The Sixth International Symposium on Operations Research and Its Applications (ISORA'06)
Xinjiang, China, August 8–12, 2006


# An Improved Approximation Algorithm for the Disjoint 2-Catalog Segmentation Problem*

Houchun Zhou[1,2,†] Dexue Zhang[1]

[1] Dept. of Math., Linyi Teachers College, Linyi, Shandong 276005, China
[2] School of Math. and Computer Science, Nanjing Normal University, Nanjing, Jiangsu 210097, China

**Abstract** For the disjoint 2-catalog segmentation problem (may be inequivalent), we propose a improved polynomial-time randomized approximation algorithm, and obtain a performance ratios $\rho$ which is not less than 0.5 for a wide range of this problem. As a result, the 0.699-approximation algorithm for the disjoint equivalent 2-catalog segmentation problem can be obtained.

**Keywords** disjoint 2-catalog segmentation; approximation algorithm; semidefinite programming

## 1 Introduction

Given a set $I$ of $n$ items and a family $S = \{S_1, S_2, \cdots, S_p\}$ of subsets of $I$, the generalized 2-catalog segmentation problem is to find $C_1, C_2 \subseteq I$ such that $|C_1| \le r_1, |C_2| \le r_2$ and the sum $\sum_{i=1}^{p} \max\{|S_i \bigcap C_1|, |S_i \bigcap C_2|\}$ is maximized. When $r_1 = r_2 = r$, it is the famous 2-catalog segmentation problem introduced by Kleinberg *et al* [1]. In [1], they presented a trivial 0.5-approximation algorithm for the 2-catalog segmentation problem, and pointed that how to improved the 0.5-approximation algorithm is a open problem. They showed that this 2-catalog segmentation problem is NP-hard, even under the assumption that the size of the collection $I$ is $2r$ and each $S_i$ contains at most 2 elements.

In this paper, we first introduce the disjoint 2-catalog segmentation problem which is a special case of the generalized 2-catalog segmentation problem, then give a polynomial-time randomized approximation algorithm, and obtain a performance guarantee of $\rho$ which is not less than 0.5 for a wide range of the problem. As a special case of the problem, the 0.699-approximation algorithm for the disjoint equivalent 2-catalog segmentation problem can be obtained.

*This work was supported by National Natural Science Foundation of China No.10231060, Important Project Foundation of Linyi Normal University and Academic Creative Project Foundation of Jiangsu Province.

†Email: zhouhouchun@163.net

## 2 The disjoint 2-catalog segmentation problem

The disjoint 2-catalog segmentation problem can be described as the following graph theoretic problem: given an undirected 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 $|X_1| = r + k$, $|X_2| = r - k$ and the quantity $d(X_1,Y_1) + d(X_2,Y_2)$ is maximized, where $d(X_i,Y_i)$ denotes the number of edges with one end in $X_i$ and the other in $Y_i$, $i = 1,2$, and $k$ is a positive integer, $0 \le k < r$.

If $k = 0$, i.e., $|X_1| = |X_2| = r$, it just is the disjoint equivalent 2-catalog segmentation problem [1].

Now, the disjoint inequivalent 2-catalog segmentation problem can be described as the following problem (P):

$$\begin{aligned}
\max \quad & d(X_1,Y_1) + d(X_2,Y_2) \\
\text{s.t.} \quad & X = X_1 \cup X_2, Y = Y_1 \cup Y_2 \\
& X_1 \cap X_2 = \phi, Y_1 \cap Y_2 = \phi, \\
& |X_1| = r - k, |X_2| = r + k.
\end{aligned}$$

where $k$ is a given positive integer $0 \le k < r$.

Let $n = 2r + p$ and $S^{n-1}$ be the unit sphere in $R^n$, and let $v_1$, $v_2$, $\cdots$, $v_{2r}$, $w_1$, $w_2$, $\cdots$, $w_p$ be vectors constrained to be in $S^{n-1}$. According to the similar analysis of Kleinberg[1] and Goemans et.al.[4], this problem can be relaxed to the following problem (SDP):

$$\begin{aligned}
\max \quad & \frac{1}{2}\sum_{i=1}^{2r}\sum_{j=1}^{p} \omega_{ij}(1 - v_i^T w_j) \\
\text{s.t.} \quad & v_i, w_j \in S^{n-1}, \\
& \sum_{i,j=1}^{2r} v_i^T v_j = 4k^2.
\end{aligned}$$

where $\omega_{ij} = 1$ if edge $(i,j) \in E$, otherwise, $\omega_{ij} = 0$.

This (SDP) problem is equivalent to the following semidefinite program (SDP):

$$\begin{aligned}
\max \quad & \frac{1}{4}\sum_{i=1}^{2r}\sum_{j=2r+1}^{n} \omega_{ij}(1 - X_{ij}) \\
\text{s.t.} \quad & ee^T \cdot X = 4k^2 \\
& X_{jj} = 1, j = 1,2,\cdots,n, X \succeq 0.
\end{aligned}$$

Here, the unknow $X \in R^{n\times n}$ is a symmetric matrix, $\cdot$ is the matrix inner product $Q \cdot X =$ trace $(QX)$, and $X \succeq 0$ means that $X$ is a positive semidefinite, $e = (1,\cdots,1,0,\cdots,0) \in R^n$ is a vector of $R^n$ whose fore $2r$ components are ones and others are zero. Obviously, $(SDP)$ is a relaxation of $(P)$, hence we have $\omega^* \le \omega^*_{SDP}$, where $\omega^*$ is the optimal value of (P) and $\omega^*_{SDP}$ is the optimal value of (SDP).

This semidefinite program can be solved. Now, we present a randomized algorithm for (P) by using the random rounding methods, and obtain a $\rho$-approximation algorithm which $\rho$ is not less than 0.5 for a wide range of the problem $(P)$.

# 3 Algorithm

In this section, we give a improved polynomial-time randomized approximation algorithm for the disjoint 2-catalog segmentation problem.

**Step 1.** Solve the problem $(SDP)$ to obtain : $v_1^*, v_2^*, \cdots, v_{2r}^*, w_1^*, \cdots, w_p^* \in S^{n-1}$, denote by $\bar{X} = (v_1^*, \cdots, v_{2r}^*, w_1^*, \cdots, w_p^*)$.

**Step 2.** Generates a random vector $u$ from a multivariate normal distribution with 0 mean and covariance matrix a convex combination of $\bar{X}$ and $X_0$, i.e.,

$$u \in N(0, \theta\bar{X} + (1-\theta)X_0).$$

**Step 3.** Let $\tilde{X}_1 = \{v_i | u^T v_i^* \geq 0, i = 1, 2, \cdots, 2r\}$, $\tilde{X}_2 = X \setminus \tilde{X}_1$; $\tilde{Y}_1 = \{w_j | u^T w_j^* \geq 0, j = 1, 2, \cdots, p\}$, $\tilde{Y}_2 = Y \setminus \tilde{Y}_1$.

**Step 4.** Suppose $|\tilde{X}_1| \geq r-k$. Let $X_1$ consist of $r-k$ elements of $\tilde{X}_1$ having the highest number of neighbors in $\tilde{Y}_1$ and $X_2$ consist of the remaining $(|\tilde{X}_1| - (r-k))$ vertices of $\tilde{X}_1$ together with $\tilde{X}_2$. Let $Y_1$ be the elements of $Y$ having more neighbors in $X_1$ than in $X_2$; and $Y_2 = Y \setminus Y_1$.

Where $0 \leq \theta \leq 1$, and

$$X_0 = \begin{pmatrix} 1 & \frac{2k^2-r}{r(2r-1)} & \cdots & \frac{2k^2-r}{r(2r-1)} \\ \frac{2k^2-r}{r(2r-1)} & 1 & \cdots & \frac{2k^2-r}{r(2r-1)} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{2k^2-r}{r(2r-1)} & \frac{2k^2-r}{r(2r-1)} & \cdots & 1 \end{pmatrix}$$

**Remark 1.** In algorithm *Step 2*, if we take $k = 1$, i.e., $u \in N(0, \bar{X})$, this algorithm was used to solve the Max-Bisection problem by Frieze and Jerrum[2], and also was a approximation algorithm for the disjoint 2-catalog segmentation problem[5], the performance ratios $\rho$ of this algorithm can be seen from Table 1 for the range of $0 \leq k \leq 0.2\ r$ (see Table 1).

**Remark 2.** In algorithm *Step 2*, if we take $k = 0$, i.e., $u \in N(0, X_0)$, we have a trivial 0.5-approximation algorithm when $r$ large enough.

**Remark 3.** In the following, we will select a reasonable value of $\theta$, such that the performance ratios $\rho$ of the algorithm is large as possible as, so that we can get more efficient algorithms.

# 4 Analysis of the algorithm

Let $S$ denote the segmentation: $X = X_1 \cup X_2$, $Y = Y_1 \cup Y_2$, $\tilde{S}$ denote the segmentation: $X = \tilde{X}_1 \cup \tilde{X}_2$, $Y = \tilde{Y}_1 \cup \tilde{Y}_2$. Let $\omega(S) = d(X_1, Y_1) + d(X_2, Y_2)$, $K^* = |X_1||X_2| = r^2 - k^2$.

Define random variables $\omega(\tilde{S})$, $\tilde{K}$ and $Z$ as follows:

$$\omega(\tilde{S}) = d(\tilde{X}_1, \tilde{Y}_1) + d(\tilde{X}_2, \tilde{Y}_2),$$
$$\tilde{K} = |\tilde{X}_1||\tilde{X}_2| = |\tilde{X}_1|(2r - |\tilde{X}_1|),$$
$$Z = \omega(\tilde{S})/\omega^*_{SDP} + \tilde{K}/K^*.$$

Clearly, the vertex swapping procedure in algorithm *Step 4* has the following property:

**Lemma 1.** *If $\tilde{X}_1 \geq r-k$, then*

$$\frac{\omega(S)}{r-k} \geq \frac{\omega(\tilde{S})}{|\tilde{X}_1|}.$$

**Lemma 2.** *Let $\alpha_0 = 0.878567$, we have*

*(1) $E[\omega(\tilde{S})] \geq \alpha_0 \omega^*_{SDP} \geq \alpha_0 \omega^*$,*

*(2) $E[\tilde{K}] \geq \alpha_0 K^*$,*

*(3) $E[Z] \geq 2\alpha_0$.*

**Proof.** (1) The proof can be seen in [4] (Goemans and Williamson, Theorem 2.3).
(2) By the similar analysis of Goemans and Williamson [4], we have

$$\begin{aligned}
E[\tilde{K}] \geq & \frac{\alpha_0}{4} \sum_{i=1}^{2r} \sum_{j=1}^{2r} (1 - \bar{X}_{ij}) \\
= & \frac{\alpha_0}{4} (4r^2 - \sum_{i=1}^{2r} \sum_{j=1}^{2r} \bar{X}_{ij}) \\
= & \frac{\alpha_0}{4} (4r^2 - ee^T \cdot \bar{X}) \\
= & \frac{\alpha_0}{4} (4r^2 - 4k^2) \\
= & \alpha_0 (r^2 - k^2) \\
= & \alpha_0 K^*.
\end{aligned}$$

(3) From (1) and (2), we get

$$E[Z] = \frac{E[\omega(\tilde{S})]}{\omega^*_{SDP}} + \frac{E[\tilde{K}]}{K^*} \geq 2\alpha_0.$$

$\square$

Similar to the discussion of Yinyu Ye [3], by selecting a reasonable value of $\theta$, we hope to provide the following two new inequalities:

$$E[\omega(\tilde{S})] = E[\frac{1}{4} \sum_{i=1}^{2r} \sum_{j=2r+1}^{n} \omega_{ij} (1 - \bar{X}_{ij})] \geq \alpha \cdot \omega^*. \tag{1}$$

and

$$E[\tilde{K}] = E[\frac{1}{4}\sum_{i=1}^{2r}\sum_{j=1}^{2r}(1-\bar{X}_{ij})] \geq \beta \cdot K^*. \qquad (2)$$

such that $\alpha$ would be slightly less than $0.878567, \beta$ would be significant greater than 0.878567, but we could give a better bound than that in[5](or Remark 1) for $\omega(\tilde{S})$.

Then, we introduce a new random variable

$$Z(\sigma) = \frac{\omega(\tilde{S})}{\omega^*_{SDP}} + \sigma\frac{\tilde{K}}{K^*}$$

where $\sigma$ is a parameter and $\sigma \geq 0$.

**Lemma 3.** *If (1), (2) hold, then*

$$E[Z(\sigma)] \geq \alpha + \sigma\beta.$$

**Theorem 4.** *Assume (1), (2) hold, then, for any given*

$$\sigma \geq \frac{(r^2-k^2)\alpha}{4r^2-(r^2-k^2)\beta},$$

*if random variable $Z(\sigma) \geq \alpha + \sigma\beta$, then*

$$\omega(S) \geq \frac{2(\sqrt{\sigma(\alpha+\sigma\beta)(r^2-k^2)} - r\sigma)}{r+k}\omega^*.$$

*In particular, if*

$$\sigma = \frac{\alpha}{2\beta}[\frac{r}{\sqrt{r^2-(r^2-k^2)\beta}} - 1],$$

*then*

$$\omega(S) \geq \frac{\alpha(r - \sqrt{r^2-(r^2-k^2)\beta})}{\beta(r+k)}\omega^*.$$

**Proof.** Let $\omega(\tilde{S}) = \lambda\omega^*_{SDP}, |\tilde{X}_1| = 2\delta r$. From lemma 1, we have

$$\omega(S) \geq \frac{r-k}{|\tilde{X}_1|}\omega(\tilde{S}) = \frac{\lambda(r-k)}{2\delta r}\omega^*_{SDP}.$$

By the hypothesis of $Z(\sigma)$ and Lemma 2, we have

$$\alpha + \sigma\beta \leq Z(\sigma) = \frac{\omega(\tilde{S})}{\omega^*_{SDP}} + \frac{\tilde{K}}{K^*} = \lambda + \frac{4\sigma\delta(1-\delta)r^2}{r^2-k^2}.$$

Hence, we obtain

$$\lambda \geq \alpha + \sigma\beta - \frac{4\sigma\delta(1-\delta)r^2}{r^2-k^2}.$$

Then we have

$$\begin{aligned}\omega(S) \geq & \frac{(r-k)}{2\delta r}(\alpha+\sigma\beta-\frac{4\sigma\delta(1-\delta)r^2}{r^2-k^2})\omega^*_{SDP} \\ = & \frac{(\alpha+\sigma\beta)(r^2-k^2)-4\sigma\delta(1-\delta)r^2}{2r(r+k)\delta}\omega^*_{SDP} \\ \geq & \frac{2(\sqrt{\sigma(\alpha+\sigma\beta)(r^2-k^2)}-r\sigma)}{r+k}\omega^*.\end{aligned}$$

The last inequality follows from simple calculus that $\delta = \sqrt{(r^2-k^2)(\alpha+\sigma\beta)}/2\sqrt{\sigma}$ yields the minimal value for $((\alpha+\sigma\beta)(r^2-k^2)-4\sigma\delta(1-\delta)r^2)/2r(r+k)\delta$ when $0<\delta\leq 1$.

In particular, substitute $\sigma = \frac{\alpha}{2\beta}[\frac{r}{\sqrt{r^2-(r^2-k^2)\beta}}-1]$ into the first inequality, we have the second inequality. □

## 5 The lower bounds of $\alpha,\beta$ and $\rho(\alpha,\beta,\varepsilon)$

**Lemma 5.** *For any $-1\leq x\leq 1$ and $0\leq\theta\leq 1, \varepsilon = k/r$, the function*

$$f(x) = \frac{1-\frac{2}{\pi}arcsin(\theta x+(1-\theta)\varepsilon^2)}{1-x}$$

*attains its minimal value $f(x_1^*)$ at $x_1^* = -0.8258$.*

$$g(x) = \frac{2}{\pi}\frac{arcsin(\theta)-arcsin(\theta x+(1-\theta)\varepsilon^2)}{1-x}$$

*attains its minimal value $g(x_2^*)$ at $x_2^* = -0.5779$.*

**Proof.** The proof can be obtained by a simple computing. □

**Theorem 6.** *When r is large enough, $X=\theta\bar{X}+(1-\theta)X_0$, then, (1) holds for $\alpha(\theta,\varepsilon)$, and, (2) holds for $\beta(\theta,\varepsilon)$, i.e.,*

$$E[\omega(\tilde{S})] = E[\frac{1}{4}\sum_{i=1}^{2r}\sum_{j=2r+1}^{n}\omega_{ij}(1-X_{ij})] \geq \alpha(\theta,\varepsilon)\cdot\omega^*. \tag{3}$$

*and*

$$E[\tilde{K}] = E[\frac{1}{4}\sum_{i=1}^{2r}\sum_{j=1}^{2r}(1-X_{ij})] \geq \beta(\theta,\varepsilon)\cdot K^*. \tag{4}$$

*where $\alpha(\theta,\varepsilon) = f(x_1^*)$, $x_1^* = -0.8258$, $\beta(\theta,\varepsilon) = 1-\frac{2}{\pi}arcsin(\theta)+g(x_2^*)$, $x_2^* = -0.5779$.*

**Proof.** The proof is similar to that of Theorem 1 in [3]. □

Table 1

| $\varepsilon = k/r$ | 0.2 | 0.15 | 0.1 | 0.05 | 0.03 | 0.01 | 0.001 | 0.0001 |
|---|---|---|---|---|---|---|---|---|
| $\rho$ | 0.497 | 0.540 | 0.579 | 0.616 | 0.631 | 0.644 | 0.650 | 0.651 |

Table 2

| $\varepsilon$ | $\theta$ | $\alpha(\theta,\varepsilon)$ | $\beta(\theta,\varepsilon)$ | $\rho(\alpha,\beta,\varepsilon)$ |
|---|---|---|---|---|
| 0.2 | 0.89 | 0.8333246 | 0.9600386 | 0.5208535 |
| 0.15 | 0.89 | 0.8343090 | 0.9609402 | 0.5689984 |
| 0.10 | 0.89 | 0.8350140 | 0.9615863 | 0.6164200 |
| 0.05 | 0.89 | 0.8354387 | 0.9619742 | 0.6607969 |
| 0.03 | 0.89 | 0.8355283 | 0.9620570 | 0.6770782 |
| 0.01 | 0.89 | 0.8355736 | 0.9620984 | 0.9627290 |
| 0.001 | 0.89 | 0.8355791 | 0.9621035 | 0.9687216 |
| 0.0001 | 0.89 | 0.8355792 | 0.9621036 | 0.6993526 |

Denote $\varepsilon = k/r$, and the performance guarantee

$$\rho(\alpha,\beta,\varepsilon) = \frac{\alpha(r - \sqrt{r^2 - (r^2 - k^2)\beta})}{\beta(r+k)} = \frac{\alpha(1 - \sqrt{1 - (1-\varepsilon^2)\beta})}{\beta(1+\varepsilon)}.$$

By computing, the ideal $\theta$ value almost is 0.89, we compute the performance ratios $\rho(\alpha,\beta,\varepsilon)$ of the improved algorithm for the range of $0 \le k \le 0.2r$ (see Table 2).

From Table 1, it can be seen that we have efficient polynomial-time approximation algorithms for a larger range of $k$, and let $k = 0$, we can obtain a 0.699-approximation algorithm for the disjoint equivalent 2-catalog segmentation problem.

## References


[1] J. Kleinberg, C. Papadimitrion and P. Raghavan. Segmentation problems. *STOC 98*, pp. 473–482, 1998.

[2] A. Frieze and M. Jerrum. Improoved algorithms for max-$k$-cut and bisection. *Proc. of 4th IPCO*, LNCS 920, Springer-Verlag, 1–13, 1995.

[3] Y. Y. Ye. A 0.699-approximation algorithm for max-bisection. *Math. Program*, 90, 101–111, 2001.

[4] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefiinite programming. *J. of ACM*, 42, 1115–1145, 1995.