> 课程链接：CS224W: Machine Learning with Graphs
> 课程视频：【课程】斯坦福 CS224W: 图机器学习 (2019 秋 | 英字)

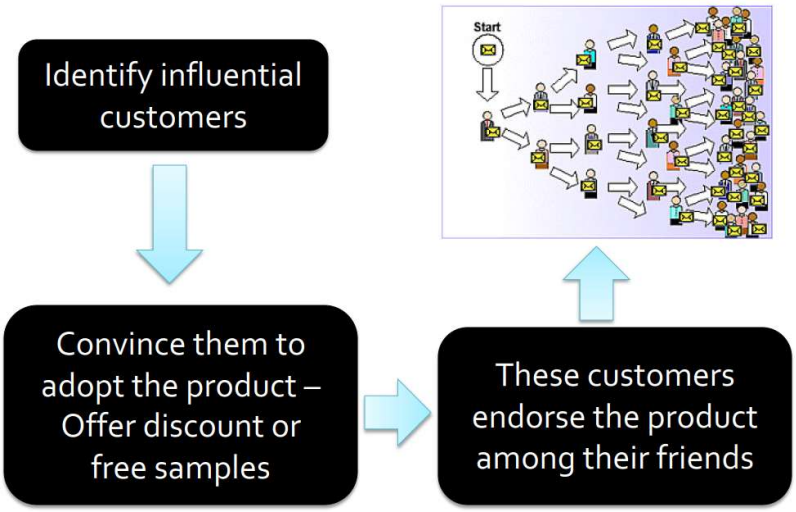
目录

这节课的内容理论性比较强，理解起来比较难。这里给两篇论文帮助理解：

- 社交网络中影响最大化的研究与应用
- Maximizing the Spread of Influence through a Social Network

1. 前言——Viral Marketing 病毒式营销

很多时候，我们都会因为朋友的安利而去购买某些产品。



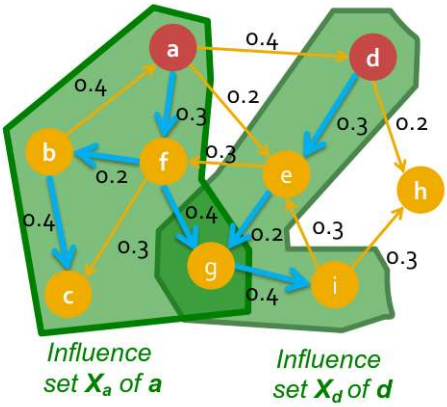
一个很形象的例子就是王妃效应(Kate Middleton effect)——凯特王妃的穿着往往会引领时尚圈的潮流。那么，我们怎么在一张网络中找到这样的Kate呢？

这个问题被定义为 **Influence Maximization（社交网络影响力最大化）** 问题——在给定的初始网络中给定初始活跃节点的个数k，影响力最大化问题即找到固定个数的活跃节点集，通过**特定的传播模型**进行传播，使得最终活跃节点数目达到最大化。

2. Influence Maximization问题

Problem: (*k* is a user-specified parameter)

- **Most influential set of size *k*:** set ***S*** of ***k*** nodes producing largest **expected** cascade size ***f(S)*** if activated [Domingos-Richardson '01]
- **Optimization problem:** $\max_{S \text{ of size } k} f(S)$



Why "expected cascade size"? X_a is a result of a random process. So in practice we would want to compute X_a for many random realizations and then maximize the "average" value $f(S)$. For now let's ignore this nuisance and simply assume that each node u influences a set of nodes X_u

$$f(S) = \frac{1}{|I|} \sum_{\text{Random realizations } i} f_i(S)$$

下面介绍两个经典的传播模型：

Linear Threshold Model 线性阈值模型

在线性阈值模型中，每个节点v都有一个影响阈值 $\theta_v \in U[0,1]$，这个阈值在0到1的范围内均匀、随机地选取，一旦确定在传播中就不再改变。

对于节点v来说，它的每个邻居节点w对它都由一个影响因素$b_{v,w}$，且$\sum_{w \text{ neighbour of } v} b_{v,w} \leq 1$。当周围邻居对该节点的影响超过它的影响阈值时，即$\sum_{w \text{ neighbour of } v} b_{v,w} \geq \theta_v$时，该节点被激活。

Independent Cascade Model 独立级联模型

- Directed finite $\boldsymbol{G = (V, E)}$
- Set $\boldsymbol{S}$ starts out with new behavior
  - Say nodes with this behavior are "**active**"
- Each edge $\boldsymbol{(v, w)}$ has a probability $\boldsymbol{p_{vw}}$
- If node $\boldsymbol{v}$ is active, it gets **one** chance to make $\boldsymbol{w}$ active, with probability $\boldsymbol{p_{vw}}$
  - Each edge fires at most once

Influence Maximization问题的难点

- **Problem: Most influential set of *k* nodes:**
  set ***S*** on ***k*** nodes producing largest expected cascade size ***f(S)*** if activated
- **The optimization problem:**

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

这个问题是一个NP问题。——Influence maximization is NP-complete。但是，我们可以用**近似算法**进行求解。

3. 爬山算法——基于次模函数的研究方法

我们先来看一下爬山算法（Greedy Hill Climbing algorithm）。爬山算法能够达到63%的精度。

- **Input:**
  Influence set X_u of each node u: $X_u = \{v_1, v_2, \dots\}$
  - That is, if we activate u, nodes $\{v_1, v_2, \dots\}$ will eventually get active
- **Algorithm:** At each iteration i activate the node u that gives largest marginal gain: $\max_u f(S_{i-1} \cup \{u\})$

贪心算法：$S: GA(G,k)$

输入：G 是传播消息的网络，用图来表示

k 是选择作为初始种子集合的大小

输出：S 是大小为k的初始种子集合

```
1  Set initial seeds set  S = ∅
2  while( |S| < k ){
3      For every node  v  in set  V − S {
4          Calculate  Δ_v = σ(S ∪ {v}) − σ(S)
5      }
6      Choose the node  v = arg max_v Δ_v  as the next seed
7  }
8  return  S
```

影响最大化的爬山算法是一种基于次模函数的研究方法——PPT中有详细的证明。或者可以看论文：An analysis of approximations for maximizing submodular set functions。

次模函数（Submodular functions）——一个集合函数，随着输入集合中元素的增加，增加单个元素到输入集合导致的函数增量的差异减小。次模函数满足下面这个公式：

$$f\left(S\cup\{v\}\right)-f\left(S\right)\geq f\left(T\cup\{v\}\right)-f\left(T\right)$$

> 它的每一次选择都能提供最大影响增量的节点，最后，通过一系列局部最优解的组合来得到全局最优解的近似值。它的优点是精确度相对较高，但是也存在着严重的效率问题，即算法时间复杂度高，执行时间长，导致在节点规模巨大的大规模社交网络中无法实际使用。——江禹. 社交网络中影响最大化的研究与应用[D].

4. 改进：Sketch-based Algorithms

参考论文：Sketch-based influence maximization and computation: Scaling up with guarantees

这部分我看起来确实很吃力，所以 - 推荐一份更认真的笔记：斯坦福CS224W图机器学习Lecture 14-influence学习笔记

以及，推荐一本书：大数据网络传播模型和算法