Adaptive Influence Maximization in Dynamic Social Networks

背景

随着信息科学的发展，社交网络成为交换观点和信息的传播平台。这体现在很多领域，如市场营销策略，人类行为分析和谣言阻塞等。为了表示传播过程，大量的模型被研究提出。其中最基本的两个传播模型线性阈值和独立级联，在该模型上有很多的相关工作关注影响力最大化问题。

但是这些传播模型没有考虑到传播过程中一些不确定的性质：在真实社交网络中种子节点不能确定被成功激活，用户之间的关系度数在频繁变化。这篇文章就是针对动态社交网络来设计影响力最大化传播模型。

解决问题

- 解决了现存的传播模型不能捕获现实社交网络中的动态方面的问题

创新之处

- 设计了新的传播模型动态独立级联模型（DIC）用于捕获真实社交网络中的动态方面。
- 基于DIC模型，构建了两个自适应种子选择策略的贪心算法A-Greedy和H-Greedy，来选取最优种子集合。

相关工作

- Kempe等人形式化的表述了影响力传播最大化的组合优化，提出了$1-1/e$近似率的贪心算法，基于IC和LT；Long和Wong从最小化观点研究该问题；Du等人提出了连续的传播模型来研究IM问题。
- 自适应种子选择策略是一种随机优化框架，Asadpour等人在独立随机变量集合的幂集上研究分析随机子模最大化问题；Golovin等人的研究表明贪心算法的自适应版本依然可以实现可证明的性能保证；Seeman等人考虑二阶段选择种子节点的策略。

解决方法

DIC模型

通常的离散传播模型，初始种子节点被激活，然后每回合被激活的节点保持激活状态不变，尝试去激活它的邻居，只尝试一次。

在DIC模型中，每个节点u都存在一个服从伯努利分布的随机变量，$X_u=1$表示表示节点u作为一个种子节点被成功激活，然后节点u都有一次机会通过边(u,v)去激活它的邻居节点v，概率为随机变量$X_{(u,v)}$。假设X_e服从一个确定的域为D_e的离散分布f_e，$d_e^i\in[0,1]$为D_e中第i个的值。X_e的值保持未知，直到节点u被激活。

在DIC模型中，图$G=(V,E,F_V,F_E)$，$F_V=\{f_u|u\in V\}$和$F_E=\{f_e|e\in E\}$是X_u和X_e的分布集合，N是节点的数量，$B\leq N$是成本。

自适应播种策略

该策略分为两个过程：播种过程和传播过程，播种过程单位为步数，传播过程单位为回合。

- 播种模式

  一个播种模式$A=(a_1,\dots,a_N)$，其中a_i表示播种过程每步选择种子节点的数量，$\sum a_i\leq B$，一个播种策略$S_A=(s_1,\dots,s_N)$是A的集合序列，$s_i\in V$，$|s_i|=a_i$。

假设模式$A_0 = (a_1, \ldots, a_n)$，当$1 \leq i \leq B$时$a_i = 1$，$i > B$时，$a_i = 0$。模式$A*$为，每次播种一个节点，直到传播停止即没有节点再被激活为止，然后再播种下一个节点。

- 传播策略

  在DIC模型上，用S_A^G表示模式A在图G上的播种策略，DIC模型是一个概率模型，目标函数是最终激活节点数量的期望值，用$E[S_A^G]$表示。

  如果对于i，$s_i = \emptyset$，但是不存在任何边(u, v)，使得u在第$i-1$个回合被它的邻居节点或者作为种子节点被激活，那么称S_A^G等待空回合。文章假设任何策略都不会等待一个或者多个空回合。

目标是，在成本约束条件下，在任何DIC网络G上，寻找一个模式A和一个策略S_A^G，使得$E[S_A^G]$最大化，即AIM问题。

贪心算法

辅助图

将DIC网络G通过一个辅助图$c - G = (V_c, E_c)$来表示，V_c包含了$N \cdot B + N$个节点，分为$N+1$个集合，表示为$V_c^i (0 \leq i \leq N)$，其中$|V_c^0| = N$，$|V_c^i| = B \quad (i > 0)$，$V_c^0 = \{v_{0,1}, \ldots, v_{0,N}\}$，$V_c^i = \{v_{i,1}, \ldots, v_{i,B}\}$。$E_c$包含了两部分$E_c^1$和$E_c^2$，$E_c^1$是$V_c^0$和$V_c^i$之间的边集，$E_c^2$是$V_0^i$之间的边集。

A-Greedy

A-Greedy的主要思想是，在DIC网络G的辅助图$c - G$下，每次选择一个影响数量期望值最大的节点作为种子节点，直到达到成本值或者没节点再被影响为止。

H-Greedy

H-Greedy考虑真实社交网络幂律分布的特性，它的思想包括两步：

1. 通过蒙特卡罗模拟，计算出每个节点影响数量的期望 $E[H(v)]$ ，所有节点影响数量平均值期望 $E[\sum_{v \in V} H(v)/N]$ 和对应标准差 $std[\sum_{v \in V} H(v)/N]$ 。
2. 然后运行A-Greedy，忽略 $E[H(v)]$ 小于 $\sum_{v \in V} H(v)/N$ 分布的1 sigma的边界值的节点，从而起到加速算法的作用。

实验

数据集

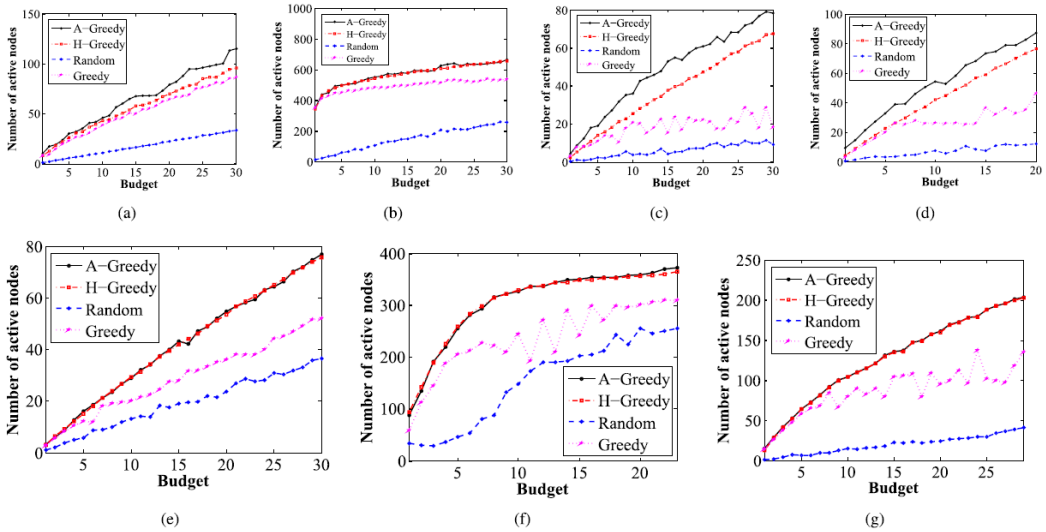
- Hep：物理领域的共同作者数据集，15K个节点，58K条有向边。
- Wiki：维基百科投票数据集，8.6K个节点，103K条有向边。
- PL：一个合成的幂律分布网络，2.5K个节点，26K条有向边。

设置

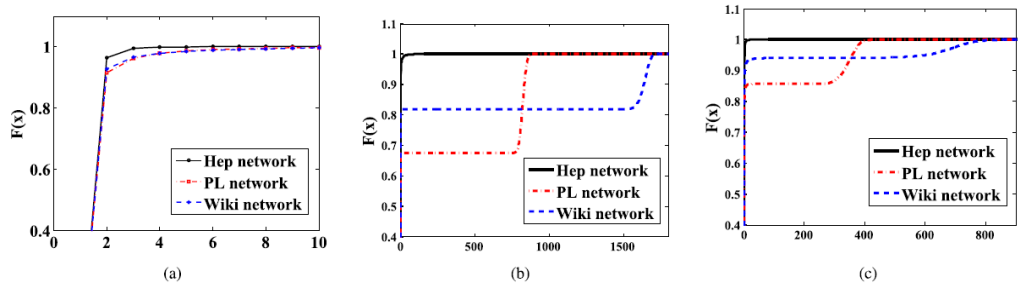
每条边的传播概率服从三个分布：F1，固定0.01；F2，平均值为0.01的指数分布；F3，服从均匀离散分布$\{0.1, 0.01, 0.001\}$。每个节点的激活概率$Prob[X_u = 1]$设置为1和0.5。

成本设置为10到30，蒙特卡罗模拟10K次。

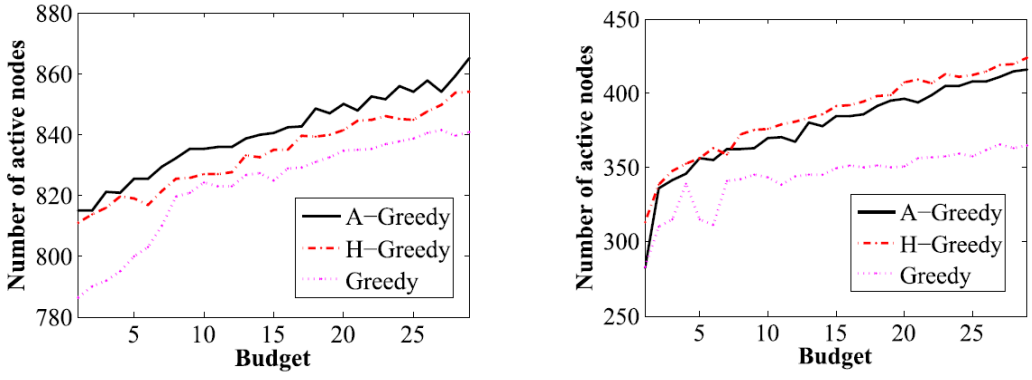
结果



上述几个图都是在不同数据集上，不同激活概率所产生的结构。激活概率为1，即对应传统IC模型，图中Greedy算法曲线稳定的对应这种情况，而在DIC模型上，Greedy算法曲线不稳定。在所有情况下，A-Greedy和H-Greedy都比Greedy结果要好。



上述图示H-Greedy在不同传播分布下，激活概率为1所产生的结果。纵坐标代表百分比，横坐标代表$E[H(v)]$的值。这三个图表明了$E[H(v)]$到达某个数量的百分比，因此来证明在H-Greedy算法上设置的1-sigma原则是可取的。



Parameter Setting	H-Greedy (ms)	A-Greedy (ms)
$\mathscr{F}^2$ & Prob$[X_u = 1] = 1$ on PL	14977	51485
$\mathscr{F}^2$ & Prob$[X_u = 1] = 1$ on Wiki	87412	268499
$\mathscr{F}^3$ & Prob$[X_u = 1] = 1$ on PL	981	11931
$\mathscr{F}^3$ & Prob$[X_u = 1] = 1$ on Wiki	31247	44625

上图在F2和F3两个传播分布上产生的结果，用来对比A-Greedy和H-Greedy算法。两者相比结果相差不大。然后上表为运行时间，可以看到H-Greedy效率明显比A-Greedy要高。因此作者开发的H-greedy算法具有较高的效率和有效性。

H2 思考

作者考虑自适应种子选择策略，并且考虑社交网络的动态性和节点度数的幂律分布特性。社交网络的动态性是之前研究所忽略的特性，再就是幂律分布特性，可以利用该性质来减少搜索次数，加速算法。自适应策略也是一种比较有效的种子选择策略。这种策略和这两种特性都可以应用于之前的一些研究上，可做适当的扩展研究。