A Note on a Maximum k-Subset Intersection Problem

Eduardo C. Xavier *
Institute of Computing
University of Campinas – UNICAMP
Campinas, SP, Brazil
eduardo@ic.unicamp.br

March 12, 2012

Abstract

Consider the following problem which we call Maximum k-Subset Intersection (*MSI*): Given a collection $\mathcal{C} = \{S_1, \ldots, S_m\}$ of m subsets over a finite set of elements $\mathcal{E} = \{e_1, \ldots, e_n\}$, and a positive integer k, the objective is to select exactly k subsets $S_{j_1}, \ldots, S_{j_k}$ whose intersection size $|S_{j_1} \cap \ldots \cap S_{j_k}|$ is maximum. In [2], Clifford and Popa studied a related problem and left as an open problem the status of the MSI problem. In this paper we show that this problem is hard to approximate.

Key Words: Approximation algorithms, Combinatorial problems, Subset Intersection

1 Introduction

In this paper we study the following problem: Given a collection $\mathcal{C} = \{S_1, \ldots, S_m\}$ of m subsets over a finite set of elements $\mathcal{E} = \{e_1, \ldots, e_n\}$, and a positive integer k, the objective is to select exactly k subsets $S_{j_1}, \ldots, S_{j_k}$ from $\mathcal{C}$ whose intersection size $|S_{j_1} \cap \ldots \cap S_{j_k}|$ is maximum. We call this problem Maximum k-Subset Intersection (*MSI*), which was left as an open problem by Clifford and Popa [2].

In this paper we present an inapproximability result for the MSI problem presenting a reduction from the Maximum Edge Biclique (*MEB*) problem. The MEB problem can be stated as follows: Given a bipartite graph $G = (V_1, V_2, E)$, the problem is to find a biclique $K_{x,y}$ subgraph of G whose number of edges xy is maximum.

The MEB problem was shown to be NP-hard by Peteers [5]. Later, Ambuhl et al in [1], proved that the MEB problem does not admit a $1/N^{\epsilon'}$ approximation, where ϵ' is a constant and N is the number of vertices, under the standard assumption that SAT has no probabilistic algorithm that runs in time $2^{n^{\epsilon}}$, where n is the instance size and $\epsilon > 0$ can be made arbitrarily close to 0. They showed the following result:

Theorem 1 (Ambuhl et al [1]) *Let $\epsilon > 0$ be an arbitrarily small constant. Assume that SAT does not have a probabilistic algorithm that decides whether a given instance of size n is satisfiable in time $2^{n^{\epsilon}}$. Then there is no polynomial (possibly randomized) algorithm for Maximum Edge Biclique that achieves an approximation ratio of $1/N^{\epsilon'}$ on graphs of size N, where ϵ' depends only on ϵ.*

In this work we show an inapproximability result for the MSI problem using the inapproximability result of Theorem 1.

*Supported by CNPq

The MEB problem has applications in community detection [3] and in bioinformatics [4], among others. The biclustering problems involved in such applications can also be tackled as a MSI problem. Generally, we have in such applications a set of individuals/genes and associated interests/conditions. The main objective is to find a set of individuals/genes with the largest number of interests/conditions in common.

In Section 2 we present a Turing reduction showing the hardness of the MSI problem, and in Section 3 we prove the inapproximability of the MSI problem by showing that if there is an α-approximation algorithm for the MSI problem, then there is also an α-approximation algorithm for the MEB problem.

2 Hardness Result

In this section we present a Turing reduction from the MEB problem to the MSI problem, by presenting a polynomial time algorithm that can be used to solve the MEB problem if the MSI problem is solvable in polynomial time.

Theorem 2 *MSI is NP-hard.*

Proof. Let $G = (V_1, V_2, E)$ be an instance for the MEB problem, where $V_1 = \{v_1, \ldots, v_{n_1}\}$ and $V_2 = \{u_1, \ldots, u_{n_2}\}$. Create an instance for the MSI problem as follows: let the set of elements be the set V_2, i.e, $\mathcal{E} = V_2$, and for each vertex $v_i \in V_1$ create a set $v_i = \{u_j \in V_2 : (v_i, u_j) \in E\}$, i.e, this set contains all vertices of V_2 that are adjacent to v_i. The collection of subsets is $\mathcal{C} = \{v_1, \ldots, v_{n_1}\}$.

Considering the construction above, we claim that for any given biclique subgraph $K_{x,y}$ of G, there are x subsets in the corresponding instance of the MSI problem such that their intersection size is at least y. Let $V_1' \subseteq V_1$ and $V_2' \subseteq V_2$ be the vertices of the biclique $K_{x,y}$. Since every vertex in V_1' is adjacent to all vertices in V_2', then all vertices of V_2' will belong to each subset corresponding to each vertex of V_1'. The intersection of these subsets contains V_2'.

On the other hand, we claim that if we find k subsets $V_1' = \{v_1', \ldots, v_k'\}$ of maximum intersection $v_1' \cap \ldots \cap v_k' = V_2' \subseteq V_2$, then there is a biclique subgraph in G with $k|V_2'|$ edges. From the construction of the MSI instance, every vertex v_i' is adjacent to all vertices in V_2'. Then the induced subgraph given by the corresponding vertices in V_1' and V_2' form a biclique of size $k|V_2'|$.

Suppose there is a polynomial time algorithm $\mathcal{A}(\mathcal{C}, k, \mathcal{E})$ that solves the MSI problem, and returns $(\mathcal{C}', I)$, where $\mathcal{C}' \subset \mathcal{C}$ contains k subsets, and I contains the elements of the intersection of these subsets. Then Algorithm 1 solves the MEB problem.

Algorithm 1 Alg($G = (V_1, V_2, E)$)

Given G, create the collection $\mathcal{C}$, and elements $\mathcal{E}$ for the MSI problem.
Let $K_{x,y}$ be an empty biclique.
for $k = 1, \ldots, n_1$ **do**
    Let $(V_1', V_2') \leftarrow \mathcal{A}(\mathcal{C}, k, \mathcal{E})$.
    Let $K'_{x',y'}$ be the biclique subgraph of G with the corresponding vertices from (V_1', V_2').
    if $xy < x'y'$ **then**
        $K_{x,y} \leftarrow K'_{x',y'}$.
    end if
end for
Return $K_{x,y}$.

Let $K^*_{x^*,y^*}$ be an optimal solution for the MEB problem. We know that when we run $\mathcal{A}(\mathcal{C}, x^*, \mathcal{E})$, the algorithm will return a solution corresponding to vertices that form a biclique subgraph of G with at least x^*y^* edges. Since the algorithm tries all values of $k = 1, \ldots, n_1$, and returns the biclique with maximum number of edges, it will return an optimal solution.

□

3 Inapproximability Result

In this section we show that if there is an α-approximation algorithm $\mathcal{A}(\mathcal{C}, k, \mathcal{E})$ for the MSI problem then we can construct another algorithm $\mathcal{A}'$ which is an α-approximation algorithm for the MEB problem.

Lemma 3 *Let $\mathcal{A}$ be an α-approximation algorithm for the MSI problem. Then there is an α-approximation algorithm $\mathcal{A}'$ for the MEB problem.*

Proof. Let $G = (V_1, V_2, E)$ be an instance of the MEB problem, where $n_1 = |V_1|$ and $n_2 = |V_2|$. We construct an instance for the MSI problem as was done in Theorem 2.

Suppose that $K_{x,y}$ is a maximum edge biclique of G. If we construct an instance for the MSI problem as stated above, and run $\mathcal{A}(\mathcal{C}, x, \mathcal{E})$ we know that the algorithm is going to find x subsets $v_{i_1}, \ldots, v_{i_x}$, whose intersection size is at least αy. Notice that the vertices $v_{i_1}, \ldots, v_{i_x}$ from V_1 and the vertices in the corresponding intersection of their subsets, form a biclique with at least αxy edges.

Suppose we run $\mathcal{A}(\mathcal{C}, k, \mathcal{E})$, for $k = 1, \ldots, n_1$. We can then find the solution $v'_{i_1}, \ldots, v'_{i_{k'}}$ that maximizes the value $k'T$ where $T = |v'_{i_1} \cap \ldots \cap v'_{i'_k}|$, among all these executions of the algorithm. Notice that the corresponding vertices $v'_{i_1}, \ldots, v'_{i_{k'}}$ from V_1 and vertices in $v'_{i_1} \cap \ldots \cap v'_{i_{k'}}$ from V_2, form a biclique of size $k'T \geqslant \alpha xy$. Then we have an α-approximation solution for the given instance G of the MEB problem.

□

Using Theorem 1 and Lemma 3 we have the following result.

Theorem 4 *Let $\epsilon > 0$ be an arbitrarily small constant. Assume that SAT does not have a probabilistic algorithm that decides whether a given instance of size n is satisfiable in time 2^{n^ϵ}. Then there is no polynomial time algorithm for the Maximum k-Subset Intersection problem that achieves an approximation ratio of $1/N^{\epsilon'}$ where N is the size of the instance, and ϵ' depends only on ϵ.*

References

[1] Christoph Ambühl, Monaldo Mastrolilli, and Ola Svensson. Inapproximability results for maximum edge biclique, minimum linear arrangement, and sparsest cut. *SIAM J. Comput.*, 40(2):567–596, 2011.

[2] Raphaël Clifford and Alexandru Popa. Maximum subset intersection. *Inf. Process. Lett.*, 111(7):323–325, 2011.

[3] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[4] Sushmita Mitra and Haider Banka. Multi-objective evolutionary biclustering of gene expression data. *Pattern Recognition*, 39(12):2464–2477, 2006.

[5] René Peeters. The maximum edge biclique problem is np-complete. *Discrete Applied Mathematics*, 131(3):651–654, 2003.