Motivation

识别网络中有影响力的节点具有重要的实际用途。一个很好的例子是“病毒式营销”，该策略利用现有的社交网络来传播和推广产品。精心设计的病毒标记产品将识别出最具影响力的客户，说服他们采用并认可该产品，然后像病毒一样在社交网络中传播该产品。

影响力最大化的关键是如何找到最具影响力的节点集？为了回答这个问题，我们首先来看两个经典的级联模型：

- 线性阈值模型
- 独立级联模型

然后，我们将开发一种方法来找到独立级联模型中最具影响力的节点集。

Linear Threshold Model

在线性阈值模型中，我们具有以下设定：

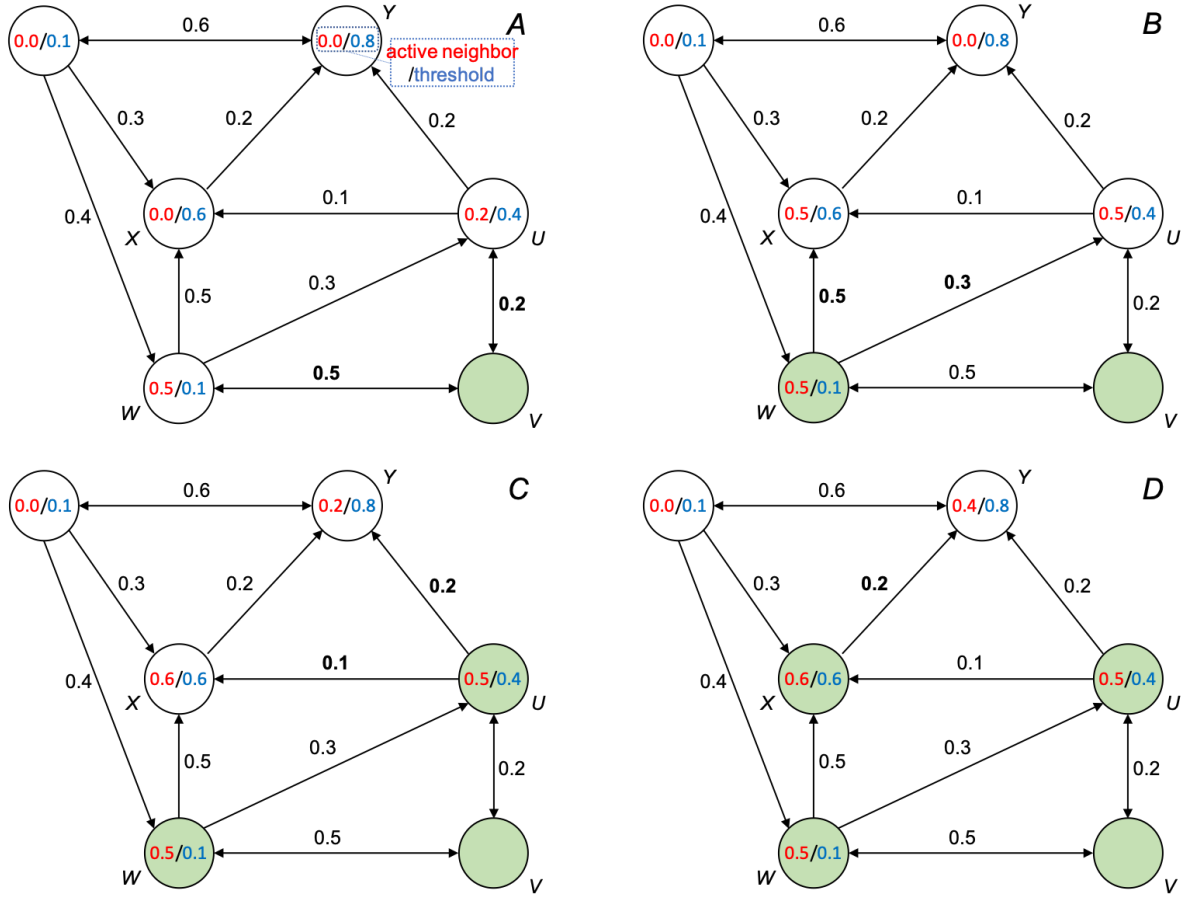
- 一个节点 v 具有随机阈值 $\theta_v \sim U[0,1]$
- 一个节点 v 受每个邻居 w 的影响 ，基于节点 v 和 w 之间的权重 $b_{v,w}$，且

$$\sum_{w \text{ neighbor of v}} b_{v,w} \leq 1$$

- 一个节点 vv 当至少 θvθv其邻居的一部分是活跃的。那是
- 当节点 v 至少 θ_v 的邻居活跃时，节点 v 才活跃，即，

$$\sum_{w \text{ active neighbor of v}} b_{v,w} \geq \theta_v$$

下图演示了该过程：



(A) 节点 V 被激活，且对 W 和 U 的影响分别为0.5和0.2 is； (B) W 被激活，对 X 和 U 的影响分别为0.5和0.3； (C) U 被激活并分别以0.1和0.2影响 X和 Y；(D) X 被激活并以0.2影响 Y，此时不能再激活任何节点；过程停止。

Independent Cascade Model

在此模型中，我们根据有向图中的概率对节点的影响（激活）进行建模：

- 给定有限图 $G = (V, E)$
- 从新的行为开始（例如采用新产品，我们说它们是活跃的）给定一个节点集 S
- 每条边 (v, w) 具有概率 p_{vw}
- 如果节点 v 被激活, 则有机会利用概率 p_{vw} 1去激活节点 w
- 通过网络传播激活

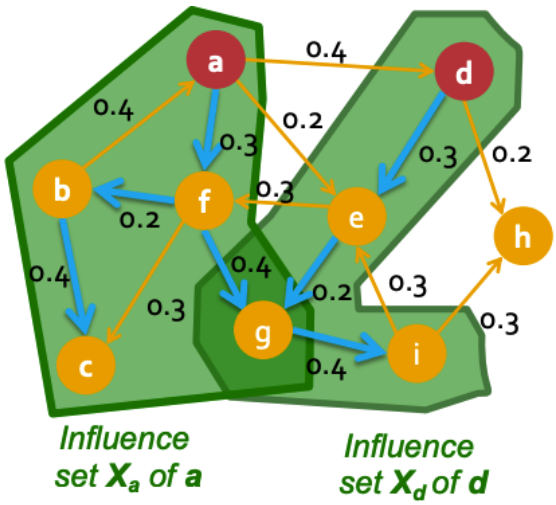
注意:

- 没个边仅建立一次
- 如果 u 和 v 都处于激活状态并且与 w 相连，哪个节点先去激活 w 并不重要

Influential Maximization (of the Independent Cascade Model)

定义

- **最具影响力的集合大小** k (k 用户定义参数) 是包含 k 个节点的集合 S 。这些节点如果被激活则会产生最大预期级联大小 $f(S)$。[为什么是"预期的级联大小"？ 由于独立级联模型的随机性，节点激活是一个随机过程，因此，$f(S)$ 是一个随机变量。在实践中，我们通常计算许多的随机模拟从而获得期望值 $f(S) = \frac{1}{|I|} \sum_{i \in I} f_i(S)$，其中 I 表示一组模拟]
- **节点 u 的影响集** X_u 是最终将被节点 u 激活的节点集合，示例如下所示



红色节点 a 和 b 都处于激活状态，两个绿色的区域分别表示由节点 a 和 b 激活的节点集，比如 X_a 和 X_b。

注意：

- $f(S)$ 是集合 X_u 的并集，即: $f(S) = |\cup_{u\in S} X_u|$。
- 如果 $f(S)$ 越大，表示集合 S 更具有影响力

问题设定

那么，有影响力的最大化问题就是一个优化问题：

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

这个问题是一个NP-hard 问题[Kempe et al. 2003]。但是，有一个贪婪近似算法（**Hill Climbing**）可以为 S 给出以下近似保证的解：

$$f(S) \geq (1-\frac{1}{e})f(OPT)$$

其中 OPT 是全局最佳解。

Hill Climbing

算法： 在每一步 i，激活并选择具有最大边际收益 $\max_u f(S_{i-1}\cup\{u\})$ 的节点 u：

- 初始 $S_0 = \{\}$
- 对于 $i = 1...k$
  - 激活节点 $u \in V \setminus S_{i-1}$ 且 $\max_u f(S_{i-1}\cup\{u\})$
  - 令$S_i = S_{i-1}\cup\{u\}$

注意： Hill Climbing 产生具有近似保证的解 $f(S) \geq (1-\frac{1}{e})f(OPT)$.

Hill Climbing 的逼近证明

单调性定义： 如果 $f(\emptyset) = 0$ 且对于所有 $S \subseteq T$，有 $f(S) \leq f(T)$，则称 $f(\cdot)$ 是单调的。

Submodular定义： 如果对于任何节点 u 和任何集合 $S \subseteq T$，有 $f(S\cup\{u\}) - f(S) \geq f(T\cup\{u\}) - f(T)$，则称 $f(\cdot)$ 是Submodular。

定理[Nemhauser et al. 1978]:（也可以看这个讲义），如果 $f(\cdot)$ 是**单调**的且为 **submodular**，则 S 可以通过贪心地添加 k 个元素最大化边际收益满足以下条件而获得：

$$f(S) \geq (1-\frac{1}{e})f(OPT)$$

基于这个定理，我们需要证明最大期望级联函数 $f(\cdot)$ 是**单调**的且为 **submodular**。

显然，函数 $f(\cdot)$ 是单调的（如果没有节点是激活的，则影响力为0，即 $f(\emptyset) = 0$，因为激活更多的节点将不会损害影响力，因此：若 $U \subseteq V$ 则 $f(U) \leq f(V)$）。所以我们只需要证明 $f(\cdot)$ 是 **submodular**。

Fact 1 of Submodular Functions: $f(S) = |\cup_{k \in S} X_k|$ 是 submodular, 其中 X_k 是一个集合。直观上，已经拥有的集合越多，将会有更少的新区域为 X_k 提供新的节点。

Fact 2 of Submodular Functions: 如果 $f_i(\cdot)$ 是 submodular 且 $c_i \geq 0$，那么 $F(\cdot) = \sum_i c_i f_i(\cdot)$ 也同样是 submodular。即submodular 函数的非负线性组合仍然是 submodular 函数。

证明 $f(\cdot)$ is Submodular： 我们在图 G 上运行了许多次模拟，对于第 i 次世界的模拟，节点 v 具有激活集 X_v^i，且 $f_i(S) = |\cup_{v \in S} X_v^i|$ 是集合 S 的级联大小。基于 **Fact 1**，$f_i(\cdot)$ 是submodular。由于 **Fact 2**，期望影响集 $f(S) = \frac{1}{|I|} \sum_{i \in I} f_i(S)$ 同样也是submodular。

在实践中评估 $f(S)$ 和 Hill Climbing 的近似保证： 如何评估 $f(S)$ 仍然是一个开放性的问题，通过对可能的世界进行多次模拟从而获得的估计值是一个很好的评估方法 [Kempe et al. 2003]：

- 通过反复模拟估计 $\Omega(n^{\frac{1}{\epsilon}})$ 个可能的世界评估 $f(S)$，其中 n 是节点的个数，ϵ 是一个很小的正实数
- 能够达到 $f(S)$ 的 $(1 \pm \epsilon)$ 逼近
- Hill Climbing 现在是一个 $(1 - \frac{1}{e} - \epsilon)$ 逼近算法

Speed-up Hill Climbing by Sketch-Based Algorithms

Hill Climbing的时间复杂度

寻找节点 u 使得 $\max_u f(S_{i-1} \cup \{u\})$ (请参见 以上算法):

- 我们需要评估每个大小为 $O(n)$ （n 是图 G 中节点的数量）的节点的影响集 X_u
- 对于每一轮评估，需要花费 $O(m)$ 的时间对所有涉及的边进行翻转（m 是图 G 中的边数)
- 我们还需要 R 来模拟来估算影响集 （R 是模拟次数/可能的世界数）

我们将需要这样做 k 次（k 为要选择的节点数）。因此，Hill Climbing 的时间复杂度是 $O(k \cdot n \cdot m \cdot R)$，这显然会很慢。我们可以使用 **sketches** [Cohen et al. 2014] 通过将评估时间从 $O(m)$ 降到 $O(1)$ 来加速评估 X_u。（除了 sketches 外，还有一些其他的建议方法可以有效地评估影响函数：approximation by hypergraphs [Borgs et al. 2012], approximating Riemann sum [Lucier et al. 2015], sparsification of influence networks [Mathioudakis et al. 2011], and heuristics, such as degree discount [Chen et al. 2009].)

单一可到达性Sketches

- 创造一个可能的世界 G^i (比如，图 G 的一种模拟使用独立的级联模型)
- 给每个节点一个统一的随机数 $\in [0, 1]$
- 计算每个节点 v 的**rank**，表示节点 v 在这个世界中可以到达的最小节点数

直观上说：如果 v 能到达大量的节点，那么他的**rank**就可能很小。因此，节点 v 的 **rank** 可以用来估计节点在 G^i 中的影响。

但是，基于单一可达性Sketches的影响估计（即单个对 G 的模拟）是不准确的。为了做出更准确的估计，我们需要基于许多模拟来构建Sketches。（这类似于对 $f_i(S)$ 的平均，但是在这种情况下，可以通过使用组合可到达性Sketches来实现）

组合可到达性 Sketches

在组合可到达性 Sketches中，我们模拟了多个可能的世界，并在所有可能的世界中保持节点 u 可以到达的最小节点值 c。

- 构造组合可到达性Sketches:
  - 产生多个可能的世界
  - 对于节点 u，给所有 (v, i) 分配一个均匀分布的随机数 $r_v^i \in [0, 1]$，其中 v 是在世界 i 中节点 u 能够达到的节点
  - 将 c 最小的 r_v^i 作为组合可到达性Sketches
- 以贪心方式运行以最大化影响力：
  - 每当贪心算法选择影响力最大的节点时，选择在其Sketches中具有最小值的节点 u
  - 当 u 被选择后，找到它的影响力集合 X_u^i，标记 (v, i) 并从其他节点的Sketches重负删除 r_v^i

注意：使用组合可达性Sketches不能为真实的预期影响提供近似保证，而是针对所考虑的可能世界提供近似保证。