Computer Science and Application 计算机科学与应用, 2020, 10(6), 1150-1157
Published Online June 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.106119

Hans汉斯

An Enumeration Algorithm of the Maximal Cluster from the Uncertain Graph Using Subgraph Partition Strategy

Meng Zhao

Yanshan University, Qinhuangdao Hebei
Email: zhaomeng_527@163.com

Received: May 20th, 2020; accepted: Jun. 2nd, 2020; published: Jun. 10th, 2020

Abstract

In order to more efficiently enumerate maximal clique from the uncertain graph, by studying the existing algorithms of the maximal clique enumeration from the deterministic and uncertain graph, and according to the relationship of the maximal clique from the deterministic graph and the α-maximal clique from the uncertain graph under the same graph structure, an efficient enumeration algorithm of the α-maximal clique from the uncertain graph based on the subgraph partition of the maximal clique from the deterministic graph with the same graph structure, called D-MULE-D, is proposed in this paper. Experimental tests were carried out on different real data sets to verify the feasibility and efficiency of the D-MULE-D algorithm, by comparing the running time of the D-MULE-D algorithm with the MULE algorithm.

Keywords

Enumeration, Deterministic Graph, Uncertain Graph, Subgraph, Clique, Pseudo-Maximal Clique, Maximal Clique, α-Maximal Clique

应用子图划分策略的不确定图极大团枚举算法

赵　孟

燕山大学，河北 秦皇岛
Email: zhaomeng_527@163.com

收稿日期：2020年5月20日；录用日期：2020年6月2日；发布日期：2020年6月10日

文章引用：赵孟. 应用子图划分策略的不确定图极大团枚举算法[J]. 计算机科学与应用, 2020, 10(6): 1150-1157.
DOI: 10.12677/csa.2020.106119

摘　要

为了更加高效地枚举出不确定图极大团，通过对现有确定图和不确定图极大团枚举算法进行研究，结合在相同图结构下确定图极大团与不确定图极大团之间的关系，提出了一种基于相同图结构确定图极大团子图划分的高效不确定图极大团枚举算法D-MULE-D。通过在不同的真实数据集上进行实验测试，对比D-MULE-D算法和MULE算法的运行时间，验证D-MULE-D算法的可行性和高效性。

关键词

枚举，确定图，不确定图，子图，团，伪极大团，极大团，α极大团

1. 引言

经过长年的研究发现，极大团枚举是一个 NP-难问题，但通过广泛的实验验证，现有的确定图极大团枚举算法都能得到最终的极大团。然而在互联网技术的不断发展下，由于数据来源的异构性和数据描述手段的局限性，造成了图数据的不确定性，确定图的极大团枚举算法已不能满足实际生活中所获取的图数据的需求，因此不确定图极大团枚举算法已成为人们日益关注的焦点。在无线传感器网络中，由于通信信号干扰，导致通信链路的不确定性，图数据中边的连通性存在一定概率，因此无线传感器网络数据就可以模拟为不确定图；再如在蛋白质分子交互网络中，由于高通量生物实验的限制，蛋白质分子间是否存在相互作用具有不确定性，通过概率来量化这种交互关系的可能性，同样可以模拟出蛋白质分子交互的不确定图等等。在现实生活中，很多领域都对不确定图极大团枚举算法有着广泛的应用，如通信网络、生物网络、社交网络及 web 分析等。

近年来，随着大数据和互联网技术的快速发展，越来越多的科研人员致力于不确定图数据挖掘技术的研究，其中，不确定图极大团枚举算法的研究更是受到大家的普遍关注与重视。目前，不确定图极大团枚举算法主要有以下几类：1) 基于 DFS 的极大团枚举算法[1]，此类算法是枚举极大团的基本方法，通过深度优先遍历所有顶点得到极大团，并应用回溯法过滤伪极大团，时间复杂度为 $O(n^2 \cdot 2^n)$；2) 基于简并顺序的极大团枚举算法[2] [3] [4] [5]，此类算法是通过贪心思想遍历图中所有顶点后得到不确定图的简并顺序，根据此顺序来处理图中顶点得到极大团，时间复杂度为 $O(n^2 \cdot 2^n)$；3) 基于顶点编号升序的极大团枚举算法[6] [7] [8]，此类算法是通过对不确定图中的顶点编号进行升序处理，通过增量计算维持两个顶点权值集合来枚举所有极大团，时间复杂度为 $O(n \cdot 2^n)$。通过比较三类极大团枚举算法，可以发现前两类算法已不适用于现实世界中超大规模的数据量，因此本文通过选择基于顶点编号升序的极大团枚举算法，并对此类算法中最典型的 MULE 算法[9]进行研究，提出了一种基于相同结构确定图极大团子图划分的高效不确定图极大团枚举算法——D-MULE-D，对于大规模数据的不确定图，应用 MULE 算法的时间复杂度是 $O(n \cdot 2^n)$，而应用 D-MULE-D 算法的时间复杂度是 $O(n \cdot 3^{n/3})$，该算法能够高效地得出最终的不确定图极大团。

本文下面的论述内容结构如下：首先通过问题定义，给出确定图极大团和不确定图极大团的定义，并根据定义，推导出在相同的结构下，确定图极大团与不确定图极大团的关系定理，并对定理正确性进行证明；其次根据推导定理，提出了基于相同结构确定图极大团子图划分的不确定图极大团枚举算法D-MULE-D，并给出了其详细的算法描述；然后通过具体实例对 D-MULE-D 算法实现进行说明；随后通过实验分析，在不同阈值 α 下，对 9 个不同领域的数据集分别应用 D-MULE-D 算法和 MULE 算法，比较两种算法的运行时间；最后通过比较得出结论。

2. 问题定义

定义 1 (确定图)：确定图通常用一个二元组 G = (V, E)表示，其中 V 是图的顶点集合，顶点 v∈V，E 是图的边集合，E = {<v, w>|v, w∈V}。其中<v, w>表示图中顶点 v 和顶点 w 是联通的。

定义 2 (团)：给定确定图 G = (V, E)，若存在顶点集合 $C \subseteq V$，且 C 中每一对顶点都存在边，那么称 C 是图 G 的一个团。

定义 3 (极大团)：给定确定图 G = (V, E)，若存在顶点集合 $M \subseteq V$，M 为团，且不存在一个顶点 $v \notin M$，使得 $M \cup v$ 是图中的团，那么称 M 是图 G 中的一个极大团。

定义 4 (不确定图)：不确定图通常用一个三元组 G = (V, E, P)表示，其中 V 是图的顶点集合，顶点 v∈V，E 是图的边集合，边 e∈E，P 是边的权值集合，$0 < p < 1$，p∈P 表示顶点间联通的概率。

定义 5 (α 团)：给定不确定图 G = (V, E, P)和阈值 α，$0 < \alpha < 1$，若存在顶点集合 $C \subseteq V$，C 中每一对顶点都存在边，且 C 的团概率大于等于 α，那么称 C 是一个 α 团。其中 C 的团概率是指 C 中顶点对在图中对应的边的权值的乘积。

定义 6 (α 极大团)：给定不确定图 G = (V, E, P)和阈值 α，$0 < \alpha < 1$，若存在顶点集合 $M \subseteq V$，M 为 α 团，且不存在一个顶点 $v \notin M$，使得 $M \cup v$ 是图中的 α 团，那么称 M 是图 G 中的一个 α 极大团。

根据以上定义，我们可以将确定图中团和极大团的定义与不确定图中 α 团和 α 极大团的定义相结合，得出以下定理：

定理 1：给定不确定图 G = (V, E, P)和阈值 α，$0 < \alpha < 1$，其任意一个 α 团 C 必然是与其结构相同的确定图 G’ = (V, E)的一个团。

定理证明

给定不确定图 G = (V, E, P)和阈值 α，$0 < \alpha < 1$，则与其相同结构的确定图 G’ = (V, E)。

对于 G 中的任意一个 α 团 C，

∵C 是 G 的 α 团；

∴根据定义 5，C 中的每一对顶点间都存在边；

∵G’与 G 结构相同，具有相同的顶点和边；

∴在 G’中，$C \subseteq V$ 且 C 中的每一对顶点间都存在边；

∴根据定义 2，C 是 G’的一个团。

定理 2：给定不确定图 G = (V, E, P)和阈值 α，$0 < \alpha < 1$，其任意一个 α 极大团 M 对于其结构相同的确定图 G’ = (V, E)，在 G’中必然存在一个极大团 $M' \supseteq M$。

定理证明

给定不确定图 G = (V, E, P)和阈值 α，$0 < \alpha < 1$，与其相同结构的确定图 G’ = (V, E)。

对于 G 中的任意一个 α 极大团 M，

∵M 是 G 的 α 极大团；

∴根据定义 6，M 是 G 的 α 团；

∵G’与 G 结构相同，具有相同的顶点和边；

∴在 G’中， $M \subseteq V$ 且 M 中的每一对顶点间都存在边；

∴根据定理 2，M 是 G’的一个团；

∴根据定义 3，在 G’中必然存在一个极大团 $M' \supseteq M$ 。

3. 算法描述

根据上面推导的定理 3，给定不确定图 G = (V, E, P)和阈值 α，$0 < \alpha < 1$，其任意一个 α 极大团 M 对于其结构相同的确定图 G’ = (V, E)，在 G’中必然存在一个极大团 $M' \supseteq M$ 。因此，如果对不确定图先进行子图划分，先得到与其结构相同的确定图的极大团，再对不确定图中这些极大团子图应用 MULE 算法，从而枚举出不确定图的 α 极大团，将大幅提高不确定图极大团枚举效率，降低时间复杂度。

基于以上不确定图子图划分策略，我们提出了一种新的不确定图极大团枚举算法 D-MULE-D，该算法先通过 Degeneracy 算法对不确定图进行子图划分，将不确定图按其相同结构的确定图划分成极大团子图集合，然后再对这些不确定图的极大团子图分别应用 MULE 算法，从而枚举出其不确定图的 α 极大团，再通过顶点规模降序过滤策略过滤掉伪 α 极大团，从而得到最终的 α 极大团集合。其中顶点规模降序过滤策略是将极大团集合根据其每个极大团内顶点个数进行降序排列，再顺序对顶点集合间进行判断，如果前者包含后者，说明后者为伪极大团，直接将伪极大团从极大团集合中删除，直到所有顶点集合间包含关系都判断完结束。该策略能够尽早发现伪极大团，通过剔除伪极大团，从而减少验证极大团的次数，提高验证效率。

```
算法 D-MULE-D
D-MULE-D(G(V, E, P), α)
输入：不确定图 G=(V, E, P)和阈值 α，0 < α < 1;
输出：α 极大团 MaxC;
{
   Subgraph=Degeneracy(G(V, E));
   //通过 Degeneracy 算法对不确定图进行子图划分；
      MaxC=Φ;
      //初始化极大团集合为空集；
   For(i=1;i<=num(Subgraph),i++)
   MaxC=MaxC ∪ MULE(Subgraph[i], α);
   //对极大团子图应用 MULE 算法枚举 α 极大团；
   DescendingOrder(MaxC);
   //对 α 极大团集合按其极大团中顶点个数进行降序排列；
   For(i=1;i<=num(MaxC);i++)
      For(j=i+1;j<=num(MaxC);j++)
         If(MaxC[i] ⊇ MaxC[j])
            MaxC=MaxC-MaxC[j];
            //对极大团集合中所有极大团进行判断，过滤掉伪极大团；
   Return(MaxC);
}
```

4. 数据分析

根据以上对 D-MULE-D 算法的描述，下面我们以具体实例对算法的实现进行说明，对于给定不确定图 G，如图 1 所示：

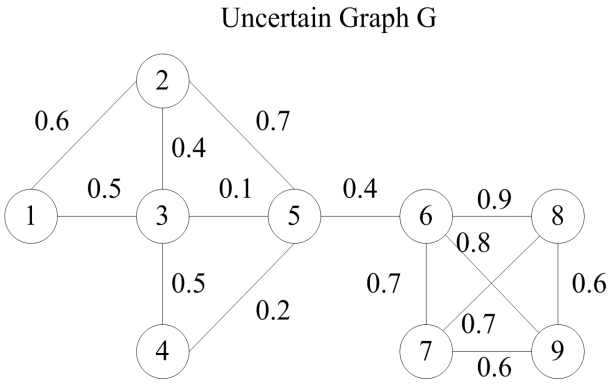

Figure 1. Uncertain graph G
图 1. 不确定图 G

设定 $\alpha = 0.1$，则对于不确定图 G 应用 D-MULE-D 算法，其算法实现如图 2~4 所示：

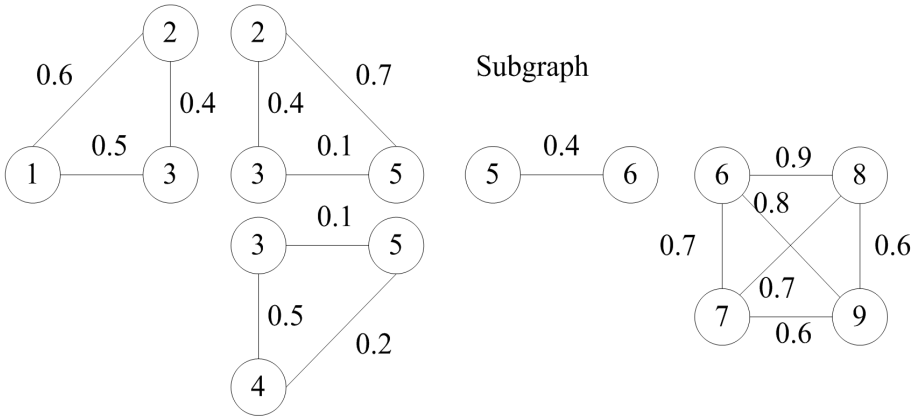

Figure 2. The subgraphs obtained by using Degeneracy algorithm
图 2. 应用 Degeneracy 算法进行子图划分

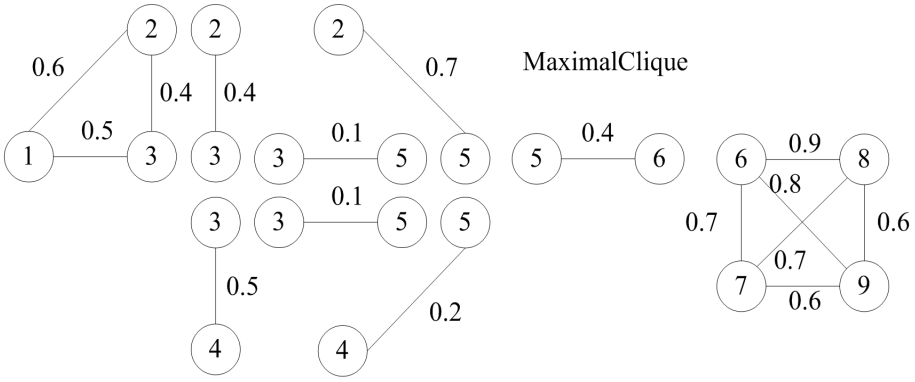

Figure 3. The MULE algorithm is applied to subgraphs respectively
图 3. 对子图分别应用 MULE 算法

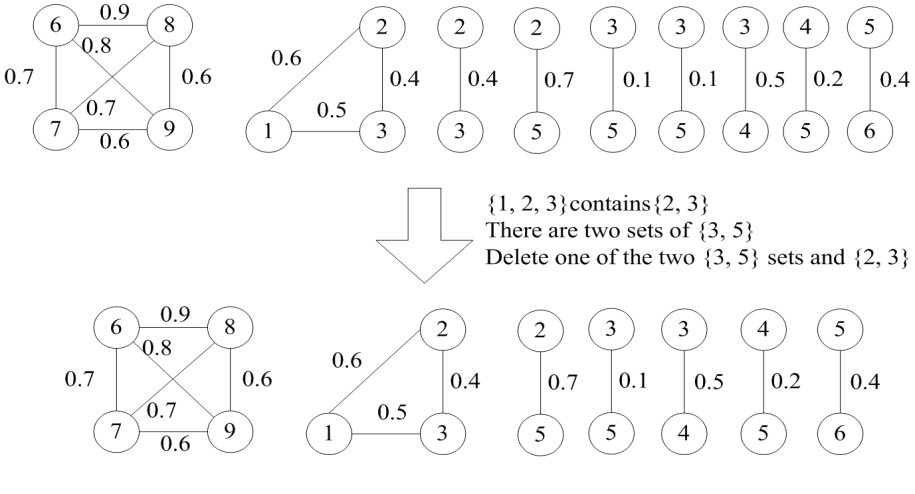

Figure 4. The vertex size descending filtering strategy is applied
图 4. 应用顶点规模降序过滤策略

5. 实验验证

实验使用的硬件平台是英特尔酷睿 i7-8700K 处理器，主频 3.70 GHz，1TB HDD 硬盘和 8 GB DDR4 内存，64 位 Windows 10 操作系统；运行环境为 Microsoft Visual Studio 2010；编程语言为 C++。本文所使用的实验数据来源于 9 个不同领域的数据集，分别是 Anthra、Mtbrv、Amaze、Kegg、Xmark、Nasa、Citeseer、Go 和 Yago 数据集。对于阈值 α = 0.2，0.4，0.6，0.8，分别应用 MULE 和 D-MULE-D 算法枚举极大团，两者的运行时间对比如图 5~8 所示。

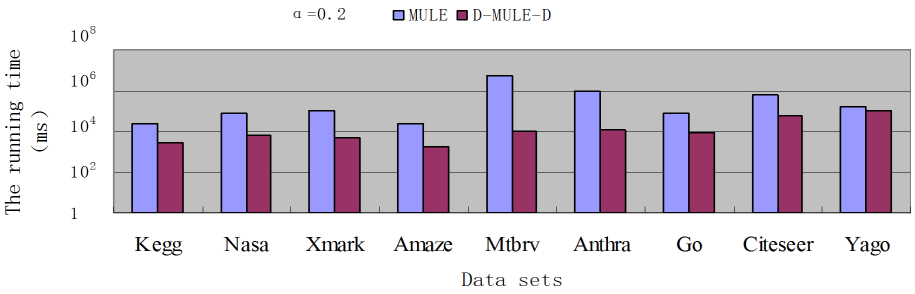

Figure 5. The running time comparison between the MULE and D-MULE-D algorithm when $\alpha = 0.2$
图 5. $\alpha = 0.2$ 时 MULE 和 D-MULE-D 算法枚举极大团运行时间对比

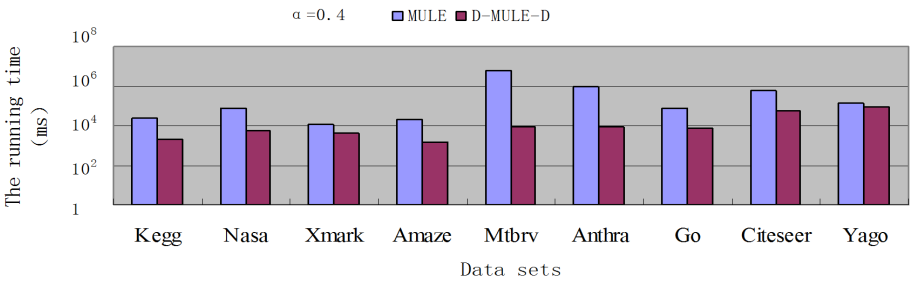

Figure 6. The running time comparison between the MULE and D-MULE-D algorithm when $\alpha = 0.4$
图 6. $\alpha = 0.4$ 时 MULE 和 D-MULE-D 算法枚举极大团运行时间对比

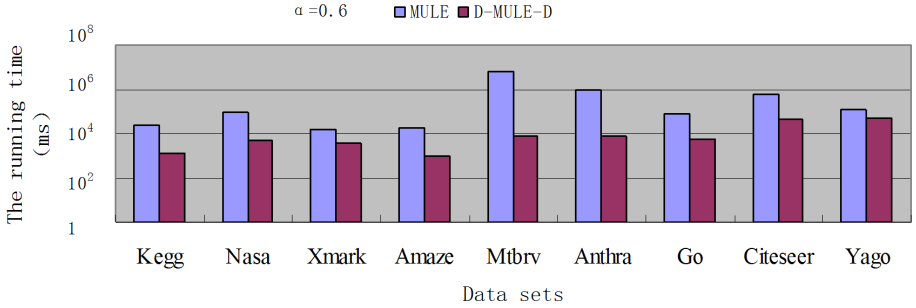

Figure 7. The running time comparison between the MULE and D-MULE-D algorithm when $\alpha = 0.6$

图 7. $\alpha = 0.6$ 时 MULE 和 D-MULE-D 算法枚举极大团运行时间对比

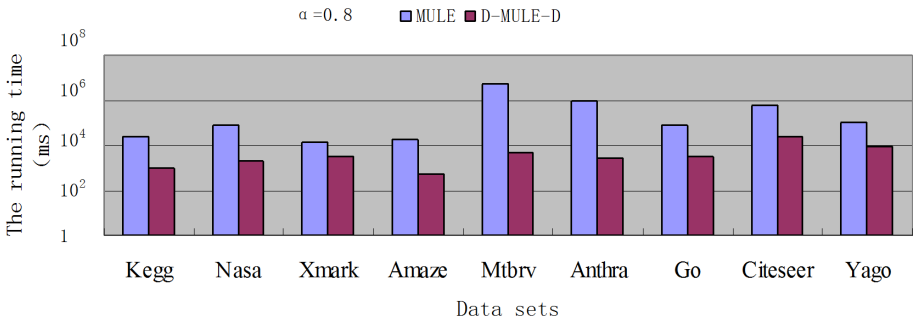

Figure 8. The running time comparison between the MULE and D-MULE-D algorithm when $\alpha = 0.8$

图 8. $\alpha = 0.8$ 时 MULE 和 D-MULE-D 算法枚举极大团运行时间对比

通过以上实验分析，我们可以发现对于大规模数据的不确定图，应用 MULE 算法的时间复杂度是 O $(n \cdot 2^n)$，而应用 D-MULE-D 算法的时间复杂度是 O $(n \cdot 3^{n/3})$，D-MULE-D 算法的运行时间要明显低于 MULE 算法。同时，对于不同的阈值 α，因 MULE 算法运行取决于顶点数量 n，其运行时间变化不大；而由于阈值 α 的增大，α 极大团会相应减少，α 极大团内顶点个数也会减少，从而检验极大性检验次数也会减少，因此 D-MULE-D 算法的运行时间会随着阈值 α 的增大而减少。

6. 结论

本文通过对不确定图极大团枚举算法 MULE 进行研究，并结合确定图与不确定图在结构相同时极大团的相互关系，提出了一种新的不确定图极大团枚举算法 D-MULE-D，该算法通过对不确定图进行极大团子图划分，再对极大团子图应用 MULE 算法枚举 α 极大团，随后对 α 极大团集合按照顶点规模倒序排列检验其极大团的正确性，尽早过滤掉其中的伪 α 极大团，从而得到最终正确的 α 极大团集合。通过实验分析，D-MULE-D 算法的运行时间要明显低于 MULE 算法，并且对于不同的阈值 α，D-MULE-D 算法的运行时间会随着阈值 α 的增大而减少。因此，D-MULE-D 算法的运行效率要明显高于 MULE 算法。

参考文献

[1] Mcauley, J. and Leskovec, J. (2014) Discovering Social Circles in Ego Networks. *ACM Transactions on Knowledge Discovery from Data*, **8**, 4-28. https://doi.org/10.1145/2556612

[2] Satuluri, V., Parthasarathy, S. and Ruan, Y. (2012) Local Graph Sparsification for Scalable Clustering. *Proceedings of the* 2011 *ACM SIGMOD International Conference on Management of Data*, Athens, June 2012, 721-732. https://doi.org/10.1145/1989323.1989399

[3] Eppstein, D., Loffler, M. and Strash, D. (2013) Listing All Maximal Cliques in Large Sparse Real-World Graphs. *ACM*

Journal of Experimental Algorithmics, **8**, 224-227. https://doi.org/10.1145/2543629

[4] Svendsen, M., Mukherjee, A.P. and Tirthapura, S. (2015) Mining Maximal Cliques from a Large Graph Using Mapreduce: Tackling Highly Uneven Subproblem Sizes. *Journal of Parallel and Distributed Computing*, **79**, 104-114. https://doi.org/10.1016/j.jpdc.2014.08.011

[5] Jin, R., Liu, L. and Aggarwal, C.C. (2011) Discovering Highly Reliable Subgraphs in Uncertain Graphs. *Proceedings of the* 17*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, 21-24 August 2011, 992-1000. https://doi.org/10.1145/2020408.2020569

[6] Khan, A., Bonchi, F., Gionis, A. and Gullo, F. (2014) Fast Reliability Search in Uncertain Graphs. *Proceedings of the* 16*th International Conference on Extending Database Technology*, New Orleans, 24-28 March 2014, 535-546.

[7] Koch, I. (2011) Enumerating All Connected Maximal Common Subgraphs in Two Graphs. *Theory Compute Science*, **250**, 1-30. https://doi.org/10.1016/s0304-3975(00)00286-3

[8] Mukherjee, A.P., Xu, P. and Tirthapura, S. (2015) Mining Maximal Cliques from an Uncertain Graph. 2015 *IEEE* 31*st International Conference on Data Engineering*, Seoul, 13-17 April 2015, 243-254. https://doi.org/10.1109/icde.2015.7113288

[9] Mukherjee, A.P., Xu, P. and Tirthapura, S. (2017) Enumeration of Maximal Cliques from an Uncertain Graph. *IEEE Transactions on Knowledge and Data Engineering*, **29**, 543-555. https://doi.org/10.1109/tkde.2016.2527643