The 2-Catalog Segmentation Problem

Yevgeniy Dodis* Venkatesan Guruswami* Sanjeev Khanna†

1 Introduction

We study the 2-CATALOG SEGMENTATION problem: Given a set I of n items and a family $\mathcal{S} = \{S_1, S_2, ..., S_p\}$ of subsets of I, find $C_1, C_2 \subseteq I$ such that $|C_1|, |C_2| \leq r$ and the sum $\sum_{i=1}^{p} \max\{|S_i \cap C_1|, |S_i \cap C_2|\}$ is maximized. The problem was recently introduced by Kleinberg *et al* [2] and is motivated by several applications to data mining and clustering operations as detailed in [2]. Under the restriction that $|S_i| = \Omega(|I|)$ for each S_i, the authors give a PTAS. But, in general, only a trivial 0.5-approximation algorithm is known – just define C_1 to be r most frequently occurring elements. The question of improving upon this factor was posed as an important open problem in [2].

We make some progress towards answering this question, under the assumption that the size of the collection I is bounded by $2r$, i.e. twice the catalog size. Our motivation comes from the following natural restriction of the 2-CATALOG SEGMENTATION problem: we are given $C_1 \cup C_2$, that is, the elements that comprise the optimal solution. How well can one approximate the optimal partition in this case? This variation of the problem appears to be a natural intermediate step towards solving the original problem and we hope that understanding it will shed some light on the structure of the general problem. By relating this problem to the well-known problem of MINIMUM BISECTION, we show that this restricted segmentation problem is **NP**-hard, even under the assumption that each S_i contains at most 2 elements. Using a semidefinite programming relaxation, similar to the one used by Frieze and Jerrum [1], we obtain a 0.565-approximation for this restriction of the problem. We obtain a ratio of 0.651 when the two catalogs are required to be disjoint; this variant is referred to as the DISJOINT 2-CATALOG SEGMENTATION problem. We also obtain some results for the *general* 2-CATALOG SEGMENTATION problem when the sets $S_i \in \mathcal{S}$ are bounded in size by a small constant.

*MIT Lab for Computer Science, 545 Technology Square, Cambridge, MA 02139. Email: {yevgen,venkat}@lcs.mit.edu

†Department of Fundamental Mathematics Research, Bell Labs, 700 Mountain Avenue, Murray Hill, NJ 07974. Email: sanjeev@research.bell-labs.com

2 Problem Statement

We study the following variations of the problem:

(A) The $2r$*-universe* case: $|I| \leq 2r$.

(B) The *disjoint* case: the catalogs are *disjoint.*

(C) The *bounded preference set* case: size of each set $S_i \in \mathcal{S}$ is bounded by a constant c.

3 The $2r$-Universe Case

Reducing from MINIMUM BISECTION, we can show:

THEOREM 3.1. *The* DISJOINT 2-CATALOG SEGMENTATION *and* 2-CATALOG SEGMENTATION *problems for the* $2r$*-universe case are* **NP***-hard even when all the sets are of size* 2*.*

We now develop our approximation algorithms.

The Disjoint Case: We view the problem as a graph-theoretic one by constructing the set-item incidence bipartite graph in the natural way. The DISJOINT 2-CATALOG SEGMENTATION problem is now equivalent to the following: given a bipartite graph $G = (X, Y, E)$ with $|X| = 2r$ and $|Y| = p$, find a partition $X = X_1 \cup X_2$ and $Y = Y_1 \cup Y_2$ such that $|X_1| = |X_2| = r$ and the quantity $[X_1 : Y_1] + [X_2 : Y_2]$ is maximized ($[A : B]$ denotes the number of edges with one end in A and the other in B). Let $\mathrm{OPT}(G)$ denote the value of the optimum solution – note that $\mathrm{OPT}(G) \geq |E(G)|/2$. Using a semidefinite programming relaxation similar to the one used by Frieze and Jerrum [1], we can obtain a 0.651-approximation for this problem. Let $n = 2r + p$ and S^{n-1} be the unit sphere in $\mathcal{R}^n$, and let $v_1, \ldots, v_{2r}, w_1, \ldots, w_p$ be vectors constrained to be in S^{n-1}. The following SDP is then a relaxation of our problem.

$$\text{Maximize } \sum_{(i,j)\in E} \frac{1 + v_i \cdot w_j}{2}$$

$$\text{s.t. } \|\sum_{i=1}^{2r} v_i\| = 0$$

After solving the above SDP (whose optimum we denote by W^*), we get a partition of $X = X_1' \cup X_2'$ and $Y = Y_1' \cup Y_2'$ by using the standard random hyperplane rounding of [3]. Assume $|X_1'| \geq |X_2'|$. Let X_1 consist of r elements of X_1' having the highest number of neighbors in Y_1' and X_2 consist of the remaining $(|X_1'| - r)$ vertices of X_1' together with X_2'. Let Y_1 be the elements of

Y having more neighbors in X_1 than in X_2; and set $Y_2 = Y \setminus Y_1$. Now X_1, X_2, Y_1, Y_2 will be our final partition. As in [1], define random variables $J = [X_1' : Y_1'] + [X_2' : Y_2']$ and $K = |X_1'||X_2'| = |X_1'|(2r - |X_1'|)$ and let $Z = J/W^* + K/r^2$. By the analysis of [3], we have $\mathbf{E}[J] \geq \alpha W^*$, where $\alpha \approx 0.878$. Similarly, using $\|\sum v_i\| = 0$, and the analysis of [1], we get $\mathbf{E}[K] \geq \alpha r^2$, and hence $\mathbf{E}[Z] \geq 2\alpha$. Also clearly $J \leq |E(G)| \leq 2\mathrm{OPT}(G) \leq 2W^*$, which together with the fact that $K \leq r^2$ yields $Z \leq 3$. Hence we will have $Z \geq 2(1-\epsilon)\alpha$ with some constant probability. Suppose, $|X_1'| = 2\delta r$ for some $1/2 \leq \delta \leq 1$. Then $Z \geq 2(1-\epsilon)\alpha$ yields $J \geq (2(1-\epsilon)\alpha - 4\delta(1-\delta))W^*$, and by our greedy choice of X_1, X_2, Y_1, Y_2, we get

$$[X_1 : Y_1] + [X_2 : Y_2] \geq [X_1 : Y_1'] + [X_2 : Y_2'] \geq \frac{r}{|X_1'|} J$$
$$\geq \frac{2(1-\epsilon)\alpha - 4\delta(1-\delta)}{2\delta} W^* \geq (0.6511 - \epsilon) W^*$$

using $\alpha = 0.878$ and minimizing the above function for $\frac{1}{2} \leq \delta \leq 1$.

Removing the Disjointness Assumption: We now relax the disjointness assumption and get a 0.565-approximation. Let $(I, \mathcal{S})$ be an instance of the segmentation problem with $|I| = 2r$ with optimum value O and optimum disjoint segmentation value O_d. Let F be the set of r most frequent elements of I, i.e. $V(F) \triangleq \sum_{i=1}^{p} |S_i \cap F|$ is maximized. If $V(F) \geq O/(2-\epsilon)$, then F together with any other r-element subset of I will give a solution that is at least a factor $1/(2-\epsilon)$ of the optimum solution. So assume $V(F) < O/(2-\epsilon)$ and let $F_1, F_2 \subseteq I$ be an optimal (possibly non-disjoint) solution, achieving value O. Then

$$\begin{aligned} O &\leq V(F_1 \cup F_2) = V(F_1) + V(F_2) - V(F_1 \cap F_2) \\ &\leq 2V(F) - V(F_1 \cap F_2) \leq \frac{2O}{2-\epsilon} - V(F_1 \cap F_2) \end{aligned}$$

and hence $V(F_1 \cap F_2) \leq \frac{\epsilon}{2-\epsilon} O$. Since the partition $(F_1, I - F_1)$ achieves a segmentation value that is at least $O - V(F_1 \cap F_2) \geq O - O\epsilon/(2-\epsilon)$, we have $O_d \geq (1 - \frac{\epsilon}{2-\epsilon})O$. Hence the algorithm which outputs the better of the two solutions: (a) the "best" one set F together with any other r element set, and (b) the 0.651-approximate solution for the disjoint case, will achieve a performance ratio of at least

$$\min_{0 \leq \epsilon \leq 1} \max \left\{ \frac{1}{2-\epsilon}, (1 - \frac{\epsilon}{2-\epsilon}) 0.651 \right\} \approx 0.565$$

Theorem 3.2. *The* 2-Catalog Segmentation *problem is 0.565-approximable when the universe size is at most twice the size of each catalog. Furthermore, if the catalogs are required to be disjoint, one can obtain a 0.651-approximation.*

4 Bounded Preference Set Case

The Densest Partition Approach: We now develop an approach that solves Disjoint 2-Catalog Segmentation (with slightly better performance than a Random Partition approach below) when every set size is bounded by 2 or 3 and the universe size is at most $2r$. The idea is based on solving the following problem, which we refer to as the Densest Partition problem: Given a graph $G = (V, E)$ on n (n even) vertices with non-negative weights on the edges, partition V into V_1, V_2 with $|V_1| = |V_2| = n/2$ such that the total weight of edges wholly within V_1 and V_2 is maximized. Note that the optimum partition is the same as the one which achieves the Minimum Bisection: the approximability of these problems, however, differ considerably. Using a semidefinite relaxation that is similar to the one used by Frieze and Jerrum for Maximum Bisection with a minor modification in the rounding stage, we can obtain a 0.54-approximation for this problem. For the obvious generalization of Densest Partition for 3-uniform hypergraphs, we can give a 0.34-approximation.

To apply this to our problem for $c = 2$, form a $2r$-vertex graph with vertex set being the universe and an edge joining two elements of the universe iff there exists a 2-element set that has precisely those two elements. The optimum segmentation value of our problem is $p + O^*$ where p is the number of sets and O^* is the value of the optimal densest partition. Using our algorithm for Densest Partition, we get approximation ratio $(p + 0.54O^*)/(p + O^*) \geq 0.77$ (since $O^* \leq p$). Similarly, we get a slightly better result for $c = 3$. Thus,

Theorem 4.1. *A 0.77 (0.78) approximation can be obtained in polynomial time for the $2r$-universe* Disjoint 2-Catalog Segmentation *with set sizes 2 (3).*

The Random Partition Approach: Let c denote an upper bound on the size of each S_i. Then the strategy of picking $2r$ most frequent elements and partitioning them randomly yields:

Theorem 4.2. *The* 2-Catalog Segmentation *problem on* general universe *is approximable within* $(1/2 + \binom{2\alpha}{\alpha}/2^{2\alpha+1})(1-\epsilon)$, *where* $\alpha = \lfloor c/2 \rfloor$, *for any* $\epsilon > 0$, *which is roughly* $\frac{1}{2} + \Theta(\frac{1}{\sqrt{c}})$ *for large* c.

References

[1] A. Frieze and M. Jerrum. Improved Algorithms for Max-k-cut and Bisection. *Proc. 4th IPCO*, LNCS 920, Springer-Verlag, 1-13.

[2] J. Kleinberg, C. Papadimitriou and P. Raghavan. Segmentation Problems. *STOC 98*, pp. 473-482.

[3] M. Goemans and D. Williamson. Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming. *J. ACM*, 42:1115–1145, 1995.