# 硕士研究生学位论文开题报告

**论文题目：时空动态的跨网络影响力最大化问题**

**一、**立题依据（包括研究目的、意义、国内外研究现状和发展趋势，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）（不少于800字）

**1.1 研究背景和意义**

近年来，随着社交网络的兴起，其在信息传播中所扮演的角色也愈加重要。我们发现越来越多的人喜欢在网络上公开他们的观点或想法。用户参与和信息共享的结合使得在线社交网络能够将信息传播到更多的人，并且传播速度甚至比大众媒体更快。

许多学者主要研究兴趣是了解社交网络中的信息传播方式以及如何找到有影响力的用户，例如，一个公司可能希望确定一小部分有影响力的用户，这样他们就可以影响他们的朋友，甚至朋友的朋友，通过口碑效应使得其他用户采用产品。在这样的情况下，公司就会考虑谁可以作为最初的体验用户。这就是所谓的影响力最大化问题，即找到一组用户使得他们在社交网络中的影响范围最广。

然而经典的影响力最大化问题存在着诸多限制。首先它基于以下假设：影响从种子集S传播到社交网络G中的最大节点需要t时间，并且t很小，这样使得G的拓扑在整个传播过程中保持不变。因此，G的拓扑结构在传播过程中是完全已知的。然而在现实网络中，影响传播可能需要相当长的时间(几天甚至几周)才能达到社交网络的最大范围。在这段时间内，网络的拓扑肯定会发生一些变化，比如新用户的加入，原本不认识的用户成为好友等。所以如何使影响力最大化问题的模型和算法能够适应网络的动态变化引起许多研究者的关注。与此同时，随着传播的进行，影响力的大小也应该随着时间的推移发生变化，如何量化影响力大小的衰减过程也是一个值得研究的问题。其次，传统的影响力最大化问题研究的是用户在单一社交网络中的影响力大小。但是在现实中，社交网络往往存在大量的重叠用户，这些用户在跨网络的信息传播方面发挥着重要作用。当用户积极参与多个网络时，他/她可以起到一个转发的作用，同时将信息传播到多个社交网络。如果我们仅仅单独考虑每个网络会大大低估一些用户的影响力。所以如何度量用户在跨网络间的影响力也是一个值得研究的问题。最后，由于社交网络庞大的规模，所以影响力最大化问题的算法应当拥有较低的时间复杂度和空间复杂度才能高效的完成任务。

**1.2 国内外研究现状和发展趋势**

Kempe[1]于2003年首先将影响力最大化建模为一个算法问题。他将一个社交网络抽象为一个图，其中V代表图G中结点的集合（用户集合），E代表图G中边的集合（用户之间的社交关系）。影响力最大化问题的目标是找到一个包含k个结点的用户集合使得它在图G中的影响范围最大。许多商家利用影响力最大化的思想对其商品进行推广，不仅节约了广告成本还使得推广效果更加出众。在影响力最大化问题的研究中，如何对信息扩散进行建模和怎样选择种子用户是现在主要的研究点。

**1）信息扩散过程模型**

对信息扩散过程进行建模是评估种子集合影响范围的重要基础[2]。最经典的两类传播模型是独立级联模型（ICM）和线性阈值模型（LTM）。

独立级联模型最早由J. Goldenberg[3]于2001年中提出。传统的独立级联模型的主要思想是当一个节点u被激活时，它会以概率p(u,v)对它未激活的出边邻居节点v尝试激活，这种尝试仅仅进行一次，而且这些尝试之间是互相独立的，即u对v的激活不会受到其他节点的影响。它的传播过程如下：

·给定初始的活跃节点集合S，当在时刻t节点u被激活后，它就获得了一次对它的邻居节点v产生影响的机会，成功的概率为p(u,v)，是随机赋予的系统参数。

·若v有多个邻居节点都是新近被激活的节点，那么这些节点将以任意顺序尝试激活节点v。如果节点u成功激活节点v，那么在t+1时刻，节点v转为活跃状态。

·在t+1时刻，节点v将对其他节点产生影响，重复上述过程。

·结束条件：网络中不存在有影响力的网络节点。

线性阈值模型最早由Granovetter 和Schelling[4]于1978年提出。传统线性阈值模型的主要思想是对图中每个节点v都分配一个激活阈值，当v的入邻居影响力大于它的阈值时，就会被激活。

传统线性阈值模型的传播过程如下：

·集合中的任意节点v随机分配阈值θ[v]∈[0,1]。只有当节点v的新处于激活状态的邻居节点对它的影响力大于该阈值时，节点v才能被激活。

·用权值b[u.v]表示节点v被它的邻居节点u的影响，∑u∈in(v)  b[u,v]≤1表示节点v的处于活跃状态的邻居节点对它的影响力之和。这里in(v)是v的入边邻居节点集合。

·给定初始的活跃节点集合S，在t时刻，所有在t-1时刻处于活跃状态的节点仍保持活跃，并且当这一时刻节点v的邻居节点的影响力之和大于节点b的阈值时，节点v被激活.

·节点v被激活后，下一时刻将对它的邻居节点产生影响，重复上述过程。

·结束条件：当网络中已存在的所有活跃节点中任意活跃节点的影响力之和都不能激活他们的处于非活跃状态的邻居节点时。

上述两个基本模型虽然可以大致描述信息在社交网络中的传播过程，但是存在很多的局限性。Liu等人[5]认为时间在影响一个用户到另一个用户的传播中起着重要的作用，用户影响另一个用户所需要的时间也是不同的，并且在现实中的营销案例中，营销者也会关心在一个固定的时间之前影响的传播范围有多广。基于此思想作者提出了时间延迟感知的独立级联模型（LAIC），在这个模型下，每个结点的状态分为三种，激活，延迟激活，非激活。对于一个在种子集合中的结点u，可以认为他在t=0时刻处于活跃态，如果它成功影响到了一个非活跃邻居节点v，那么节点v变为延迟活跃节点。在LAIC模型中，当一个节点u在t步第一次被激活时，他会在步骤对它的当前不活跃节点v以概率进行激活，其中是一个影响延迟。如果一个节点有多个邻居影响它，则只关注时间最早的那一次，之后的都会被忽略。Kim等人[6]随后提出了连续激活和时间限制的IC模型（CT-IC），该模型的传播方式为：给定一个有向图G=（V，E）每条边上都有一个概率pp0(u,v)代表节点u在某一步激活v的概率，给定给一个种子集合S，和一个时间限制T，在t=0时，初始集合中的种子被激活，随后开始进行传播。使用At表示在时间t的活跃种子，在时间t，所有属于At的节点都要以概率试图去激活它的未活跃邻居。Mohammadi A等人[7]提出了与用户在线行为相关的时间和开销限制的影响力模型（TCIO），它的主要思想如下：每条边上有一个转发概率puv，它代表了当v在线且u已经转发消息时，结点v从结点u获得消息的概率是多少。一天的时间被划分为多个时间段，一个结点可以根据其在线模式辨别在一个时间段是在线还是离线。mvt表示节点v在时间段t中的在线模式，等于0时代表离线，等于1是代表在线。当选择每个结点作为传播消息的种子结点时，它都有一个“标价”。每个转发消息的节点都可以根据其在线模式反复影响其邻居，直到其邻居受到影响转发消息或达到过期时间。Chen 等人[8]认为影响从一个人到另一个人的传播是需要二者之间“相遇”才有机会传播的。这里所谓的相遇不一定非要是两个人面对面在一起，也可以是在社交媒体上的点赞转发，于是他提出了IC-M模型。它的过程如下：在图中的每条边上会有一个相遇概率m(u,v)，它符合在[0,1]之间符合均匀分布。在第0步时，种子节点被激活，在之后的任意步骤中，一个激活的节点u会以概率m(u,v)独立的遇到他的不活跃邻居节点v。如果u与v之间首次相遇，则给u一次尝试激活v的机会，成功概率p(u,v)。如果尝试成功，v在步骤t处变得活跃，并将在t+1处开始传播影响。当所有的活动节点都与它们的邻居相遇，并且没有新的节点可以被激活时，扩散过程就停止了。Li H 等人[9] 认为经典的影响力最大化问题忽略了影响传播的时间，而影响传播的时间在现实中可能是非常重要的，在此期间，潜在的社交网络可能会进化。因此，作者正式地将这个经典的影响力最大化问题重新定义为Propagation Time-conscious Influence Maximization Problem：设G0 = (V0, E0)为t0时刻的当前网络，k为预算。Tr为影响传播时间，并且此时的G0演化到了Gr=（Vr,Er）。Propagation Time-conscious Influence Maximization的目标是在t0时刻选择一个种子集合S属于V0 ，使得在时刻tr最大。并且假设Gr的拓扑在t0时刻是不被知道的。作者假设传播符合独立级联模型（ICM），并且结合了Forest Fire model（森林火灾模型）来预测从G0到Gr的网络拓扑结构。

对传统传播模型的改进还体现在将其扩展到多个网络之中。Nguyen D T[10] 首先提出了如何评估影响同时在多个网络中的传播。他提出一种将多个网络耦合成一个新的单一网络的方案，该方案保留了每个网络的影响扩散特性。作者使用LT模型，将其扩展到多个网络。在多个网络G1…k中，信息在每个网络中分别传播，可以通过重叠的用户从一个网络转移到另一个网络。信息开始从种子用户集合S (i)开始传播。S中的所有用户都处于活动状态，其余用户都处于非活动状态，在t时刻，如果某个网络中来自其活动邻居的总影响超过其阈值，则用户u变为活动用户。在每个时间步骤之后，新的非活动用户将被激活，并继续激活其他用户。此过程将继续，直到不再激活不活动的用户为止。随后，Nguyen D T[11]又提出了在多个网络中的最小成本影响问题（LCI），该问题要求确定最终能够影响大量用户的种子用户的最小数量，作者解决了多路OSNs中的LCI问题，通过不同的耦合方案将多路网络简化为一个网络，同时保留了影响最大的传播特性。Zhan Q[12]认为在现实世界中，在线社交网络通常是异构的，用户可以通过多个渠道相互影响，并且在线社交网络还可以共享公共用户，信息可以通过这些用户在不同网络中传播。本文首次研究了部分对齐异构社交网络中的影响最大化问题，并且提出了MM模型，它基于一组网络内部和网络内部的社会元路径，从对齐的异构社交网络中提取多对齐的多关系网络(MMNs)。Zhan Q [13]认为如今的用户通常同时参与多个在线社交网络，那些加入Facebook的用户也在使用其他网络，比如Twitter，Foursquare和Instagram。如果想要研究Facebook网络中影响力问题的话，也要研究影响力从上述其他网络中的传入。Ma Y[14] 认为本地专家可以在很多应用中发挥重要作用，如解决本地信息查询、社会事件安排等。作者在多个社交网络中寻找本地专家，并且用到了两个指标：local authority和topic authority。这两个指标分别与地理位置和主题相关。Kuhnle A[15]认为影响在不用的网络中的传播模型是不一样的。作者将不同网络抽象为一层，并且提出了在不同层中传播模型可能不一样的传播方式。假设网络是两层的，并且已经预先知道了这两个网络中的重叠结点。作者首先将两个网络中的节点进行补齐，补齐后的结点是孤立节点。随后种子节点被激活，并且分别在两个网络中进行传播直到两个网络中没有节点能继续被激活，此时如果重叠节点在一个网络中被激活，那么这个节点会将激活状态转移到另一个网络中，然后两个网络继续开始影响力的传播，直到没有节点能被激活为止。Singh[16]认为在多个网络中，种子用户可以同时宣传多个产品，而非种子用户可以同时在网络上接受不同的产品。针对上述场景，作者提出了一个跨多个社交网络的多重影响最大化框架(MIM2)。

**2）种子用户选择算法**

尽管影响力最大化问题在计算上很复杂并且已经被证明是个NP-hard问题[1]，但是依然可以通过一些巧妙的算法来得到近似最优解。得到最优解的条件是影响函数要满足以下两个条件：单调性和子模性。直观上来说单调性就是在种子集合S中加入一个结点后，它的影响范围不会减少，而子模性的意思是在集合S中加入一个结点时，这个结点所带来的边际效益要小于上一个结点的加入带来的边际效益。D. Kempe等人[1]已经证明了影响函数在经典的IC模型和LT模型中是满足上述两个性质的。大量社交网络的影响力最大化算法是基于贪心算法和启发式算法。D. Kempe等人[1]提出了使用蒙特卡洛模拟来评估每个结点的边际收益，并且每次将边际收益最大的结点加入到种子集合中。这样的做法在网络比较大的时候显然是非常低效的。为了解决这个问题，J. Leskovec等人[17]提出了CELF算法来减少蒙特卡洛模拟的次数，它首先计算所有的结点的边际收益，得到初始集合S1。前一次得到的边际收益就是下一次的上界。它对余下结点按照上一次计算的边际收益的降序来进行计算，如果一个节点的边际收益值近似于上界，则将其直接加入集合中。C. Zhou等人[18]提出了UBLF，用了一种矩阵分析的方法快速获取所有结点的边际收益上界。Wang等人提出了CGA算法，它将图划分成社区后利用社区内每个节点的影响来决定选择哪些节点作为种子，它只用在局部子图中运行蒙特卡洛从而降低蒙特卡洛模拟的复杂度。上述使用蒙特卡洛模拟的方法虽然可有很好的模型通用性，但是却阻碍了性能的提升。因此，最近研究人员已经开始探索基于代理和基于草图的方法。影响力排名代理的想法非常直观。它根据一个近似用户影响的度量对图G中的所有用户进行排序，然后直接从排序中生成种子集。因此，这里最重要的挑战是得到一个好的排名指标[2]。比较经典的一个算法是W. Chen等人[19]提出的MIA算法，它的主要思想是计算节点u对植根于u的局部树状结构的影响扩散。基于草图的方法在提高理论效率的同时也可以使近似值得到保证。基于草图的方法基于特定的扩散模型预先计算多个草图，然后利用草图来评估影响扩散。它一般分为两大类，正向影响草图（FI-Sketch）和反向可达草图（RR-Sketch）。正向影响草图与蒙特卡洛方法相似。在一个图中，假设每条边上的传播概率为P1,P2…Pn。利用蒙特卡洛模拟计算一个节点的影响范围时，首先将图中所有边以1-Pk的概率移除，在这个子图中计算当前节点的影响范围。S. Cheng等人提出的staticGreedy算法建立个草图并计算影响范围。但是正向影响草图在最坏情况下的时间复杂度依然很高。反向可达草图通过选择随机结点并生成随机的RR集来估计结点的边际收益。Y. Tang等人[21]提出的TIM算法可以确定到底需要多少RR集但是对内存的消耗很大。为了减少内存浪费X. Wang等人[22]提出了BKRIS算法，它首先用启发式的方法去估计一个OPT的下限，然后从这个下限得出数量足够多的RR集。

然而上述算法都没有考虑到网络的动态性。在现实世界中，社交网络具有高度的动态性，并且随着时间的推移发展迅速[23]，因此，关于影响力结点的计算和结果很快就会过时。Zhuang[24]考虑了一个社交网络在动态的变化，但是只能在特殊时间探测到少数结点，因此，信息只在当前网络的一小部分可用。文献[25][26][27]都使用一系列网络快照来表示图在一定时间内的动态变化。

目前，几乎没有人在跨网络传播中考虑时间因素，跨网络传播中的动态问题更是几乎没有提及。由此可见，对影响在跨网络中的动态传播进行建模并且提出一种算法找到最具影响力的用户是很有必要的。

参考文献

1. D. Kempe, J. Kleinberg, and E. Tardos, “Maximizing the spread of influence through a social network,” in Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2003, pp. 137–146
2. Li, Y., Fan, J., Wang, Y., & Tan, K. L. (2018). Influence maximization on social graphs: A survey. IEEE Transactions on Knowledge and Data Engineering, 30(10), 1852-1872.
3. J. Goldenberg, B. Libai, and E. Muller, “Talk of the network: A complex systems look at the underlying process of word-ofmouth,” Marketing Lett., vol. 12, no. 3, pp. 211–223, 2001
4. M. Granovetter, “Threshold models of collective behavior,” Am.J. Sociol., vol. 83, no. 6, pp. 1420–1443, 1978.
5. Liu, Bo, et al. "Time constrained influence maximization in social networks." 2012 IEEE 12th international conference on data mining. IEEE, 2012
6. Kim, Jinha, Wonyeol Lee, and Hwanjo Yu. "CT-IC: Continuously activated and time-restricted independent cascade model for viral marketing." Knowledge-Based Systems 62 (2014): 57-68.
7. Mohammadi A, Saraee M, Mirzaei A. Time-sensitive influence maximization in social networks[J]. Journal of Information Science, 2015, 41(6): 765-778.
8. Chen, Wei, Wei Lu, and Ning Zhang. "Time-critical influence maximization in social networks with time-delayed diffusion process." Twenty-Sixth AAAI Conference on Artificial Intelligence. 2012.
9. Li H, Bhowmick S S, Cui J, et al. Time is what prevents everything from happening at once: Propagation time-conscious influence maximization[J]. arXiv preprint arXiv:1705.10977, 2017.
10. Nguyen D T, Das S, Thai M T. Influence maximization in multiple online social networks[C]//2013 IEEE Global Communications Conference (GLOBECOM). IEEE, 2013: 3060-3065.
11. Nguyen, D. T., Zhang, H., Das, S., Thai, M. T., & Dinh, T. N. (2013, December). Least cost influence in multiplex social networks: Model representation and analysis. In 2013 IEEE 13th International Conference on Data Mining (pp. 567-576). IEEE.
12. Zhan Q, Zhang J, Wang S, et al. Influence maximization across partially aligned heterogenous social networks[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2015: 58-69.
13. Zhan Q, Zhang J, Philip S Y, et al. Discover tipping users for cross network influencing[C]//2016 IEEE 17th International Conference on Information Reuse and Integration (IRI). IEEE, 2016: 67-76.
14. Ma Y, Yuan Y, Wang G, et al. Local Experts Finding Across Multiple Social Networks[C]//International Conference on Database Systems for Advanced Applications. Springer, Cham, 2019: 536-554.
15. Kuhnle A, Alim M A, Li X, et al. Multiplex influence maximization in online social networks with heterogeneous diffusion models[J]. IEEE Transactions on Computational Social Systems, 2018, 5(2): 418-429.
16. Singh, S. S., Singh, K., Kumar, A., & Biswas, B. (2019). MIM2: Multiple influence maximization across multiple social networks. Physica A: Statistical Mechanics and its Applications, 526, 120902.
17. J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, “Cost-effective outbreak detection in networks,”in Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2007, pp. 420–429
18. C. Zhou, P. Zhang, W. Zang, and L. Guo, “On the upper bounds of spread for greedy algorithms in social network influence maximization,” IEEE Trans. Knowl. Data Eng., vol. 27, no. 10, pp. 2770–2783, Oct. 2015.
19. W. Chen, C. Wang, and Y. Wang, “Scalable influence maximization for prevalent viral marketing in large-scale social networks,” in Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2010, pp. 1029–1038
20. S. Cheng, H. Shen, J. Huang, G. Zhang, and X. Cheng,“Staticgreedy: Solving the scalability-accuracy dilemma in influence maximization,” in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage., 2013, pp. 509–518.
21. Y. Tang, X. Xiao, and Y. Shi, “Influence maximization: Nearoptimal time complexity meets practical efficiency,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2014, pp. 75–86.
22. X. Wang, Y. Zhang, W. Zhang, X. Lin, and C. Chen, “Bring order into the samples: A novel scalable method for influence maximization,” IEEE Trans. Knowl. Data Eng., vol. 29, no. 2, pp. 243–256, Feb. 2017.
23. R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In Link Mining: Models, Algorithms, and Applications, pages 337-357.Springer, 2010
24. Zhuang, H., Sun, Y., Tang, J., Zhang, J., & Sun, X. (2013, December). Influence maximization in dynamic social networks. In 2013 IEEE 13th International Conference on Data Mining (pp. 1313-1318). IEEE.
25. X. Chen, G. Song, X. He, and K. Xie, “On influential nodes tracking in dynamic social networks,” in Proc. SIAM Int Conf Data Mining, 2015, pp. 613–621.
26. Xu, S., Xu, N., Zhang, J., Li, F., & Li, S. (2017, December). Seed set selection in evolving social networks. In 2017 3rd IEEE International Conference on Computer and Communications (ICCC) (pp. 2323-2328). IEEE.
27. Hafiene N, Karoui W, Romdhane L B. Influential Nodes Detection in Dynamic Social Networks[C]//International Conference on Business Information Systems. Springer, Cham, 2019: 62-73.

**二、研究内容和目标（说明课题的具体研究内容，研究目标和效果，以及拟解决的关键科学问题。此部分为重点阐述内容）（不少于2500字）**

**2.1研究目标**

**2.2 研究内容**

针对现有社交网络信息扩散模型的缺点，提出一种新的考虑时间因素的动态跨网络模型，并且提出一种适应于新模型的动态选种算法。

1）构建一个符合实际情况的跨网络的信息强度衰减模型。

目前大部分的信息扩散模型都是在独立级联模型和线性阈值模型上进行改进使得与现实中的社交网络传播方式更加相近。独立级联模型和线性阈值模型只能描述一个消息在用户与用户之间最基本的传播方式，并不是非常贴近现实。许多研究者从现实角度出发，认为一个消息传播的时间越久它能影响到别人的概率就会越小，基于这种想法研究者对独立级联模型进行改进，将用户之间固定的传播概率改为随着时间（传播跳数）增加而逐渐减小。但是他们没有对消息的影响如何随着时间增加而减小给出一个合理的解释，许多研究者只是简单的给出了一个所谓的时间延迟函数，用这个时间延迟函数去与原来的传播概率相乘就得到了新的传播概率。不同研究者对于这个延迟函数符合什么分布也存在不同的理解，但是他们都没有说明为什么这个是符合实际情况的。其次，目前几乎没人研究在跨网路传播中的信息强度衰减，所以如何度量在跨网络传播中的信息强度衰减是个重要的问题。现有的对于跨网络传播的模型也有不少，但是方法是五花八门。有的研究者将多个网络耦合成一个新的单一网络，在这个单一网络中可以保留每个原始网络中的扩散特性，他们为网络中的每个用户创建代理节点，转换顶点以及账户顶点，并将传播步骤分解为子步骤来处理多个传播。但是这样会造成边的大量冗余，在网络较大或者网络中的边数较多的情况下对存储和计算提出了更高的要求。

2）提出一个动态选种算法

**2.3拟解决的关键科学问题**

**三、研究方案设计及可行性分析（包括：研究方法，技术路线，理论分析、计算、实验方法和步骤及其可行性等）（不少于800字）**

**四、本研究课题可能的创新之处（不少于500字）**

**五、研究基础与工作条件（1.与本项目相关的研究工作积累基础 2.包括已具备的实验条件，尚缺少的实验条件和拟解决途径）（不少于500字）**