

本科生毕业设计（论文）开题报告

题　　目：动态二部图上高效极大二分团枚举算法研究及实现

计算机科学与技术学院

院　　系\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

CSIE1701

专业班级\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

俞景昳

姓　　名\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

U201713732

学　　号\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

华强胜

指导教师\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2021年3月**开题报告填写要求**

1. 开题报告主要内容：

1.课题来源、目的、意义。

2.国内外研究现况及发展趋势。

3.预计达到的目标、关键理论和技术、主要研究内容、完成课题的方案及主要措施。

4.课题研究进度安排。

5.主要参考文献。

1. 报告内容用小四号宋体字编辑，采用A4号纸双面打印，封面与封底采用浅蓝色封面纸（卡纸）打印。要求内容明确，语句通顺。
2. 指导教师评语、教研室（系、所）或开题报告答辩小组审核意见用蓝、黑钢笔手写或小四号宋体字编辑，签名必须手写。
3. 理、工、医类要求字数在3000字左右，文、管类要求字数在2000 字左右。
4. 开题报告应在第八学期第二周之前完成。

# 课题来源

该课题是本人在第八学期（大四下学期）开启的毕业设计课题，由华强胜老师指导带领研究。

图是现实世界一些应用的很自然地抽象，被广泛地应用在了诸如社交网络、网络搜索、计算生物领域。我们将重点放在可以用来为两种不同类型实体之间的交互建模的二部网络上，也就是二部图。例如，用户和新闻文章之间的关系可以被建模为二部网络图，其中用户和新闻文章是两个顶点集，如果用户已经看过一篇文章，则用户和一篇新闻文章之间存在一条边。我们研究了演化二部图中的极大二分团枚举（MBE）。现实世界的应用有部分强依赖于极大二分团，然而，在一些实际的应用中，过去的一些对于二部图极大二分团挖掘的算法有局限，不能很好的适应处理动态图的情况。

# 课题目的和意义

以前的相关研究成果中，我们已经知道从一个图中枚举所有的极大二分团问题是NP完全的，有许多研究者关注于静态图，希望挖掘出一个高效的静态图算法；但近期也有研究者将关注点放在了动态图中，并给出了一个比较简单的可以实现的算法。本课题的目的在于进一步扩展动态图的算法，主要可以分为两个方面：进一步降低时间和空间复杂度；研究其更高效的实现；或者从建模或者其他角度研究并识别出一些更高效的模型，也就是对同一问题提出其他更优秀的算法。

研究该课题的直接意义在于，它可以从一定程度上优化很多实际应用的实现。应用层面的社交网络和基础建设领域的事务型数据库[9,10,11,12]都是直接地应用对象，他们直接服务于我们生活。因此，通过解决该问题，我们可以帮助科学家们在相关领域中做出进一步的研究，并且直接应用到实际领域，起到更为切实的帮助。

# 国内外研究现况和趋势

我们先从极大二分团的静态图的算法入手。在一张静态图中，枚举极大二分团的算法已经被相关的研究证明是NP完全的[1]，其有几个研究G. Alexe等人[1]，Y.Zhang等人[2]，G.liu等人[3]针对静态图已经提出过相关的枚举极大二分团的算法。G.liu提供了当前最佳的时间复杂度。

极大二分团因为其已经被应用去解决Edge Covering问题[4]，并且成功地应用在了许多热门的领域中，比如网络社区发现[5]、蛋白质-蛋白质交互网络中的拓扑结构发现[6]、最大级联系统发育数据集发现[7]，因此显得尤为重要，也很好地说明了研究它的意义。

随着时代的发展，现如今大部分的网络（图）都是动态的，也就是说，它随着时间的改变，会不断的有新的边的加入这就给以前的算法带来了挑战，重复应用静态图的算法会极大地增加性能开销，在多数流式场景下不被接受。最近Apurba等人的论文[8]宣称其找到了一个二部图中的动态图算法，可以以较低的时间复杂度枚举其中的极大二分团，同时避免显示地枚举旧图中的极大二分团，这是一个变化敏感的算法。本课题也受其启发，想要其的算法出发，进行探索发现找出可以进行优化的目标。

# 课题目标

本课题主要关注二部图中枚举极大二分团的算法。其目标是在以下几个方面（之一或其中几个）中改进[8]中提出的算法：进一步降低时间和空间复杂度；研究其更高效的实现；或者从建模或者其他角度研究并识别出一些更高效的模型，也就是对同一问题提出其他更优秀的算法。

从另一个角度来说，本人相信自己能够在研究这个课题的过程中学习如何开展一个学术研究，积累研究经验并提高自己的算法分析设计和实现能力。该课题也是一个锻炼自己作为一个本科生从零开始进行科研的一个尝试，希望能为将来的研究生生涯起到一个入门的作用。

# 问题描述

首先进行定义：设为的顶点集，为的边集，设和分别为的顶点数和边数。让表示中与顶点相邻的顶点集。如果从上下文来看，图G的含义是清晰的，我们使用表示。对于一条边，让表示从G删除后的图,表示图添加为一组边到G后的图。对于一组边集，让表示添加(删除)到（从）之后的图。同理，对于顶点，设表示将加到后的图，对于顶点，设表示从中删除及其所有相邻边后的图。设为中顶点的最大度，为中顶点的最小度。

其次，设为一个简单的无向二部图，其顶点集划分为，和边集合。一个中的二部图是一个二分，当X中的每一个顶点连接到Y中的的每一个顶点。一个二分B被称作极大二分团，如果不存在一个二分团，使得是其真子集。我们让表示G中的极大二分团集合。

我们的目标是对于一张给定的图，和一个给定的新增边集合，去寻找一个高效地算法，在中的边不断以批或流的形式加入原图其中的时候，去维护新图中的集合。

# 研究手段

首先是文献的阅读，本报告前文中提到的数篇论文对本课题都有很大的帮助。阅读文献并从中获得启发是非常重要的一部分。

及时和指导老师华强胜华老师请教、同步进度。

研究的主要方向应该是研究如何降低算法实现的渐进时间复杂度，这一部分可以从本人过去经历的算法竞赛中获得一些启发，活用过去学过的算法设计、分析、实现相关的知识是达成这一目标的重要手段。

因为本课题的代码实现可能会很复杂，因此提前准备好开发、测试环境非常重要。这里应当发挥过去积累的工程经验，将系统、工程化的程序编写、代码管理和软件测试经验应用到这次课题中。不能因为这是一个研究类的课题就轻视它的代码复杂度（相对于工程类课题而言）。

# 风险分析

风险来源一：可能欠缺研究该课题的知识或技能，主要表现在部分未经过系统学习的科目。比如抽象代数，在讨论排列的时候，经常会涉及到群论，本课题中就采用有向排列对基因组进行建模。除此之外，可能会有其他更多的需要补充的知识。就对应措施而言，如果出现了必须要进行大量学习的知识漏洞，应当及时入手相关教材，并主动学习所欠缺的知识，必要时会与导师进行商讨。

风险来源二：测试时可能欠缺高性能计算机。本课题涉及到的实验可能是一个花费很多算力的实验，在后期进行进行测试时，为了得出特定算例的最优解，可能会依赖非多项式时间复杂度的暴力求解器。运行暴力求解器需要花费大量的计算力，因此缺乏计算力的可能性需要提前考虑。

风险来源三：存在欠缺测试数据的可能性。这个风险可以通过向其他研究人员索要数据来规避，因此有必要在研究前期积极与其他在同一导师下研究过同一问题的前辈联系。

风险来源四：进度安排难以确定。因为这是一个研究性质的课题，难以在前期就确定好具体的进度安排。这个风险可能导致的直接后果是在毕业设计结题时拿不出成果从而导致无法正常毕业，因此这个风险也是本人最重视的一个风险。本人打算采取的方案是积极与导师保持联系，随时进行交流和讨论，避免出现最后拿不出成果的情况，同时也要保留在本学期中后期更换课题的可能性。

# 课题研究进度安排

由于本课题是一个研究性的课题，难以给出一个详细的时间计划表。在研究该课题的过程中，本人预计会花大量的时间进行文献阅读和问题思考，这个过程无法给出一个确定的时间。然而仍然可以给出一个大致的进度安排如表 1所示。

表 1　课题研究进度安排表

|  |  |
| --- | --- |
| 月份 | 工作任务 |
| 2021年1月  ~2021年3月 | 接受课题，阅读导师给出的文献，准备开发测试环境 |
| 实现前人的算法，阅读代码和结果并从中获得启发 |
| 2021年3月  ~2021年4月 | 研究并改进原有算法，积极与导师和同组成员讨论 |
| 自主学习课题涉及到的陌生知识 |
| 2021年5月 | 准备课题答辩 |
| 2021年6月 | 书写最终报告，准备最终答辩 |

# 主要参考文献

1. G. Alexe, S. Alexe, Y. Crama, S. Foldes, P. L. Hammer, and B. Sime- one. Consensus algorithms for the generation of all maximal bicliques. Discrete Applied Mathematics, 145(1):11–21, 2004.
2. G. Liu, K. Sim, and J. Li. Efficient mining of large maximal bicliques. In Data warehousing and knowledge discovery, pages 437- 448. Springer, 2006.
3. Y. Zhang, C. A. Phillips, G. L. Rogers, E. J. Baker, E. J. Chesler,

and M. A. Langston. On finding bicliques in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types. BMC bioinformatics, 15(1):1, 2014.

1. F. Chung. On the coverings of graphs. Discrete Applied Mathematics, 30(2):89–93, 1980.
2. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In Proceeding of the 8th international conference on World Wide Web, pages 1481–1493, 1999.
3. D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen. Topological structure analysis of the protein protein interaction network in budding yeast. Nucleic Acids Research, 31(9):2443-2450, 2003.
4. M. J. Sanderson, A. C. Driskell, R. H. Ree, O. Eulenstein, and S. Langley. Obtain- ing maximal concatenated phylogenetic data sets from large sequence databases. Molecular Biology and Evolution, 20(7):1036–1042, 2003.
5. A. Das and S. Tirthapura, "Incremental Maintenance of Maximal Bicliques in a Dynamic Bipartite Graph," in IEEE Transactions on Multi-Scale Computing Systems, vol. 4, no. 3, pp. 231-242, 1 July-Sept. 2018, doi: 10.1109/TMSCS.2018.2802920.
6. J. Wang, J. Pei, and J. Han. Closet+: Searching for the best strategies for mining frequent closed itemsets. In Proc. of the 9th ACM SIGKDD Conference, pages

236–245, 2003.

1. M. J. Zaki and C.-J. Hsiao. Charm: An efficient algorithm for closed itemset mining.

In Proc. of SIAM International Conference on Data Mining, pages 398–416, 2002.

1. G. Liu, H. Lu, W. Lou, and J. X. Yu. On computing, storing and querying frequent patterns. In Proc. of the 9th ACM SIGKDD Conference, pages 607–612, 2003.
2. T. Uno, M. Kiyomi, and H. Arimura. Lcm ver. 2: Efficient mining algorithms

for frequent/closed/maximal itemsets. In Proc. of the ICDM 2004 Workshop on

Frequent Itemset Mining Implementations, 2004.

**华中科技大学本科生毕业设计（论文）开题报告评审表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **姓名** |  | **学号** |  | **指导教师** |  |
| **院（系）专业** | |  | | | |
| **指导教师评语**   1. 学生前期表现情况。 2. 是否具备开始设计（论文）条件？是否同意开始设计（论文）？ 3. 不足及建议。 | | | | | |
| 指导教师（签名）：  年 月 日 | | | | | |
| **教研室（系、所）或开题报告答辩小组审核意见** | | | | | |
| 教研室（系、所）或开题报告答辩小组负责人（签名）：    年 月 日 | | | | | |