武汉大学本科毕业论文（设计）

开　题　报　告

题　　目：IC 传播机制下的双层网络影响力最大模型

学生姓名：王翔

学　　号：2020302011088

所在学院：数学与统计学院

专　　业：数据科学与大数据技术

指导教师：胡捷

职　　称：讲师

1 选题描述

1.1 选题重要性

社交网络中，人们可以更加便捷的进行信息的获取与传播，随着科技的发展与在线社交技术的成熟，信息传播也变得更加迅速与广泛。

为了优化和控制信息的传播，人们提出了影响力最大化问题。随着大数据时代的到来，人们有更多的数据进行建模与研究；同时，社交网络数据庞大的规模与复杂性大大提高了分析与建模社交网络影响力最大化问题的难度。此外，建立的模型越优秀，对病毒营销，推荐系统，信息扩散，时间探测等等领域越具有指导意义。

对于影响力最大化问题，经典的优化算法是基于独立级联（IC）模型和基于线性阈值（LT）模型提出的近似比为（1-1/e）的贪心算法；后续提出的算法大致可以分为三类：基于仿真的方法、基于代理（proxy）的方法、基于草图的方法。

现在的信息传播逐渐发展成传统线下社交网络和线上社交网络并行的趋势，过去基于单层网络影响力最大化问题的研究已经不能很好的起到作用，因此需要深入对双层网络影响力最大化问题的研究。

1.2 传播模型选取

在影响力最大化问题研究中，信息传播模型的建立与选择是举足轻重的一环。采取线性阈值（LT）模型时，可以很好体现人的从众行为；但线性阈值模型的随机性完全是由节点的被影响阈值决定，一旦阈值确定，后面的传播过程就是完全确定的。因此，本课题将采用独立级联（IC）模型来进行探究。

2 发展现状

[Domingos and Richardson, 2001]、[Richardson and Domingos, 2002] 首次提出了影响力最大化问题，旨在利用数据挖掘技术来寻找最佳的病毒式营销方案，初步确定了衡量影响力最大化的客观指标。

[Kempe et al., 2003] 首次将影响力最大化问题定义为离散优化问题，并且基于独立级联（IC）模型和线性阈值（LT）模型提出了贪心算法（此算法近似比可达到（1-1/e））。由于贪心算法时间复杂度太高，不适用于大型网络求解。

[Leskovec et al., 2007] 利用独立级联(IC)模型的子模性,在传统独立级联(IC)模型的基础上改进提出 CELF 算法，不仅大大提速，还具有通用性强的特点；缺点是算法时间复杂度依然太高，不适用于大型网络求解。

[Kumar et al., 2022] 将复杂的影响模型转化为代理模型，提出了 CSR 算法；[Zhang et al., 2016] 基于投票机制提出了 VoteRank 算法，都是比较实用高效的尝试，美中不足是没有足够的理论支撑算法结果的准确性。

以上所提到的都是对单层网络的社交网络影响力最大化问题的研究，随着线上社交技术的发展与成熟，上述研究已经无法很好的模拟现代社交网络。因此，针对双层网络甚至多层网络的研究地位越来越高。

[胡斌 et al., 2011] 等人从双层网络来进行 RBF 神经网络的建模，同时采用粒子群优化算法 (Particle Swarm optimization,PSO) 进行调整和优化，得出以下结论：双层网络比单层网络更适合模拟信息传播的模型。此外，许多研究人员基于传染病动力学以及传染病模型（SIR），建立了许多信息传播模型。

事实上,无论是最近几年的 [Ye et al., 2022],[Banerjee et al., 2020]、[Li et al., 2018],还是距今有一段时间的 [Sun and Tang, 2011]、[Goyal et al., 2011]，他们的研究都是十分重要且有意义的，都能在一定程度上帮助解决他们所设的前提下的影响力最大化模型。

随着科技的进步与网络数据库的逐渐丰富，逐渐出现如 [Wen et al., 2017]（基于在线数据的研究)、[Lotf et al., 2022](基于遗传理论改进的研究)、[Fischetti et al., 2018](着眼研究传播损耗的研究)、[Akrouf et al., 2013]（采用 YouTube 和 Flickr 数据的研究）等等与时俱进的细分方向。

许多现实问题都需要利用双层网络的前提下进行分析研究，今后针对双层网络的影响力最大化研究地位将越来越高。

3 背景知识

3.1 社交网络概述

社交网络指一个图 G（V,E,P)，其中 V 为节点集，E 为边集，P 为概率集；每个用户都是一个不同的节点 v，用户之间的关系就是边 e，每条边都有一个概率 p，信息会沿着边按照对应的概率在节点之间传播。

3.2 影响力最大化问题分类

影响力最大化问题的目标是寻找社会网络中最终影响范围最大的 k 个节点，以这 k 个节点作为初始活跃节点集合（即种子集），经过影响在社会网络中的传播，最终被影响的节点个数最大。

主要分为两大部分：

（1）是给定节点数 k，选择出 k 个节点作为种子集使得种子集能影响的节点数最多。

（2）是给定所要求产生的影响力，找到满足条件的最小节点集合。

3.3 贪心算法

3.3.1 单调性与次模性

定义 3.1 (单调性). 若集合函数 $\sigma\ 2^V \to R$ 满足对任何子集 $S \subset T \subset V, \sigma(S) \leq \sigma(T)$, 则称其为单调非减的；集合函数 $\sigma\ 2^V \to R$ 满足对任何子集 $S \subset T \subset V, \sigma(S) \geq \sigma(T)$, 则称其为单调非增的。若集合函数 为单调非增的或单调非减的，则可称其为单调的。

定义 3.2 (次模性). 若集合函数 $\sigma\ 2^V \to R$ 满足对任何子集 $S \subset T \subset V$ 和任何的 $v \in V \backslash T$ 都有 $\sigma(S \cup v) - \sigma(S) \geq \sigma(T \cup v) - \sigma(T)$, 则称 σ 满足次模性

由于贪心算法的特性，模型在满足单调性和次模性时，采用贪心算法得到的解较优。

3.3.2 伪代码

贪心算法基于影响力最大化问题的单调性（Monotonicity）和次模性（Submodularity）特点，使用爬山策略（Hill Climbing），其基础算法框架如下：

事实上，我们无法得知节点集 S 的真实传播范围，往往采用蒙特卡洛方法估计它的影响范围，则真实情况下算法框架如下：

Algorithm 1: 贪心算法（k,f）：普通贪心算法

Input: k: 返回子集的大小； f: 满足单调性和次模性的集函数

Output: 选定子集

初始化 $S \leftarrow \varnothing$;

for $i = 1, 2, \ldots, k$ **do**

  $u \leftarrow \arg\max_{w \in V \setminus S} f(S \cup \{w\}) - f(S)$

  $S \leftarrow S \cup \{u\}$

end

return S

Algorithm 2: MC-贪心算法（k,f）：蒙特卡洛贪心算法

Input: G: 带有 IC 或 LT 模型的弧权值的社交图谱； k: 种子集大小

Output: 选定种子集

初始化 $S \leftarrow \varnothing$;

for $i = 1, 2, \ldots, k$ **do**

  $u \leftarrow \arg\max_{w \in V \setminus S} MC - Esitmate(S \cup \{w\}, G)$

  $S \leftarrow S \cup \{u\}$

end

return S

Function *MC-Estimate(S,G)*:

  $count \leftarrow 0$

  for $j = 1, 2, \ldots, R$ **do**

    $simulate\ diffusion\ process\ on\ graph G\ with\ seed\ set\ S$

    $n_a \leftarrow the\ number\ of\ activate\ nodes\ after\ the\ diffusion\ ends$

    $count \leftarrow count + n_a$

  end

  return $count/R$

end

3.4 传播模型

在此介绍两种经典的模型：独立级联（IC）模型和线性阈值（LT）模型。

3.4.1 独立级联（Independent Cascade, IC）模型

独立级联模型是一个概率模型，其中每一个节点有两种状态：活跃状态与非活跃状态，处于非活跃状态的节点有概率被处于活跃状态的邻居节点影响激活，从而进入活跃状态；已经进入活跃状态的节点无法恢复非活跃状态。在传播最开始时，只有种子集里的节点处于活跃状态；在 t 时刻为活跃状态的 u 节点将在 $t+1$ 时刻试图激活邻居节点 v，激活成功的概率为 $p_{u,v}$；u 节点对它的所有邻居节点都有且只有一次激活机会，无论激活成功与否，u 节点都丧失了再次激活同一节点的能力。这一类节点被称为无影响力的活跃节点。

3.4.2 线性阈值（Linear Threshold Model, LT）模型

线性阈值模型给每个节点赋予了一个阈值，该阈值反映对应节点的受影响的难易程度。与节点 v 相邻的节点 w 以非负的权重对节点 v 产生影响，并且 v 的所有邻居 w 的之 $b_{v,w}$ 和小于等于 1。对于一个处于未活跃状态的节点 v，只有当它的活跃邻居节点的影响力之和大于等于其阈值，节点 v 才会被激活，即网络中个体的决策依赖于其所有邻居节点的决策。且节点 v 的活跃邻居节点可以多次参与激活 v。

3.4.3 对比

两种传播模型都在一定程度上体现了信息扩散的特点，并且都引入了随机性，但是有着明显差别。

线性阈值模型：

- 以非活跃的节点为中心 (receiver-centered)，通过观察一个节点的所有邻居节点来判断这个节点是否被激活
- 一个节点激活与否取决于该节点的所有邻居节点
- 当模型中节点的阈值确定下来后，整个传播过程也确定下来，不再存在随机性

独立级联模型：

- 以活跃节点为中心 (sender-centered)，当一个节点处于活跃状态后将会以一定概率去激活每一个邻居节点

- 一个节点独立地尝试激活每一个邻居节点
- 整个信息传播过程是随机的

4 项目特色与创新点

- 探究双层社交网络上的影响力最大化问题
- 采用独立级联模型为传播机制
- 结合贪婪算法进行近似最优解

5 进度安排

- 2024 年 1 月 16 日–2024 年 2 月 10 日:

  学习社交网络影响力最大模型、独立级联模型及基础的贪心算法并写成读书笔记。

- 2024 年 2 月 10 日–2024 年 2 月 20 日:

  进行基于仿真、基于代理（proxy）、基于草图的三大类改进算法的学习，将改进算法与基础贪心算法比对，尝试编写相关代码并写成读书笔记。

- 2024 年 2 月 20 日–2024 年 3 月 2 日:

  进行网络模型学习与选取，编程进行单层社交网络影响力最大化模型实验并写成实验报告。

- 2024 年 3 月 3 日–2024 年 3 月 24 日:

  进行网络模型学习与选取，编程进行双层社交网络影响力最大化模型实验并写成实验报告。

- 2024 年 3 月 25 日-2024 年 3 月 31 日:

  改进先前实验，并写成实验报告。

- 2024 年 4 月 1 日–2024 年 4 月 7 日:

  分析总结整理研究过程资料。

- 2024 年 4 月 8 日–2024 年 4 月 23 日：

  以之前的读书笔记、实验报告为基础，完成毕业论文初稿。

- 2024 年 4 月 24 日–2024 年 4 月 30 日：

  修改初稿，完成毕业论文。

导师签名：

年　　月　　日

参考文献

[Akrouf et al., 2013] Akrouf, S., Meriem, L., Yahia, B., and Eddine, M. N. (2013). Social network analysis and information propagation: A case study using flickr and youtube networks. *International Journal of Future Computer and Communication*, 2(3):246–252.

[Banerjee et al., 2020] Banerjee, S., Jenamani, M., and Pratihar, D. K. (2020). A survey on influence maximization in a social network. *Knowledge and Information Systems*, 62:3417–3455.

[Domingos and Richardson, 2001] Domingos, P. and Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66.

[Fischetti et al., 2018] Fischetti, M., Kahr, M., Leitner, M., Monaci, M., and Ruthmair, M. (2018). Least cost influence propagation in (social) networks. *Mathematical Programming*, 170(1):293–325.

[Goyal et al., 2011] Goyal, A., Bonchi, F., and Lakshmanan, L. V. (2011). A data-based approach to social influence maximization. *arXiv preprint arXiv:1109.6886*.

[Kempe et al., 2003] Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146.

[Kumar et al., 2022] Kumar, S., Gupta, A., and Khatri, I. (2022). Csr: A community based spreaders ranking algorithm for influence maximization in social networks. *World wide web*, 25(6):2303–2322.

[Leskovec et al., 2007] Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2007). Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429.

[Li et al., 2018] Li, Y., Fan, J., Wang, Y., and Tan, K.-L. (2018). Influence maximization on social graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1852–1872.

[Lotf et al., 2022] Lotf, J. J., Azgomi, M. A., and Dishabi, M. R. E. (2022). An improved influence maximization method for social networks based on genetic algorithm. *Physica A: Statistical Mechanics and its Applications*, 586:126480.

[Richardson and Domingos, 2002] Richardson, M. and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70.

[Sun and Tang, 2011] Sun, J. and Tang, J. (2011). A survey of models and algorithms for social influence analysis. *Social network data analytics*, pages 177–214.

[Wen et al., 2017] Wen, Z., Kveton, B., Valko, M., and Vaswani, S. (2017). Online influence maximization under independent cascade model with semi-bandit feedback. *Advances in neural information processing systems*, 30.

[Ye et al., 2022] Ye, Y., Chen, Y., and Han, W. (2022). Influence maximization in social networks: theories, methods and challenges. *Array*, page 100264.

[Zhang et al., 2016] Zhang, J.-X., Chen, D.-B., Dong, Q., and Zhao, Z.-D. (2016). Identifying a set of influential spreaders in complex networks. *Scientific reports*, 6(1):27823.

[胡斌 et al., 2011] 胡斌, 王敬志, and 刘鹏 (2011). 基于双层网络的混合 pso 算法的 rbf 建模. *西南科技大学学报*, 26(2):78–81.