大规模不确定图上的 Top-k 极大团挖掘算法

邹兆年 朱 鎔

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 该文研究了从不确定图中挖掘出前 k 个出现概率最高的极大团的问题,提出了一种基于划分的高效并行算法.在该算法中,输入的大规模不确定图首先被划分为若干互不重叠的规模较小的子图,每个子图通过扩展邻居结点信息成为扩展子图.而后,应用改进后的分支界限搜索策略,并行挖掘各个扩展子图,以得到局部 top-k 结果.最后,归并所有的局部 top-k 结果,得到全局 top-k 极大团.同时,该文还提出了两种预处理策略,以提高算法效率.并且严格证明了算法的正确性.在多组不确定图数据集上的实验结果表明,算法具有很高的效率和很好的实用性.

关键词 不确定图;top-k 极大团;图划分算法;扩展子图

中图法分类号 TP311 **DOI 号** 10.3724/SP.J.1016.2013.02146

Mining Top-k Maximal Cliques from Large Uncertain Graphs

ZOU Zhao-Nian ZHU Rong

(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

Abstract This paper investigates the problem of mining k top-ranked vertex subsets in an uncertain graph which have the highest probabilities of being maximal cliques in practice. A decomposition-based algorithm taking advantage of parallelism is proposed to solve the problem on large uncertain graphs. In the algorithm, the input large uncertain graph is firstly decomposed into some disjoint subgraphs in much smaller scale by an efficient graph division algorithm, and subgraphs are then extended to be extension subgraphs by bringing some adjacent vertices in. Then, local top-k results of maximal cliques are obtained on each extension subgraph in parallel by a mining algorithm adopting branch and bound search method. Finally, the local results are merged together to get the final top-k maximal cliques in the global input uncertain graph. Moreover, we provide two preprocessing methods to improve the algorithm's efficiency by decreasing the input graph size. It is proved that the algorithm could guarantee the completeness and correctness of the final mining results. Also, extensive experiments are made across several uncertain graph datasets to evaluate our algorithm. It's shown that the algorithm is both effective and efficient. Experiment results verify that this algorithm is of great significance in real applications.

Keywords uncertain graph; top-k maximal cliques; graph division; extension subgraph

1 引 言

近年来,随着图数据模型的迅速发展,多种实际应用都使用图来建模和表示数据,如社交网络[1]、Web 网络[2]、生物蛋白质网络[3]等.通过图模型挖掘数据中的知识具有重要的研究意义和应用价值.因此,图数据挖掘已经成为一个重要的研究方向并

收稿日期:2013-06-26;最终修改稿收到日期:2013-08-29.本课题得到国家自然科学基金(61173023)、中央高校基本科研业务费专项资金(HIT.NSRIF.201180)资助.邹兆年,男,1979 年生,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为数据库、数据挖掘.E-mail:znzou@hit.edu.cn.朱 鎔,男,1992 年生,硕士研究生,主要研究方向为图数据管理、数据库.

且大量算法[4-5]被提出. 其中，极大团枚举问题（简称MCE）是一个十分重要的问题. 在图中，团是一个任意两顶点均邻接的顶点集合，极大团是指不是任何其它团的子集的团. 极大团枚举问题就是挖掘给定图中全部极大团. 由于具有最佳的连通性，团和极大团通常被视为图中稠密子结构的核心部分，蕴藏着集中的知识信息. 因此，MCE 被广泛应用于实际应用中，如社交网络中的社区发现[1]、Web 网络中的拓扑分析[2]等. 同时，MCE 问题和图论中的图染色问题[6]、最大独立集问题[7]等理论问题具有密切联系.

长期以来，人们对于 MCE 问题进行了深入研究. 文献[5,7-11]中提出了一些具有代表性的算法. 虽然理论上 MCE 是一个 NP-难问题[7]，但实验表明上述算法在实际的稀疏图中表现良好[5]. 然而，现有的 MCE 算法所关注的图模型都是准确和完备的确定图. 在现实应用中，图数据往往带有内生的不确定性，称为不确定图. 例如，在无线传感器网络中，我们用点表示传感器结点，用边表示结点间的通信链路. 由于能量耗尽，网络中的某些结点可能失效，同时由于通信干扰，结点之间的通信链路可能受到影响[12]. 因此，整个网络中的顶点和边都以一定的概率存在，整体的图结构具有不确定性.

在不确定图中挖掘极大团同样具有重要意义[13]. 例如挖掘生物蛋白质网络中的极大团可以预测不同蛋白质在功能上的密切联系. 但是由于缺乏对不确定性的考虑，现有的 MCE 算法在不确定图上并不适用. 假设在某个极大团中，顶点和边的存在概率均很低，那么其作为极大团结构真实存在的概率很小，不具备现实的应用价值. 因此，不确定图上的 MCE 算法必须建立合适的评价指标，区分不同极大团的出现可能并找出其中存在概率较高的极大团. 在我们的前期工作[13]中，作者提出了极大团概率的概念，用于表示不确定图中某个顶点集合作为极大团真实出现的概率. 同时，文中提出了挖掘不确定图中 top-k 极大团概率的分支界限搜索算法. 虽然该算法适用于不确定图上的极大团挖掘，但在处理大规模不确定图时面临挑战. 为解决此问题，本文将其进行拓展并提出一种基于划分的并行算法，用于不确定大图中的极大团发掘.

首先，本文采用文献[14]中提出的一种高效图划分算法将一个不确定大图划分为若干个规模较小的不确定子图，并在各个子图上并行处理. 为了保证结果正确和完整，对划分得到的子图进行适当扩展，得到扩展子图. 而后在每个扩展子图上，将文献[13]中的分支界限算法加以改动以挖掘其中的 top-k 极大团，称为局部结果. 最后，归并所有的局部结果得到全局的 top-k 极大团结果. 另外，我们还提出了两种预处理策略，通过减小输入图的规模来提高算法效率. 为了测试算法的正确性和效率，我们在多组不确定图数据上进行了实验. 实验结果表明，算法具有很高的效率，并且挖掘结果具有很高的实用价值.

在本文第 2 节介绍本文相关工作；第 3 节给出问题的形式化描述；第 4 节介绍基于划分的并行算法；第 5 节给出实验结果与分析；第 6 节总结本文工作.

2　相关工作

极大团枚举问题的相关算法研究为时已久，文献[15]中给出了关于 MCE 问题典型算法的详细综述. 其中主要的工作包括文献[8]中提出的基于子问题规约的算法、文献[9]中基于深度优先搜索的算法、文献[11]中基于并行搜索树的算法等. 文献[7]中提出了一种在多项式时间间隔内连续产生极大团的算法. 文献[16]对其进行了改进和提高. 在文献[5]中，作者提出了一种高效的外存算法，可通过图中局部顶点的信息计算得到全局的极大团结果，具有很好的并行性和扩展性. 但是，这些算法均没有考虑图数据中的不确定性，不能从概率的角度给出出现可能较大的极大团. 因此，上述 MCE 算法并不适合于不确定图上的极大团挖掘.

不确定图的相关研究工作最近开展较多，提出了大量的模型和挖掘算法. 文献[4,13,17]给出了用概率加权形式表示的不确定图模型. 文献[4,17]给出了概率语义下不确定图中挖掘频繁子模式的定义与方法. 作为本文的前期工作，文献[13]中给出了不确定图上 top-k 极大团问题的形式化定义和一种分支界限搜索算法. 为了解决其在处理大图数据时面临的挑战，我们将以此为基础提出一种新的并行算法.

为了适应大图数据处理，文中采用了图划分的策略. 早期基于边加权方式[18]的图划分算法不够高效，而基于层次聚类和基于最大-最小割[19]的算法只在特定图中效果较好. 文献[14]提出了基于模块度的划分算法，其时间复杂度达到了 $O(|V|\cdot\log^2(|V|))$，非常适用于处理大规模图数据. 因此，本文将采用此算法来划分不确定大图.

3 问题定义

本节介绍一种不确定图模型，并形式化描述不确定图上的 top-k 极大团挖掘问题.

3.1 不确定图模型

不确定图是一个四元组 $\mathcal{G}=(V,E,P_V,P_E)$，其中 $G=(V,E)$ 是一个无向确定图；$P_V:V\rightarrow(0,1]$ 是一个函数，它为每个顶点 v 赋予存在概率 $P_V(v)$，用于表示顶点 v 实际存在的概率；$P_E:E\rightarrow(0,1]$ 也是一个函数，它为每条边 $e=(u,v)$ 赋予条件存在概率 $P_E(e)$，表示在顶点 u 和 v 存在的条件下，边 e 的存在概率. 因此，若 $P_V(v)=1$，则顶点 v 一定存在；若 $P_E(e)=1$，则边 e 在两个端点存在的条件下一定存在. 于是，一个确定图等价于一个所有顶点的存在概率均为 1 且所有边的条件存在概率也均为 1 的特殊不确定图.

由于不确定图中的顶点和边是否真正存在是不确定的，因此一个不确定图实际对应着一组可能的确定图结构. 我们称不确定图 $\mathcal{G}=(V,E,P_V,P_E)$ 导出确定子图 $G=(V',E')$，记为 $\mathcal{G}\Rightarrow G$，当且仅当 G 满足：

1. $\{v\mid v\in V,\ P_V(v)=1\}\subseteq V'\subseteq V$,
2. $\{e=(u,v)\mid u,v\in V',P_E(e)=1\}\subseteq E'\subseteq E$.

其中条件 1 说明 V' 是顶点集 V 的子集，并且确定存在的顶点（即 $P_V(v)=1$ 的顶点）一定要出现在 V' 中. 条件 2 说明 G 中所有边的端点均在顶点子集 V' 中，并且在 V' 存在的条件下，确定存在的边（即 $P_E(e)=1$ 的边）一定要出现在 E' 中.

在本文中，我们假设不确定图中各个顶点的存在概率相互独立，同时各条边的条件存在概率也相互独立. 那么，不确定图 $\mathcal{G}=(V,E,P_V,P_E)$ 导出确定子图 $G=(V',E')$ 的概率为

$$P(\mathcal{G}\Rightarrow G)=\prod_{v\in V'}P_V(v)\cdot\prod_{v\in V\setminus V'}(1-P_V(v))\cdot\prod_{(u,v)\in E'}P_E(u,v)\cdot\prod_{(u,v)\in(E\cap(V'\times V'))\setminus E'}(1-P_E(u,v)) \tag{1}$$

例如，图 1(a) 给出了一个不确定图 $\mathcal{G}$. 图 1(b) 给出了 $\mathcal{G}$ 的可能导出的一个确定子图 G. 容易验证，$P(\mathcal{G}\Rightarrow G)\approx 1.0838\times10^{-4}$.

我们用 $\Omega(\mathcal{G})$ 表示不确定图 $\mathcal{G}$ 导出的所有确定子图的集合. 可以看出，函数 $p(G)=P(\mathcal{G}\Rightarrow G)$ 是 $\Omega(\mathcal{G})$ 上的一个概率质量函数. 为了表述方便，我们将确定图 (V,E) 称为不确定图 $\mathcal{G}=(V,E,P_V,P_E)$ 对应的主确定图，记为 $\hat{\mathcal{G}}=(V,E)$.

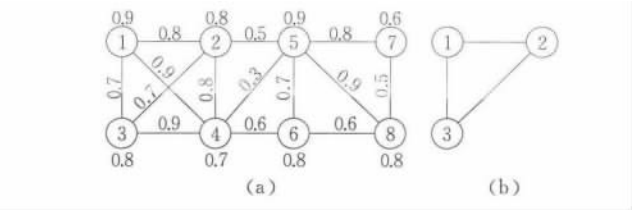

图 1 不确定图及其导出的确定图

3.2 问题描述

给定确定图 $G=(V,E)$ 和顶点子集 $C\subseteq V$，如果对于任意两个不同顶点 $u,v\in C$，均有 $(u,v)\in E$，则称 C 为团. 如果 C 是团，并且 C 不是其它任何团的子集，则称 C 为极大团. 团 C 中所包含的顶点数量称为 C 的大小，记作 $|C|$. 例如，在图 1(a) 中的不确定图 $\mathcal{G}$ 的主确定图 $\hat{\mathcal{G}}$ 中，顶点集 $\{1,2,3\}$ 是一个大小为 3 的团，顶点集 $\{5,6,8\}$ 则是一个极大团.

给定不确定图 $\mathcal{G}=(V,E,P_V,P_E)$ 和顶点子集 $C\subseteq V$，我们用 $mcliq(C)$ 表示 C 在 $\Omega(\mathcal{G})$ 的所有确定图中成为极大团的事件. 因此，事件 $mcliq(C)$ 发生的概率等于 C 作为极大团出现的确定图的导出概率之和，即：

$$P(mcliq(C))=\sum_{G\in\Omega(\mathcal{G}),C\text{在}G\text{中是极大团}}P(\mathcal{G}\Rightarrow G) \tag{2}$$

为简便起见，我们将 $P(mcliq(C))$ 称为 C 的极大团概率. $P(mcliq(C))$ 越高，表示顶点子集 C 在实际中以极大团结构出现的可能性越大. 在实际应用中，人们往往只关注给定不确定图中极大团概率前 k 个大的顶点子集，称为不确定图上的 top-k 极大团. 因此，不确定图上的 top-k 极大团问题描述如下：

输入：不确定图 $\mathcal{G}=(V,E,P_V,P_E)$，正整数 k 和 s.

输出：包含 k 个顶点集合的集族 F，对于任意顶点子集 $C\in F$，均有 $|C|\geq s$；且对任何顶点集合 $C'\notin F$，$|C'|\geq s$，均有 $P(mcliq(C))\geq P(mcliq(C'))$.

文献[13]证明了不确定图上的 top-k 极大团问题是一个 NP-难问题.

4 算法描述

本节提出一种基于划分的并行挖掘算法，用于挖掘不确定大图中的 top-k 极大团. 首先，4.1 节回顾经典的分支界限搜索算法[13]. 而后，4.2 节给出基于划分的算法的细节. 在 4.3 节中，我们介绍了两种提高算法效率的预处理策略.

4.1 分支界限搜索算法

我们首先回顾文献[13]中提出的用于挖掘不确定图中的 top-k 极大团的分支界限搜索算法. 为了

计算极大团概率，我们引入团概率的概念. 在不确定图$\mathcal{G}=(V,E,P_V,P_E)$中，用 $cliq(C)$表示顶点子集 $C\subseteq V$在$\Omega(\mathcal{G})$的所有确定图中成为团的事件. 与极大团概率类似，我们称 $cliq(C)$为 C 的团概率，定义为

$$P(cliq(C))=\sum_{G\in\Omega(\mathcal{G}),C\text{在}G\text{中是团}}P(\mathcal{G}\Rightarrow G) \quad (3)$$

在不确定图$\mathcal{G}$中，如果顶点子集 C 在主确定图$\hat{\mathcal{G}}$中不是团，那么 C 在 $\Omega(\mathcal{G})$中所有确定图中均不是团. 因此，我们只考虑在$\hat{\mathcal{G}}$中成为团的顶点集合. 下述命题 1 和 2 给出了在多项式时间内计算团概率 $P(cliq(C))$和极大团概率 $P(mcliq(C))$的方法.

命题 1.　给定不确定图$\mathcal{G}=(V,E,P_V,P_E)$，对于主确定图$\hat{\mathcal{G}}$中任意团 C，C 的团概率为

$$P(cliq(C))=\prod_{v\in C}P_V(v)\cdot\prod_{u,v\in C,u\neq v}P_E(u,v) \quad (4)$$

命题 2.　给定不确定图$\mathcal{G}=(V,E,P_V,P_E)$，对于主确定图$\hat{\mathcal{G}}$中任意团 C，设 $C_1,C_2,\cdots,C_m$是$\hat{\mathcal{G}}$中比 C 多出一个顶点的团，那么 C 的极大团概率为

$$P(mcliq(C))=P(cliq(C))\prod_{i=1}^{m}\left(1-\frac{P(cliq(C_i))}{P(cliq(C))}\right) \quad (5)$$

有关命题 1 和 2 的详细证明可以在文献[13]中得到. 下面给出分支界限搜索算法[13]. 设$\prec$是顶点集 V 上的一个偏序关系，根据$\prec$可将$\hat{\mathcal{G}}$中所有的团组织成一颗搜索树. 其中搜索树的根节点为空集$\varnothing$，树中的每个结点对应于一个团 C，其儿子结点为比 C 多一个顶点的团 C'，并且对$\{w\}=C'\backslash C$和任意顶点 $v\in C$，均有 $v\prec w$. 具体算法描述如下.

算法 1.　分支界限搜索算法.

输入：不确定图$\mathcal{G}$，正整数 k 和 s

输出：包含 k 个顶点集的集族 F

S1：初始化 F 为空集，用于存放已得到的部分 top-k 概率的极大团. 初始化优先级队列 Q 用于存放待搜索的团，其中 Q 中元素按照团概率递减排列，F 中元素按照极大团概率递增排列.

S2：对任意顶点 $v\in V$，根据式(4)计算单顶点团的团概率 $P(cliq(\{v\}))$，而后将$\{v\}$插入 Q 中.

S3：重复下列步骤，直到 Q 为空或者 Q 中最大元素的团概率小于 F 中最小元素的极大团概率：

1. 取 Q 队首元素 C，按照式(5)计算 $P(mcliq(C))$.

2. 若$|C|\geq s$，则更新 F，即如果$|F|<k$，则将 C 直接插入 F；否则，若 $P(mcliq(C))$大于 F 中最小元素的极大团概率，则将 F 中最小元素弹出，插入 C.

3. 枚举 C 在搜索树中的儿子结点，根据式(4)计算其团概率并插入 Q 中.

S4：输出 F 中的全部元素，即为 top-k 概率的极大团.

文献[13]中描述了算法的更多细节步骤. 对于图 1(a)中的不确定图，设 $k=5$，$s=3$，则其 top-5 极大团为 $\{5,6,8\}$，$\{1,2,4\}$，$\{1,3,4\}$，$\{5,7,8\}$ 和 $\{2,3,4\}$.

4.2　基于划分的算法

本节提出了一种基于划分的 top-k 极大团并行挖掘算法，该算法的基本过程如下：给定不确定图$\mathcal{G}=(V,E,P_V,P_E)$和正整数 k 和 s，首先将顶点集 V 划分为 N 个互不相交的顶点子集 $V_1,V_2,\cdots,V_N$，使得$\bigcup_{i=1}^{N}V_i=V$，其中 N 表示并行机器的数量. 这样，我们可得到由 $V_1,V_2,\cdots,V_N$ 导出的不确定子图 $\mathcal{G}_1$，$\mathcal{G}_2,\cdots,\mathcal{G}_N$. 为了保证结果的正确性，我们对子图 $\mathcal{G}_1$，$\mathcal{G}_2,\cdots,\mathcal{G}_N$进行扩展，得到 N 个不确定扩展子图$\mathcal{G}_1^+$，$\mathcal{G}_2^+,\cdots,\mathcal{G}_N^+$. 然后将$\mathcal{G}_1^+,\mathcal{G}_2^+,\cdots,\mathcal{G}_N^+$分布于 N 台机器上，并使用改进后的分支界限搜索算法在扩展子图 $\mathcal{G}_1^+,\mathcal{G}_2^+,\cdots,\mathcal{G}_N^+$上并行挖掘，分别得到每个$\mathcal{G}_i^+$上的局部 top-$k$ 极大团 F_i. 最后，归并所有的局部结果 F_1，$F_2,\cdots,F_N$，得到$\mathcal{G}$中的 top-k 极大团 F.

下面，我们首先介绍图划分算法，然后给出子图扩展方法并证明其正确性，而后介绍经过改进的分支界限搜索算法，最后给出整体算法.

4.2.1　图划分算法

给定不确定图$\mathcal{G}=(V,E,P_V,P_E)$和正整数 N，我们将 V 划分为 N 个互不相交的子集 $V_1,V_2,\cdots,V_N$，使得$\bigcup_{i=1}^{N}V_i=V$. 注意到实际上任何满足此要求的划分策略在此处均适用，然而不同划分策略的代价截然不同. 直观上，我们希望划分得到的子图“高内聚，低耦合”，即同一划分内部的边尽可能多，而连接不同划分的边尽可能少. 同时对于大规模图，划分算法的效率也至关重要. 因此，本文选用了文献[14]中的算法，其时间复杂度为 $O(|V|\cdot\log^2(|V|))$.

在该算法中，为衡量划分质量，作者提出了模块度的概念. 对划分 $V_1,V_2,\cdots,V_N$，其模块度$\mathcal{Q}$定义为

$$\mathcal{Q}=\sum_{i=1}^{N}(e_{ii}-a_i^2) \quad (6)$$

此处 $e_{ii}=\delta(V_i,V_i)/m$，$a_i=\sum_k\delta(V_i,V_k)/m$，其中 m 是图中边的数量，$\delta(V_i,V_j)$表示连接划分 V_i和 V_j 的边的数量. 因此，e_{ii}表示划分 V_i内部边的比例，而 a_i表示和 V_i中顶点邻接的边所占比例. 模块度$\mathcal{Q}$越高，算法的划分效果越好. 为此，算法采用迭代方式进行处理. 初始时，每个顶点都是一个单独的划分. 然后在每一轮迭代中，选取能够使模块度提高最大的两个划分进行合并，直到划分数量到达 N 为止. 在该算法中，为了能够快速地找到最大的 $\Delta\mathcal{Q}$，算法

采用了高效的数据结构进行优化.此处,我们给出该算法的一个概括性描述.

算法 2. 图划分算法.

输入:不确定图$\mathcal{G}=(V,E,P_V,P_E)$,划分数量 N

输出:N 个不相交集合 $V_1,V_2,\cdots,V_N$,并有 $\bigcup_{i=1}^{N}V_i=V$

S1:初始化一个 $|V|$ 行的 $\Delta\mathcal{Q}$ 矩阵,存储划分 V_i 和 V_j 合并时的模块度变化值 $\Delta\mathcal{Q}(i,j)$.将 $\Delta\mathcal{Q}$ 的每一行都同时组织成一个最大堆和一棵平衡树.初始化一个全局最大堆 H,将 $\Delta\mathcal{Q}$ 的每行最大元素连同位置信息 (i,j) 一同压入 H 中.

S2:若划分规模大于 N 则继续,否则转向 S4.

S3:取 H 的堆顶元素 $\Delta\mathcal{Q}(i,j)$,将 $\Delta\mathcal{Q}$ 矩阵的 i 和 j 行合并到 i 行,合并后的 i 行包含原 i 和 j 行中所有顶点.更新矩阵 i 行的元素值,并更新最大堆和平衡树,而后更新全局最大堆 H.返回 S2.

S4:输出 $\Delta\mathcal{Q}$ 矩阵中每行所代表的顶点集为划分结果 $V_1,V_2,\cdots,V_N$.

文献[14]中给出了矩阵元素初始值和更新值的详细计算公式.同时文中给出了算法时间复杂度的详细分析过程.

4.2.2 子图扩展

在不确定图$\mathcal{G}$中,顶点子集 $S\subseteq V$ 导出的不确定子图为 $\mathcal{G}[S]=(V_S,E_S,P_V,P_E)$,其中 $V_S=S,E_S=(S\times S)\cap E$.对划分结果 $V_1,V_2,\cdots,V_N$,其对应的不确定导出子图为 $\mathcal{G}_1=\mathcal{G}[V_1],\mathcal{G}_2=\mathcal{G}[V_2],\cdots,\mathcal{G}_N=\mathcal{G}[V_N]$.但直接在 $\mathcal{G}_i$ 上运行挖掘算法并不能保证最终结果的正确性.若某个团 C 中的顶点被划分到若干不同的划分中,那么在每个划分导出的子图中,C 均不会被搜索和计算到,最终结果中和 C 一样的跨划分的团将全部丢失.例如在图 1(a)中,若 $V_1=\{1,2,3,4,6\},V_2=\{5,7,8\}$,那么团 $\{5,6,8\}$ 将从最终结果中丢失.因此,我们必须对子图进行扩展,使得每个团都包含在至少一个扩展子图中.

为了保证挖掘结果的完整性,我们需要将划分 V_i 的邻居信息包含到子图中.对于划分 V_i,其邻居顶点集定义为 $N(V_i)=\{v\mid(u,v)\in E,u\in V_i,v\notin V_i\}$.我们用 $\mathcal{G}_i^+$ 表示由 $V_i\cup N(V_i)$ 导出的不确定子图,称为划分 V_i 对应的扩展子图.下面我们证明这种扩展方式的正确性,首先我们通过定理 1 证明每个团 C 都包含在至少一个扩展子图 $\mathcal{G}_i^+$ 中.

定理 1. 对于 $\hat{\mathcal{G}}$ 中的任意团 C,一定存在一个划分 V_i,使得 $C\subseteq V_i\cup N(V_i)$ 且 $C\cap V_i\neq\varnothing$.

证明. 对于任意团 $C\subseteq V$ 和顶点 $v\in C$,设 V_i 为 v 所属划分,显然 $C\cap V_i\neq\varnothing$.由于 C 是团,对任意其它顶点 u,u 和 v 必邻接.因此,若 u 不在 V_i 中,则 u 一定在 $N(V_i)$ 中.故存在划分 V_i 使得 $C\subseteq V_i\cup N(V_i)$. 证毕.

定理 1 保证了挖掘结果的完整性.下面的定理 2 保证了结果的正确性,即在扩展子图上得到的极大团概率可以和在全图中计算得到的相同.

定理 2. 对于 $\hat{\mathcal{G}}=(V,E)$ 中的任意团 C,若 $C\subseteq V_i\cup N(V_i)$ 且 $C\cap V_i\neq\varnothing$,则 C 在 $\mathcal{G}_i^+$ 中的极大团概率与在 $\mathcal{G}$ 中的极大团概率相等.

证明. 根据式(5),若要计算 C 的极大团概率,必须计算 C 和比 C 仅多一个顶点的团 C_i 的团概率.根据式(4),若计算 C 和 C_i 的团概率,则要知道 C 和 C_i 中所有点和边的概率信息.由于 $C\subseteq V_i\cup N(V_i)$,因此在 $\mathcal{G}_i^+$ 中可以直接计算 $P(cliq(C))$.对于 C_i,设 $\{v_i\}=C_i\backslash C$,由于 C_i 为团且 $C\cap V_i\neq\varnothing$,于是有 $v_i\in V_i\cup N(V_i)$,因此可以计算 $P_{(cliq(C_i))}$.综上所述,可知在 $\mathcal{G}_i^+$ 中可以计算 $P(mcliq(C))$ 并且和在 $\mathcal{G}$ 中计算的结果相同. 证毕.

定理 2 保证了在扩展子图 $\mathcal{G}_i^+$ 上挖掘得到的至少包含一个 V_i 中顶点的极大团是正确的.综合定理 1 和 2,我们可以先将图划分,而后挖掘所有扩展子图 $\mathcal{G}_i^+$ 上至少包含一个 V_i 中顶点的极大团.下面的命题说明所有扩展子图中的 top-k 极大团归并以后,可以得到正确的全局 top-k 极大团.

命题 3. 设 $C_1,C_2,\cdots,C_k$ 是 $\mathcal{G}$ 中的 top-k 极大团,对于每一个 C_i,一定存在某个扩展子图 $\mathcal{G}_j^+$ 使得 C_i 是 $\mathcal{G}_j^+$ 上的 top-k 极大团.

证明. 根据定理 1,必定存在 $\mathcal{G}_j^+$ 使得 C_i 为其中的团,并且有 $C_i\cap V_j\neq\varnothing$.根据定理 2,其在 $\mathcal{G}_j^+$ 上的极大团概率可以计算,并且和全局极大团概率相等.由于其极大团概率在全局是 top-k 的,在 $\mathcal{G}_j^+$ 中必定也是 top-k 的.因此命题得证. 证毕.

综合上述,我们可以通过图划分和图扩展的方式得到准确完备的全局 top-k 极大团.考虑图 1(a),将图划分为 $V_1=\{1,2,3,4\}$ 和 $V_2=\{5,6,7,8\}$,其扩展子图在图 2 中给出,每个子图的邻居顶点用白色标出.

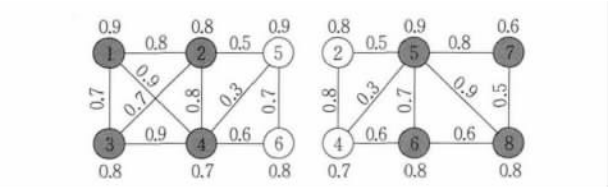

图 2 扩展子图示例

4.2.3 分支界限搜索算法改进

为了满足图划分和子图扩展方式下挖掘极大团的要求,我们要对扩展子图 $\mathcal{G}_i^+$ 上的分支界限搜索算法加以改进,以满足每个局部 top-k 极大团中包含

至少一个划分 V_i 中结点. 同时，注意到，对于跨划分的团的搜索存在重复计算的冗余情况. 例如图 2 中的团 $\{2,4,5\}$，其在 $\mathcal{G}_1^+$ 和 $\mathcal{G}_2^+$ 中均会被计算和搜索. 为了保证正确性和消除冗余，我们对分支搜索的策略加以改进，下面命题说明任何一个全局 top-k 极大团都可以仅被搜索一次.

命题 4.　设 C 是 $\mathcal{G}$ 中的 top-k 极大团，那么有且只有一个划分 V_i 使得 $C\cap V_i\neq\varnothing$，C 是 $\mathcal{G}_i^+$ 上的 top-k 极大团，并且对任意 $j<i$，V_j 内不包括 C 中的任何顶点.

证明.　我们分两种情况证明此命题.

在第 1 种情况中，团 C 仅属于某个划分 V_i. 根据命题 3，C 必是 $\mathcal{G}_i^+$ 上的 top-k 极大团. 同时对其它任意划分 V_j，均有 $C\cap V_j=\varnothing$.

在第 2 种情况中，团 C 被划分到多个划分 V_{k_1}，V_{k_2}，…，V_{k_M} 中. 不妨设 $k_1<k_2<\cdots<k_M$. 对任意 V_{k_j}，$1\leq j\leq M$ 均有 $C\cap V_{k_j}\neq\varnothing$. 根据定理 1 和 2，团 C 在其中任意一个划分中均可计算得到正确的极大团概率. 再根据命题 3，团 C 是全局的 top-k 极大团，那么团 C 在任意一个划分所对应扩展子图上也是 top-k 极大团. 在此，令 $i=k_1$，由于 k_1 最小，因此不存在 $j<i$ 并且 V_j 包含 C 中顶点. 在并行搜索过程中，我们只需在 $\mathcal{G}_{k_1}^+$ 中搜索团 C，在其它划分对应的扩展子图上忽略团 C 即可.　　证毕.

根据命题 4，对于扩展子图 $\mathcal{G}_i^+$，可以要求算法在搜索每个团在搜索树中的子结点时只考虑向当前团中增加 V_{i+1}，V_{i+2}，…，V_N 中的顶点. 这样一个全局 top-k 极大团就不会在不同划分中被重复搜索. 同时在算法初始阶段，可以仅将 V_i 中的顶点放入搜索队列. 通过设定偏序关系，使得 $\forall u\in V_i$，$v\in N(V_i)$ 均有 $u\prec v$，这样就可以保证所有被搜索到的团中都包含至少一个 V_i 中顶点. 下面我们给出改进后的分支界限搜索算法.

算法 3.　改进分支界限搜索算法.

输入：不确定子图 $\mathcal{G}_i^+$，正整数 k 和 s

输出：k 个顶点集的集族 F_i

S1：初始化 F_i 为空集. 初始化优先级队列 Q. 其中 Q 中元素按照团概率递减排列，F_i 中元素按照极大团概率递增排列.

S2：对所有顶点 $v\in V_i$，计算 $P_{(cliq(\{v\}))}$，而后将 $\{v\}$ 插入 Q 中.

S3：重复下列步骤直到 Q 中没有元素或者 Q 中最大元素的团概率小于 F 中最小元素的极大团概率：

1. 取 Q 队首元素 C，计算极大团概率 $P(mcliq(C))$.

2. 若 $|C|\geq s$，则更新 F_i，即如果 $|F_i|<k$ 直接插入 F_i；否则若 $P(mcliq(C))$ 大于 F_i 中最小元素的极大团概率，则将 F_i 中最小元素弹出并插入 C.

3. 枚举 C 的子树结点，如果新增结点 $u\in V_j$ 且 $j\geq i$，则计算其团概率并插入 Q 中.

S4：输出 F_i 中的全部元素，即为 $\mathcal{G}_i^+$ 上的 top-k 极大团.

4.2.4　算法小结

至此，可以将本小节的各部分算法加以总结得到最终算法.

算法 4.　基于划分的并行算法.

输入：不确定图 $\mathcal{G}$，正整数 k 和 s

输出：k 个顶点集的集族 F

S1：对于 $\mathcal{G}$ 划分得到顶点集合 V_1，V_2，…，V_N，经过子图扩展后得到扩展子图 $\mathcal{G}_1^+$，$\mathcal{G}_2^+$，…，$\mathcal{G}_N^+$.

S2：对于 $\mathcal{G}_i^+(1\leq i\leq N)$ 并行处理，运用改进的分支界限搜索算法得到局部结果 F_i.

S3：将得到的 $F_i(1\leq i\leq N)$ 按照极大团概率使用选择树或者最小堆归并，得到最终的全局 top-k 极大团结果 F.

4.3　预处理策略

为了提高分支界限的搜索效率，我们给出两种预处理搜索策略：度过滤和边过滤.

度过滤方法如下：由于团大小至少是 s，那么一个度数小于 $s-1$ 的顶点肯定不会出现在最终结果中. 因此，我们可以迭代地移除图中所有度数小于 $s-1$ 的顶点，直至图中不包含度数小于 $s-1$ 的顶点为止.

边过滤的方法如下：对于边 $e=(u,v)$，如果 u 和 v 能够形成团，则共同邻接 u 和 v 的顶点个数不能小于 $s-2$. 否则，边 (u,v) 可以移除. 因此，我们可以迭代地移除所有不满足条件的边.

显然，度过滤和边过滤的方法对于不确定图 $\mathcal{G}$ 及其子图都适用. 两者可以相互配合直到两者条件均满足为止. 在算法 4 中，我们可以先对 $\mathcal{G}$ 进行预处理而后进行划分. 在子图挖掘前，可以对不确定子图进行预处理以提高算法效率.

5　实　验

为了验证本文算法的效率和正确性，我们进行了大量实验. 实验中我们测试了算法在不同输入参数下的性能表现以验证算法效率. 同时，我们测试了算法的预处理策略以及分支界限搜索算法的优化效果. 算法采用 C++ 编程实现，实验环境为 PC 机，2 GHz 处理器，2 GB 内存，运行 Windows 7 操作系统. 实验采用了下列 3 种数据集.

真实数据集. 实验中采用欧洲分子生物学实验室 STRING (http://string-db.org) 提供的真实不确定图数据. 数据集是一个蛋白质网络，其中顶点代

表蛋白质，具有确定性. 边代表不同蛋白质之间存在相互作用的概率，通过生物实验加以测定. 整个网络包括 6865 个顶点和 70288 条边.

合成数据集. 实验中的合成数据集采用确定图数据加上人工标注概率的方法生成. 真实的图数据是一个大规模的 PGP 网络，包括 10680 个结点和 24317 条边. 我们为其中所有的点和边随机生成存在概率，概率取值为随机生成的 0.000～0.999 之间的三位有效小数.

人造数据集. 我们随机生成了一规模人造网络，其包括 10^4 个顶点. 首先随机连接不同顶点，生成 10^5 条边. 而后为所有的顶点和边随机生成存在概率，概率取值为随机生成的 0.000～0.999 之间的三位有效小数.

5.1 算法参数实验

我们首先测试算法在不同输入参数情况下的性能表现，包括不同的划分数目 N、不同的结果集大小 k 和不同的最小团大小限制 s. 在测试某一参数时，我们固定其它参数，并给出算法的运行时间关于参数的变化关系，其中 Protein 曲线表示真实蛋白质数据集，PGP 曲线表示合成数据集，Artificial 代表人造数据集.

首先设定参数 $s=3$，$k=50$，测试划分数目 N 从 2～15 变化时算法的运行时间.

从图 3 可以看出，随着划分数目的增多，算法运行时间 t 首先呈现明显的下降趋势，在达到某阈值 N_c 后，算法运行时间基本保持稳定.

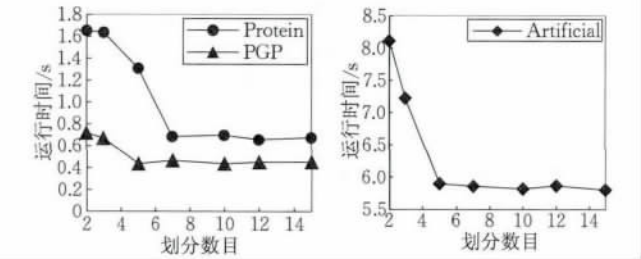

图 3 不同划分数目下的运行时间

此种实验现象的原因在于，在基于划分的算法中，算法的执行时间和划分与扩展后得到的不确定扩展子图规模相关. 当划分数量增加时，划分得到的顶点集规模逐渐减少. 在划分数量较少时，划分得到的顶点集合规模较大，故其增加邻居顶点的数量较少. 因此，此时不确定子图扩展前后规模变化不大，由于子图规模不断变小，算法运行时间快速下降. 随着划分数目增大，顶点集数目变小，新增加邻居结点的概率不断增大，经过扩展后子图规模基本保持稳定. 因此在达到某一阈值 N_c 后，算法运行时间保持稳定.

这一实验结果说明通过提高划分数目来降低算法运行时间的方法存在下界，其在实际应用中具有重要意义. 对于给定的数据集，我们只需通过实验测定 N_c 值，而后使用稍大于 N_c 的机器数量并行运行算法即可获得稳定的运行时间，大大节约了计算资源.

而后，通过对比可以看出采用本文中基于划分的算法在运行时间上的优势. 传统的算法包括穷举搜索算法、面向全图的分支界限搜索算法两种. 从理论上看，穷举搜索算法其枚举数量和图的顶点规模成指数关系，具有极高的时间和空间复杂度. 在图规模较大时，穷举法在有限计算时间内无法得到结果，因此不具有实用价值. 在 $N=1$ 时，本文算法即为面向全图的分支界限搜索算法. 在表 1 中，我们对比了全图分支界限搜索法和基于划分算法的稳定运行时间. 可以看出，基于划分算法其运行时间只有传统算法的 40%左右. 说明通过图划分的手段减小了计算规模，显著降低了运行时间.

表 1 算法时间对比

算法类型	运行时间/s		
	Protein	PGP	Artificial
分支界限	1.688	0.852	8.931
基于划分	0.676	0.437	5.753

而后设定参数 $s=3$，$N=10$，测试结果集大小 k 从 10～120 变化时算法的运行效率.

由图 4 看出，算法运行时间 t 关于结果集大小 k 呈现接近于线性的增长趋势. 分析算法过程可知这一趋势符合算法特性. 在算法中，结果集 F 可用最小堆实现. 初始时构建最小堆复杂度为 $O(k\cdot\log(k))$，而后向堆中删除或者插入一个元素复杂度为 $O(\log(k))$. 同时，增大 k 会减少最小概率导致削弱剪枝效果. 因此，算法时间对 k 呈现接近线性的增长趋势. 对于正常的 top-k 查询，这一效率足够满足需要.

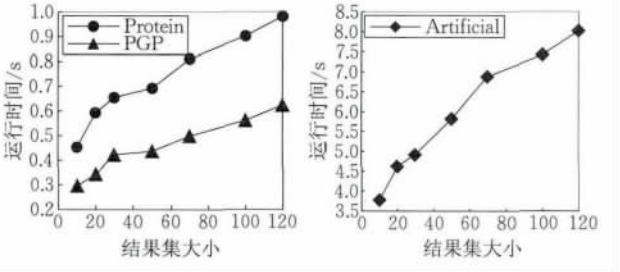

图 4 不同结果集大小下的运行时间

下面我们考察算法效率和最小团大小限制 s 的关系. 设定 $N=10$，$k=50$，令 s 由 3～7 增长. 通过测试算法运行时间可看出，在图 5 中，时间 t 和 s 呈现

明显的指数增长关系. 在分支界限搜索算法中，我们将图组织成一颗搜索树，算法的搜索规模等价于计算树中所有小于 s 层的结点概率. 随着 s 增大这一数字以指数形式增长，因此算法时间表现出相同的趋势. 在一般的 top-k 查询中，s 取值不大，因此对于算法效率影响有限.

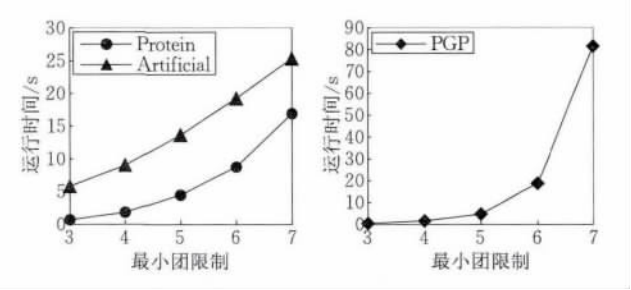

图 5　不同团大小限制下的运行时间

综合参数实验可以看出，选取合适的划分数目 N，对于一般的查询参数 k 和 s，算法具有较高的效率. 一般情况下，对于顶点规模在 10^4 左右的不确定图，算法处理时间在 10 s 以内. 其中，稠密图如人造数据集比稀疏图如合成数据集处理时间多出数倍，因此算法需要采取适当的预处理策略尽量降低输入图的规模.

5.2　算法优化措施实验

测试不同优化措施对于算法性能的影响，包括预处理策略(度过滤和边过滤)和分支界限搜索算法的消除冗余搜索的效果.

首先测试两种预处理策略度过滤和边过滤的优化效果. 预处理流程为迭代使用度过滤消除顶点和使用边过滤去除边. 在测试中，我们选定 $s=5$，$N=10$，变化结果集大小 k 值，对比观察有无预处理策略下算法的运行时间差异. 在图 6 中，我们给出了真实蛋白质图数据中预处理措施的优化效果. 对比算法运行时间可以发现，经过预处理后算法运行时间只有原先的 60%～70%. 证明预处理措施对提高算法效率有很好的效果.

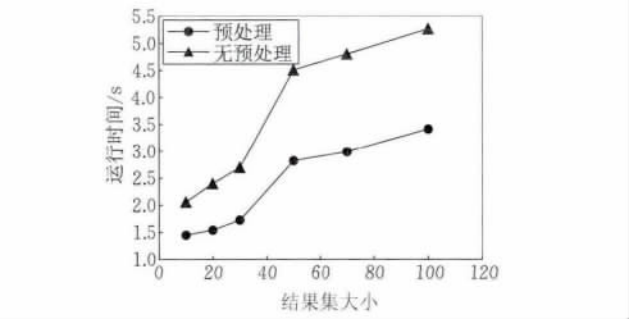

图 6　预处理测试的优化效果

同样，我们测试了改进分支界限算法消除冗余搜索的效果. 设定 $s=5$，$N=10$，变化结果集大小 k 值，我们对比算法是否采用冗余消除搜索算法的时间差异.

图 7 中显示了在 Protein 数据集上测试结果，可以看出使用冗余消除搜索算法时间降低程度在 10%以内. 这说明经过划分算法后跨划分的 top-k 极大团数量有限，因此冗余消除效果并不明显，但由于不增加程序实现难度，在实际应用中此方法可以采用. 在其它数据集中的类似结果，由于篇幅限制不再列出.

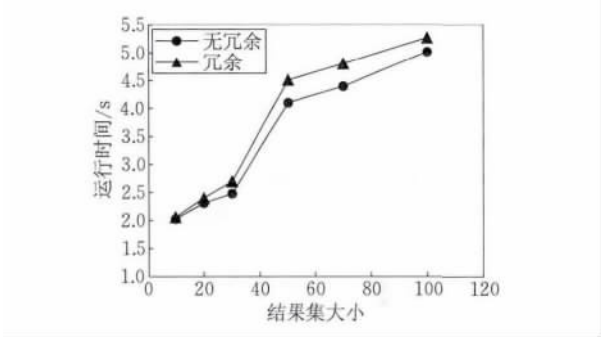

图 7　消除冗余搜索的优化效果

5.3　算法结果应用实例

此处，我们给出一个本文算法的应用实例，针对实验中 Protein 数据集得到的 top-5 极大团如表 2 所示. 在 top-1 结果中，我们认为有 99%以上的可能 F2，FGB 和 FGG 这 3 种蛋白质分子存在相互作用. 其中，FGB 和 FGG 蛋白质属于凝血酶. 根据最新研究成果，F2 蛋白和 FGX 系列蛋白共同作用，在治疗先天性凝血酶原缺陷症和遗传性纤维蛋白原血症中有重要作用[20]. 因此，针对蛋白质网络的极大团挖掘分析，可以帮助发现未知的蛋白质作用途径，为生物学实验和药物开发提供方向和指导. 对于化合物分子网络的分析可以为化合物性质分析、化学合成提供帮助.

表 2　top-5 蛋白质极大团

序号	蛋白质	概率
1	F2 FGB FGG	0.991026
2	ABI1 BAIAP2 WASF2	0.991017
3	ABI1 BAIAP2 WASL	0.984079
4	CDK8 CYCC SKD	0.977139
5	DOCK1 ELMO1 RHOG	0.966249

6　结　论

本文提出了一种在大规模不确定图上挖掘

top-k 极大团的并行算法.该算法采用的图划分策略使得我们可以在划分子图上并行地挖掘局部极大团.该算法的子图扩展策略保证了对局部挖掘结果进行合并可以得到正确的全局挖掘结果.实验结果表明,该并行挖掘算法具有很好的执行效率,与传统算法相比显著提高了挖掘效率.

参 考 文 献

[1] Xiao D, Du N, Wu B, et al. Community ranking in social network//Proceedings of the 2nd International Multi-Symposiums on Computer and Computational Sciences. Lowa, USA, 2007: 322-329

[2] Kumar R, Raghavan P, Rajagopalan S, et al. Trawling the web for emerging cyber-communities. Computer Networks, 1999, 31(11): 1481-1493

[3] Yu H, Paccanaro A, Trifonov V, et al. Predicting interactions in protein networks by completing defective cliques. Bioinformatics, 2006, 22(7): 823-829

[4] Zou Zhao-Nian, Li Jian-Zhong, Gao Hong, et al. Mining frequent subgraph patterns from uncertain graphs. Journal of Software, 2009, 20(11): 2965-2976(in Chinese)
(邹兆年,李建中,高宏等.从不确定图中挖掘频繁子图模式.软件学报,2009,20(11):2965-2976)

[5] Cheng J, Ke Y, Fu A W C, et al. Finding maximal cliques in massive networks by h*-graph//Proceedings of the 2010 International Conference on Management of Data. Indianapolis, USA, 2010: 447-458

[6] Byskov J M. Algorithms for k-colouring and finding maximal independent sets//Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms. Baltimore, USA, 2003: 456-457

[7] Tsukiyama S, Ide M, Ariyoshi H, et al. A new algorithm for generating all the maximal independent sets. SIAM Journal on Computing, 1977, 6(3): 505-517

[8] Akkoyunlu E A. The enumeration of maximal cliques of large graphs. SIAM Journal on Computing, 1973, 2(1): 1-6

[9] Tomita E, Tanaka A, Takahashi H. The worst-case time complexity for generating all maximal cliques and computational experiments. Theoretical Computer Science, 2006, 363(1): 28-42

[10] Modani N, Dey K. Large maximal cliques enumeration in sparse graphs//Proceedings of the 17th ACM Conference on Information and Knowledge Management. Napa Valley, USA, 2008: 1377-1378

[11] Du N, Wu B, Xu L, et al. A parallel algorithm for enumerating all maximal cliques in complex network//Proceedings of the 6th IEEE International Conference on Data Mining. HongKong, China, 2006: 320-324

[12] Eschenauer L, Gligor V D. A key-management scheme for distributed sensor networks//Proceedings of the 9th ACM Conference on Computer and Communications Security. Washington, USA, 2002: 41-47

[13] Zou Z, Li J, Gao H, et al. Finding top-k maximal cliques in an uncertain graph//Proceedings of the 26th IEEE International Conference on Data Engineering. Long Beach, USA, 2010: 649-652

[14] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. Physical review E, 2004, 70(6): 066111

[15] Bomze I M, Budinich M, Pardalos P M, et al. Handbook of Combinatorial Optimization: The maximum clique problem. Netherlands: Kluwer Academic Publishers, 1999

[16] Makino K, Uno T. Algorithm Theory-SWAT2004: New Algorithms for Enumerating All Maximal Cliques. Berlin: Springer, 2004

[17] Zou Z, Gao H, Li J. Discovering frequent subgraphs over uncertain graph databases under probabilistic semantics//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 633-642

[18] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs. Bell System Technical Journal, 1970, 49: 291-307

[19] Ding C H Q, He X, Zha H, et al. A min-max cut algorithm for graph partitioning and data clustering//Proceedings of the 2001 IEEE International Conference on Data Mining. San Jose, USA, 2001: 107-114

[20] Emeis J J, Jirouskova M, Muchitsch E M, et al. A guide to murine coagulation factor structure, function, assays, and genetic alterations. Journal of Thrombosis and Haemostasis, 2007, 5(4): 670-679

ZOU Zhao-Nian, born in 1979, Ph.D., assistant professor. His research interests include databases and data mining.

ZHU Rong, born in 1992, M.S. candidate. His research interests include graph data management and databases.

Background

With the emergence of the massive data and the development of data mining, graphs have been used widely in modelling and representing various kinds of scientific data. Due to the existence of noise, incompleteness and inaccuracy, graph data tends to be uncertain in practice. In regardless of the probabilistic semantics in uncertain graph data, mining uncertain graph data greatly differs from the traditional technologies.

In this paper, we investigate on an important and typical problem in uncertain graph data mining, which is the maximal clique enumeration problem. In consideration of uncertain nature, each vertex set has a probability to be a maximal clique in all implicated subgraphs of a given uncertain graph. Thus, we may just focus on the k top-ranked sets of vertices in terms of their maximal clique probabilities, which is called the top-k maximal cliques in short.

In our prior work, we have researched on this problem have proposed a branch-and-bound algorithm to find top-k maximal cliques in the input uncertain graph. We give out the concept of *maximal clique probability*, that is, the probability that a set of vertices is a real maximal clique. The branch-and-bound algorithm find k top-ranked sets of vertices in terms of their maximal clique probabilities. However, this algorithm doesn't take advantages of the parallelism, so it doesn't scale on very large uncertain graphs. To solve this problem, we extend our branch-and-bound algorithm and propose a partition-based algorithm which can be easily parallelized to mine top-k maximal cliques from large uncertain graphs. This partition-based algorithm leverages the structure independency between different groups of cliques in practical uncertain graphs. The output of the algorithm is correct and complete. The experimental results also verify that the algorithm outperforms the branch-and-bound algorithm proposed in our prior work.

This work supported in part by National Natural Science Foundation of China under Grant No. 61173023 and the Fundamental Research Funds for the Central Universities under Grant No. HIT. NSRIF. 201180.