TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 01/09 pp777–784
DOI: 10.26599/TST.2021.9010051
Volume 27, Number 5, October 2022

Approximation Algorithm for the Balanced 2-Correlation Clustering Problem

Sai Ji, Dachuan Xu, Donglei Du, Ling Gai*, and Zhongrui Zhao

Abstract: The Correlation Clustering Problem (CorCP) is a significant clustering problem based on the similarity of data. It has significant applications in different fields, such as machine learning, biology, and data mining, and many different problems in other areas. In this paper, the Balanced 2-CorCP (B2-CorCP) is introduced and examined, and a new interesting variant of the CorCP is described. The goal of this clustering problem is to partition the vertex set into two clusters with equal size, such that the number of disagreements is minimized. We first present a polynomial time algorithm for the B2-CorCP on M-positive edge dominant graphs ($M \geqslant 3$). Then, we provide a series of numerical experiments, and the results show the effectiveness of our algorithm.

Key words: balanced clustering; k-correlation clustering; positive edge dominant graphs; approximation algorithm

1 Introduction

Clustering problems arise in many applications, such as machine learning, computer vision, data mining and data compression, and have been widely studied[1–7].

In this study, we focus on the classical clustering problem, i.e., the Correlation Clustering Problem (CorCP), which was introduced by Bansal et al.[8] and has applications in data mining and machine learning. The input to the CorCP is a complete graph $G = (V, E)$, where V and E are the sets of vertices and edges in the graph, respectively. Moreover, each edge marked as positive or negative. The goal is to partition set V into several clusters, such that edges within clusters are mostly positive and edges between clusters are mostly negative. However, there is not necessarily a perfect partition for an instance. Let each positive edge whose two endpoints belong to the same cluster and each negative edge whose two endpoints belong to different clusters be an agreement. Similarly, let each positive edge whose two endpoints belong to different clusters and each negative edge whose two endpoints belong to the same cluster be a disagreement. Based on the purpose of the CorCP, it has two versions: minimizing disagreements and maximizing agreements. In the "minimizing disagreements" version, the goal is to partition set V into disjoint clusters so as to minimize the number of disagreements. In the "maximizing agreements" version, the goal is to partition set V into several disjoint clusters so as to maximize the number of agreements. In this study, we only focus on the "minimizing disagreements" version of the CorCP and its variants. The CorCP mentioned below belongs to the "minimizing disagreements" version and will not be emphasized in this paper.

Bansal et al.[8] first proved that the CorCP is NP-hard and APX-hard. People usually examine this problem by designing approximation algorithms[9–14]. Bansal

• Sai Ji and Zhongrui Zhao are with Department of Operations Research and Information Engineering, Beijing University of Technology, Beijing 100124, China. E-mail: jisai@amss.ac.cn; zhaozhongrui@emails.bjut.edu.cn.
• Dachuan Xu is with Beijing Institute for Scientific and Engineering Computing, Beijing University of Technology, Beijing 100124, China. E-mail: xudc@bjut.edu.cn.
• Donglei Du is with Faculty of Business Administration, University of New Brunswick, Fredericton, NB E3B 5A3, Canada. E-mail: ddu@unb.ca.
• Ling Gai is with Glorious Sun School of Business and Management, Donghua University, Shanghai 200051, China. E-mail: lgai@dhu.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2021-02-13; revised: 2021-05-27; accepted: 2021-06-28

et al.[8] provided a 17 433-approximation algorithm based on the combination technique. Charikar et al.[15] proposed a very natural linear programming of the CorCP and proved that the integer gap of the Linear Programming (LP) is 2. Then, based on the aboved LP and a region growth method, a 4-approximation algorithm was proposed, which significantly improves the approximation ratio of the algorithm proposed by Bansal et al.[8] As of now, the citation rate of this paper has reached 571, which is probably attributed to two main reasons: first, the algorithm is simple and has great innovation, and second, it has an important reference value when solving other combinatorial optimization problems. The best approximation algorithm is a 2.06-approximation algorithm, which was proposed by Chawla et al.[16] in 2015.

The CorCP has some limitations. To make the CorCP more effective and deal with the actual problems, some interesting variants of the CorCP have been widely studied, such as min-max CorCP[17, 18], CorCP in data streams[19], fair CorCP[20], capacitated CorCP[21], CorCP with a fixed number of clusters[11], and CorCP with noisy input[13, 22].

Balanced clustering problems arise in many applications, such as wireless sensor networks, routing and resource allocation. These problems have been widely studied[23–26]. The fair CorCP and capacitated CorCP have limitations on the proportion of vertices of different types in each cluster and on the number of vertices in each cluster, respectively. Accordingly, we propose and study the Balanced CorCP (BCorCP), which has a limitation on the ratio of the number of vertices in each cluster to the number of all vertices. The goal of the problem is to partition the vertices into several clusters with equal size so as to minimize the number of disagreements. In this paper, we examine a special case, i.e., Balanced 2-CorCP (B2-CorCP), which returns two clusters with equal size. In sum, this study has two main contributions:

(1) We developed a $(3, 24)$-balanced approximation algorithm for the B2-CorCP on M-positive edge dominant graphs ($M \geqslant 3$).

(2) We conducted numerical experiments and presented their results to show the effectiveness of our algorithm.

The remainder of this paper is structured as follows: In Section 2, we present definitions and the formulation of the B2-CorCP. In Sections 3 and 4, we discuss the approximation algorithm and theoretical analysis, respectively. In Section 5, we present the numerical experiments. Finally, in Section 6, we provide the conclusions of the study.

2 Definition and Formulation of B2-CorCP

In this section, we present the definitions and the formulation of the B2-CorCP used in this study.

Definition 1 (BCorCP) Given a labeled complete graph $G = (V, E)$, the goal is to partition the set V into several clusters with equal size, such that the number of disagreements is minimized.

Definition 2 (B2-CorCP) Given a labeled complete graph $G = (V, E)$, the goal is to partition the set V into two clusters with equal size, such that the number of disagreements is minimized.

Definition 3 ((α, β)-balanced approximation algorithm) ALG is an (α, β)-balanced approximation algorithm if, for any instance I, it returns a solution $\mathcal{C}^I = \{V_1^I, V_2^I, \ldots, V_k^I\}$ that satisfies the following properties:

(1) $\max\{|V_1^I|, \ldots, |V_k^I|\} \leqslant \alpha \min\{|V_1^I|, \ldots, |V_k^I|\}$.

(2) $ALG(I) \leqslant \beta \times OPT(I)$, where $ALG(I)$ and $OPT(I)$ are the objective function value of the solution returned by the algorithm ALG of Instance I and the objective function value of the optimal solution of Instance I, respectively.

Definition 4 (Positive edge dominant graph) Let E^+ be the set of positive edges and E^- be the set of negative edges in graph $G = (V, E)$. Graph G is a positive edge dominant graph if $|E_y^+| \geqslant |E_y^-|$ holds for each vertex $y \in V$, where $E_y^+ := \{(w, y) \in E^+ : w \in V\}$ and $E_v^- := \{(w, y) \in E^- : w \in V\}$.

Definition 5 (M-positive edge dominant graph) Graph $G = (V, E)$ is an M-positive edge dominant graph if

$$\inf_{y \in V} \frac{|E_y^+|}{|E_y^-|} \geqslant M.$$

For each edge $(w, y) \in E$, we introduce a 0−1 variable x_{wy} to represent whether vertices w and y belong to the same cluster. Variable $x_{wy} = 0$ if vertices w and y belong to the same cluster, and $x_{wy} = 1$ otherwise. Based on the above 0−1 variables, we can formulate the B2-CorCP as follows:

$$\min \quad \sum_{(w,y)\in E^+} x_{wy} + \sum_{(w,y)\in E^-} (1 - x_{wy}),$$

$$\text{s. t.} \quad x_{wy} + x_{yz} \geqslant x_{wz}, \quad \forall w, y, z \in V;$$

$$\sum_{y\in V}(1-x_{wy}) = |V|/2,\ \ \forall y \in V;$$
$$x_{yy} = 0, \quad \forall y \in V;$$
$$x_{wy} \in \{0,1\}, \quad \forall w, y \in V \tag{1}$$

The objective function has two parts: the first part is the number of disagreements derived from the positive edges, and the second part is the number of disagreements derived from the negative edges. Constraints in the formulation have three types: The first one ensures that we can obtain a feasible clustering of the CorCP. The second one ensures that we can exactly obtain two clusters of equal size. The third one is the natural constraint. By relaxing the 0−1 variables, we can determine the LP relaxation of Formula (1),

$$\begin{aligned}\min \quad & \sum_{(w,y)\in E^+} x_{wy} + \sum_{(w,y)\in E^-}(1-x_{wy}),\\ \text{s. t.} \quad & x_{wy} + x_{yz} \geqslant x_{wz}, \quad \forall w, y, z \in V;\\ & \sum_{y\in V}(1-x_{wy}) = |V|/2, \quad \forall y \in V;\\ & x_{yy} = 0, \quad \forall y \in V,\\ & 0 \leqslant x_{wy} \leqslant 1, \quad \forall w, y \in V \end{aligned} \tag{2}$$

3 Algorithm

In this section, we present our algorithm for the B2-CorCP on M-positive edge dominant graphs ($M \geqslant 3$), as shown in Algorithm 1, which mainly consists of two phases: Phase 1 (Steps 1−7) is a computational process based on the optimal fractional solution x^*. For each vertex y, we first sort the vertices in V in non-decreasing order by the value of x^* and let T_y be the set of the first half of the vertices according to the above order. Then, we compute the average value Avg_y of the vertices in T_y to vertex y. Phase 2 (Steps 8−14) is a clustering process. First, we select a center vertex $cen(V)$, and then cluster the vertices by comparing the value $Avg_{cen(V)}$ with a given threshold. If $Avg_{cen(V)}$ is greater than or equal to the threshold, then we make set $T_{cen(V)}$ a cluster and $V\backslash T_{cen(V)}$ another one. Otherwise, we let all vertices that are less than or equal to $1/2$ away from vertex $cen(V)$ be a cluster and the remaining vertices be a cluster. The second type of cluster does not essentially provide a feasible clustering, but we can determine the multiple relationships between the number of vertices in the two clusters.

Algorithm 1 Threshold-based algorithm

Input: A labeled M-positive edge dominant complete graph ($M \geqslant 3$).

Output: A partition V_1 and V_2 of V.

1: Initialize $V_1 = V_2 = \varnothing$.
2: Obtain the optimal solution x^* by solving Formula (2).
3: **for** each vertex $y \in V$ **do**
4: Sort vertices in V in non-decreasing order of x^*.
5: Let T_y be the set of the first half of vertices in V sorted by above order.
6: Denote
$$Avg_y := \frac{\sum_{w\in T_y} x^*_{wy}}{|T_y|}.$$
7: **end for**
8: Select vertex $cen(V)$ with the minimum $Avg_{cen(V)}$.
9: **if** $Avg_{cen(V)} \geqslant 1/4$ **then**
10: Update $V_1 := T_{cen(V)}$ and $V_2 := V/V_1$.
11: **else**
12: Update $V_1 := \{w \in V : x^*_{cen(V)w} \leqslant 1/2\}$ and $V_2 := V/V_1$.
13: **end if**
14: **return** V_1 and V_2.

4 Analysis

Recall Algorithm 1, We have $V_1 = T_{cen(V)}$ or $V_1 := \{w : x^*_{cen(V)w} \leqslant 1/2, w \in V\}$. Next, we analyze the upper bounds on the number of disagreements generated in the above two cases in Subsections 4.1 and 4.2, respectively.

4.1 Case 1: $V_1 = T_{cen(V)}$

In this case, we have $|V_1| = |V_2|$ and

$$Avg_v = \frac{\sum_{t\in T_v} x^*_{vt}}{|T_v|} \geqslant \frac{1}{4},\ \forall v \in V \tag{3}$$

Lemma 1 For any set $A \subseteq V$ with $|A| = |V|/2$, we have

(i) $\dfrac{\sum_{w\in A} x^*_{wy}}{|A|} \geqslant \dfrac{1}{4},\ \forall y \in V;$

(ii) $\dfrac{\sum_{w\in A} x^*_{wy}}{|A|} \leqslant \dfrac{3}{4},\ \forall y \in V.$

Proof

(i) From Formula (2), for each vertex $y \in V$, we have

$$\sum_{w\in V}(1-x^*_{wy}) = |V|/2 \Rightarrow \sum_{w\in V} x^*_{wy} = |V|/2 \Rightarrow$$
$$\frac{\sum_{w\in V} x^*_{wy}}{|V|} = \frac{1}{2} \tag{4}$$

Combining Formulas (3) and (4) and the definition of T_y, for each vertex $y \in V$, we have

$$\frac{1}{4} \leqslant \frac{\sum_{w\in T_y} x^*_{wy}}{|T_y|} \leqslant 1.$$

Therefore, for any set $A \subseteq V$ with $|A| = |V|/2$, we have

$$\frac{\sum_{w\in A} x^*_{wy}}{|A|} \geqslant \frac{1}{4}.$$

(ii) From Formula (4) and the definition of T_y, we have

$$\sum_{w\in V_2} x^*_{wy} = \sum_{w\in V} x^*_{wy} - \sum_{w\in V_1} x^*_{wy} \leqslant \frac{N}{2} - \frac{1}{4}\cdot|V_1| \leqslant \frac{3|V|}{8},$$

which implies that

$$\frac{\sum_{w\in V_2} x^*_{wy}}{|V_2|} \leqslant \frac{3}{4}.$$

Therefore, for any set $A \subseteq V$ with $|A| = |V|/2$, we have

$$\frac{\sum_{w\in A} x^*_{wy}}{|A|} \leqslant \frac{3}{4}.$$

Lemma 1 is concluded. ■

There are two types of disagreements:

(1) Edges $(w, y) \in E^+$ with $w \in V_1$ and $y \in V_2$, the number of these disagreements can be analyzed by Lemma 2.

(2) Edges $(y, z) \in E^-$ with $y, z \in V_i$ and $i = 1, 2$, the number of these disagreements can be analyzed by Lemma 3.

Lemma 2 For each vertex $y \in V_1$, the number of disagreements derived from the positive edges $(w, y), w \in V_2$ is no more than $4\sum_{(w,y)\in E^+} x^*_{wy}$.

Proof Graph $G = (V, E)$ is an M-positive edge dominant complete graph ($M \geqslant 3$), and for each vertex $y \in V_1$, inequality $|E^+_y| \geqslant 3|V|/4$ holds. Let A be any subset of E^+_y with $|A| = |V|/2$, then the disagreements derived from the positive edges $(w, y), w \in V_2$ is no more than $|A|$. Recall (i) of Lemma 1, we can obtain the following inequality:

$$|(w, y) \in E^+ : w \in V_2| \leqslant |A| \leqslant 4 \sum_{(w,y)\in E^+} x^*_{wy}.$$

The lemma is concluded. ■

Lemma 3 For each vertex $y \in V_i, i = 1, 2$, the number of disagreements derived from edges $(y, z) \in E^-$, with $y, z \in V_i$ and $i = 1, 2$, is no more than $4\sum_{(y,z)\in E^+,} x^*_{yz}$.

Proof We take $i = 1$ as an example. For each vertex $y \in V_1$, the number of disagreements derived from negative edges $(y, z), z \in V_1$, can be bounded by $|E^-_y|$. Similar to the proof of Lemma 2, let A be any subset of E^+_y with $|A| = |V|/2$. Combining (i) of Lemma 1 and the fact that $|E^-_y| \leqslant |V|/4$. Then we can obtain the following inequalities:

$$|(y, z) \in E^-, y, z \in V_1| \leqslant |E^-_y| < |A| \leqslant 4 \sum_{(y,z)\in E^+} x^*_{yz}.$$

Lemma 2 is concluded. ■

Theorem 1 If $|V_1| = |T_{cen(V)}| = |V_2|$, then the number of disagreements occurring due to the partition is no more than $24\sum_{(w,y)\in E^+} x^*_{wy}$.

Proof The number of disagreements equals

$$\sum_{y\in V_1} |(w, y) \in E^+ : w \in V_2| + |(y, z) \in E^- : y, z \in V_i, i = 1, 2|,$$

and it is less than

$$\sum_{y\in V_1} |(w, y) \in E^+ : w \in V_2| + \sum_{y\in V_i, i=1,2} |(y, z) \in E^- : z \in V_i|.$$

From Lemmas 2 and 3, we have

$$\sum_{y\in V_1} |(w, y) \in E^+ : w \in V_2| + \sum_{y\in V_i, i=1,2} |(y, z) \in E^- : z \in V_i| \leqslant 4 \sum_{y\in V_1,(w,y)\in E^+} x^*_{wy} + 4 \sum_{y\in V_1,(w,y)\in E^+} x^*_{wy} + 4 \sum_{y\in V_2,(w,y)\in E^+} x^*_{wy} \leqslant 24 \sum_{(w,y)\in E^+} x^*_{wy}.$$

Theorem 1 is concluded. ■

4.2 Case 2: $V_1 := \{y: x^*_{cen(V)y} \leqslant 1/2, y \in V\}$

In this case, we have

$$Avg_{cen(V)} = \frac{\sum_{y\in T_{cen(V)}} x^*_{cen(V)y}}{|T_{cen(V)}|} < \frac{1}{4} \quad (5)$$

$$\frac{\sum_{y\in V\setminus T_{cen(V)}} x^*_{cen(V)y}}{|V\setminus T_{cen(V)}|} > \frac{3}{4} \quad (6)$$

Lemma 4 If $V_1 := \{y : x^*_{cen(V)y} \leqslant 1/2, y \in V\}$, then we have

$$\min\{|V_1|, |V_2|\} < \max\{|V_1|, |V_2|\} < 3\min\{|V_1|, |V_2|\}.$$

Proof We consider Lemma 4 from the following cases:

(1) $\max_{y\in T_{cen(V)}} x^*_{cen(V)y} > 1/2$. If $|V_1| \leqslant |V|/4$, then we have $|T_{cen(V)}/V_1| = |V|/2 - |V_1| \geqslant |V|/4$ and

$$Avg_{cen(V)} = \frac{\sum_{y\in T_{cen(V)}} x_{cen(V)y}}{|T_{cen(V)}|} \geqslant \frac{\sum_{y\in T_{cen(V)}\setminus V_1} x_{cen(V)y}}{|T_{cen(V)}|} \geqslant \frac{\frac{1}{2}\cdot\frac{|V|}{4}}{\frac{|V|}{2}} \geqslant \frac{1}{4},$$

which contradicts with Formula (5). Therefore, we have $|V|/4 < |V_1| \leqslant |V|/2$, and hence $|V_1| < |V_2| < 3|V_1|$.

(2) $\max_{y\in T_{cen(V)}} x^*_{cen(V)y} \leqslant 1/2$. If $|V_2| \leqslant |V|/4$, then we have $|V_1\backslash T_{cen(V)}| \geqslant |V|/4$ and

$$\frac{\sum_{y\in V\backslash T_{cen(V)}} x^*_{cen(V)y}}{|V\backslash T_{cen(V)}|} = \frac{\sum_{y\in V_1\backslash T_{cen(V)}} x^*_{cen(V)y} + \sum_{y\in V_2} x^*_{cen(V)y}}{|V\backslash T_{cen(V)}|} \leqslant \frac{\frac{1}{2}|V_1\backslash T_{cen(V)}| + |V_2|}{|V\backslash T_{cen(V)}|} = \frac{\frac{1}{2}\left(\frac{|V|}{2} - |V_2|\right) + |V_2|}{|V\backslash T_{cen(V)}|} = \frac{\frac{|V|}{4} + \frac{1}{2}|V_2|}{|V\backslash T_{cen(V)}|} \leqslant \frac{\frac{|V|}{4} + \frac{|V|}{8}}{\frac{|V|}{2}} \leqslant \frac{3}{4},$$

which contradicts with Formula (6). Therefore, we have $|V|/4 < |V_2| \leqslant |V|/2$, and hence $|V_2| < |V_1| < 3|V_2|$.

Combining the above two cases, we conclude Lemma 4. ■

Lemma 5 The average distance of the vertices in V_1 to vertex $cen(V)$ satisfies

$$\frac{\sum_{y\in V_1} x^*_{cen(V)y}}{|V_1|} \leqslant \frac{1}{3}.$$

Proof We consider the following cases:

(1) If $|V_1| \leqslant |V_2|$, the lemma is evident and we omit the proof.

(2) If $|V_2| \leqslant |V_1| < 3|V_2|$, then we have $|V|/2 < |V_1| < 3|V|/4$. Therefore

$$\frac{\sum_{y\in V_1} x^*_{cen(V)y}}{|V_1|} \leqslant \frac{\frac{1}{4}\cdot\frac{|V|}{2} + \frac{1}{2}\cdot\frac{|V|}{4}}{\frac{3|V|}{4}} \leqslant \frac{1}{3}.$$

Lemma 5 is concluded. ■

Based on Lemmas 4 and 5 and Ref. [15], we can analyze the upper bound on the number of disagreements by Lemmas 6–8.

Lemma 6 The upper bound on the number of disagreements derived from the positive edges satisfies

(1) for each positive edge $(w, y), w \in V_1, y \in V_2$, with $x^*_{cen(V)y} \geqslant 2/3$. The number of disagreement generated by edge (w, y) can be bounded by $6x^*_{wy}$;

(2) for each $y \in V_2$, if $1/2 < x^*_{cen(V)y} < 2/3$, then the total number of disagreements $(w, y) \in E^+, w \in V_1$, can be bounded by

$$6\left[\sum_{(w,y)\in E^+, w\in V_1} x^*_{wy} + \sum_{(w,y)\in E^-, w\in V_1} (1 - x^*_{wy})\right].$$

Proof The proof is based on Ref. [15]. We omit the proof. ■

Lemma 7 The upper bound on the number of disagreements derived from the edges $(y, z) \in E^-$ with $y, z \in V_1$ satisfies

(1) for each edge $(y, z) \in E^-$ with $y, z \in V_1$, if $x^*_{cen(V)y}$ and $x^*_{cen(V)z} \leqslant 1/3$, then the number of disagreement derived from edge (y, z) can be bounded by $3(1 - x^*_{yz})$.

(2) for each vertex $y \in T_{cen(V)}$, if $1/3 < x^*_{cen(V)y} \leqslant 1/2$, then the number of disagreements derived from the edges $(y, z) \in E^-$ with $x^*_{cen(V)z} < x^*_{cen(V)y}$ can be bounded by

$$6\sum_{(y,z)\in E^+, x^*_{cen(V)z} < x^*_{cen(V)y}} x^*_{yz} + 6\sum_{(y,z)\in E^-, x^*_{cen(V)z} < x^*_{cen(V)y}} (1 - x^*_{yz}).$$

Proof The proof is based on Ref. [15]. We omit the proof. ■

In the following, we analyze the upper bound on the number of disagreements derived from edges in $(y, z) \in E^-$, where $y, z \in V_2$.

Lemma 8 The upper bound on the number of disagreements derived from edges $(y, z) \in E^-$ with $y, z \in V_2$ has the following two properties:

(1) For each vertex $y \in V_2$, if $x^*_{cen(V)y} \geqslant 2/3$, then the number of disagreements derived from the negative edges $(y, z) \in E^-, z \in V_2$, is no more than $6\sum_{(w,y)\in E^+, w\in V_1} x^*_{wy}$;

(2) For each vertex $y \in V_2$, if $1/2 \leqslant x^*_{cen(V)y} < 2/3$, the total disagreements derived from edges $(y, z) \in E^-, z \in V_2$, is no more than

$$6\left[\sum_{(w,y)\in E^+, w\in V_1} x^*_{wy} + \sum_{(w,y)\in E^-, w\in V_1} (1 - x^*_{wy})\right].$$

Proof Recall graph is an M-positive edge dominant complete graph ($M \geqslant 3$) and Lemma 7. For each $y \in V_2$, the number of disagreements is equal to $|\{(y, z) \in E^-, z \in V_2\}|$, which is less than $|\{(w, y) \in E^+, w \in V_1\}|$.

Combined with Lemma 6, we obtain Lemma 8. ■

Combining Lemmas 4–8, we obtain Theorem 2.

Theorem 2 If $V_1 := \{y : x^*_{cen(V)y} \leqslant 1/2, y \in V\}$, then we have

$$\max\{|V_1|, |V_2|\} < 3\min\{|V_1|, |V_2|\},$$

and the upper bound on the number of the disagreements is bounded by

$$12\left[\sum_{(w,y)\in E^+} x^*_{wy} + \sum_{(w,y)\in E^-} (1 - x^*_{wy})\right].$$

Combined with Theorems 1 and 2, we derive the main results of this study.

Theorem 3 Algorithm 1 is a $(3, 24)$-balanced approximation algorithm for the B2-CorCP on M-positive edge dominant graphs ($M \geqslant 3$).

5 Experiment

In this section, we explain how to generate M-positive edge dominant graphs and present the results of our numerical experiments.

5.1 Generation of M-positive edge dominant graphs

Let integer N be the number of vertices in an M-positive edge dominant graph. Moreover, let $[N] = 1, 2, \ldots, N$, for a given N, we generate an M-positive edge dominant graph ($M \geqslant 3$) based on the following steps:

(1) Generate a matrix $D \in \mathbf{R}^{N\times N}$, $d_{i,j} \in (0, 1)$, $i, j \in [N]$.

(2) For each $d_{i,j} \in D$, let

- $d_{i,j} = 1$, if $i \neq j$ and $d_{i,j} > 0.5$;
- $d_{i,j} = -1$, if $i \neq j$ and $d_{i,j} < 0.5$;
- $d_{i,j} = 0$, if $i = j$;
- $d_{i,j} = d_{j,i}$,

where i and j represent two vertices. Variable $d_{i,j} = 1$ if edge (i, j) is a positive edge, and variable $d_{i,j} = -1$ if edge (i, j) is a negative edge.

(3) For each row d_i in matrix D, $i \in [N]$, let P_i be the number of variables d_{ij}, $j \in [N]$, which is equal to 1, and N_i be the number of variables d_{ij}, $j \in [N]$, which is equal to -1. Perform the following steps from $i = 1$ to $i = k$: If $P_i \geqslant 3N_i$, then each variable stays the same. Otherwise, let the first $\left\lceil \dfrac{3N_i - P_i}{4} \right\rceil$ variables equal to -1 be reset as 1 from the variables and $d_{i,j} = d_{j,i}$.

(4) Take the lower or upper triangular matrix of D to represent the M-positive edge dominant graphs, because D is a symmetric matrix.

5.2 Results

We present Table 1 by taking different values of N. Opt is the objective function value of the optimal solution returned by Formula (2); Alg is the objective function value of the solution obtained by the Algorithm 1; $Appro$ is the ratio of *Alg* to *Opt*.

As shown in Table 1, although there are two cases in which Algorithm 1 outputs V_1 and V_2, the second case, as shown by the data, is rarely seen, and most of the time, we have $|V_1| = |V_2|$.

Furthermore, although the theoretical approximation ratio is 24, the actual ratio *Alg* to *Opt* is between 1 and 2 in the actual data, and hence our algorithm is expected to perform much better for real-life instances than what the worst-case approximation ratio suggests. As shown in Fig. 1, the ratio *Alg* to *Opt* (vertical axis) remains relatively stable with the increase in data points (horizon axis).

Table 1 Numerical experiment results for different N.

N	Opt	Alg	$Appro$	$\lvert V_1\rvert$	$\lvert V_2\rvert$
10	21.667	26	1.200	6	4
16	52.000	52	1.000	8	8
20	87.357	109	1.248	10	10
26	143.231	180	1.257	13	13
30	187.043	229	1.224	15	15
36	271.290	315	1.161	18	18
40	338.781	417	1.231	20	20
46	445.747	527	1.182	23	23
50	521.095	600	1.152	25	25
56	661.881	850	1.284	28	28
60	759.070	934	1.230	30	30
66	924.500	1098	1.187	33	33
70	1028.520	1244	1.209	35	35
76	1230.660	1435	1.166	38	38
80	1355.300	1560	1.154	40	40

6 Discussion and Conclusion

In this paper, we introduce the B2-CorCP and provide a $(3, 24)$-balanced approximation algorithm for the B2-CorCP on M-positive edge dominant graphs ($M \geqslant 3$). In sum, we have the following research directions for the B2-CorCP in the future:

- In this paper, we focus on the B2-CorCP on M-positive edge dominant graphs for $M \geqslant 3$. The case for $1 \leqslant M < 3$ is still open.
- The extension of the LP-rounding technique to obtain similar results for the B2-CorCP on general graphs will be discussed in future works.

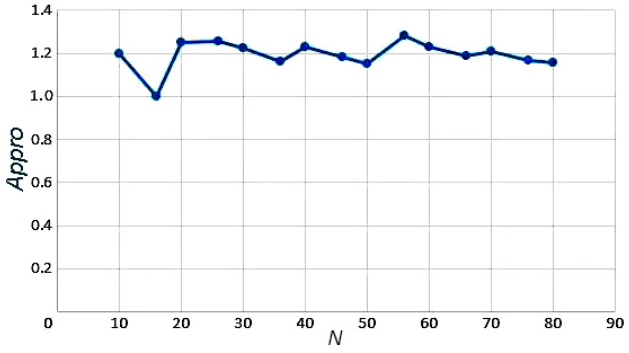


Fig. 1 Tendency of *Appro*.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (Nos. 12131003, 12101594, 11771386, 11728104, and 11201333), the Beijing Natural Science Foundation Project (No. Z200002), the China Postdoctoral Science Foundation (No. 2021M693337), and the Natural Sciences and Engineering Research Council of Canada (NSERC) (No. 06446).

References

[1] D. Arthur and S. Vassilvitskii, k-means++: The advantages of careful seeding, in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, Philadelphia, PA, USA, 2007, pp. 1027–1035.

[2] S. Ahmadian, A. Norouzi-Fard, O. Svensson, and J. Ward, Better guarantees for k-means and euclidean k-median by primal-dual algorithms, in *Proc. 58th Annu. IEEE Symp. Foundations of Computer Science*, Berkeley, CA, USA, 2017, pp. 61–72.

[3] J. Castro, S. Nasini, and F. Saldanha-Da-Gama, A cutting-plane approach for large-scale capacitated multi-period facility location using a specialized interior-point method, *Mathemat. Programm.*, vol. 163, nos. 1&2, pp. 411–444, 2017.

[4] S. Li and O. Svensson, Approximating k-median via pseudo-approximation, *SIAM J. Comput.*, vol. 45, no. 2, pp. 530–547, 2016.

[5] Y. Tian, R. Q. Zheng, Z. L. Liang, S. N. Li, F. X. Wu, and M. Li, A datadriven clustering recommendation method for single-cell RNA-sequencing data, *Tsinghua Science and Technology*, vol. 26, no. 5, pp. 772–789, 2021.

[6] Y. C. Xu, D. C. Xu, D. L. Du, and D. M. Zhang, Approximation algorithm for squared metric facility location problem with nonuniform capacities, *J. Supercomput.*, doi: 10.1016/j.dam. 2019.03.013.

[7] X. Zhao, Z. D. Wang, L. Gao, Y. L. Li, and S. J. Wang, Incremental face clustering with optimal summary learning via graph convolutional network, *Tsinghua Science and Technology*, vol. 26, no. 4, pp. 536–547, 2021.

[8] N. Bansal, A. Blum, and S. Chawla, Correlation clustering, *Mach. Learn.*, vol. 56, nos. 1–3, pp. 89–113, 2004.

[9] F. Bonchi, A. Gionis, and A. Ukkonen, Overlapping correlation clustering, *Knowl. Inform. Syst.*, vol. 35, no. 1, pp. 1–32, 2013.

[10] E. D. Demaine, D. Emanuel, A. Fiat, and N. Immorlica, Correlation clustering in general weighted graphs, *Theoret. Computer Sci.*, vol. 361, nos. 2&3, pp. 172–187, 2006.

[11] I. Giotis and V. Guruswami, Correlation clustering with a fixed number of clusters, in *Proc. 17th Annu. ACM-SIAM Symp. Discrete Algorithms*, Miami, FL, USA, 2006, pp. 1167–1176.

[12] N. Ailon, N. Avigdor-Elgrabli, E. Liberty, and A. Van Zuylen, Improved approximation algorithms for bipartite correlation clustering, *SIAM J. Comput.*, vol. 41, no. 5, pp. 1110–1121, 2012.

[13] C. Mathieu and W. Schudy, Correlation clustering with noisy input, in *Proc. 21th Annu. ACM-SIAM Symp. Discrete Algorithms*, Austin, TX, USA, 2010, pp. 712–728.

[14] E. Achtert, C. Bohm, J. David, P. Kröger, and A. Zimek, Global correlation clustering based on the Hough transform, *Stat. Anal. Data Min.*, vol. 1, no. 3, pp. 111–127, 2008.

[15] M. Charikar, V. Guruswami, and A. Wirth, Clustering with qualitative information, *J. Comput. Syst. Sci.*, vol. 71, no. 3, pp. 360–383, 2005.

[16] S. Chawla, K. Makarychev, T. Schramm, and G. Yaroslavtsev, Near optimal LP rounding algorithm for correlation clustering on complete and complete k-partite graphs, in *Proc. 47th Annu. ACM Symp. Theory of Computing*, New York, NY, USA, 2015, pp. 219–228.

[17] S. Ahmadi, S. Khuller, and B. Saha, Min-max correlation clustering via multiCut, in *Proc. Int. Conf. Integer Programming and Combinatorial Optimization*, Ann Arbor, MI, USA, 2019, pp. 13–26.

[18] G. J. Puleo and O. Milenkovic, Correlation clustering and biclustering with locally bounded errors, *IEEE Trans. Inform. Theory*, vol. 64, no. 6, pp. 4105–4119, 2018.

[19] K. J. Ahn, G. Cormode, S. Guha, A. McGregor, and A. Wirth, Correlation clustering in data streams, *Algorithmica*, vol. 83, no. 7, pp. 1980–2017, 2021.

[20] S. Ahmadian, A. Epasto, R. Kumar, and M. Mahdian, Fair correlation clustering, in *Proc. 23rd Int. Conf. on Artificial Intelligence and Statistics*, Palermo, Italy, 2020, pp. 4195–4205.

[21] G. J. Puleo and O. Milenkovic, Correlation clustering with constrained cluster sizes and extended weights bounds, *SIAM J. Optimizat.*, vol. 25, no. 3, pp. 1857–1872, 2015.

[22] K. Makarychev, Y. Makarychev, and A. Vijayaraghavan, Correlation clustering with noisy partial information, in *Proc. 28th Annu. Conf. Computational Learning Theory*, Paris, France, 2015, pp. 1321–1342.

[23] P. Kuila and P. K. Jana, Approximation schemes for load balanced clustering in wireless sensor networks, *J. Supercomput.*, vol. 68, no. 1, pp. 87–105, 2014.

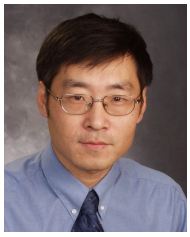
[24] B. Behsaz, Z. Friggstad, M. R. Salavatipour, and R. Sivakumar, Approximation algorithms for min-sum k-clustering and balanced k-median, *Algorithmica*, vol. 81, no. 3, pp. 1006–1030, 2019.

[25] J. M. Hendrickx and J. N. Tsitsiklis, Convergence of type-symmetric and cut-balanced consensus seeking systems, *IEEE Trans. Automat. Control*, vol. 58, no. 1, pp. 214–218, 2013.

[26] M. Zhao, Y. Y. Yang, and C. Wang, Mobile data gathering with load balanced clustering and dual data uploading in wireless sensor networks, *IEEE Trans. Mobile Comput.*, vol. 14, no. 4, pp. 770–785, 2015.

Sai Ji received the MS degree from Hunan Normal University, China in 2016. Currently, she is a PhD candidate at Beijing University of Technology. Her research interests include combinatorial optimization, approximation algorithm, and machine learning.

Dachuan Xu received the PhD degree from Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China in 2002. Currently he is a professor at Beijing University of Technology. His research interests include combinatorial optimization, robust optimization, game theory, machine learning, statistical optimization, and supply chain management.



Donglei Du received the PhD degree from Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China in 1996. Currently he is a professor at University of New Brunswick, Canada. His research interests include combinatorial optimization, robust optimization, approximation algorithm, supply chain management, and quantitative investment management.

Ling Gai received the PhD degree from Zhejiang University, China in 2007. Currently she is an associate professor at Donghua University. Her research interests include combinatorial optimization, approximation algorithm, behavioral operations research, and algorithmic game theory.

Zhongrui Zhao received the BEng degree from Qingdao University of Technology, China in 2020. Currently he is a master student at Beijing University of Technology. His research interests include combinatorial optimization, approximation algorithm, and submodular optimization.