布者数量
  - 传播媒介
    - 例如谣言是通过个体之间的口口相传(mouth to mouth)

目　录

- 计算传播学基本原理
- 社会影响力和同质性
- 信息传播模型
- 影响力最大化
- 热门话题分析与预测

计算传播学简介

- Computational Communication
  - 计算社会学的重要分支，与计算新闻学、网络科学等密切相关，是典型的交叉学科研究
  - 关注人类传播行为的可计算性基础　研究对象
  - 以网络分析、文本挖掘、数据采集、数学建模等为主要分析工具来大规模地收集并分析人类传播行为数据研究工具
  - 挖掘人类传播行为背后的模式和法则，分析模式背后的生成机制与基本原理 研究目标
  - 可被广泛地应用于数据新闻、舆情监测、计算广告等场景　应用场景

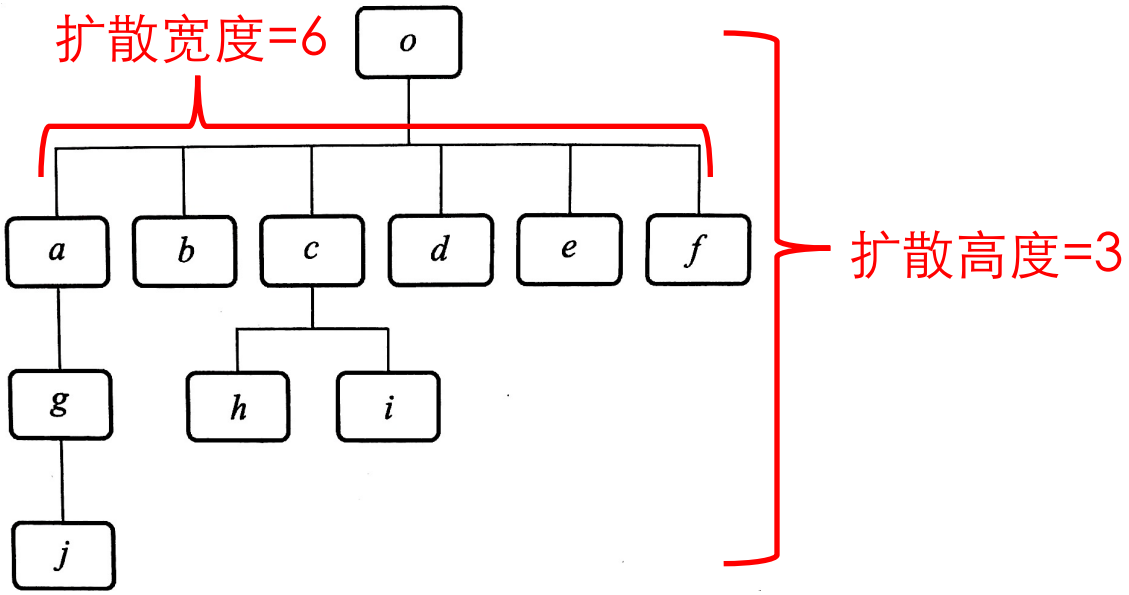
计算传播学简介

- 信息传播的测量维度
  - 扩散规模：信息扩散的数量（回帖量、转帖量、回帖/转帖人数等）
  - 扩散率/转发率：$D(i)/S(i)$

转发信息i的用户数

看到信息i的用户数

  - 扩散网络



计算传播学简介

- 信息传播的测量维度
  - 级联率(Cascade Ratio)： $N(u,i)/S(i)$

跟随用户u转发信息i的数量(用户数)

信息i的扩散规模(总转发量)

  - 扩散时间
  - 扩散速度
    - 单位统计时间内的扩散数量
    - 每一步扩散所消耗的时间
  - 爆发性
    - 峰值比率 PF=peak/S
    - 变异系数
    - 峰值时间：从信息出现到扩散峰值之间的时间
  - 持续性
    - 衰减时间：从峰值时间到信息扩散达到扩散规模的75%所需时间

计算传播学简介

- 信息传播的测量维度
  - 爆发性

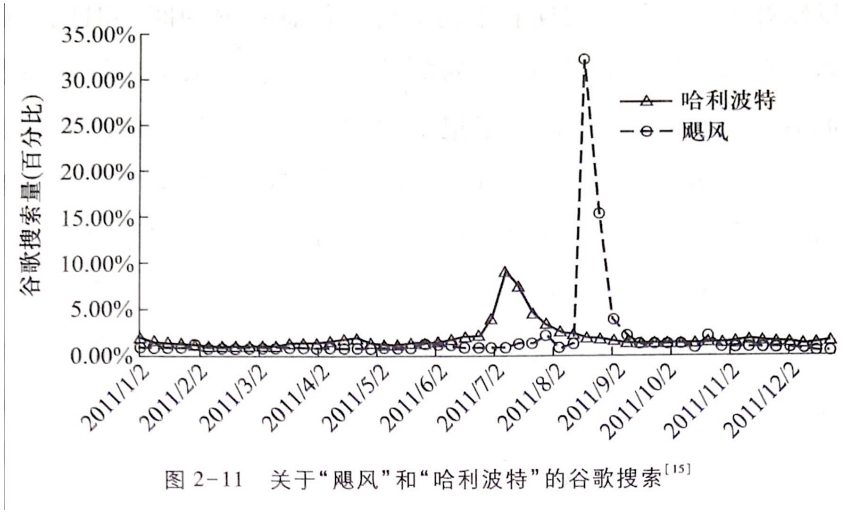
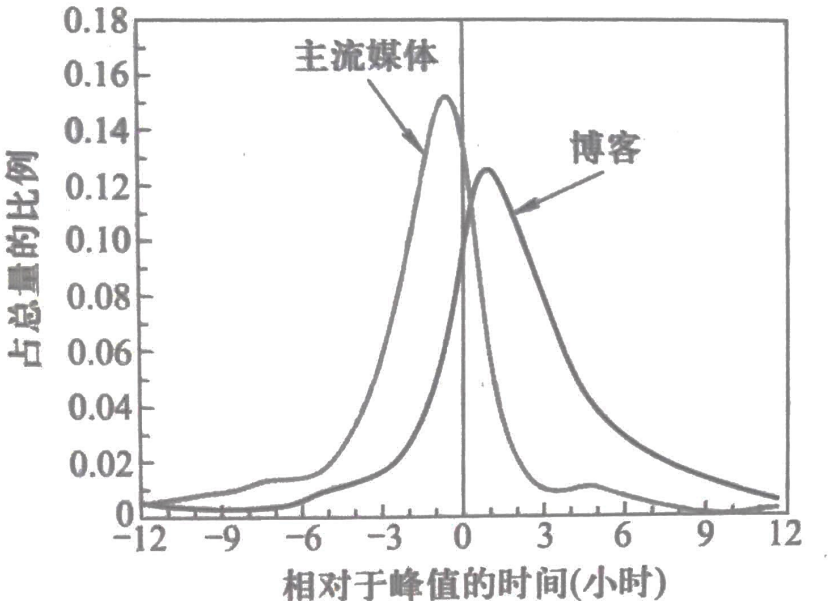


图 2-11　关于“飓风”和“哈利波特”的谷歌搜索[15]



统计发现：主流媒体和博客对某一话题的注意力峰值存在2.5小时的时间差

结点传播力测量

- 意见领袖在信息传播中的作用
  - Katz和Lazarsfield假设：信息从大众媒介先传递到意见领袖，由意见领袖再讲信息传递给他们所影响的人二级传播
- 社交用户影响力的量化方法
  - 激发回复：用户发布一条帖子后得到的回复数
  - 激发对话：用户发布一条帖子所激发的他人相互讨论
  - 语义扩散：用户帖子中使用的词语被他人沿用

结点传播力测量

- 基于结点度的测量
  - 无向图

    $C(v)=d$
  - 有向图

    $C(v)=d_{in}$ 声望

    $C(v)=d_{out}$ 合群性

    $C(v)=d_{in}+d_{out}$

    } 两者有何区别?

结点传播力测量

- 结点特征向量中心性
  - 核心思想：通过结合无向图中邻居结点（或有向图的入度邻居）的重要性来概括（某个结点的）度中心性
  - 特征向量中心性计算公式：

$$C_e(v_i) = \frac{1}{\lambda}\sum_{j=1}^{n} A_{j,i}\, C_e(v_j) \qquad (1)$$

其中，结点j是结点i的入度邻居，λ是固定常量， $\mathbf{A}$是邻接矩阵

  - 假设$\mathbf{C_e}$是所有结点中心性（值）的向量，则3.1式可表示为：

$$\lambda \boldsymbol{C}_e = \boldsymbol{A}^T \boldsymbol{C}_e$$

因此，$\mathbf{C_e}$是$\mathbf{A^T}$（无向图中，$\mathbf{A^T} = \mathbf{A}$）的特征向量， λ则是对应的特征值

  - 每个节点的中心值最好都>0，因此应寻找各维度均>0的特征向量
  - 通过Perron-Frobenius定理，可以通过求解$\mathbf{A}$的最大特征值对应的特征向量，即得到图中所有结点的特征向量中心性

结点传播力测量

- Katz中心性
  - 核心思想：由3.1式可知，有向图中的入度邻居可以贡献中心性（值），而没有出边的结点则无法传递中心性，为避免没有入度邻居的结点中心性为0，加入一个偏差项β，使得每个结点的中心性值都至少含有β
  - Katz中心性计算公式

$$C_k(v_i) = \alpha \sum_{j=1}^{n} \boldsymbol{A}_{j,i}\, C_k(v_j) + \beta \qquad (2)$$

  - 假设$\mathbf{C}_k$是所有结点Katz中心性（值）的向量，则(2)式可表示为：

$$\mathbf{C}_k = \alpha \mathbf{A}^{\mathrm{T}} \mathbf{C}_k + \beta \mathbf{1}$$

或

$$\mathbf{C}_k = \beta (\boldsymbol{I} - \alpha \mathbf{A}^{\mathrm{T}})^{-1} \cdot \mathbf{1} \qquad (3)$$

  - 当α=0，所有结点的中心性值为β；
  - 当$\alpha = \frac{1}{\lambda}$时（λ是$\mathbf{A}^{\mathrm{T}}$最大特征值），根据特征值定义，有$\det(\boldsymbol{I} - \alpha \mathbf{A}^{\mathrm{T}})=0$，使得$\boldsymbol{I} - \alpha \mathbf{A}^{\mathrm{T}}$不可逆，中心性值出现偏差，因此一般取$\alpha < \frac{1}{\lambda}$

结点传播力测量

- PageRank中心性
  - 核心思想：参照之前中心性计算方法，有向图中，一个中心性很高的结点（权威结点）如有很多出边，则其较高的中心性都会传递给他的出度邻居，与实际情况不太相符（例如名人的朋友并不都是名人）。因此，入度邻居贡献过来的中心性应当除以该邻居的出度，于是有：
  - PageRank中心性计算公式

Google的制胜之道

$$C_p(v_i) = \alpha \sum_{j=1}^{n} \boldsymbol{A}_{j,i} \frac{C_p(v_j)}{d_o} + \beta \qquad (4)$$

其中，d_o是结点 v_j的出度

  - 同样可以将(4)式可表示为：

$$\boldsymbol{C}_p = \alpha \boldsymbol{A}^T \boldsymbol{D}^{-1} \boldsymbol{C}_p + \beta \mathbf{1}$$

或

$$\boldsymbol{C}_p = \beta (\boldsymbol{I} - \alpha \boldsymbol{A}^T \boldsymbol{D}^{-1})^{-1} \cdot \mathbf{1}$$

其中，$\boldsymbol{D}$是度对角矩阵

  - 类似Katz中心性，当λ是$\boldsymbol{A}^T \boldsymbol{D}^{-1}$ 的最大特征值时，一般取$\alpha < \frac{1}{\lambda}$

结点传播力测量

- 中间/中介中心性
  - 核心思想：体现结点在连接其他结点时的重要性，因此可以计算其他结点间的最短路径中通过该结点的路径数目
  - 计算公式：

$$C_b(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}} \qquad (5)$$

信息传递路径

其中，σ_{st}是结点s到t的最短路径数，而$\sigma_{st}(i)$是经过结点i的最短路径数

  - 常常需要对中间中心性进行归一化处理，除以最大值即可（任意结点的中间中心性最大值是多少？）

$$2\binom{n-1}{2} = (n-1)(n-2)$$

  - 无向图中，由于s到t和t到s可视为不同的路径，所以(5)式的计算要x2
  - 社会学研究意义
    - 中间中心性高的结点是社交网络信息传播的关键人物

结点传播力测量

- HITS中心性
  - HITS算法由John Kleinberg于1997年提出
  - 核心思想
    - Web网络中的各页面都有权威值（authority，用a(i)表示）和中枢值（hub，用h(i)表示）
    - 权威网页：与某个主题相关的高质量网页，入链数量（入度）很大，如谷歌和百度的首页
    - 中枢网页：指向很多高质量（权威）网页的网页，如hao123
    - a(i)和h(i)是相互增强关系
  - 权威值和中枢值的迭代计算
    - 每个结点的a(i)和h(i)初始都为1
    - $a(i) = \sum_{j=1}^{n} h(j)$, j是i的入度邻居
    - $h(i) = \sum_{j=1}^{n} a(j)$, j是i的出度邻居
    - a(i)和h(i)归一化后再重复迭代计算，直至收敛

搜索引擎应当尽量返回
符合查询条件的权威网页

结点传播力测量

- 群体度中心性
  - 群体外部的结点连接到群体内部结点的数目
- 群体中间中心性

$$C_b^g(S) = \sum_{s \neq t, s \notin S, t \notin S} \frac{\sigma_{st}(S)}{\sigma_{st}}$$

$\sigma_{st}(S)$是结点s到t经过集合S中结点的最短路径数目

- 群体接近中心性

$$C_c^g(S) = 1 / \left(\frac{1}{|V-S|} \sum_{i \notin S} l_{S,i} \right)$$

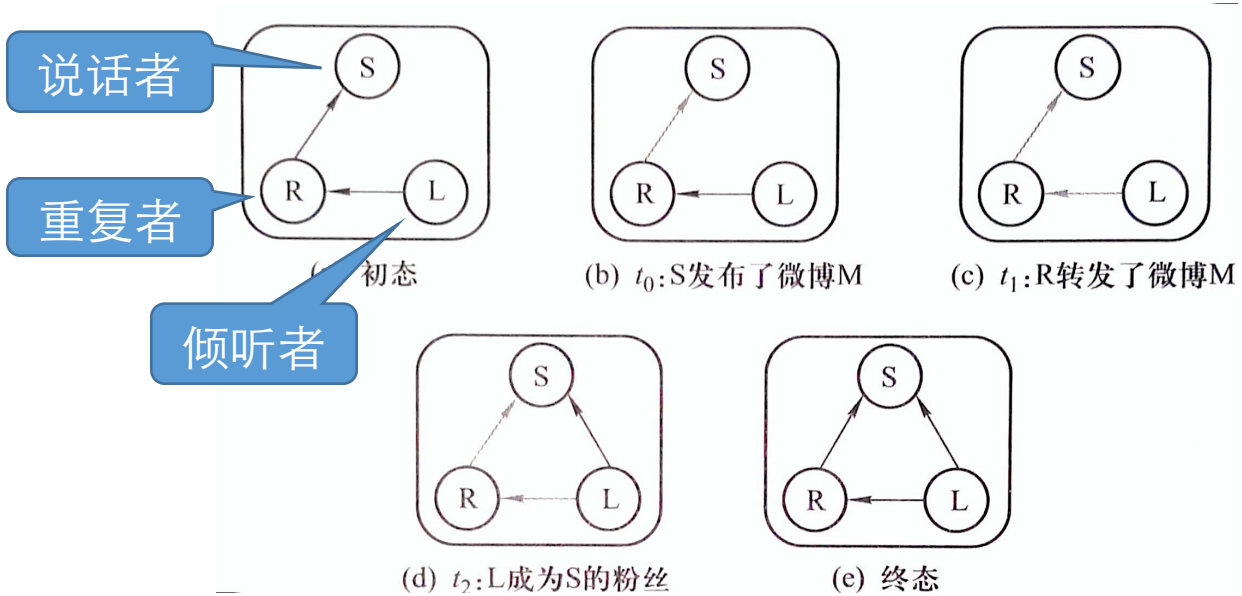
$l_{S,i}$是群体S到S外结点i的最短路径长度，可以在S中寻找距离i最近或最远的结点，用这两点间的最短路径长度来表示，也可以计算S中所有结点到i的最短路径长度，取平均值来表示$l_{S,i}$

信息传播对于社交网络结构演化的作用

- 网络上的动力学
  - 假定网络结构不变的情况下结点或连边的状态变化
- 网络动力学
  - 网络拓扑结构的变化

两种动力往往相互作用和影响

- 信息传播往往会导致社交网络结构向良性的方向发展



“微博-转发-追随”事件TRF体现了传递性(三元闭包)原理，对增加社交网络的密度具有重要影响

目　录

- 计算传播学原理
- 社会影响与同质性
- 信息传播模型
- 影响力最大化
- 热门话题分析与预测

社会影响

- “影响力是在没有明显的强制措施和直接命令的情况下影响他人的行为或能力”——韦氏字典
- 什么是社会影响力?
  - 产生于社会媒体中的影响力，很大程度上表现为社会好友（邻居）对个体的影响
  - 动因：人们因社交压力（需要与自己的朋友保持一致）而改变自己的行为，即一个人更容易受自己好友的影响做相似/相同的动作

社会影响力度量

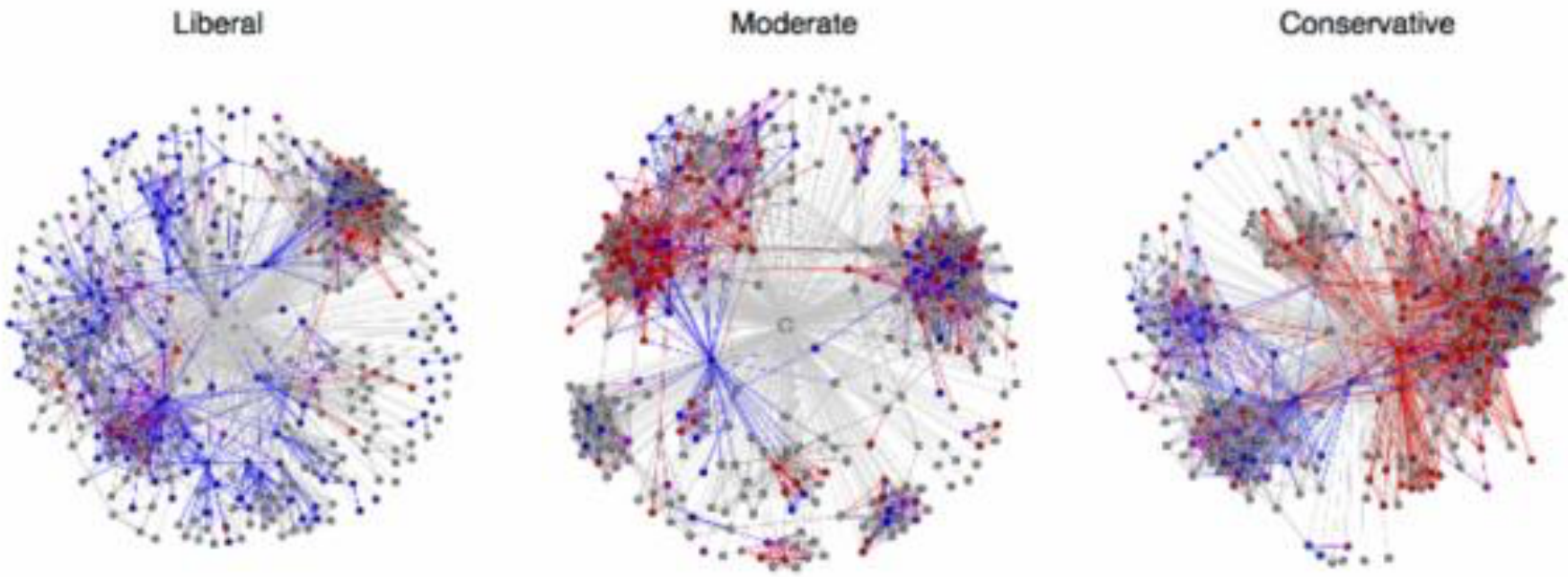
- 基于预测的方法
  - 通过度量网络中结点的中心性值来预测该结点的影响力
- 基于观察的方法
  - 不同的场景中有不同的度量方法
    - 将个体作为榜样：受榜样的个人魅力、学识等属性影响的观众人数来量化影响力，如明星、学者、领导人等
    - 个体是被传播的对象：传播的级数、受众数量/比例等来量化影响力，如一条新闻、流行病、产品等
    - 个体的参与提升了某事物或行为的价值：购买了某商品的用户会提升该商品的价值（使其更容易被其他人接受），所以用价值的提升量/提升率来量化影响力，如第一个购买传真机的人因为没有发送传真的对象而显得价值不大，而第二个购买传真机的人则提升了传真机的价值

同质性

- 什么是同质性（homophily）？
  - 人和自己朋友之间往往具有相似性的特点
  - 具有相似特点的人更容易成为朋友（建立连接）
- 思想来源
  - “相似性带来友谊”——柏拉图
  - “人们喜欢与自己相似的人”——亚里士多德
  - “物以类聚、人以群分”
- 现实案例
  - 1、两个人经过共同的朋友介绍而相互认识
  - 2、两个在同一所学校或者就职于同一家公司的人成为了朋友

同质性

- 典型的同质性社交网络连接图



选择与社会影响

- 同质性现象的背后机制
  - 选择：人们倾向于与其相似的人形成友谊，如同族群内
    - 个体特征主导社会网络连接的形成
  - 社会影响：人们需要与自己的朋友保持一致而改变自己的行为，即一个人更容易受自己好友的影响做相似的动作
    - 已存在的社会网络连接将会改变个体(可变)的特征

两者相反

- 同质性是同配网络形成的主要原因
  - 同配(assortative)网络中，相似个体比不相似个体更容易形成连接

选择与社会影响

- 研究案例
  - 为研究人类行为的动因，Robert B. Cairns针对青少年吸毒问题做了调查研究，想弄清到底是选择因素还是社会影响因素对青少年吸毒行为更大些，从而找到更有效的戒毒方法
  - 如果主要是因为社会影响造成的，则可以选定一个吸毒特定学生群体，通过强制他们停止吸毒来使这些学生的朋友也停止吸毒
  - 但如果主要是因为选择造成的，则以上的方法还可行么?

选择与社会影响同样也对社会网络中的信息传播产生影响

选择与社会影响的大数据实证

- 实证案例：维基页面编辑(人)之间的相似行为
  - 定义编辑之间的相似性：

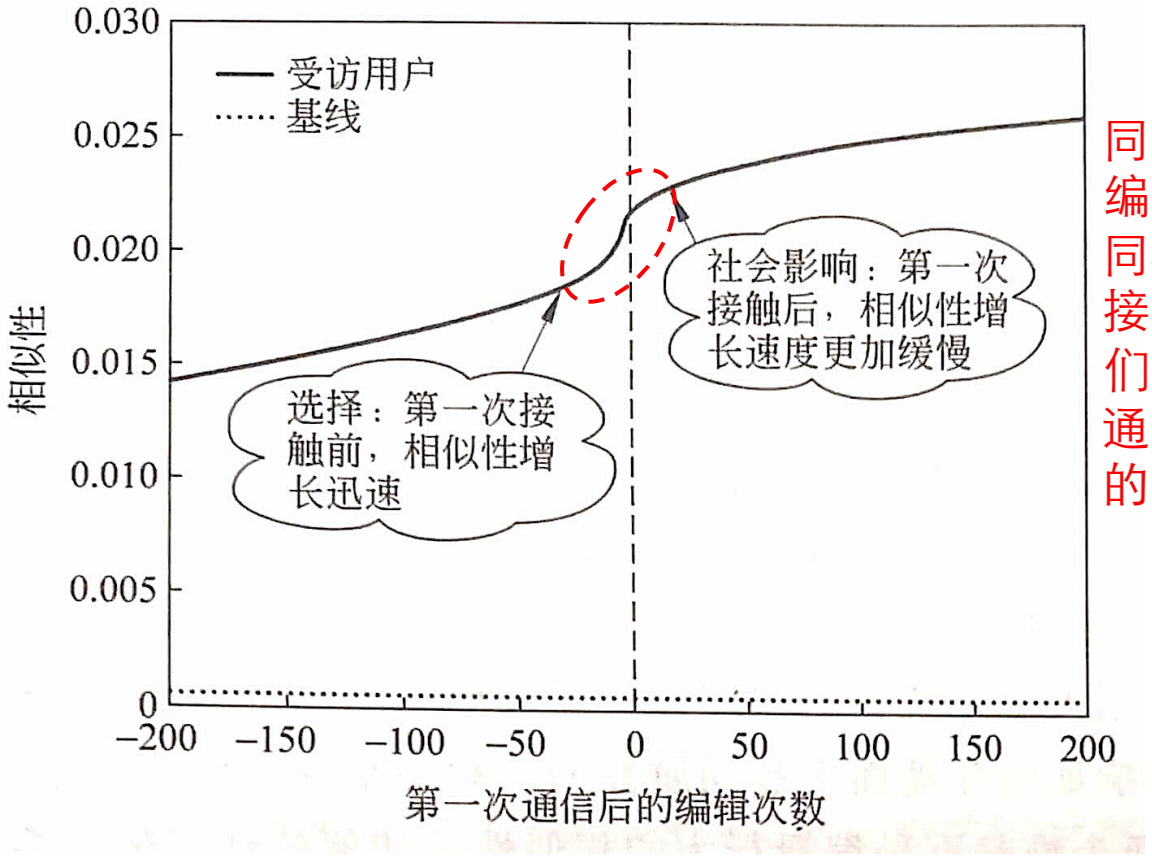
$$sim(i,j) = \frac{D(i) \cap D(j)}{D(i) \cup D(j)}$$

$D(i)$ ∪是编辑i编辑过的页面集合

  - 定义建立社会连接：编辑之间产生通信联系
  - 结果表现形式：记录相似性随时间变化的情况

选择与社会影响的大数据实证

- 实证案例：维基页面编辑(人)之间的相似行为



同质性的表现是因为编辑们倾向与编辑过同样页面的人形成连接(选择)，还是因为他们被引导去编辑那些通信过的人所编辑过的页面(社会影响)?

目　录

- 计算传播学原理
- 社会影响力与同质性
- 信息传播模型
- 影响力最大化
- 热门话题分析与预测

社会网络的信息传播模型

- 社会网络的信息传播模型：
  - 羊群效应（Herd Behavior）
  - 信息级联（Information Cascade）
  - 线性阈值模型（Linear Threshold Model）
  - 创新扩散（Innovation Diffusion）
  - 流行病/传染病模型（Epidemic Model）

羊群效应

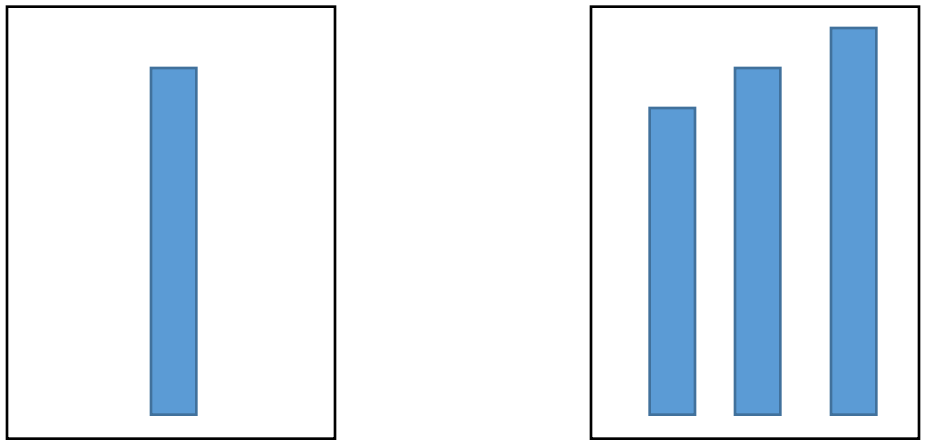
- 定义
  - 个体观察所有其他人行为后，采取与其一致的行为效应

随大流/跟风

- 历史
  - 最早由英国外科医生威尔弗雷德·特罗特于1916年提出，该现象不仅在羊群、牛群动物的集体行为中常见，在体育赛事、示威游行、宗教集会等人类活动中也普遍观察到
- 传播实验条件：
  - 个体的选择决定需要遵从一定的顺序
  - 个体是在掌握一定的信息后才做出决定
  - 个体之间无法传递信息
    - 一个人无法知道他人掌握的信息，但是可以通过观察他人行为来推断他人掌握的信息

羊群效应

- 著名的羊群效应实验
  - 食客对餐馆的选择
  - Solomon Asch (1956)实验



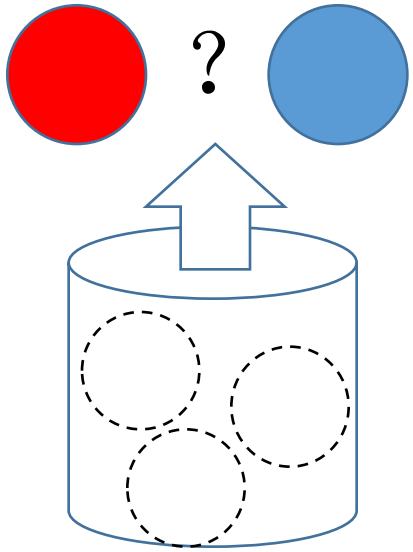
右图中的哪一条和左图中等高?

当周围人都给出正确答案时，只有3%的参与者回答错误；
当周围有人给出错误答案时，32%的参与者会回答错误

羊群效应的贝叶斯建模

- 取球猜色实验（Anderson&Holt1996）
  - 每个容器中有三个球，只有红和蓝两种颜色每个学生走上台从容器中取出一个球，再猜测容器中是红球还是蓝球多，并将猜测结果写在黑板上，其他学生只能看到黑板却看不到他取出球的颜色，然后其他学生再一个个上来做同样的动作。
- 实验的基本结论
  - 只要前两个学生猜测的颜色一致，从第三个学生开始，无论他取出球的颜色是红是蓝，都会猜前两人颜色的球占多数，于是羊群效应产生。

为什么？



羊群效应的贝叶斯建模

- 贝叶斯建模
  - 假设B代表蓝球多，R代表红球多，o=B代表（观察到）取出的是蓝球，o=R代表取出的是红球
  - 根据实验条件，有

P(B)=P(R)=1/2, P(O=B|B)=P(O=R|R)=2/3

  - 根据贝叶斯规则，有

$$P(B|O=B)=\frac{P(O=B|B)P(B)}{P(O=B)}$$

且 $P(O=B)=P(O=B|B)P(B)+P(O=B|R)P(R)=\frac{2}{3}*\frac{1}{2}+\frac{1}{3}*\frac{1}{2}=\frac{1}{2}$

  - 因此，$P(B|O=B)=\frac{\frac{2}{3}*\frac{1}{2}}{\frac{1}{2}}=\frac{2}{3}$

所以，第一个学生取出是蓝球则猜蓝球多，取出是红球则猜红球多。

羊群效应的贝叶斯建模

- 贝叶斯建模（续）
  - 如果第一个学生猜蓝球多，第二个学生取出的是蓝球，则猜蓝球多。此时，若第三个学生取出的是红球，则根据贝叶斯规则，有

$$P(B|O=\{B,B,R\})=\frac{P(O=\{B,B,R\}|B)P(B)}{P(O=\{B,B,R\})}$$

且 $P(O=\{B,B,R\}|B)=\frac{2}{3}*\frac{2}{3}*\frac{1}{3}=\frac{4}{27}$

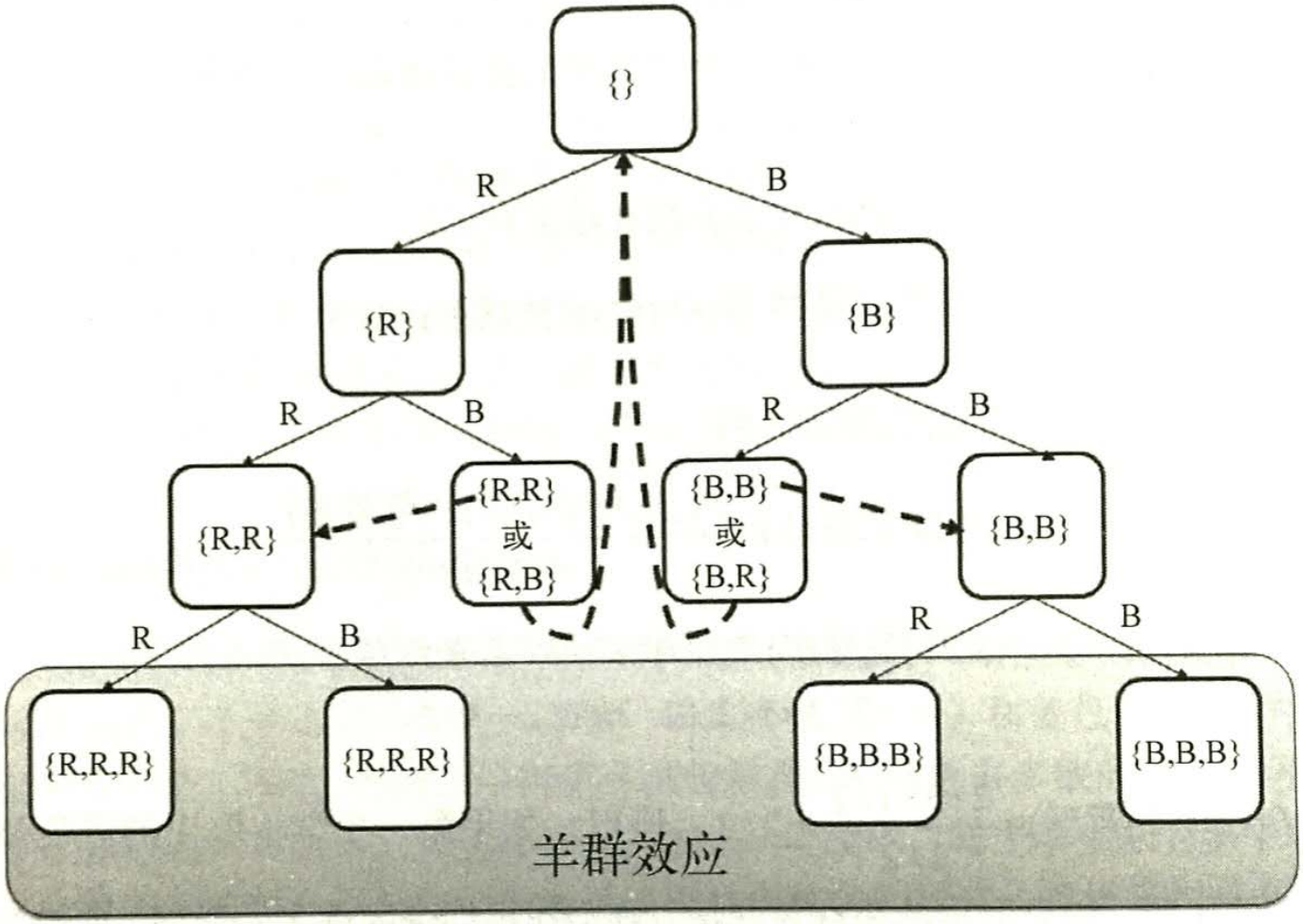
$$P(O=\{B,B,R\})=P(O=\{B,B,R\}|B)P(B)+P(O=\{B,B,R\}|R)P(R)$$

$$=\frac{2}{3}*\frac{2}{3}*\frac{1}{2}+\frac{1}{3}*\frac{1}{3}*\frac{1}{2}=\frac{1}{9}$$

因此，$P(B|O=\{B,B,R\})=\frac{\frac{4}{27}+\frac{1}{2}}{\frac{1}{9}}=2/3$，根据此计算结果，第三个即便取出红球，也会猜测蓝球多。

  - 此后的学生无论取出什么颜色的球，由于根据黑板上列出的前序猜测值（可能与实际值不符合），蓝球多的概率始终大于1/2，所以都会猜测是蓝球多。于是，羊群效应产生。

羊群效应的贝叶斯建模



容器实验。矩形框表示学生写在黑板上的预测值，箭头上方的值表示学生的观察值，矩形框里是根据条件概率计算得到最可能的情况

羊群效应

- 对羊群效应的干预手段
  - 羊群效应使得群体中的个体随时间推移达到一致性共识/选择，要停止这种效应扩散就需向要做出选择的个体提供更多的信息
    - 例如取球猜色实验中，让每个人再把自己取出球的实际颜色写下来

信息级联

- 定义
  - 指信息或决策在一群个体中扩散的过程，例如社会网络中，个体经常转发（或采取相同决策）邻近好友的发布内容
- 传播实验的条件
  - 个体存在于一个社会网络中
  - 个体只能观察到邻近好友的决策行为

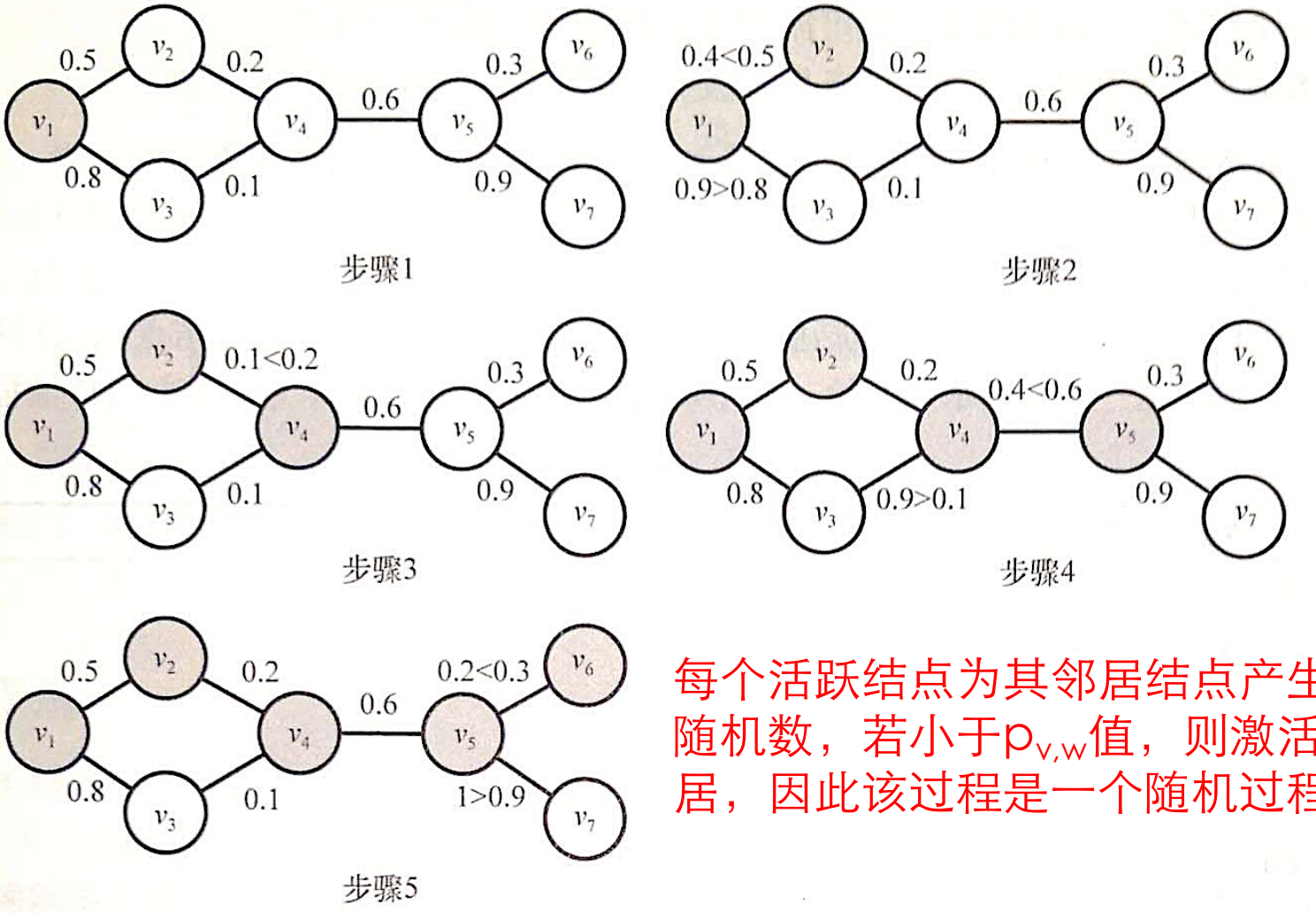
用户可获得的信息比羊群效应实验更少

体现了“社会影响”的因素

独立级联模型（ICM）

- 由Kempe等人在2003年提出，模型假设如下：
  - 网络图中的结点表示行为的执行者，边表示邻居关系
  - 每个结点只有活跃或不活跃两种状态，被激活的结点表示该结点采纳了某种行为/创新；活跃结点可视为信息发布者，被激活结点可视为接收者
  - 一个结点在t时刻被激活后便可以在t+1时刻激活它的一个邻居结点，激活概率为$p_{v,w}$
  - 激活是一个渐进过程，结点只能从不活跃跳转到活跃状态，反之不行

独立级联模型（ICM）



每个活跃结点为其邻居结点产生一个随机数，若小于$p_{v,w}$值，则激活该邻居，因此该过程是一个随机过程

ICM模拟。边上的数字代表$p_{v,w}$。当存在不等式时，激活条件满足。不等式的左侧是产生的随机数，右侧是$p_{v,w}$

级联范围最大化

- 目标：初始时用最少的激活结点达到最终时激活最多的结点
- 应用价值：产品营销，用最低的广告代价/预算（选择最少的初始客户）达到最大的广告效果（最终购买产品的客户最多）
- 算法
  - 形式化定义：S表示初始激活结点集合，f(S)表示最终能激活的结点集合，则算法目标是，对于给定的预算k，找到合适的S，使得|S|=k，|f(S)|最大
  - 由于f(S)是次模函数，该问题是NP难问题
  - 因此，可使用贪心算法构建S以得到近似的最优解

级联范围最大化

算法7.2　级联传播最大化——贪心算法

输入：传播图$G(V, E)$，预算k

1: **return** 种子结点集S

2: $i = 0$;

3: $S = \{\}$;

4: **while** $i \neq k$ **do**

5: $v = \arg\max_{v \in V \setminus S} f(S \cup \{v\})$;

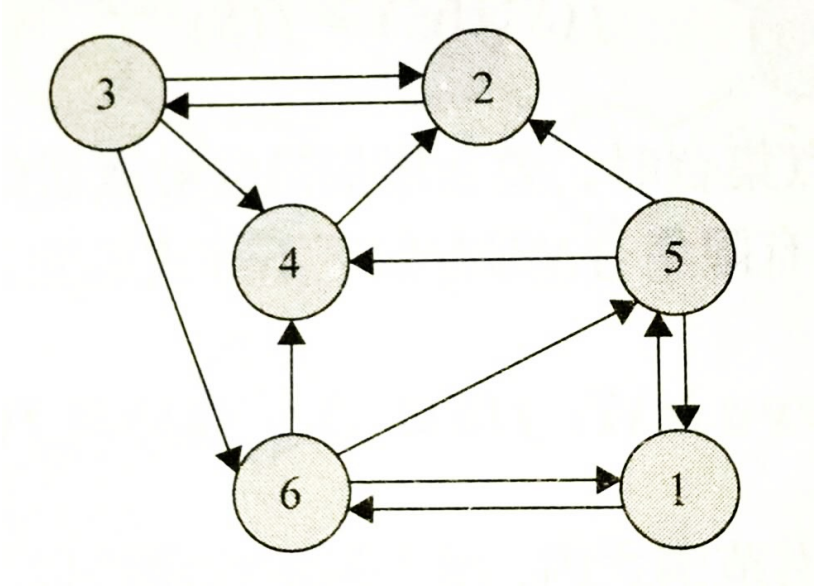
或v=arg $\max_{v \in V \setminus S} f(S \cup \{v\}) - f(s)$

6: $S = S \cup \{v\}$;

7: $i = i + 1$

8: **end while**

9: 返回S



问题：假设激活条件为|i-j|≡2(mod3)，求解k=2时的S

解：根据激活条件，可以推导出

|f({1})|=|f({6})|=4，|f({4})|=2，

|f({2})|=|f({3})|=|f({5})|=1

因此，第一步选取1或6激活，则最终会激活{1,2,4,6}，剩下结点3和5做第二步选择；

由于|f({6,3})|=|f({6,5})|=5，所以选3或5都可

信息级联

- 对信息级联的干预手段
  - 限制信息（决策）发布者的出链
  - 限制接收者的入链
  - 降低网络结点（个体用户）的激活概率$p_{v,w}$，

线性阈值模型

- 结点v的邻居u能激活v的概率$p_{u,v}$满足

$$\sum_{u \in N(v)} p_{u,v} \leq 1$$

- 结点v被激活的阈值为θ_v
- 在时刻t，一个未激活结点v同时受到其所有已激活邻居的影响，如果$\sum_{u \in N(v)} p_{u,v} \geq \theta_v$，则v被激活

可以看作是ICM模型的变种

创新扩散

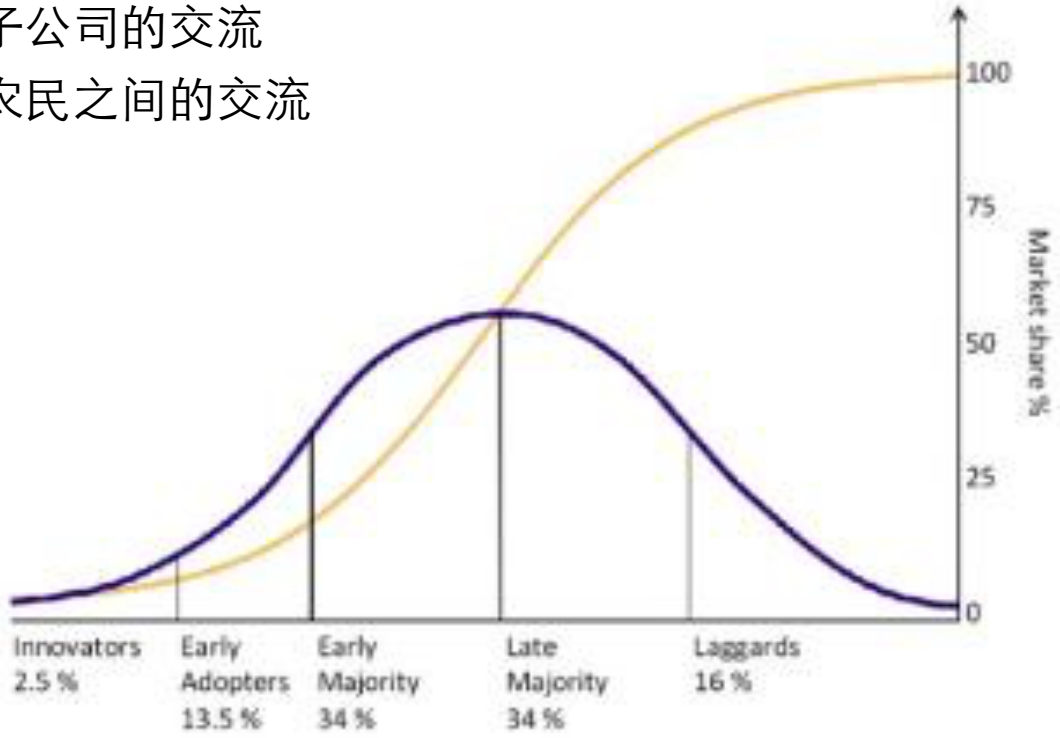
- 何为创新?
  - 被个体或团体看作新事物的一个观念、行为或实体
  - 示例：一首歌曲、一段视频、一条新闻…
  - 特征
    - 高度可见（可实验性）
    - 有相对优势
    - 与其所处的社会文化规范相容
    - 复杂度不能过高
- 创新扩散理论
  - 该理论阐述创新扩散的本质、原因和过程，并涉及扩散的受众和扩散速度
  - 创新扩散模型是对创新采用的各类人群进行研究归类的一种模型，其指导思想是在创新面前，部分人会比另一部分人思想更开放，愿意采纳创新

创新扩散

- 创新传播的五个步骤：
  - 认知（knowledge）
  - 说服（persuasion）
  - 决定（decision）
  - 实施（implementation）
  - 确认（confirmation）
- 创新扩散的五类受众：
  - 创新者(Innovator)
  - 早期采用者(Early Adopter)
  - 早期大众(Early Majority)
  - 晚期大众(Late Majority)
  - 落后者(Laggard)

创新扩散

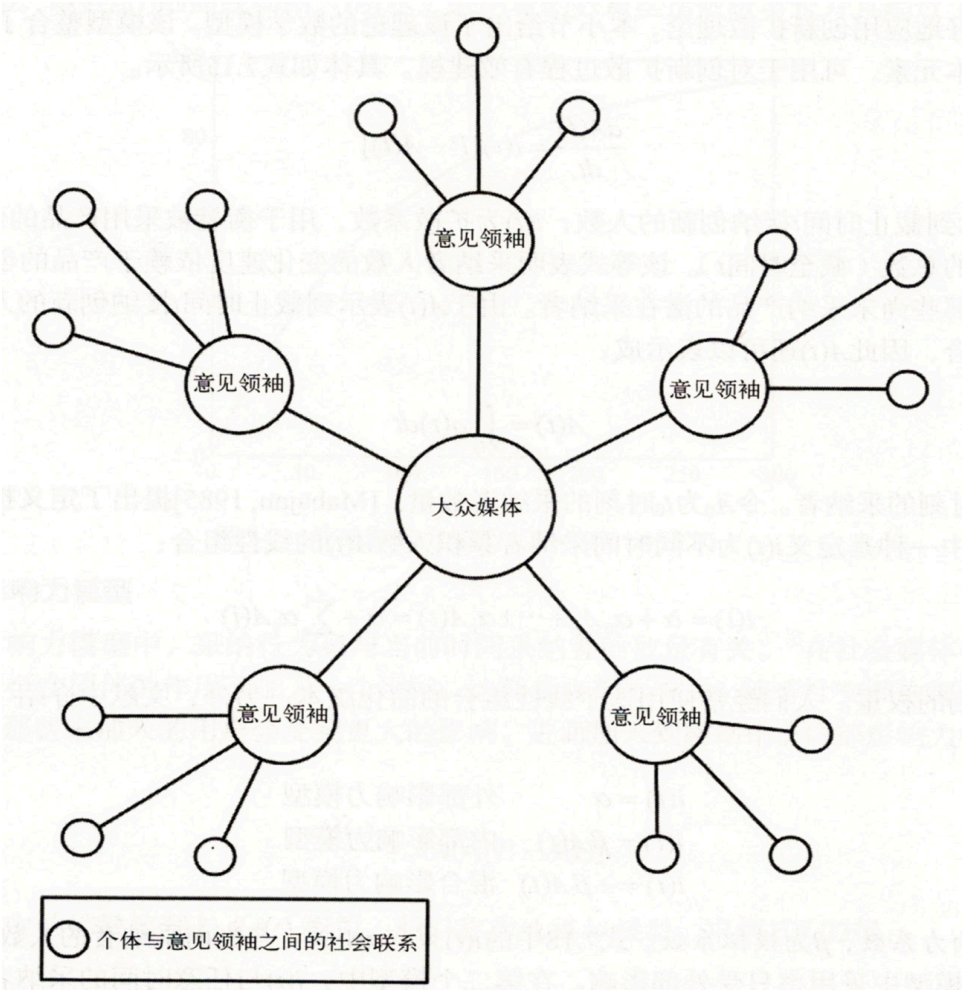
- 受众类型
  - 由Ryan和Gross在1943年调研爱德华州农民对杂交玉米新种子采用情况后得出的模型
  - 鉴于新种子无法繁殖且价格昂贵，农民大多不愿采用新种子，他们获取关于新种子信息的渠道包括：
    - 与种子公司的交流
    - 部分农民之间的交流



受众类型与S型采用累积曲线图

创新扩散

- Katz两级传播模型
  - 用以表示大众传播中信息传递的两级（多级）传播模型



创新扩散

- Rogers创新扩散过程
  - 了解阶段：接触新鲜事物，但知之甚少
  - 兴趣阶段：发生兴趣，并寻求更多的信息
  - 评估阶段：联系自身需求，考虑是否采纳
  - 试验阶段：观察是否适合自己的情况
  - 采纳阶段：决定在大范围内实施

创新扩散的数学建模

- 采用创新的人数变化趋势可以表示为：

$$\frac{\Delta A(t)}{\Delta t} = i(t)[P - A(t)]$$

采纳人数变化速度依赖于产品的创新度

$A(t)$为截止到t时刻采用创新的总人数，P为潜在的采用者总数，$i(t)$是创新扩散系数，可衡量被采用创新（产品）的创新度

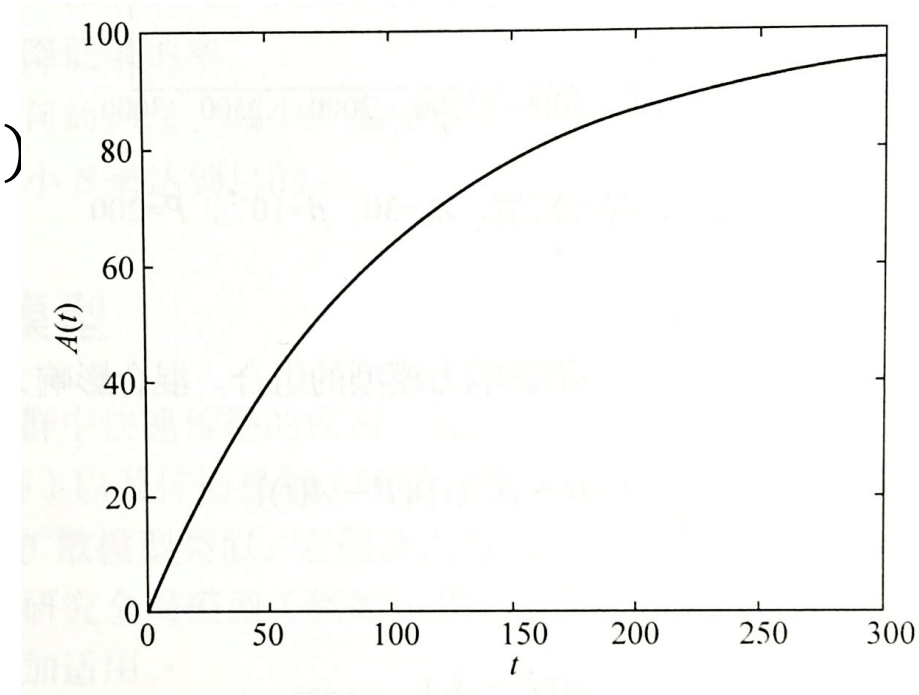
- 根据不同模型类型，$i(t)$可以按如下几种方式计算：
  - $i(t) = \alpha$，外部影响力模型
  - $i(t) = \beta A(t)$，内部影响力模型
  - $i(t) = \alpha + \beta A(t)$，混合影响力模型

创新扩散的数学建模

- 外部影响力模型
  - 创新（产品）本身的创新度（内容质量等）决定了新的采用人数的规模
    - 例如新闻本身的重要/新颖程度决定了阅读人数
  - 采用创新的人数增长率可以表示为：

$$\frac{\Delta A(t)}{\Delta t} = \alpha[P - A(t)]$$

求解可得：$A(t) = P(1 - e^{-\alpha t})$



P=100，α=0.01

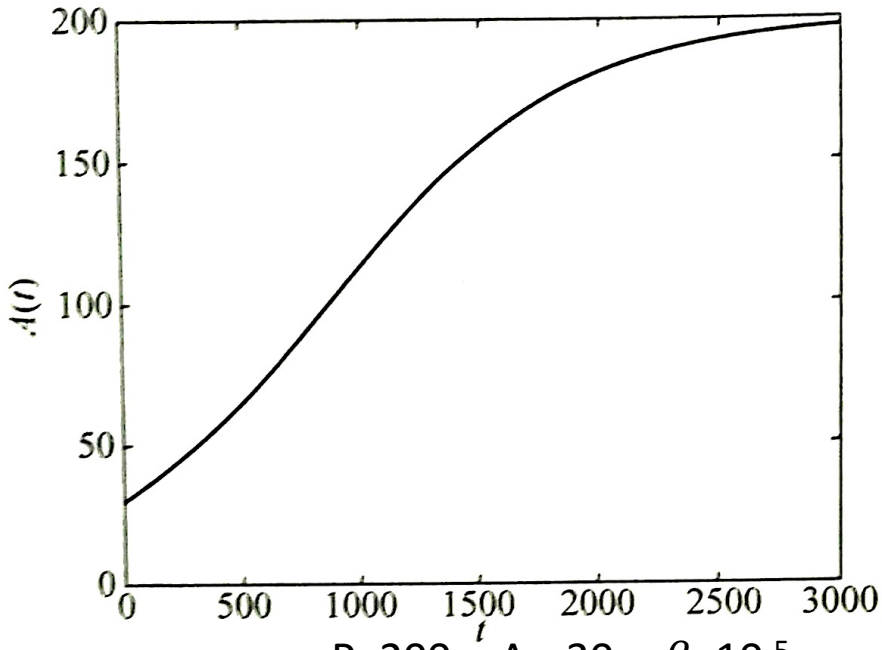
创新扩散的数学建模

- 内部影响力模型
  - 对创新（产品）本身的采纳行为只与当前时间的采纳者数量有关，也称为纯模拟模型
    - 例如某人会在已经加入某社交群体的朋友影响下也加入该群体
  - 采用创新的人数增长率可以表示为：

$$\frac{\Delta A(t)}{\Delta t} = \beta A(t)[P - A(t)]$$

- 求解可得：$A(t) = \dfrac{P}{1+\frac{P-A_0}{A_0}e^{-\beta P(t-t_0)}}$

A_0是初始时刻采用创新的人数



P=200，A_0=30，β=10^{-5}

创新扩散的数学建模

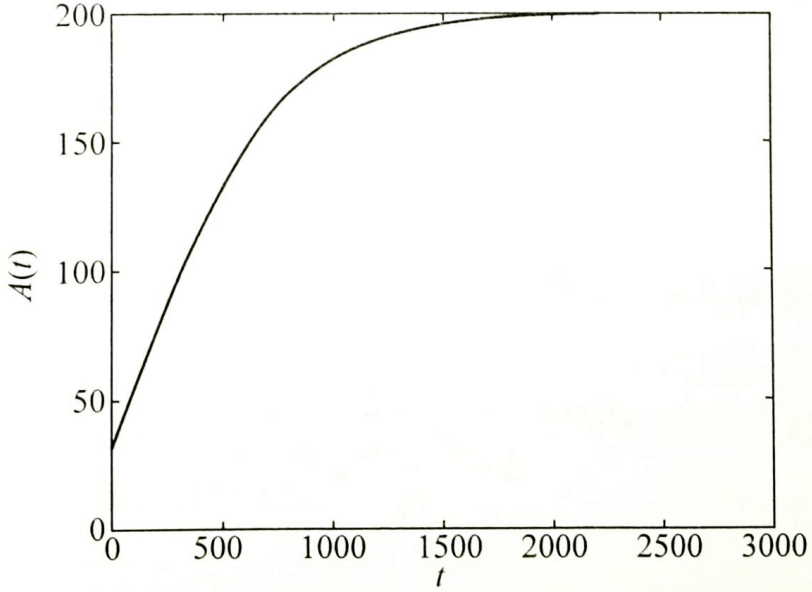
- 混合影响力模型
  - 采用创新的人数增长率可以表示为：

$$\frac{\Delta A(t)}{\Delta t} = (\alpha + \beta A(t))[P - A(t)]$$

- 求解可得：

$$A(t) = \frac{P - \dfrac{\alpha(P - A_0)}{\alpha + \beta A_0} e^{-(\alpha+\beta P)(t-t_0)}}{1 + \dfrac{\beta(P - A_0)}{\alpha + \beta A_0} e^{-(\alpha+\beta P)(t-t_0)}}$$

A_0是初始时刻采用创新的人数



混合影响力模型，P=200，β=10^{-5}，A_0=30，α=10^{-3}

创新扩散

- 对创新扩散的干预手段
  - 限制（新）产品或用户的分布
  - 降低用户对产品的兴趣
  - 减少用户之间的沟通

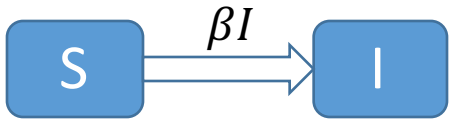
流行病模型

- 流行病的三要素
  - 病原体（传播的疾病）
  - 宿主（人、动物、植物等）
  - 传播/感染机制（呼吸、饮用水、性行为等）
- 流行病模型假设人与人之间关系未知（即存在一个隐性网络），因此更加注重研究全局模型，而不是观察具体的传播路径（不像羊群效应和信息级联）
- 流行病模型不仅在医学领域大量应用，在互联网中计算机病毒传播研究中也有非常重要作用
- 流行病的研究方法
  - 接触网络：通过观察宿主如何接触来设计描述流行病在人群网络中发生的方法
  - 全混合：只分析传染者感染、恢复的速度而不考虑接触网络的相关信息

本课程关注

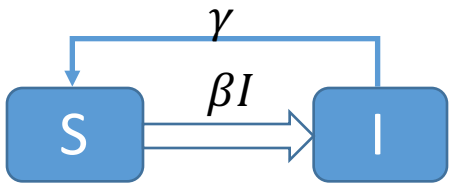
流行病模型

- 模型定义
  - 群体总人数为N，若不考虑出生和死亡，则N不变
  - 群体中每个个体处于三种状态之一：
    - 易感染（当前时刻的人数用$S(t)$表示，简写为S）
    - 被感染（当前时刻的人数用$I(t)$表示，简写为I）
    - 已康复/死亡（当前时刻的人数用$R(t)$表示，简写为R）
- 模型类型
  - SI模型：

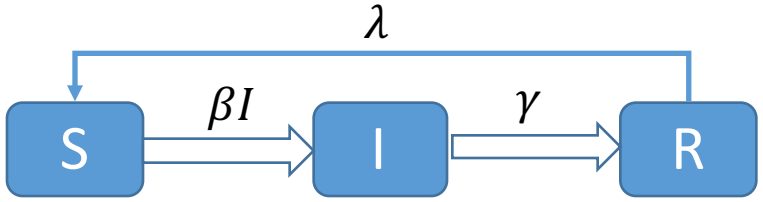


  - SIR模型：

  - SIS模型：



  - SIRS模型：

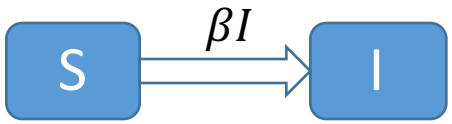


SI模型

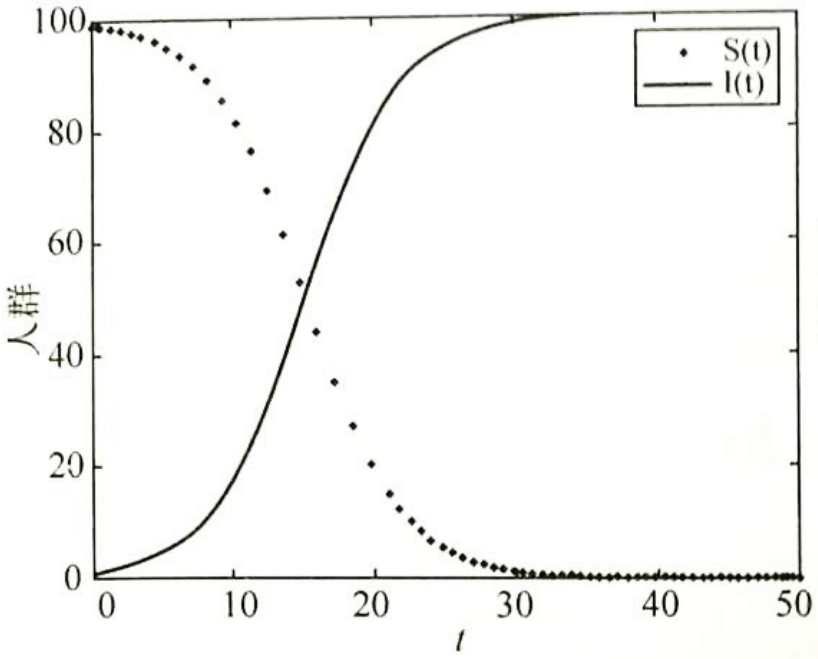
- 该模型是最简单的流行病模型，假设被感染者不能被治愈，即没有R，随着时间推移，S不断减小直至0，I不断增大直至N
- 假设β为接触概率，则被感染者平均会接触βN个人，考虑全部N个人中只有S个可能会在下一时刻被感染，所以每个被感染者将会感染βS个人
- 考虑到当前时刻共有I个人已经被感染，所以有

$$\frac{dI}{dt} = \beta IS,\ \frac{dS}{dt} = -\beta IS$$

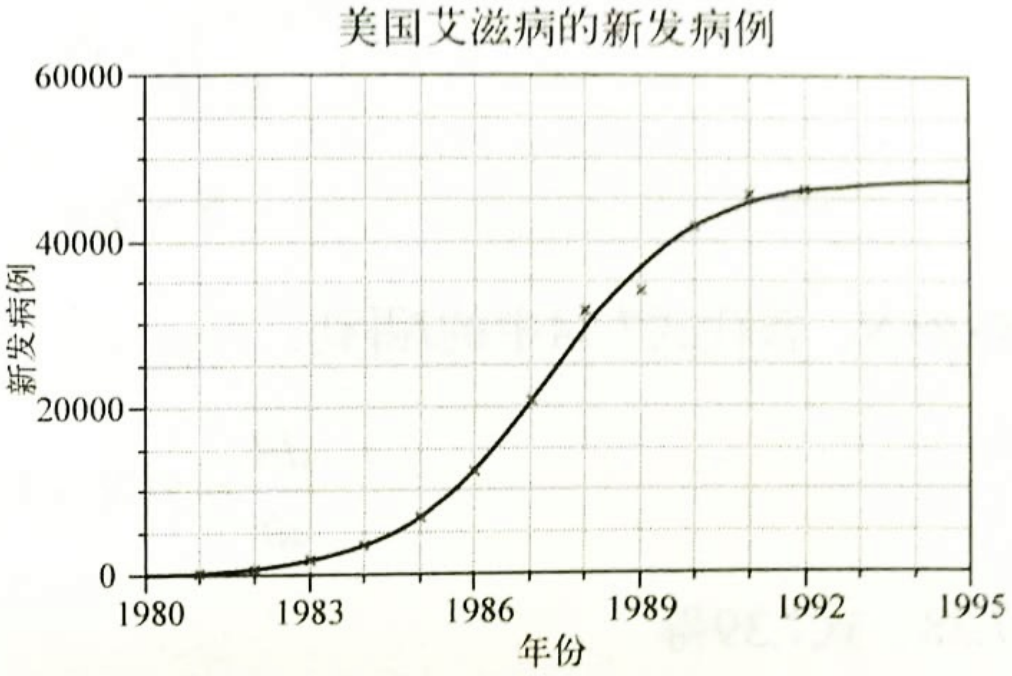
- 由于$S = N - I$，所以有 $\frac{dI}{dt} = \beta I(N - I)$



SI模型



(a) SI模型仿真



(b) HIV/AIDS感染人群

SIR模型

- 感染者经过一段时间后会康复（存在R），且不会被再次感染
- 仍然假设β为接触概率，γ为单位时间dt内的康复概率，参照SI模型，则有

$$\frac{dS}{dt} = -\beta IS$$

$$\frac{dI}{dt} = \beta IS - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

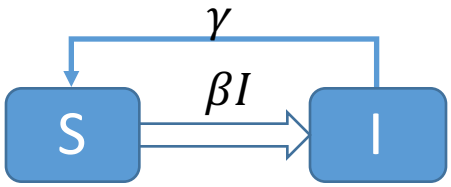
如果考虑有部分感染者会死亡，则模型要更复杂

SIS模型

- 是SI模型的扩展，不同在于康复者会马上全部再变成易感染者
- 参照SI和SIR建模过程，不考虑R的存在（因为马上会变成S），则有

$$\frac{dS}{dt} = -\beta IS + \gamma I$$

$$\frac{dI}{dt} = \beta IS - \gamma I$$



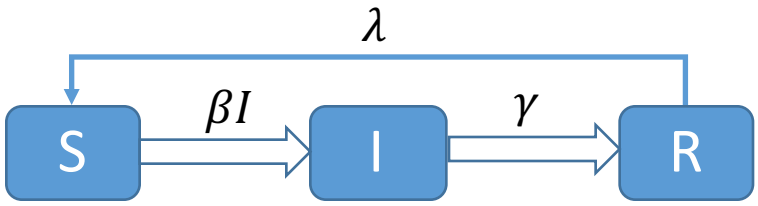
SIRS模型

- 是SIR模型的扩展，不同在于康复者只有一部分会再变成易感染者
- 假设λ 是康复者变成易感染者的概率，则有

$$\frac{dS}{dt} = -\beta IS + \lambda R$$

$$\frac{dI}{dt} = \beta IS - \gamma I$$

$$\frac{dR}{dt} = \gamma I - \lambda R$$



流行病模型

- 对流行病的干预手段
  - 对社交频繁的人接种疫苗
  - 对整群体内96%的随机个体接种疫苗达到群体免疫
  - 或者先对群体内30%的随机个体接种疫苗，再找到他们擅于交际的朋友进行接种，从而达到群体免疫
  - 对被感染者和易感染者进行免疫隔离

其他传播模型

- 博弈论模型
  - 用户根据个人利益最大化来主导自己的行为（进行选择）
- 马尔科夫随机场模型
- 谣言传播模型
- 竞争性的信息扩散模型
- 跨异构社交网络的扩散模型

目　录

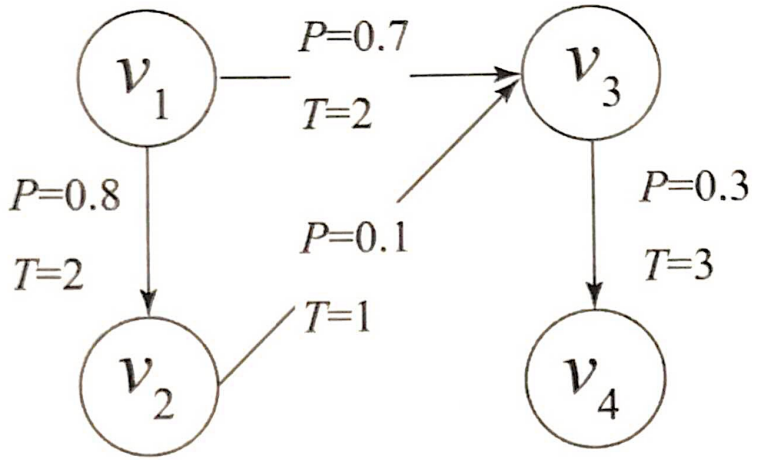
- 计算传播学原理
- 社会影响力与同质性
- 信息传播模型
- 影响力最大化
- 热门话题分析与预测

影响力最大化

- 问题定义
  - 目标是要找到一组数量为k的种子结点，使得信息传播结束后网络中活跃的结点数最多 类似于级联范围最大化
- 应用场景
  - 如何选择一个新产品的试用用户群，这些用户通过社会影响让更多的用户也去使用这款产品
  - 如何选择潜在的客户打广告也可看成是该问题
- 已被证明是一个NP-hard问题

影响力的新鲜度衰减

- 网络中传播的信息产生的时间越久(出现的频次越高)，用户转发的概率越低——新鲜度衰减



在有向边上影响概率为 P、时延为 T 的社交网络有向图

- 如果不考虑新鲜度衰减，假设传播的种子集合是{v_1,v_2}，则v_3被激活的概率是0.1+(1-0.1)x0.7
- 如果考虑新鲜度衰减，在v_2尝试激活了v_3后，v_3被v_1激活的概率应该少于(1-0.1)x0.7

由于结点之间的相互影响要考虑时延，因此相关算法比较复杂，影响力传播范围可以通过建立传播路径(基于某种算法)来估计

影响力的新鲜度衰减

- 新鲜度衰减函数
  - 令TP_n是一个结点被n个结点尝试激活后被成功激活的概率，p_n是一个结点被第n个结点尝试激活后被成功激活的概率，则TP_n和TP_{n-1}之间的关系模型为：

$$TP_n=TP_{n-1}+(1-TP_{n-1})\times p_n$$

$$p_n=(TP_n-TP_{n-1})/(1-TP_{n-1})$$

  - 新鲜度衰减函数可以表示为$f(n)=p_n/p_{n-1}$， $p_1=TP_1$是所有结点在第一次被尝试激活后就成功激活的概率

影响力最大化基本算法

- 启发式算法
  - 基于High Degree
    - 选择度数最多的那些结点
  - 基于Distance Centrality
    - 选择那些与其他结点的平均距离最短的结点
  - Random算法
    - 随机选择k个结点

选出的结点不一定最优，因为他们有可能都处于一个团体中（富人俱乐部）

影响力最大化基本算法

- 贪心算法
  - 假设当前已经选出的结点集合为A，选择下一个结点时，检查A以外的每一个结点u，计算AU{u}的影响力（添加边际收益），选择能产生最大影响力的u加入A
  - 不断重复上述过程，直至|A|=k

影响力最大化改进算法

- 改进的贪心算法
  - 原始的贪心算法要计算A以外所有结点带来的边际收益增长，代价太大
  - 通过一些改进的算法能够对新种子候选集进行剪枝（例如边际收益的增长要大于一个阈值），从而降低代价
  - 例如R-Greedy，动态优化

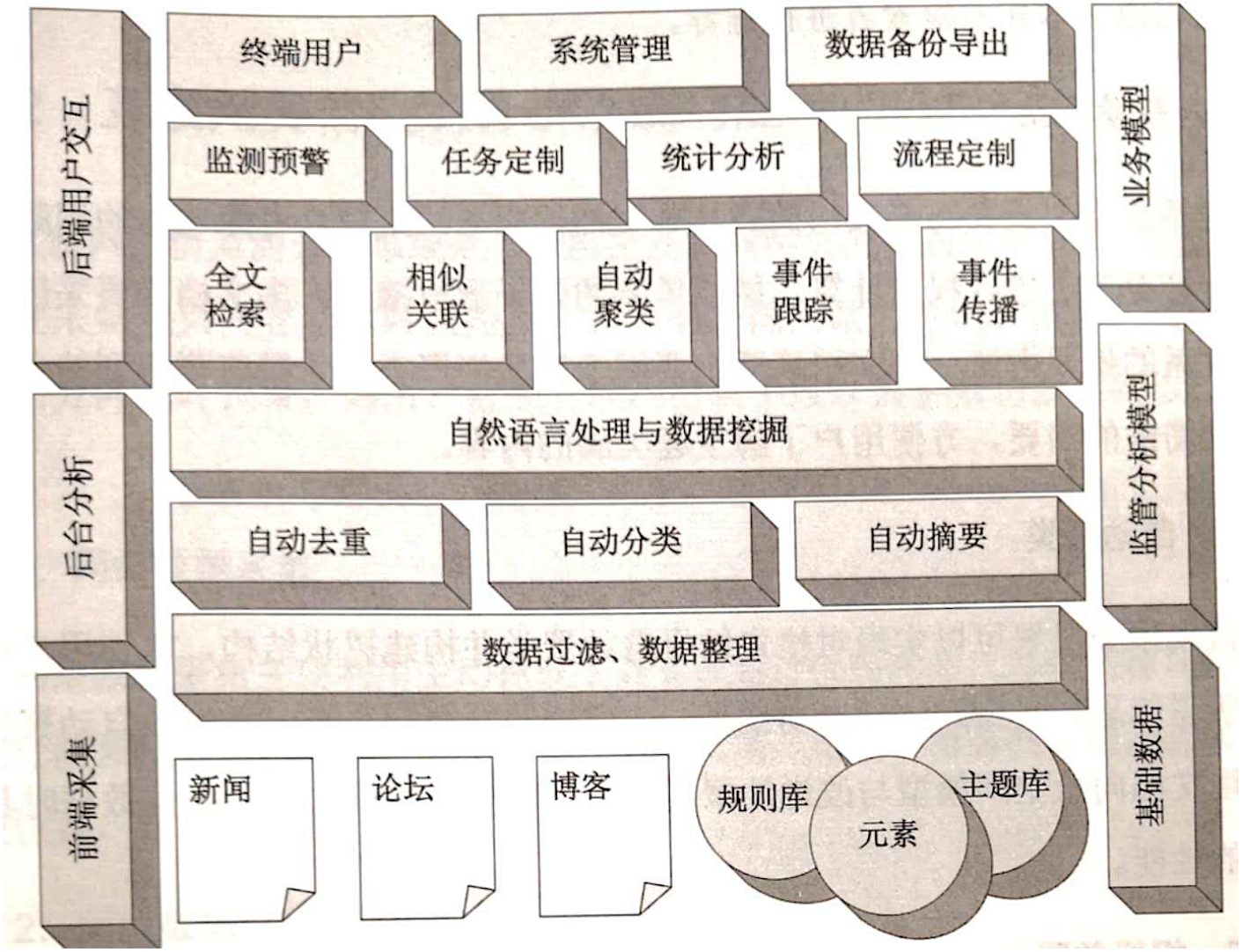
信息覆盖最大化

- 信息覆盖与信息传播
  - 线性阈值和独立级联模型中，活跃结点指接受消息并传播的结点，非活跃结点是不会传播的结点
  - 非活跃结点还可以细分为消息结点和真正的非活跃结点，消息结点是能看到信息的结点
  - 对应到微博网络中，一个用户发表微博后，其所有粉丝都是消息结点，转发他微博的粉丝才是活跃结点；一个结点要成为消息结点，必须他的邻居（关注）中至少有一个是活跃结点
  - 信息覆盖最大化对比影响力最大化，都是寻找k个种子结点，但是信息覆盖最大化目标是能产生最多的活跃结点和消息结点
  - 独立级联模型、贪心算法、启发式算法等经过修改后仍可以用于解决信息覆盖最大化的问题

目　录

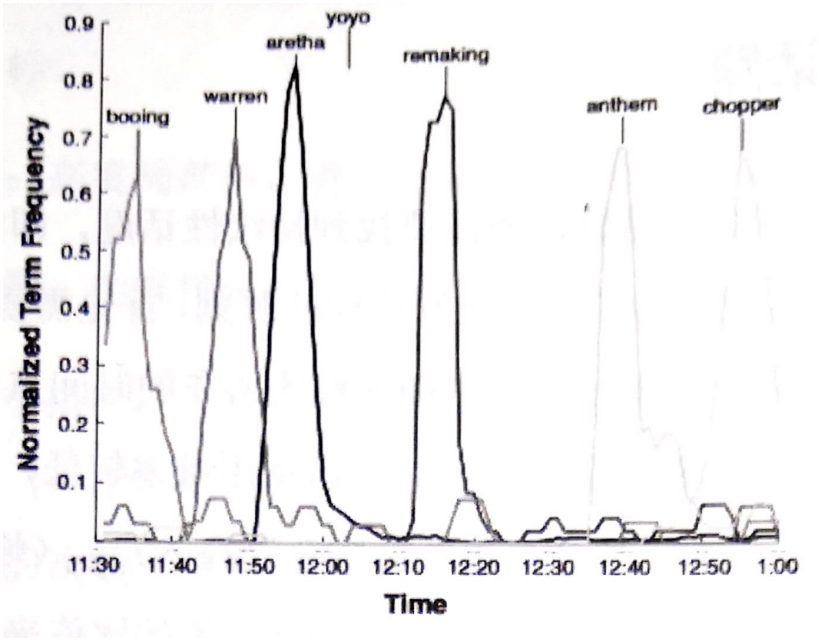
- 计算传播学原理
- 社会影响力与同质性
- 信息传播模型
- 影响力最大化
- 热门话题分析与预测

网络舆情监测系统

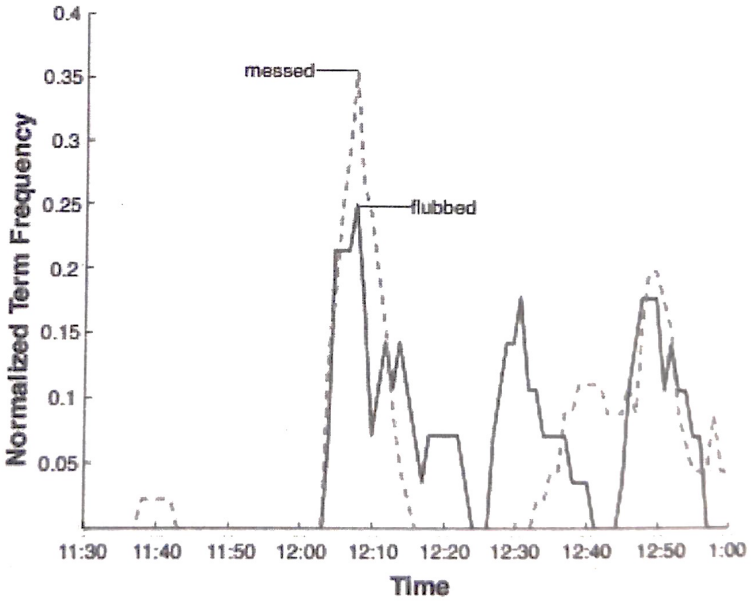

后端用户交互
终端用户
系统管理
数据备份导出
业务模型
监测预警
任务定制
统计分析
流程定制
全文检索
相似关联
自动聚类
事件跟踪
事件传播
后台分析
自然语言处理与数据挖掘
监管分析模型
自动去重
自动分类
自动摘要
数据过滤、数据整理
前端采集
新闻
论坛
博客
规则库
元素
主题库
基础数据

热门话题分析

- 热门话题的特征
  - 高度局部性：在某一时段内有较高的讨论热度
  - 持续性：在相对较长的时段内都能维持一定的讨论热度



局部热门话题



持续性热门话题

热门话题分析

- 话题热度值的量化
  - 基于TF-IDF模型
    - 选定代表话题的关键词
    - 采集不同时间片（通常较短）内的讨论帖（文档）
    - 用文档中各话题关键词的TF-IDF值来量化当前时间片中话题的热度

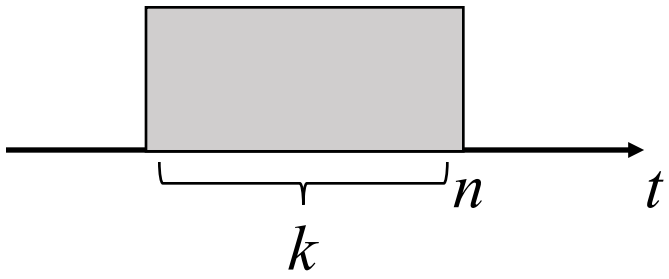
热门话题预测

- 话题趋势预测
  - 用$f(i)$表示某个时间片的话题热度值(关键词的TF-IDF值)
  - 假设移动平均值的窗口大小为k(个时间片)，则话题在第n个时间片的移动平均值为：

$$MA(n,k)=\frac{\sum_{i=n-k+1}^{n} f(i)}{k}$$

如果$n<k$，则

$$MA(n,k)=\frac{\sum_{i=1}^{n} f(i)}{k}$$

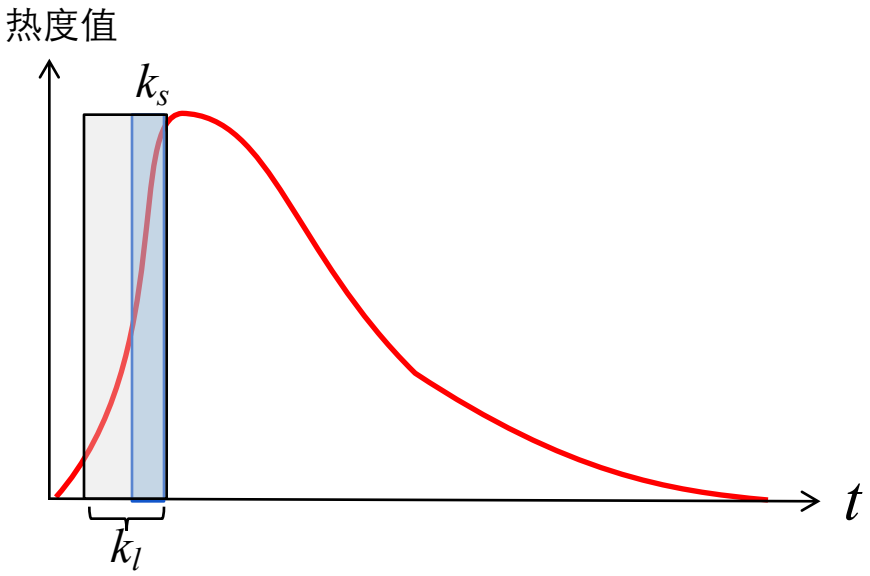


热门话题预测

- 话题趋势预测
  - 假设k_s和k_l分别表示较短和较长的时间窗口，则话题在时间片n上的趋势动力为：

$$TM(n) = MA(n, k_s) - MA(n, k_l)$$

  - $TM(n)$为正则表明话题热度在持续上升，否则在下降，值越大表明上升/下降的趋势越明显(坡度越明显)



热门话题预测

- 影响话题趋势变化的因素
  - 关键用户(权威用户)参与讨论往往让话题变得更热门
  - 一个新话题的关键词已经在其他热门话题内频繁出现，也可能让新话题很快变成热门
  - 话题热度的衰竭是因为讨论的人减少(都去讨论新的热门话题了)，整个网络中参与讨论的人数是有限的，因此不同话题间的热度会相互影响

热门话题预测

- 影响用户转发(讨论)微博的因素[1]
  - General：帖子的新鲜度(已发表时间的长短)
  - Recent：发帖人是否是最近交互过的好友
  - Topic：帖子主题是否和该用户的兴趣相近
  - Profile：发帖子的兴趣主题是否和该用户相近

Model	General	Recent	Topic	Profile
	10.2%	16.0%	51.6%	22.2%

帖子内容(主题)是否相关比时间因素造成的新鲜度衰减更重要

[1] Sofus A. Macskassy and Matthew Michelson. Why Do People Retweet? Anti-Homophily Wins the Day! （基于Twitter数据的分析结果）

课后作业

- 数据描述
  - 包含8000多个微博用户id及其标签列表、标签的权重分数
  - 每个人的标签列表包含通过某种算法生成的20个标签，该算法也产生了标签权重，刻画了标签对用户的特征描述能力
- 作业要求
  - 利用LSA、LDA等主题模型并结合KNN、K-Means等分/聚类算法将这些用户进行分组
  - 对每组用户进行群体画像
    - 可基于TF-IDF等标签权重计算方法为每组用户选出最具代表性的标签集合
- 提交内容
  - 6月6日（周三）前上交实验报告，包括算法/模型介绍、实验结果及分析