Web信息处理与应用

第十六节

社交传播导论

徐童

2023.12.18

- **社会网络的巨大潜力**

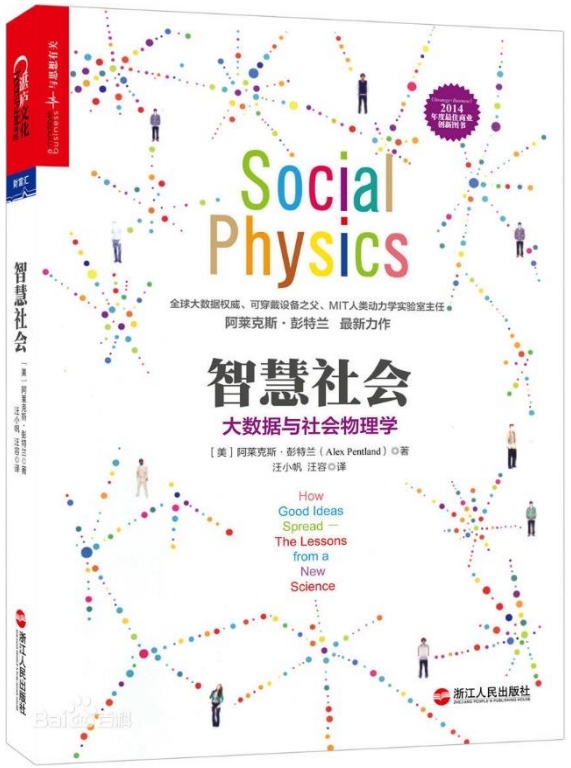
- 2009年，**“寻找红气球”**挑战赛
  - 美国国防部高级研究计划署（DARPA）为了纪念互联网诞生40周年所举办。
  - 他们在全美各地布设了10个红气球，能用最短时间找到全部气球坐标的个人或组织，将获得4万美元的高额奖金。
  - 美国国家地理空间情报局（NGA）的一位高级分析员将之称为“传统的情报收集方法无法解决”的难题。

1969年，ARPANet诞生

2009年，DARPA的红气球

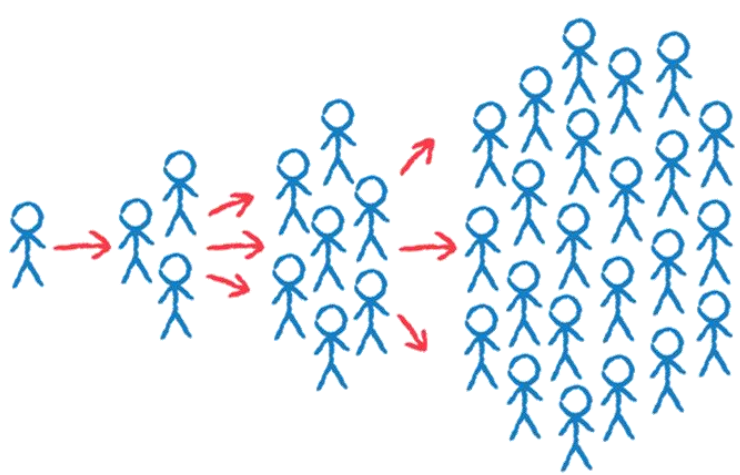
- **社会网络：新的信息传播渠道**

- 然而，难题的快速解决出乎意料
  - 最后，来自MIT的团队仅用8小时52分41秒，就将10个红气球的坐标全部标示完毕。
    - 团队的领导阿莱克斯·彭特兰（Alex Pentland）是全球知名的计算机科学家，《智慧社会》一书的作者
  - 据说，该团队所动员的总人数大约为两百万！
    - 如何实现如此快速而大规模的动员？



- **社会网络：新的信息传播渠道**

- 成功的秘诀：社会网络 + 巧妙的激励机制
  - 据称，彭特兰在短短数小时内便动态组建了一支成员多达5000人的团队，这5000名队员中的每个人又平均通知了400名朋友。
  - 如何实现动员激励？
    - 不仅奖励正确告知气球地点的人，还奖励那些把找到气球的人成功介绍给团队的人。
      - 类似传销的链式奖励规则（手动狗头）。

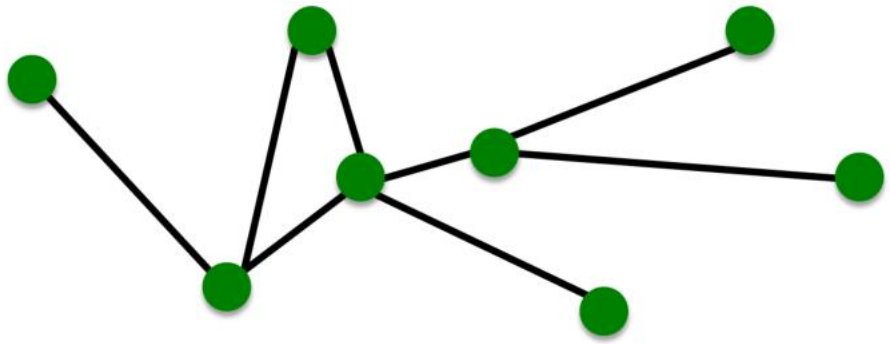


本节目录

- **基本概念**
- 节点角色
- 社会网络中的传播
  - 基本传播模型
  - 传播最大化问题
  - 衍生传播问题

- **社会网络的基本元素**

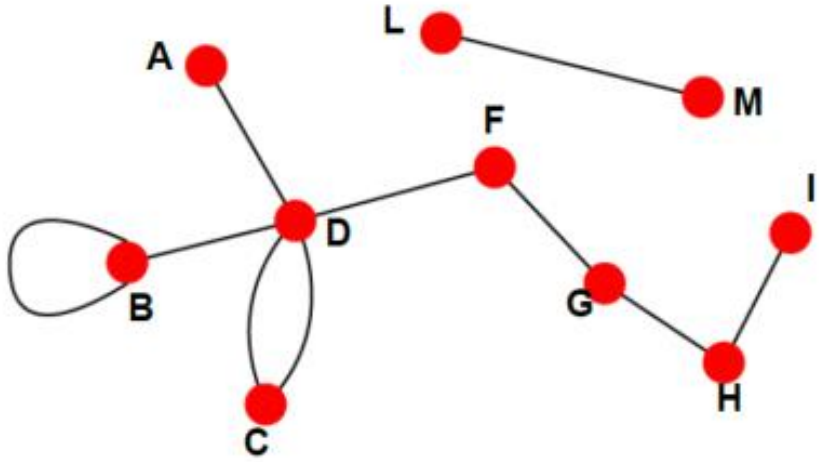
- 从数学抽象的角度，可以将社会网络表示为图（Graph）的结构
  - 节点（Node/Vertex）：用于表示网络中的实体，如社会网络中的人
  - 边（Link/Edge）：用于描述网络中的关系，如人们之间的社交关系



- **社会网络的基本元素：有向/无向边**

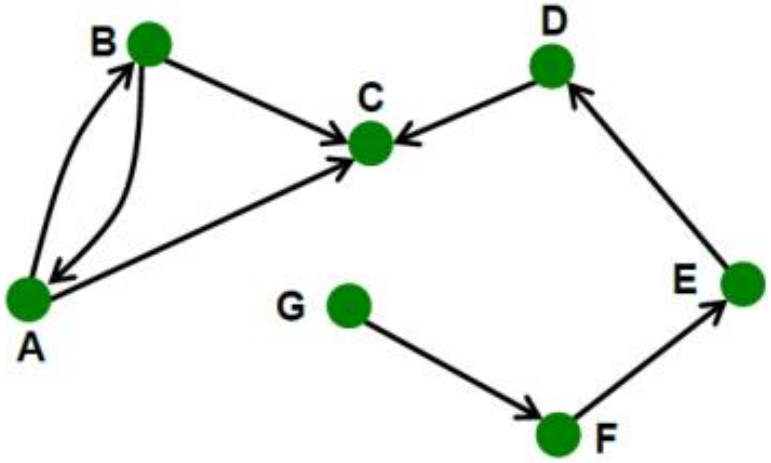
- 网络中的边可能有向，也可能无向，各自表达不同含义

无向边：或双向边、表示对称关系



例如：朋友、合作关系

有向边：表示非对称关系



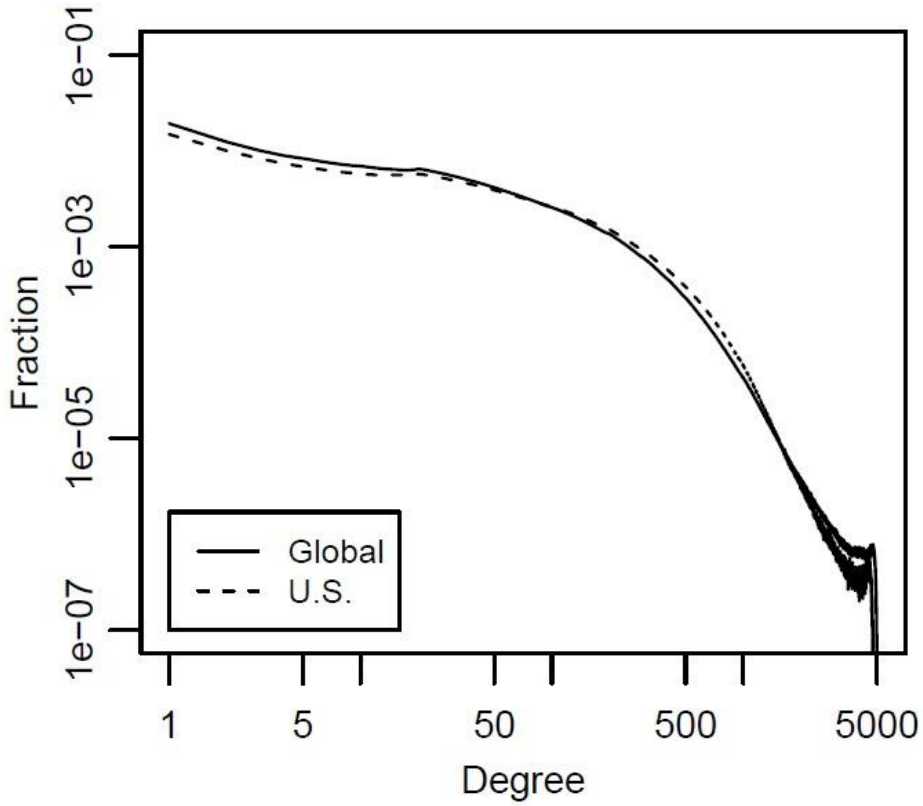
例如：通讯、关注关系

- **社会网络的基本元素：邻居、出入度**

- 对于任意节点而言，与其直接相连的节点被称作“邻居” （Neighbor）
  - 对节点 v 而言，往往采用 N(v) 表示其邻居集合
  - 在有向网络中，入边邻居和出边邻居集合应加以区分
- 同时，节点所连边的数量（或邻居的数量）被称作“度” （Degree）
  - 对节点 v 而言，往往采用 d_v 表示其度数
  - 同样，出度与入度应加以区分
    - 显然，对于一个网络而言，节点出度之和*等于*入度之和*等于*边数

- **社会网络的基本元素：邻居、出入度**

- 一般而言，真实网络中的节点度数往往符合幂律（Power-law）分布
  - 少数节点拥有大多数的边
  - 这些少数节点即形成了所谓“影响力节点”，也就是俗称的“大V”

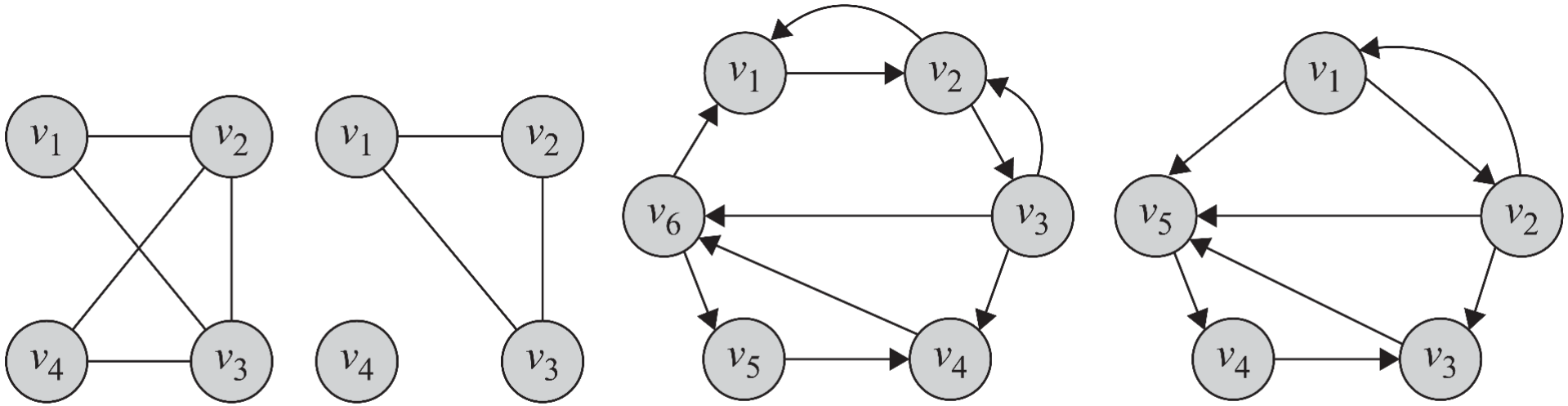


- **社会网络的基本元素：连通性**

- 两个节点是连通的（Connected），当且仅当节点之间存在一条路径（Path）
  - 注意：这并不意味着两个节点之间是直接相连的
- 进而，一个图是连通的，当且仅当图中任意两个节点都是连通的
  - 对于有向图而言，有强 / 弱连通的区别
    - 任意两个节点之间存在双向的连通路径，即为强连通
    - 忽略方向的前提下任意两个节点之间存在一条路径，即为弱连通
  - 如果存在两个节点无法连通，则图是不连通的（Disconnected）

- **社会网络的基本元素：连通性**

- 图的连通性（Connectivity）的实例

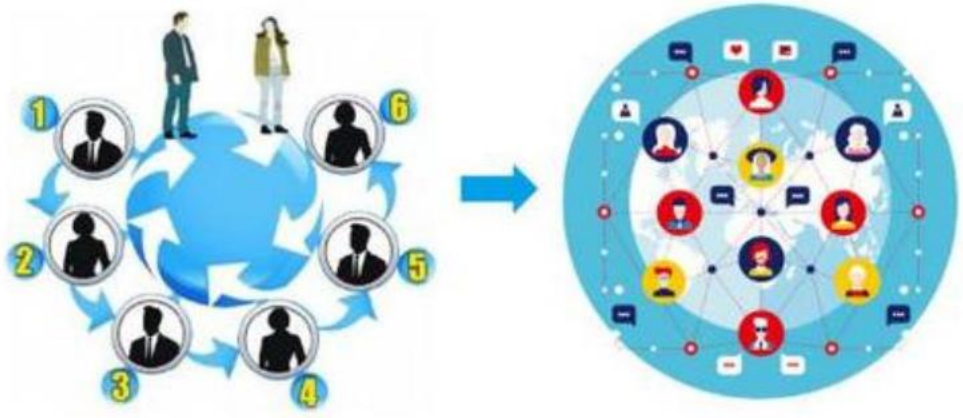


(a) Connected (b) Disconnected (c) Strongly connected (d) Weakly connected

(d)中仅存在单向连通路径

- **社会网络的基本元素：连通性**

- 有关连通性的拓展阅读：小世界网络（Small World）
  - 茫茫人海，两个互不相识的人是否相互连通？
  - 如果想要互相认识，中间需要经过几个人？



六度空间理论　　六度人脉社交系统

- **社会网络的基本元素：连通性**

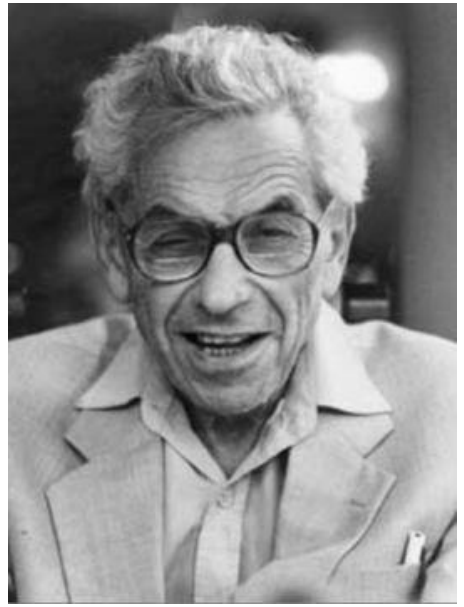
- 六度空间理论，源自Stanley Milgram于1967年的著名实验
  - 设计：观察需要经过多少中间人，才能使信息从随机起点到达特定终点
  - 实验规则如下：
    - 参与人只能向信件转发给熟人，并请他继续转发
    - 参与者需力争让信件尽快达到目的地（有选择性地挑选转发人！）
  - 结果：平均通过5位中间人的转发可以抵达目标

- **社会网络的基本元素：连通性**

- 小世界现象的拓展阅读：
  - Milgram实验在更大规模数据集上是否仍具有指导意义？
  - 2006年，微软的研究人员基于MSN数据，进行了类似的尝试
    - 结果发现，在将近2亿用户的网络中，平均6.6人即可使将近2000亿的配对产生关联
    - 超过87%的配对在7次以内即可产生关联

- **社会网络的基本元素：连通性**

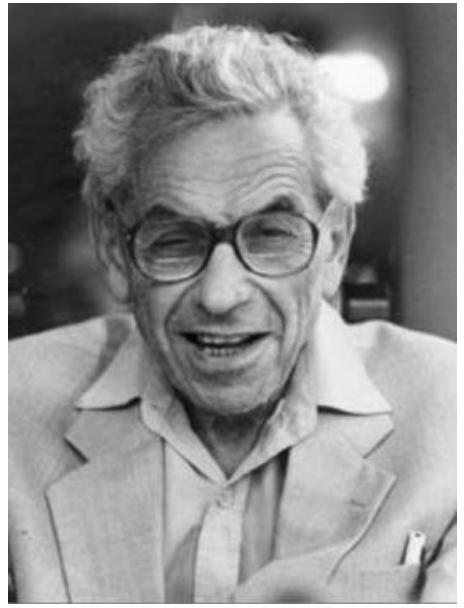
- 拓展阅读：专业领域（更严格的关系筛选）是否同样存在小世界现象？
  - 保罗·埃尔德什（Paul Erdős），匈牙利数学家
    - 一生发表论文1475篇，与511人合作，被誉为“最高产的数学家”
  - 埃尔德什数（Erdős Number）
    - 描述学者与埃尔德什“合作距离”的一种方式
      - 埃尔德什本人，Erdős Number = 0
      - 直接与本人合作，Erdős Number = 1
      - 与其合作者合作， Erdős Number = 2，以此类推



六度空间

- **社会网络的基本元素：连通性**

- 拓展阅读：专业领域（更严格的关系筛选）是否同样存在小世界现象？
  - 埃尔德什数（Erdős Number）
    - 统计显示，菲尔茨奖得主的Erdős Number中位数约为 3
    - 由于跨领域合作，许多非数学家也具有Erdős Number
      - 顺便提（吹）一句，本人的Erdős Number = 4



Tong Xu +1

MR Erdos Number = 3

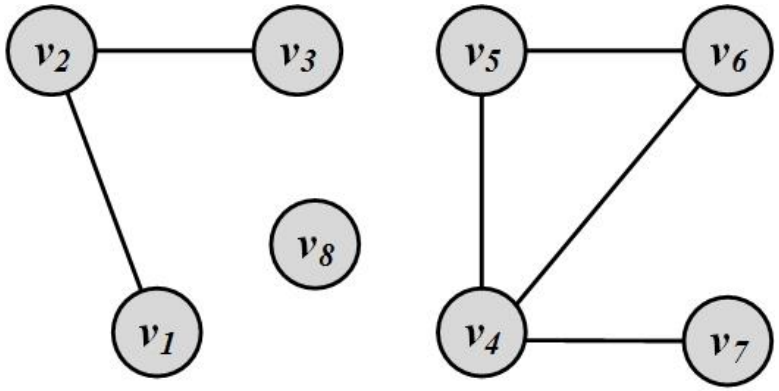
En-Hong Chen	coauthored with	Minming Li	MR2539922
Minming Li	coauthored with	Ronald Lewis Graham	MR2318683
Ronald Lewis Graham	coauthored with	Paul Erdős[1]	MR0592420

- **社会网络的基本元素：连通组件**

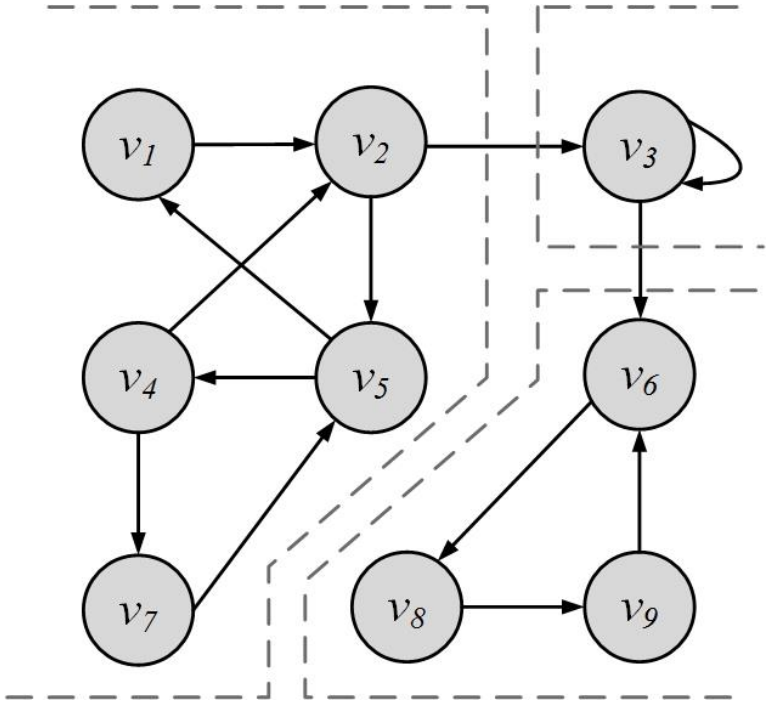
- 对于一个图而言，其中的一个连通组件（Component），即一个连通的子图
  - 换言之，该子图的任意两个节点之间都是连通的
  - 在有向图中，我们将强连通组件定义为该有向图的一个强连通子图
    - 即任意两个节点之间存在双向连通路径
    - 相应的，弱连通子图对应着弱连通路径

- **社会网络的基本元素：连通组件**

- 连通组件（Component）的实例



3 components



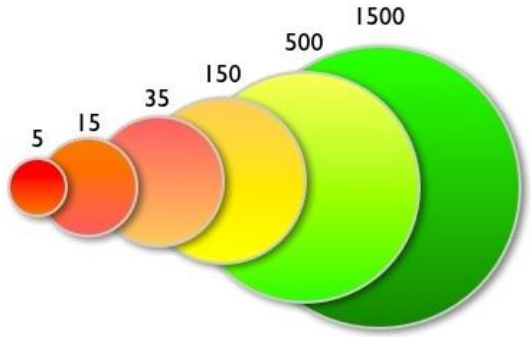
3 Strongly-connected components

- **社会网络的基本元素：连通组件**

- 有关连通组件的拓展阅读：邓巴数与150原则
  - 150原则：人类智力所允许拥有稳定社交网络的规模大约是150人。
  - 该原则由英国牛津大学的人类学家罗宾·邓巴（Robin Dunbar）在20世纪90年代提出，源于以下观察：
    - 古代：兄弟会（Anabaptist）的不成文规定：每当聚居人数超过150人的规模，他们就把聚居点变成两个，再各自发展
    - 动物社会中也有类似的社群与分裂现象，例如：蜘蛛猿通常形成2只到17只的子群，但每个子群往往只持续一两个小时

- **社会网络的基本元素：连通组件**

- 有关连通组件的拓展阅读：邓巴数与150原则
  - 邓巴数的成因：人们需要通过合作来发挥潜能，但过大规模的网络将导致沟通效率的下降，最终导致团队的分裂
    - 这个过程，无疑是连通组件拆分为更多子连通组件的过程
    - 某种意义上说，150可视作一个“维持社交关系”的人数上限
      - 例如，早期手机通讯录与社交软件的好友上限往往即为150
      - 现在，人数/群用户数虽然更为扩大，但仍有上限

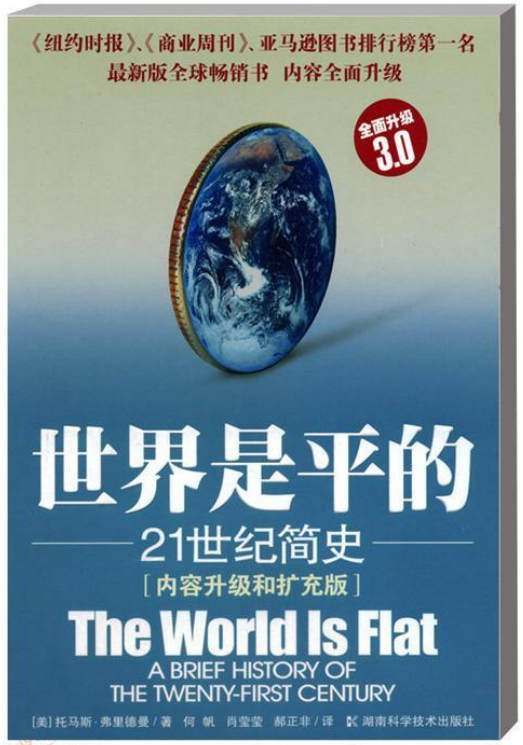


本节目录

- 基本概念
- **节点角色**
- 社会网络中的传播
  - 基本传播模型
  - 传播最大化问题
  - 衍生传播问题

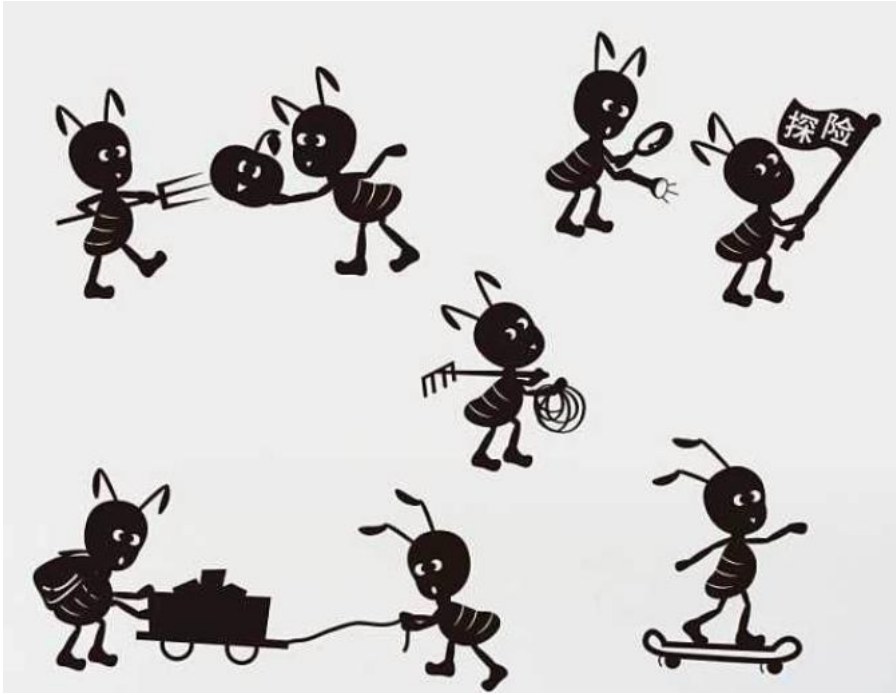
- **节点之间是平等的吗?**

- 托马斯·弗里德曼的知名著作《世界是平的》曾经指出:“个人透过全球化进程获得权力”，并指出这一过程与科技发展如网络密切相关
- 然而，这是否意味着网络中的节点是平等的?
  - 回顾开头的例子：保罗·里维尔的成功秘诀
    - “每来到一个城镇，都确切知道……当地民兵组织的领导人是谁，谁是该镇的首要人物”



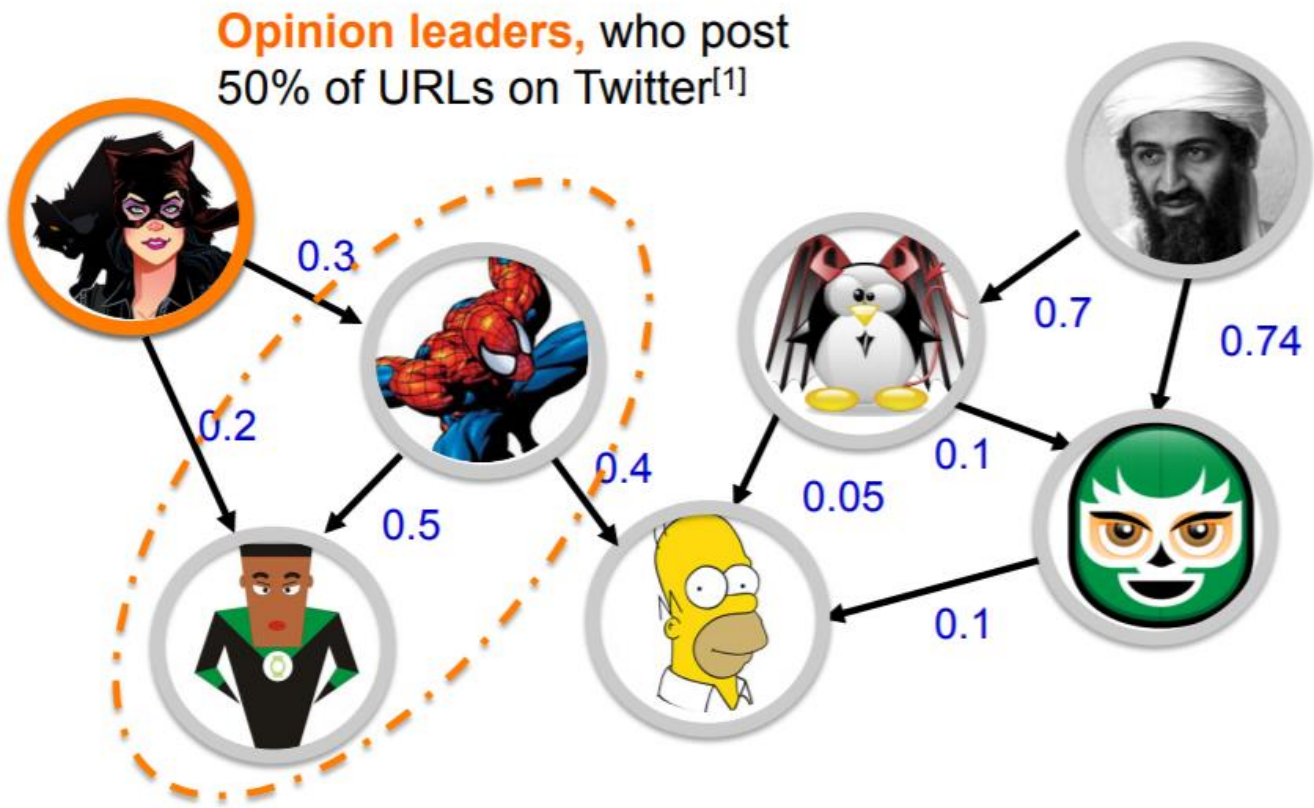
- **节点的角色不同**

- 研究者也发现，在同一个社会网络中，不同节点往往也扮演不同角色
  - 例如，一个部门，往往有以下分工：
    - 部门经理：负责领导部门
    - 技术专家：负责提供技术指导
    - 项目经理：负责外部需求沟通
    - ……



- **节点的角色不同**

- 第一类常见的节点角色，即所谓的“意见领袖”（Opinion Leader）

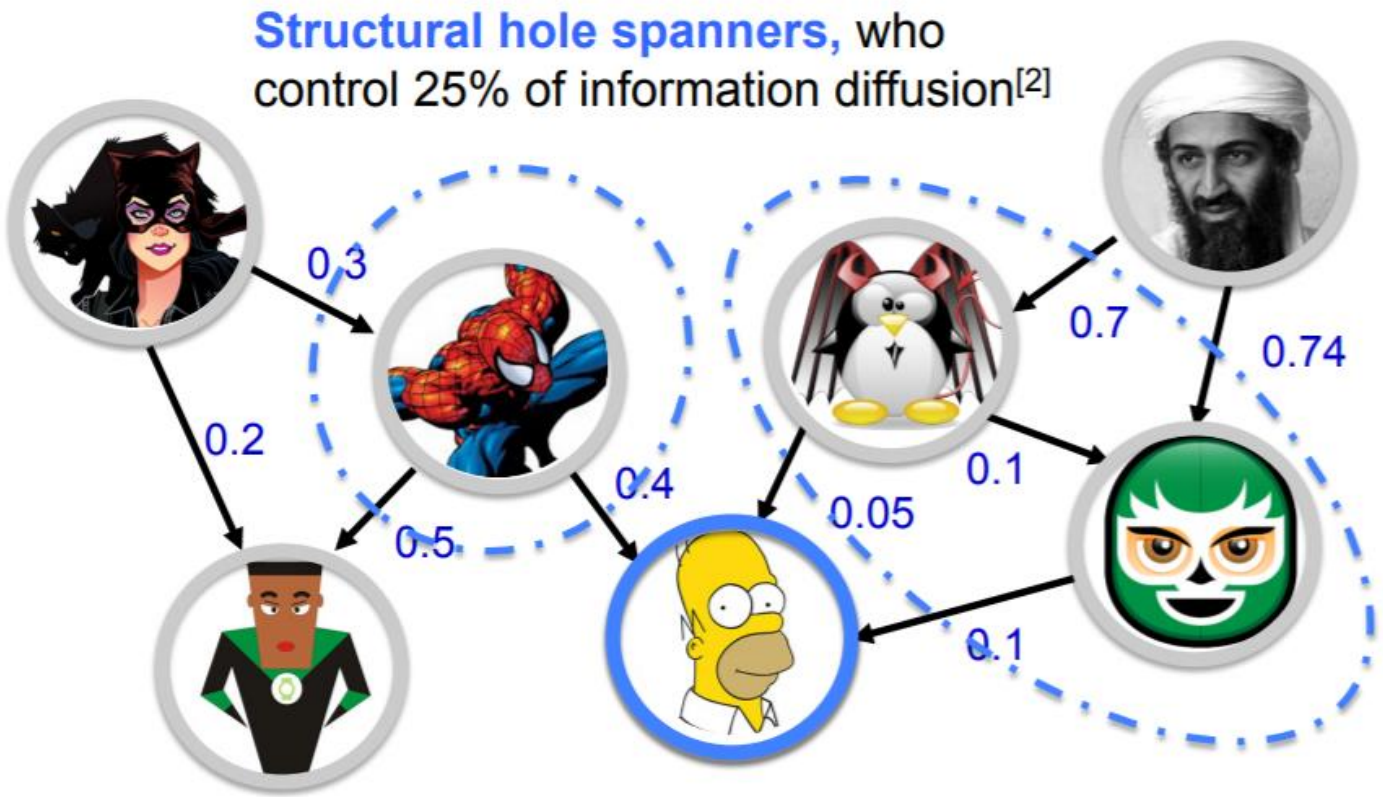


- **节点的角色不同**

- 意见领袖的挖掘，可参考HITS算法及两类网页的区分
  - 权威（Authority）网页与枢纽（Hub）网页的区分
  - 权威网页：指某个领域或某个话题相关的高质量网页
    - 如科研领域的中科院之声，视频领域的优酷与爱奇艺等
  - 中心网页：类似中介，指向了很多高质量的权威网页
    - 如“hao123”，各个浏览器自带的首页（手动滑稽）

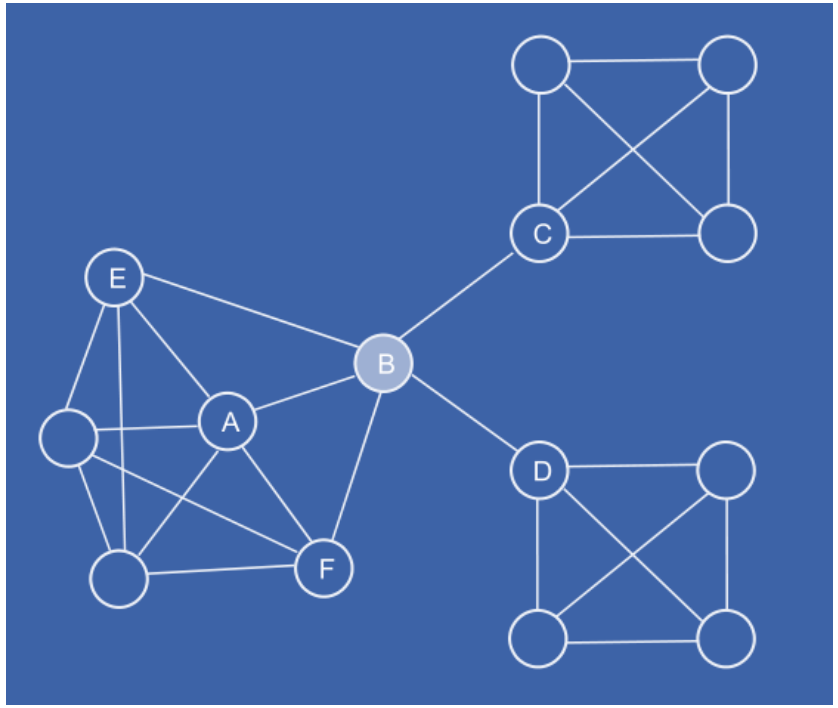
- **节点的角色不同**

- 第二类常见的节点角色，即所谓的“结构洞”（Structural Hole）



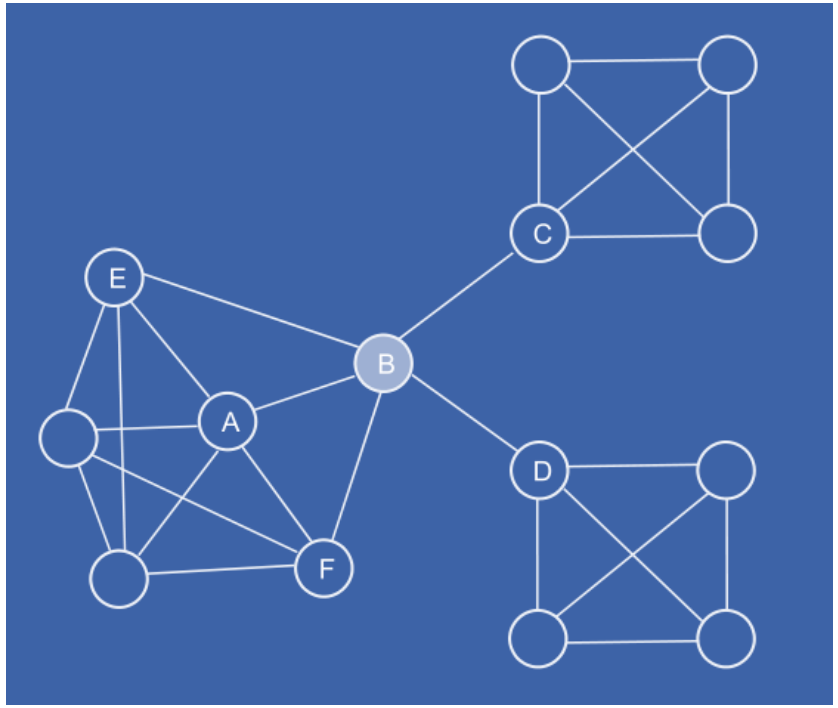
- **节点的角色不同**

- 结构洞的主要作用，在于为组织引入外部的信息
  - 例如，不同部门、不同社团之间的信息沟通：“项目经理”的角色
  - 因此，一种直观的结构洞的判定方法为：
    - 如果一个节点，移除该节点就会使网络变成多个连通组件，则该节点即为一个结构洞



- **节点的角色不同**

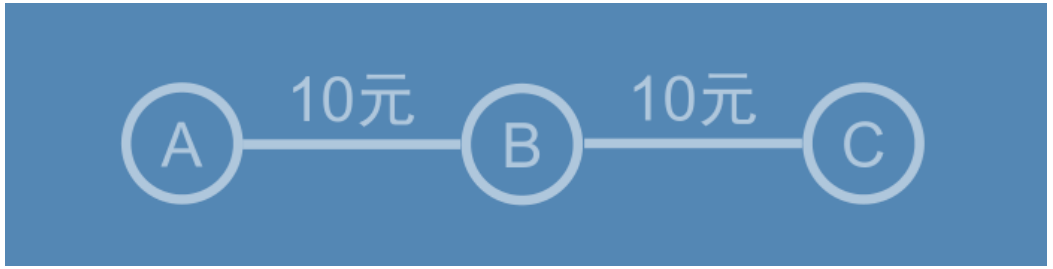
- 结构洞的主要作用，在于为组织引入外部的信息
  - 通过连通组件判定，假设可能过强，计算也较为不便
  - 另一种衡量方式：聚集系数
    - 某个节点的聚集系数为：它的任意两个好友也互为好友的概率（比重）
    - 显然，聚集系数越低，该节点作为中介的作用也就越大



- **节点的角色不同**

- 结构洞的“意义”：各方沟通的桥梁
  - 相应的，这种“意义”也成为了结构洞的“权力”
    - 例如，网络交换实验中，结构洞可能获得更大收益
    - ➢ 体现出结构洞具有排他性，该权力来自于“选择的自由”

往往以平分收场



如果仅可与邻居交涉，则B收益更大

本节目录

- 基本概念
- 节点角色
- **社会网络中的传播**
  - **基本传播模型**
  - 传播最大化问题
  - 衍生传播问题

- **社会网络中广泛存在的相互影响**

- 现代版的“三人成虎”：你为什么在仰望天空?
  - Milgram（还记得他不？）在20世纪60年代的实验
    - ➢ 当一个人在街头仰望天空的时候，少数路人会停下来一起盯着天空看
    - ➢ 当15个人一起仰望天空的时候，45%的路人会停下来
      - ✓ 从众现象普遍存在!

- **信息级联现象**

- 信息级联（Information Cascade），直译为“信息瀑布”
  - “Cascade”，指像瀑布一样从高处倾泻下来
  - 这个词用来描述人们在信息流中的“从众”行为，非常生动形象
  - 信息级联现象，可以描述人们从他人行为中获取消息/进行决策的这一过程
    - 这个过程具有鲜明的“传播”特性

• 信息传播的元素

- 一般而言，我们将信息传播过程中涉及的元素归为以下三类：
  - 发送者（Sender），也称作信息源（Source）或“种子节点”（Seed）
    - 指在信息传递开始时拥有信息的那一小部分用户集合
  - 接收者（Receiver），指作为潜在传播目标的广大用户集合
    - 接收者集合的规模要远大于前者，且不同发送者的目标集合存在重叠
  - 媒介（Medium），指传播过程发生的平台
    - 例如，寻找红气球比赛中的社交媒体 / 论坛等

- **信息级联中的基本假设**

- 首先，信息级联发生在一张有向图上
  - 对于无向图，可以将其连边转化为双向边进行处理
  - 对于网络中的节点来说，这些边就是信息传递的媒介
- 其次，每个节点仅能将信息传递给与其直接相连的节点
  - 例如，大V可以将信息传递给其粉丝，但不能传递给未关注他/她的人
    - 信息传递的局部可达性！

- **信息级联中的基本假设**

- 特别需要注意的是，网络中节点的状态是二元（Binary）的
  - 激活（Active/Activated）：表示节点已经收到了这一信息
  - 未激活（Inactive）：表示节点尚未接收到这一信息
    - 不存在薛定谔的状态！

➢ 一个小问题：什么情况下算是激活？

- 接收到信息，并且尝试将信息传给别人才叫激活
  - 两个动作缺一不可（是否合理？）

- **信息级联中的基本假设**

- 特别需要注意的是，网络中节点的状态是二元（Binary）的
- 已激活的节点才具备激活其他节点的能力
  - 而且，激活能力有一定的**时限**！
    - 传播中存在着时间“轮次”的概念
    - 类似于传染病模型的设定
    - 这个设定是否普遍合理？如何确定一个合理的时限？
      - 后面我们会展示去除这一约束的特殊模型

- **信息级联中的基本假设**

- 特别需要注意的是，网络中节点的状态是二元（Binary）的
- 激活是**不可逆**的过程
  - 可以从未激活到激活，但不能从激活退回未激活
  - 这个约束又是否普遍合理？
    - 核心争议在于：是否接受信息的“二次传播”
    - 后面我们同样会展示去除这一约束的特殊模型

- **基本模型（1）独立级联模型**

- 独立级联模型（Independent Cascade Model）
  - “独立”体现在，每次激活都是一次独立事件，相互不产生影响
    - 激活的尝试相当于一次以特定概率抛硬币的过程
  - 同时，每个已激活节点，只有一次机会尝试激活他/她的未激活邻居节点
    - 一旦尝试失败，不会再有第二次尝试机会

- **基本模型（1）独立级联模型**

- 独立级联模型（Independent Cascade Model）中的重要概念：轮次
  - 如果某个节点在第 t 轮被激活，那么，他仅有一次机会，即仅能在 t+1 轮，尝试激活他所有未被激活的邻居节点
    - t = 1时，仅有种子节点可以尝试激活其他节点
  - 对于节点 v 而言，他激活邻居节点 w 的概率采用 P_{vw} 表示
    - 以 P_{vw} 为概率进行抛硬币
  - 整个传播过程直到所有节点都被激活，或没有新节点可以被激活为止

- **基本模型（1）独立级联模型**

- 独立级联模型（Independent Cascade Model）中的重要概念：轮次
  - 对于节点 v 而言，他激活邻居节点 w 的概率采用 P_{vw} 表示
    - 以 P_{vw} 为概率进行抛硬币
    - P_{vw} 的取值方式：
      - 基本传播模型里，为简化考虑，一般将P_{vw}设为1/N，N为w节点的入边的数量
      - 当然，也有实现确定带权图的做法（如后续的例子）
      - 此外，也可以基于主题等因素对P_{vw}进行扩展

- **基本模型（1）独立级联模型**

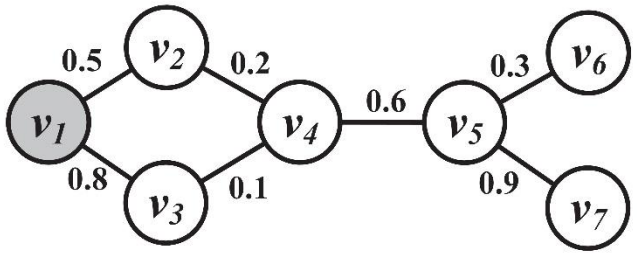
Algorithm 7.1 Independent Cascade Model (ICM)

Require: Diffusion graph $G(V, E)$, set of initial activated nodes A_0, activation probabilities $p_{v,w}$

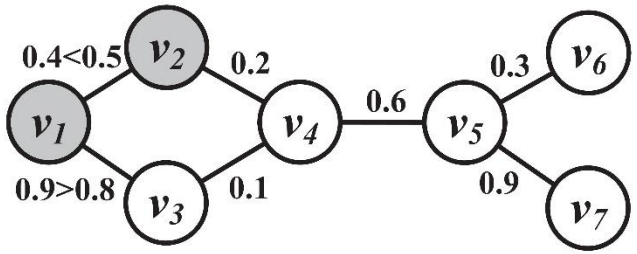
return Final set of activated nodes A_∞
i = 0;
while $A_i \neq \{\}$ **do**

  $i = i + 1$;
  $A_i = \{\}$;
  for all $v \in A_{i-1}$ **do**
    for all w neighbor of v, $w \notin \cup_{j=0}^{i} A_j$ **do**
      rand = generate a random number in [0,1];
      if rand < $p_{v,w}$ **then**
        activate w;
        $A_i = A_i \cup \{w\}$;
      end if
    end for
  end for
end while
$A_\infty = \cup_{j=0}^{i} A_j$;
Return A_∞;

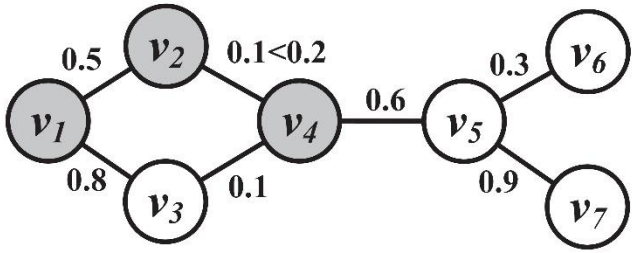
- **基本模型（1）独立级联模型**



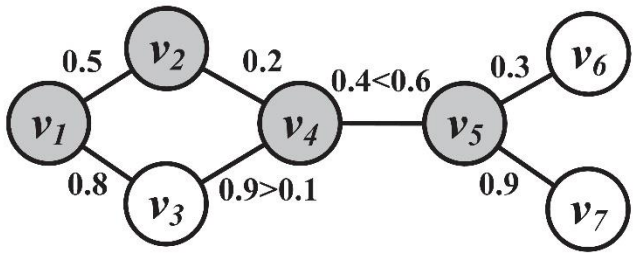
Step 1



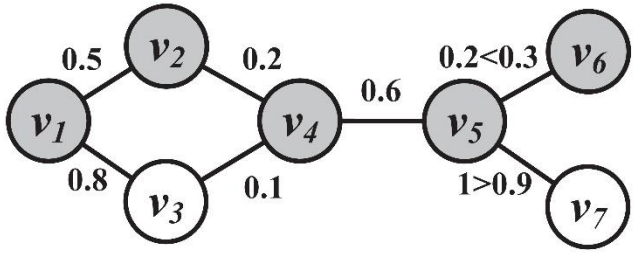
Step 2



Step 3



Step 4



Step 5

- **基本模型（2）线性阈值模型**

- 线性阈值模型（Linear Threshold Model）
  - 另一种视角：将信息传递过程视作多人影响的叠加过程
  - 一个用户会被某个信息激活，如果来自他已激活邻居的影响超过某个阈值

$$\sum_{v_j \in N_{\text{in}}(v_i), v_j \in A_{t-1}} w_{j,i} \geq \theta_i$$

    - 阈值预先设定，往往为从[0,1]均匀分布中随机抽取的一个数值
      - 更复杂的情况：可以根据用户对信息的兴趣等决定

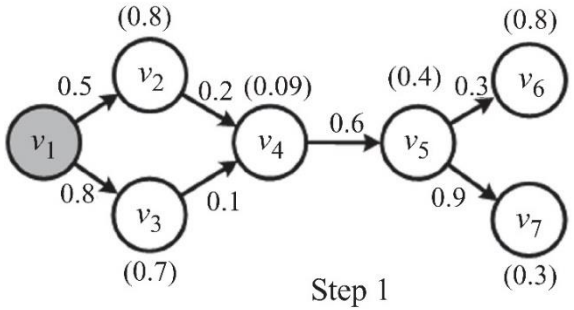
- **基本模型（2）线性阈值模型**

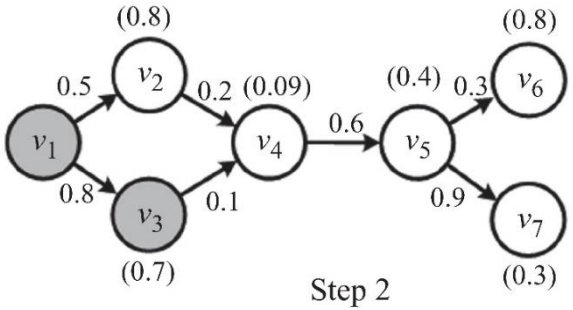
Algorithm 8.1 Linear Threshold Model (LTM)

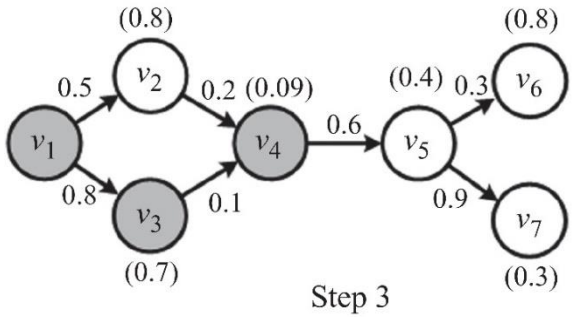
Require: Graph $G(V, E)$, set of initial activated nodes A_0

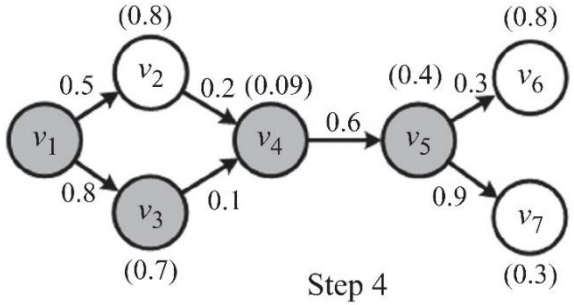
1: **return** Final set of activated nodes A_∞
2: i=0;
3: Uniformly assign random thresholds θ_v from the interval [0, 1];
4: **while** $i = 0$ or $(A_{i-1} \neq A_i, i \geq 1)$ **do**
5: $A_{i+1} = A_i$
6: inactive $= V - A_i$;
7: **for all** $v \in$ inactive **do**
8: **if** $\sum_{j \text{ connected to } v, j \in A_i} w_{j,v} \geq \theta_v$. **then**
9: activate v;
10: $A_{i+1} = A_{i+1} \cup \{v\}$;
11: **end if**
12: **end for**
13: $i = i + 1$;
14: **end while**
15: $A_\infty = A_i$;
16: Return A_∞;

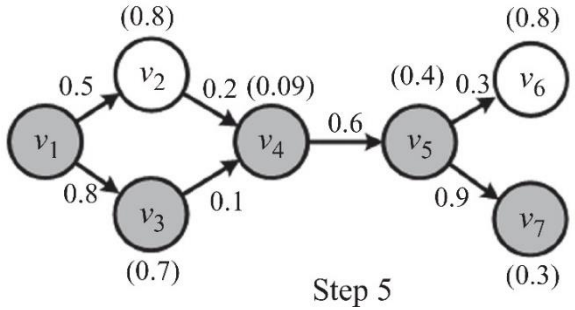
- **基本模型（2）线性阈值模型**











• 基本模型（2）线性阈值模型

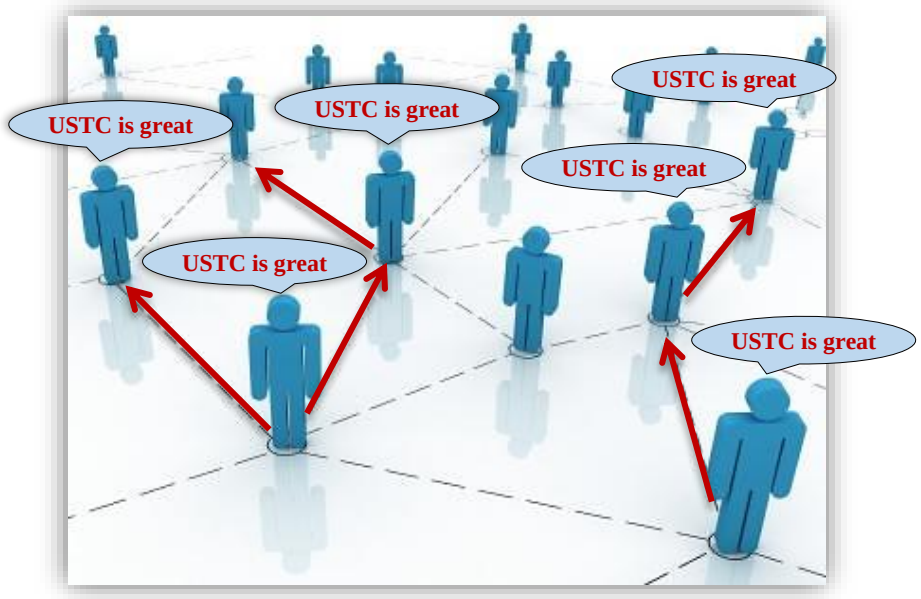
- 线性阈值模型与独立级联模型的区别：随机性
  - 对于独立级联模型来说，其随机性在于抛硬币的过程
    - 因此，独立级联模型是完全随机过程，每一次的结果可能都不相同
    - 一般需要重复多次以确定个体节点被激活的可能性
  - 对于线性阈值模型来说，其随机性在于边权重/阈值的确定
    - 如果采用启发式方法确定边权/阈值，则该方法结果完全由方法设计决定
    - 一旦确定边权/阈值（无论何种方式），其结果具有唯一性

本节目录

- 基本概念
- 节点角色
- **社会网络中的传播**
  - 基本传播模型
  - **传播最大化问题**
  - 衍生传播问题

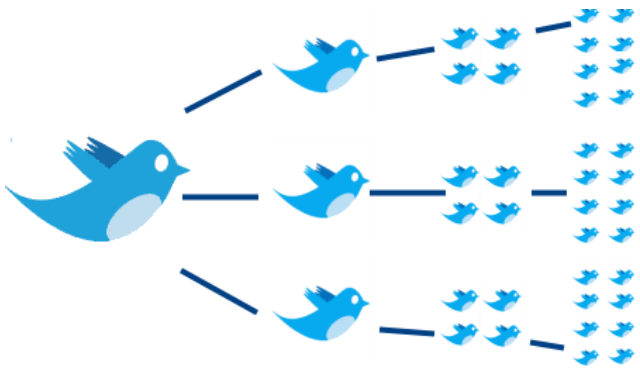
- **社会网络中的信息传播**

- 口碑营销（Word of Mouth）
  - 信息传播的过程，核心在于信息的接受
    - 传统的信息传播建模，往往将信息传播与信息接受合二为一，传播即视作接受了信息
    - 因此，信息传播分析在市场营销领域有着大量的应用



- **信息传播最大化问题**

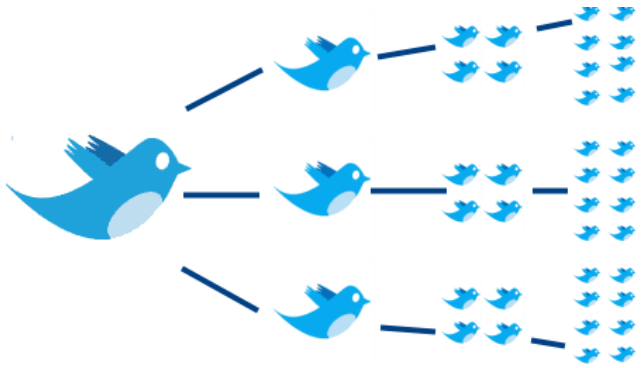
- 为什么会有传播最大化问题?
  - 口碑营销的常见应用：通过优惠来吸引潜在客户
    - 例如，通过发放优惠券 / 赠品的方式来扩大客户群
  - 然而，商家的预算是有限的
    - 因此，往往仅能通过收买少数用户来扩散消息
    - 这个时候，选择目标用户就至关重要!



• 信息传播最大化问题

- 传播最大化问题（Propagation Maximization）的定义
- 假定初始的种子节点集合为 S，预期激活的节点集合为 *f(S)*
- 信息传播最大化的目的，在于限定 S 集合的前提下，最大化 *f(S)* 的规模
  - 常见的约束为限定 S 集合的规模，即 |S|
  - 如果 S 集合中的节点价值不等，则可将约束进一步扩展

Y Yang, et al., Continuous Influence Maximization: What Discounts Should We Offer to Social Network Users?, SIGMOD 2016



- **信息传播最大化的启发式方法**

- 解决传播最大化问题的启发式思路：寻找网络中最具影响力的节点
  - 例如，如果你想宣传自家商品，找网红带货是个不错的手段
  - 找到影响力节点后，由他们发起信息传播
  - 问题在于，谁是最具影响力的节点?

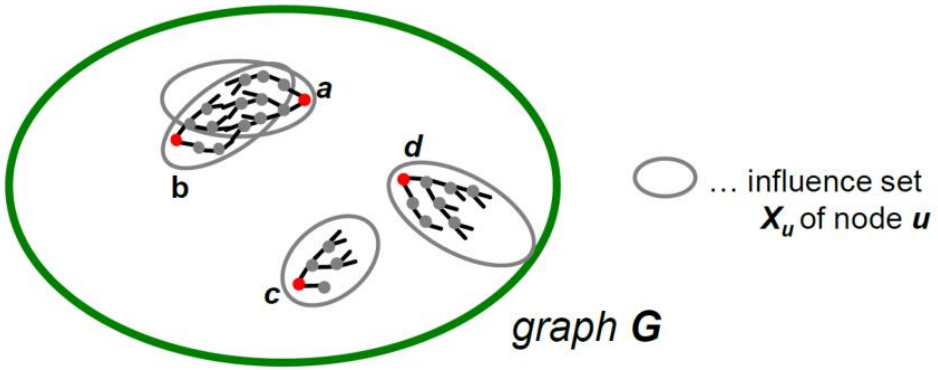
- **信息传播最大化的启发式方法**

- 启发式方法（1）PageRank及其衍生模型
  - 在网页排序部分，我们曾经介绍过，PageRank可用来衡量网页权威性
    - 因此，PageRank及其各种衍生算法如HITS都可以采用

- 启发式方法（2）核心性（Centrality）度量
  - 用于衡量网络中最重要的节点。常见核心性度量如度（Degree）、紧密度（Closeness）、介数（Betweenness）等

- **信息传播最大化的启发式方法**

- 启发式方法（3）计算单个节点所能够激活的邻居数量，再进行排序
- 上述启发式方法，在寻找“最具影响力的节点”时可行
- 然而，在确定影响力节点集合时不可行
  - 节点的影响范围之间可能存在重叠
    - 在单个节点影响力够强的情况下，没必要重叠“双保险”



- **一般化的信息传播最大化方法**

- 基于前述的传播模型（例如：ICM / LTM），对于给定 S 集合计算 *f(S)*
- 此时，传播最大化问题可转化为一个带约束的最优化问题
  - 以 *f(S)* 为目标函数，找到一个集合 S，使得 *f(S)* 的期望最大
  - 同时，S 应满足 |S| = k （即只选择 k 个节点作为初始节点集合）

$$\max_{S \text{ of size } k} f(S)$$

  - 注意，是期望最大（因为整个过程是随机过程）

- **一般化的信息传播最大化方法**

- *f(S)*的一些有趣的性质：子模特性（Submodularity）
- 1. *f(S)* 函数是非负的（显而易见）
- 2. *f(S)* 函数是单调非减的，即 $f(S+v) \geq f(S)$
  - 也很好理解，新增加一个节点，至多不增加新激活，不至于减少
- 3. *f(S)* 函数是具有子模特性（Submodularity）的，即：
  - 对于任何集合对S，T，且满足S ⊆ T时，给定节点 v，有

$$f(S \cup \{v\}) - f(S) \quad \geq \quad f(T \cup \{v\}) - f(T),$$

- **信息传播最大化问题的解决方案**

- 基于前述问题定义，我们有坏消息，也有好消息
- 坏消息：在ICM / LTM等模型定义下，传播最大化问题是个NP难问题
  - 简要证明思路：将这两个模型定义下的传播最大化问题，归约为集合覆盖（Set Cover）和节点覆盖（Vertex Cover）问题
    - 详细证明可参见如下论文：

D. Kempe, et al., Maximizing the Spread of Influence through a Social Network, KDD 2003

- **信息传播最大化问题的解决方案**

- 基于前述问题定义，我们有坏消息，也有好消息
- 好消息：由于 $f(S)$ 函数具有子模特性，我们可以采用贪心算法近似求解
  - 以空集合为起点，即初始 S = ∅
  - 经过 k 次迭代，每次选择最大化 $f(S \cup \{v\}) - f(S)$ 的节点 v
  - 效果如何？论文证实贪心算法可以实现至少 $(1 - 1/e)$ 的近似效果
    - 这就意味着，贪心法所得 S 可以激活至少 63% 最优解能激活的节点数

G. Nemhauser, et al. An analysis of the approximations for maximizing submodular set functions. Mathematical Programming, 14(1978), 265–294.

- **信息传播最大化问题的解决方案**

- 基于贪心算法的传播最大化问题求解伪代码：

Algorithm 7.2 Maximizing the spread of cascades – Greedy algorithm

Require: Diffusion graph $G(V, E)$, budget k

1: **return** Seed set S (set of initially activated nodes)
2: $i = 0$;
3: $S = \{\}$;
4: **while** $i \neq k$ **do**
5: $v = \arg\max_{v \in V \setminus S} f(S \cup \{v\})$;
or equivalently $\arg\max_{v \in V \setminus S} f(S \cup \{v\}) - f(s)$
6: $S = S \cup \{v\}$;
7: $i = i + 1$;
8: **end while**
9: Return S;

本节目录

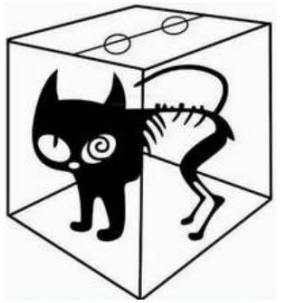
- 基本概念
- 节点角色
- **社会网络中的传播**
  - 基本传播模型
  - 传播最大化问题
  - **衍生传播问题**

- **独立级联模型的局限性**

- 独立级联模型具有易于求解，假设直观的优点，但也存在一些缺点
  - 小问题 1：每个节点只有一次传播信息的机会，是否过于苛刻？
    - 实际情况下，只要信息还在，就可以持续输出影响
  - 小问题 2：节点状态未必二元化，也难以获得清晰明确的激活轮次
    - ICM 适合类似微博等具有明确转发记录的场景
    - 然而，很多场景下并没有明确的信息传播轨迹
      - 节点是否真的被激活？何时被激活？无法回答

- **稳定状态传播模型**

- 独立级联模型的松弛版本：稳定状态传播（Steady State Spread，SSS）
  - 对ICM的改动体现在以下两点
    - 节点状态不再二元化，而是引入一个变量表示当前被激活的概率
      - 薛定谔的节点出现了！
    - 如果被激活概率不为0，则节点可以持续对外输出信息 / 影响



C Aggarwal, et al., On Flow Authority Discovery in Social Networks, SDM 2011

- **稳定状态传播模型**

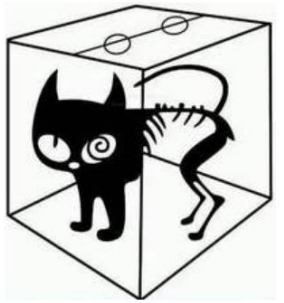
- 独立级联模型的松弛版本：稳定状态传播（Steady State Spread，SSS）
  - 稳定状态传播模型的核心公式：

$$1-\pi(i)=\prod_{l\in N(i)}(1-\pi(l)\cdot p_{li})$$

注意此处不同

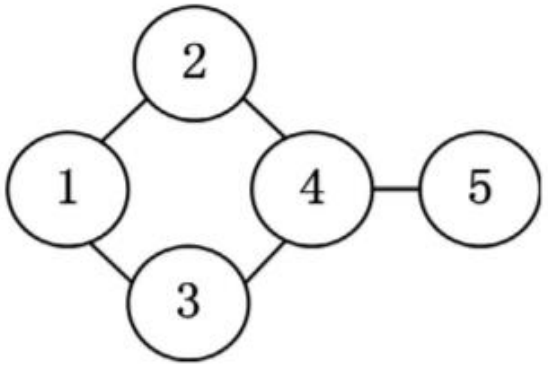
$\Pi(l)$ 表示 l 节点的当前状态，即被激活的概率

当 $\Pi(l)=0$时，显然不影响邻居节点激活状态

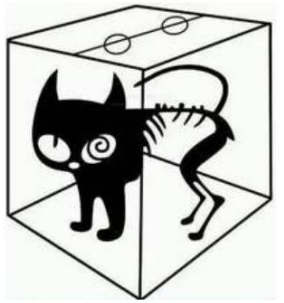


- **稳定状态传播模型**

- 稳定状态传播的一个隐患：不合理的反向传播
  - 我们观察形如下图的网络结构，假设信息源为节点 1

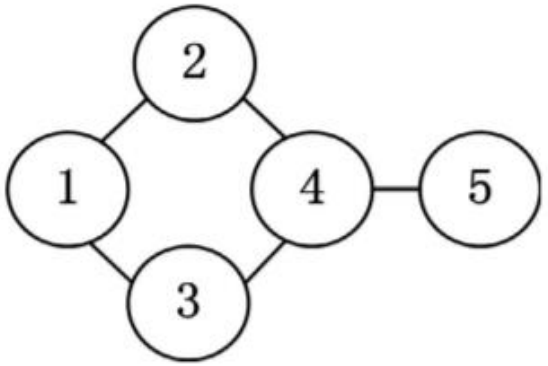


  - 在此种情况下，信息只可能途经节点 4 到达节点 5
  - 在ICM中，节点5被激活意味着节点4已被激活，但SSS呢?

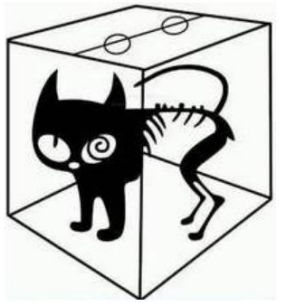


- **稳定状态传播模型**

- 稳定状态传播的一个隐患：不合理的反向传播
  - 由于SSS只有薛定谔的激活状态，节点4、5都只有激活概率

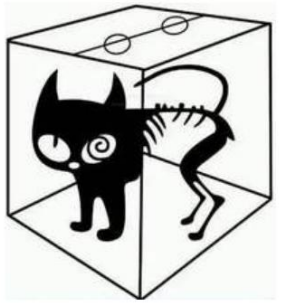


  - 在此种情况下，将出现节点 5 反向提升节点 4 激活概率的现象
    - 然而，这在道理上是说不通的（子节点激活父节点？）



- **稳定状态传播模型**

- 稳定状态传播的一个隐患：不合理的反向传播
  - 一种解决方法是遍历网络，找到所有类似这样的子节点
    - 但这种做法显然过于费时费力，且结构特性难以判断
  - 另一种：启发式方法，限制每个节点的迭代次数，超过一定轮次阈值，则该节点被激活的概率 $\pi(i,t)$ 不再更新
    - 阈值大致根据节点到信息源的最长 / 短路径决定



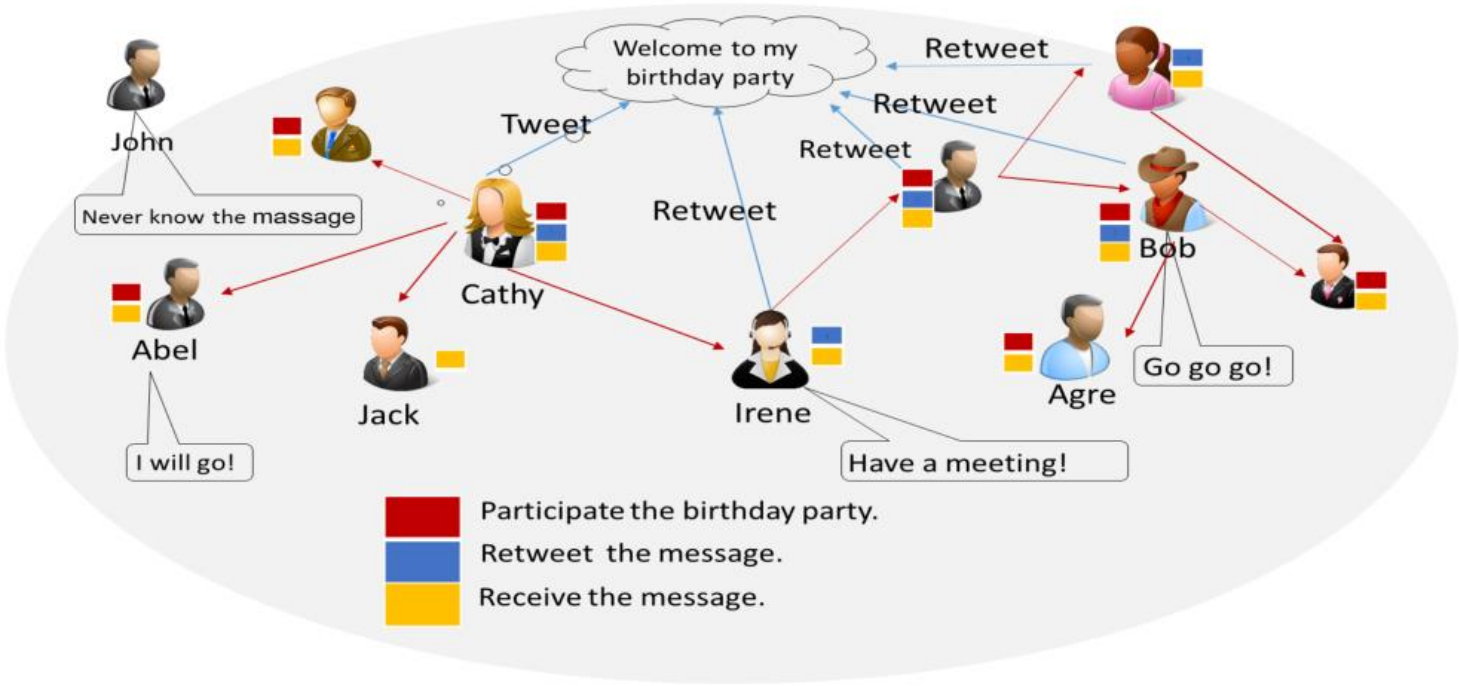
Y Yang, et al., On Approximation of Real-World Influence Spread, ECML-PKDD 2012

- **信息级联的另一个小问题**

- 独立级联模型在场景应用上还有另一个局限性：信息传播与接收的捆绑
  - 回想一下，ICM / LTM等模型在假设有一个前提，一个转发行为 = 传播 + 接受，两者缺一不可
    - 但实际上，两者不可混为一谈
    - 例如，有些人可能看到并接受了信息，但由于种种原因并没有转发；而另一些转发的人可能实际上并没有接受信息

- **信息级联的另一个小问题**

- 信息传播与接收不一定捆绑的一个实例



- **信息级联的衍生模型**

- 实现信息传播与接收的解绑，先从目标函数改起
  - 第一种思路：考虑信息覆盖问题，即信息覆盖了多大的人群
  - 核心假设：如果某个节点被激活，那么他的所有邻居都被覆盖
  - 由此，衍生出信息覆盖最大化问题

$$\arg\max_{S} \; F(S) = E(|I(S)|) + E(|\bigcup_{a \in I(S)} N(a)|)$$

$$s.t. \quad |S| = k$$

邻居部分

- **信息级联的衍生模型**

- 信息覆盖最大化问题同样具有信息传播最大化问题的性质
  - 因此，可以采用类似的贪心算法加以求解
  - 详细证明与算法可参见如下论文：

    Z Wang, et al., Maximizing the Coverage of Information Propagation in Social Networks, IJCAI 2015

  - 然而，这篇论文仍然有个较强假设：收到信息 = 接受信息
    - 实际上，大多数我们收到的信息都被忽略了

- **信息级联的衍生模型**

- 更进一步实现信息传播与接收的解绑，修改模型框架
  - 第二种思路：单独对信息接受过程进行建模

$$F(S) = \mathbf{Adopt}(S) = \sum_{u \in V} [f_u(A_u)]$$

  - 其中，引入函数 $f_u(A_u)$，描述 u 节点接受信息的概率
    - 显然，这一概率与有多少个邻居已经接受了信息正相关
    - A_u 的定义：$u \cup N^{in}(u) \cap \mathbf{Active}(S)$

- **信息级联的衍生模型**

- 更进一步实现信息传播与接收的解绑，修改模型框架
  - 第二种思路：单独对信息接受过程进行建模
  - 由此，衍生出第二个新问题：信息接受最大化问题
    - 即，什么种子节点集合会导致接受信息的节点数期望最大
    - 该问题在$f_u(A_u)$符合一定特性时，同样与信息传播最大化具有类似属性
    - 相关详细证明与算法可参见如下论文：

T Jin, et al., Maximizing the Effect of Information Adoption: A General Framework, SDM 2018

- **信息级联的衍生模型**

- 事实上，信息接受最大化问题可视作一个更为一般化的框架
  - 当满足如下条件时，该问题可退化为信息传播最大化问题：

$$f_v(A_v) = \begin{cases} 1 & \text{if } v \in A_v \\ 0 & \text{if } v \notin A_v. \end{cases}$$

  - 当满足如下条件时，该问题可退化为信息覆盖最大化问题：

$$f_v(A_v) = \begin{cases} 1 & \text{if } A_v \neq \emptyset \\ 0 & \text{if } A_v = \emptyset. \end{cases}$$

本章小结

社会网络

- 社会网络的基本元素与概念
- 常见的特殊节点角色：意见领袖、结构洞
- 基本传播模型
  - 独立级联模型
  - 线性阈值模型
- 传播最大化问题
- 衍生传播模型与传播问题

tongxu@ustc.edu.cn