

# A Note on the Inapproximability of Correlation Clustering

Jinsong Tan \*

## Abstract

We consider inapproximability of the correlation clustering problem defined as follows: Given a graph  $G = (V, E)$  where each edge is labeled either "+" (similar) or "-" (dissimilar), correlation clustering seeks to partition the vertices into clusters so that the number of pairs correctly (resp. incorrectly) classified with respect to the labels is maximized (resp. minimized). The two complementary problems are called MAXAGREE and MINDISAGREE, respectively, and have been studied on complete graphs, where every edge is labeled, and general graphs, where some edge might not have been labeled. Natural edge-weighted versions of both problems have been studied as well. Let  $\mathcal{S}$ -MAXAGREE denote the weighted problem where all weights are taken from set  $\mathcal{S}$ , we show that  $\mathcal{S}$ -MAXAGREE with weights bounded by  $O(|V|^{1/2-\delta})$  essentially belongs to the same hardness class in the following sense: if there is a polynomial time algorithm that approximates  $\mathcal{S}$ -MAXAGREE within a factor of  $\lambda = O(\log |V|)$  with high probability, then for any choice of  $\mathcal{S}'$ ,  $\mathcal{S}'$ -MAXAGREE can be approximated in polynomial time within a factor of  $(\lambda + \epsilon)$ , where  $\epsilon > 0$  can be arbitrarily small, with high probability. A similar statement also holds for  $\mathcal{S}$ -MINDISAGREE. This result implies it is hard (assuming  $\mathcal{NP} \neq \mathcal{RP}$ ) to approximate unweighted MAXAGREE within a factor of  $80/79 - \epsilon$ , improving upon a previous known factor of  $116/115 - \epsilon$  by Charikar et. al. [4].<sup>1</sup>

**Keywords:** Correlation Clustering, Inapproximability, Randomized Rounding, Graph Algorithm

## 1 Introduction

Motivated by applications of document clustering, Bansal, Blum and Chawla [2] introduced the correlation clustering problem where for a corpus of documents, we represent each document by a node, and an edge  $(u, v)$  is labeled "+" or "-" depending on whether the two documents are similar or dissimilar, respectively. The goal of correlation clustering is thus to find a partition of the nodes into clusters that agree as much as possible with the edge labels. Specifically, there are two complementary problems. MAXAGREE aims to maximize the number of agreements: the number of + edges inside clusters plus the number of - edges across clusters; on the other hand, MINDISAGREE aims to minimize the number of disagreements: the number of + edges across different clusters plus the number of - edges inside clusters. Correlation clustering is also viewed as a kind of agnostic learning problem [9] and seems to have been first studied by Ben-Dor et al. [3] with applications in computational biology; Shamir et al. [10] were the first to formalize it as a graph-theoretic problem, which they called Cluster Editing. Since Bansal et al.'s independent introduction of this problem [2], it has been studied quite extensively in recent years [1, 4, 5, 6, 7, 11].

---

\*Department of Computer & Information Sciences, University of Pennsylvania, Philadelphia, PA 19104. Email: jinsong@seas.upenn.edu

<sup>1</sup>Throughout the paper, when we talk about approximation factors we adopt the convention of assuming the factor is greater than 1 for both maximization and minimization problems.

MAXAGREE and MINDISAGREE have been studied on complete graphs, where every edge is labeled, and general graphs, where some edge might not have been labeled. The latter captures the case where a judge responsible for producing the labels is unable to tell if certain pairs are similar or not. Also, it is natural for the judge to give some ‘confidence level’ for the labels he produces; this leads to the natural edge-weighted versions, which we call  $\mathcal{S}$ -MAXAGREE and  $\mathcal{S}$ -MINDISAGREE respectively, indicating the edge weights are taken from set  $\mathcal{S}$ .

The various versions of correlation clustering are fairly well studied. For complete unweighted case, Bansal *et al.* [2] gave a PTAS for MAXAGREE and Charikar *et al.* [4] gave a 4-approximation for MINDISAGREE and showed APX-hardness. For general weighted graphs, an  $O(\log n)$ -approximation algorithm was also given in [4] for MINDISAGREE, and algorithms with the same approximation factor were also obtained independently by Demaine and Immorlica [5], and Emanuel and Fiat [6]; a  $\frac{1}{0.7664}$ -approximation algorithm was given for MAXAGREE in [4], and this was improved by Swamy [11] with a  $\frac{1}{0.7666}$ -approximation algorithm.

In this paper, we focus on the general graph case. Our main contribution is to show  $\mathcal{S}$ -MAXAGREE (resp.  $\mathcal{S}$ -MAXAGREE) with absolute values of weights bounded by  $O(|V|^{1/2-\delta})$  belongs to the same hardness class in the following sense: if there is a polynomial time algorithm that approximates  $\mathcal{S}$ -MAXAGREE (resp.  $\mathcal{S}$ -MAXAGREE) within a factor of  $\lambda = O(\log |V|)$  with high probability, then for any choice of  $\mathcal{S}'$ ,  $\mathcal{S}'$ -MAXAGREE (resp.  $\mathcal{S}$ -MAXAGREE) can be approximated in polynomial time within a factor of  $(\lambda + \epsilon)$ , for any constant  $\epsilon > 0$ , with high probability. This result implies it is hard (assuming  $\mathcal{NP} \neq \mathcal{RP}$ ) to approximate unweighted MAXAGREE within a factor of  $80/79 - \epsilon$ , improving upon a previous known factor of  $116/115 - \epsilon$  by Charikar, Guruswami and Wirth [4].

**Theorem 1** ([4]) *For every  $\epsilon > 0$ , it is  $\mathcal{NP}$ -hard to approximate the weighted version of MAXAGREE within a factor of  $80/79 - \epsilon$ . Furthermore, it is  $\mathcal{NP}$ -hard to approximate the unweighted version of MAXAGREE within a factor of  $116/115 - \epsilon$ .*

## 2 Definitions and Notations

We give definitions and notations in this section.

**Definition 1** ( $\mathcal{S}$ -MAXAGREE) *A MAXAGREE problem is called  $\mathcal{S}$ -MAXAGREE if all edge weights are taken from set  $\mathcal{S}$ . An element in  $\mathcal{S}$  can be either a constant or some function in the size of the input graph.*

$\mathcal{S}$ -MINDISAGREE is defined likewise. We assume 0 is always an element in  $\mathcal{S}$  as we are interested in the problem on general graphs in this paper. Assigning weight 0 to non-edges allows us to view any general graph as a complete one.

**Definition 2** ( $N$ -fold Roll) *Given a graph  $G = (V, E)$  where  $V = \{v_1, v_2, \dots, v_n\}$ . Let  $(N - 1)$  be a multiple of  $(n - 1)$ , an  $N$ -fold roll (denoted by  $G^N$ ) of  $G$  is created by embedding multiple copies of  $G$  into an  $N$  by  $n$  grid where there are  $N$  parallel copies of  $V$  and a node  $v_{ij}$  corresponds to  $v_j$  in the  $i$ th copy of  $V$ .*

*Edges of  $G^N$  are created as follows. For any pair of nodes  $v_{i_1 j_1}$  and  $v_{i_2 j_2}$ , where  $i_1, i_2 \in \{1, 2, \dots, N\}$ ,  $j_1, j_2 \in \{1, 2, \dots, n\}$ . Define the ‘wrapped-around’ vertical distance of the two nodes*

$$d(v_{i_1 j_1}, v_{i_2 j_2}) = \begin{cases} (i_2 - i_1 \bmod N) & (j_1 \leq j_2) \\ \infty & (\text{otherwise}) \end{cases}$$

*A pair  $(v_{i_1 j_1}, v_{i_2 j_2})$  is called a grid-bone if and only if*

- 1)  $j_1 \neq j_2$ ; and
- 2)  $\frac{d(v_{i_1 j_1}, v_{i_2 j_2})}{j_2 - j_1} \in \{0, 1, \dots, \frac{N-1}{n-1}\}$ .

A grid-bone  $(v_{i_1 j_1}, v_{i_2 j_2})$  is an edge identical to  $(v_{j_1}, v_{j_2})$  (resp. non-edge), depending on whether  $(v_{j_1}, v_{j_2})$  is an edge (resp. non-edge) in  $G$ . All non-grid-bone pairs  $(v_{i_1 j_1}, v_{i_2 j_2})$  are non-edges.

Note by construction  $G^N$  consists of exactly  $N(\frac{N-1}{n-1} + 1) > \frac{N^2}{n}$  duplicates of  $G$ . It is conceptually easier to see this by indexing each duplicate with pair  $(i, c)$ , where  $i \in \{1, 2, \dots, N\}$  indexes the  $N$  parallel copies of  $V$  and  $c \in \{0, 1, \dots, \frac{N-1}{n-1}\}$  can be thought of as the ‘slope’ of the grid-bones in this copy. More precisely, duplicate  $(i, c)$  consists of nodes

$$\{v_{(i \bmod N)1}, v_{(i+c \bmod N)2}, v_{(i+2c \bmod N)3}, \dots, v_{(i+(n-1)c \bmod N)n}\}$$

For our purpose that will be evident in the rest of the paper and for the sake of simpler analysis, we assume w.l.o.g. that there are exactly  $\frac{N^2}{n}$  duplicates of  $G$ . Note this can be thought of as erasing all edges on (any) excessive  $N(\frac{N-1}{n-1} + 1) - \frac{N^2}{n}$  duplicates.

In this construction, we obtain  $\frac{N^2}{n}$  disjoint duplicates of  $E$  from just  $N$  disjoint duplicates of  $V$ , this asymptotic gap is crucial in our proof of the main technical results (i.e. Lemma 3 and 4). We will discuss why we need this gap in the proof of Lemma 3.

**Definition 3** ( $\mathcal{S}$ -to- $\{-\alpha, 0, \beta\}$  randomized rounding)

**Input:** An instance of  $\mathcal{S}$ -MAXAGREE ( $\mathcal{S}$ -MINDISAGREE) on general graph  $G = (V, E)$ , where w.l.o.g. it is assumed that  $\gamma \leq 1, \forall \gamma \in \mathcal{S}$ ; and  $\alpha, \beta \geq 1$ .

**Output:** The same graph with the following randomized rounding. For each edge of weight  $\gamma > 0$  (resp.  $\gamma < 0$ ), round  $\gamma$  to either 0 or  $\beta$  (resp. either  $-\alpha$  or 0) independently and identically at random with expectation being  $\gamma$ .

Denote by  $w(\cdot)$  the weight function before rounding, and  $w'(\cdot)$  the one after rounding. We slightly abuse notation here by allowing both weight functions to take edges and clusterings as parameter. For a clustering  $C$ , denote by  $w'_\gamma(C)$  the total post-rounding weight of  $C$  contributed by former- $\gamma$ -edges.

**Definition 4** (Contributing) Given an  $\mathcal{S}$ -MAXAGREE (resp.  $\mathcal{S}$ -MINDISAGREE) instance and a clustering  $C$ , we call an edge  $(i, j)$  of weight  $\gamma$  a contributing edge iff  $\gamma > 0$  (resp.  $\gamma < 0$ ) and  $(i, j)$  is inside a cluster of  $C$ , or  $\gamma < 0$  (resp.  $\gamma > 0$ ) and  $(i, j)$  is cross different clusters of  $C$ .

### 3 Main Theorems

Given an  $\mathcal{S}$ -MAXAGREE (resp.  $\mathcal{S}$ -MINDISAGREE) instance, first construct an  $N$ -fold roll  $G^N = (V^N, E^N)$ , and then apply the  $\mathcal{S}$ -to- $\{-\alpha, 0, \beta\}$  randomized rounding on  $G^N$ . If we solve the  $\{-\alpha, 0, \beta\}$  instance on  $G^N$ , the solution clustering  $C$  implies a total of  $\frac{N^2}{n}$  (not necessarily distinct) ways to cluster  $G$ , one for each of the  $\frac{N^2}{n}$  duplicates of  $G$ . To see this, note  $C$  is simply a partition of  $V^N$ , and this partition induces a partition, thus a clustering, on each of the  $\frac{N^2}{n}$  duplicates of  $G$ . We call each of these clusterings a *candidate solution* to the initial  $\mathcal{S}$ -MAXAGREE (resp.  $\mathcal{S}$ -MINDISAGREE) instance on  $G$  and denote them as  $C_1, C_2, \dots, C_{\frac{N^2}{n}}$ .

Note although these  $\frac{N^2}{n}$  duplicates of  $G$  share nodes of  $G^N$ , their edge sets are disjoint. In fact, these  $\frac{N^2}{n}$  duplicates of  $E$  form a partition of  $E^N$ . Lemma 1 is immediate.

**Lemma 1** For both  $\mathcal{S}$ -MAXAGREE and  $\mathcal{S}$ -MINDISAGREE,  $w(C) = \sum_{i=1}^{N^2/n} w(C_i)$ .

Our next lemma says that if an edge is not contributing before rounding, it must not be contributing after rounding. Therefore, to calculate the weight of  $C$  both before and after the rounding, we only need to concern ourself with the same set of edges.

**Lemma 2** For both  $\mathcal{S}$ -MAXAGREE and  $\mathcal{S}$ -MINDISAGREE, let  $E(C)$  be the set of contributing edges of  $C$  before randomized rounding is applied to  $G^N$ , i.e.  $w(C) = \sum_{e \in E(C)} w(e)$ . Then after rounding, the new weight of  $C$  is still a summation over the same set of edges, i.e.  $w'(C) = \sum_{e \in E(C)} w'(e)$ .

**Proof.** This follows from the observation that positive edges are rounded to have either positive or zero weights, and negative edges are rounded to have either negative or zero weights.  $\square$

We are now ready to give our main technical result in Lemma 3. We concern ourself only with  $\mathcal{S}$ -MAXAGREE here; a similar result holds for  $\mathcal{S}$ -MINDISAGREE and is given in Lemma 4.

**Lemma 3** Given an  $\mathcal{S}$ -MAXAGREE instance  $G = (V, E)$ , let  $G^N = (V^N, E^N)$  be the  $N$ -fold roll of  $G$  with  $\mathcal{S}$ -to- $\{-\alpha, 0, \beta\}$  randomized rounding applied. If

1.  $\alpha + \beta = O((Nn)^{1/2-\delta})$ , where  $\delta \in (0, \frac{1}{2}]$ ; and
2. there is a  $\lambda$ -approximation algorithm for  $\{-\alpha, 0, \beta\}$ -MAXAGREE, where  $\lambda = O(\log n)$

then for any arbitrarily small number  $\epsilon > 0$  there exists a polynomial time algorithm that approximates  $\mathcal{S}$ -MAXAGREE within a factor of  $(\lambda + \epsilon)$  with probability at least  $\frac{1}{2}$ .

**Proof.** For any  $\gamma \in \mathcal{S}$ , let  $X^{(\gamma)}$  denote the random variable representing the new weight of a former- $\gamma$ -edge after rounding. Define random variable  $Y^{(\gamma)} = X^{(\gamma)} - \gamma$ ; clearly  $E[Y^{(\gamma)}] = 0$ . Note it is assumed w.l.o.g. that  $|\gamma| \leq 1, \forall \gamma \in \mathcal{S}$ .

Suppose there is a polynomial time algorithm  $\mathcal{A}$  that approximates  $\{-\alpha, 0, \beta\}$ -MAXAGREE within a factor of  $\lambda$ , we can then run  $\mathcal{A}$  on  $G^N$ , the output clustering  $C_2^*$  corresponds to  $\frac{N^2}{n}$  ways to cluster  $G$  (not necessarily all distinct). Let  $C_1^*$  be the most weighted among these  $\frac{N^2}{n}$  clusterings of  $G$ , in the rest of the proof we show that with high probability,  $C_1^*$  is a  $(\lambda + \epsilon)$ -approximation of  $\mathcal{S}$ -MAXAGREE on  $G$  for any fixed  $\epsilon$ .

Denote by  $\mathbb{E}$  the bad event that  $C_2^*$  does not imply a  $(\lambda + \epsilon)$ -approximation on  $G$ , i.e.  $C_1^*$  is not a  $(\lambda + \epsilon)$ -approximation. Let  $C'$  be an arbitrary clustering of  $G^N$  that does not imply a  $(\lambda + \epsilon)$ -approximation on  $G$ . Denote by  $\mathbb{E}(C')$  the event that  $C'$  becomes a  $\lambda$ -approximation on  $G^N$  after rounding. Since there are at most  $(Nn)^{Nn}$  distinct clusterings of  $G^N$ , by union bound we have  $Pr\{\mathbb{E}\} \leq e^{Nn \ln Nn} \cdot Pr\{\mathbb{E}(C')\}$ . (We note that the randomness of event  $\mathbb{E}(C')$  comes from the randomized rounding and the randomness of event  $\mathbb{E}$  comes from both the randomized rounding and the internal randomness of  $\mathcal{A}$ .)

Let the weight of an optimal clustering  $U$  of  $G$  be  $K$ , denote by  $U^N$  the corresponding duplication clustering in  $G^N$ . That is,  $U^N$  has the same number of clusters as  $U$ , and there is a one-to-one mapping between the two sets of clusters such that a node  $v_j$  is in a cluster of  $U$  if and only if all its  $N$  duplicates,  $v_{1j}, v_{2j}, \dots, v_{Nj}$ , are in the corresponding cluster of  $U^N$ . We now proceed to prove that event  $\mathbb{E}(C')$  happens with negligible probability. Before delving into the details, we first offer a high level discussion of the idea behind the proof.

**Intuition Behind the Proof.** Since  $U$  is an optimal clustering of  $G$ , by Lemma 1 it is easy to see that  $U^N$  is an optimal clustering of  $G^N$  before randomized rounding and its weight is  $\frac{KN^2}{n}$ .  $C'$  is an arbitrary but fixed clustering. Since it does not imply a  $(\lambda + \epsilon)$ -approximation on  $G$ , it must be the case that before rounding the weight of  $C'$  on  $G^N$  is less than  $\frac{KN^2}{(\lambda + \epsilon)n}$ . Since  $\epsilon$  is a fixed constant, this leaves a gap between  $\frac{KN^2}{(\lambda + \epsilon)n}$  and  $\frac{KN^2}{\lambda n}$ . By Lemma 2 the expectation of the new weight of  $U^N$  is  $\frac{KN^2}{n}$  and that of  $C'$  is at most  $\frac{KN^2}{(\lambda + \epsilon)n}$ . Therefore for the bad event  $\mathbb{E}(C')$  to happen either  $C'$  has to be really lucky in the rounding so that its new weight ends up hitting as high as  $\frac{KN^2}{\lambda n}$ , or  $U^N$  has to be really unlucky in the rounding so that its new weight ends up touching as low as  $\frac{\lambda KN^2}{(\lambda + \epsilon)n}$ , or mostly likely some sort of combination of the two. Whichever case

happens, the common thing shared is that one has to rely on pure chance to close the gap. And we show that by setting  $N = \text{poly}(n)$  sufficiently large, this happens with negligible probability. In fact, the probability of  $\mathbb{E}(C')$  is so small that even  $(Nn)^{Nn}$  times of it is still negligible.

We now resume the proof. For any  $\gamma \in \mathcal{S}$ , and a clustering  $C$  of  $G^N$ , denote by  $E(C, \gamma)$  the set of former- $\gamma$ -edges that are contributing in  $C$  before rounding. If  $\mathbb{E}(C')$  happens, then

$$\sum_{\gamma \in \mathcal{S}} w'_\gamma(C') \geq \frac{1}{\lambda} \left( \sum_{\gamma \in \mathcal{S}} w'_\gamma(U^N) \right)$$

$$\sum_{\gamma \in \mathcal{S}} |E(C', \gamma)| \cdot |\gamma| < \frac{KN^2/n}{\lambda + \epsilon} = \frac{1}{\lambda + \epsilon} \left( \sum_{\gamma \in \mathcal{S}} |E(U^N, \gamma)| \cdot |\gamma| \right)$$

where the first inequality follows because  $C'$  is a  $\lambda$ -approximation of  $G^N$ , and  $\sum_{\gamma \in \mathcal{S}} w'_\gamma(\cdot)$  is the total weight of a clustering after rounding; the second inequality follows from Lemma 1 and the fact that each of the  $\frac{N^2}{n}$  candidate solutions implied by  $C'$  has weight less than  $\frac{K}{\lambda + \epsilon}$ . Simple manipulation of the two inequalities above yields

$$S_1 - \frac{S_2}{\lambda} > \frac{\epsilon}{\lambda(\lambda + \epsilon)} \left( \sum_{\gamma \in \mathcal{S}} |E(U^N, \gamma)| \cdot |\gamma| \right) = \frac{\epsilon KN^2/n}{\lambda(\lambda + \epsilon)}$$

where  $S_1 = \left( \sum_{\gamma \in \mathcal{S}} (w'_\gamma(C') - |E(C', \gamma)| \cdot |\gamma|) \right)$  and  $S_2 = \frac{1}{\lambda} \left( \sum_{\gamma \in \mathcal{S}} (w'_\gamma(U^N) - |E(U^N, \gamma)| \cdot |\gamma|) \right)$ .

Since  $\lambda = O(\log(Nn)) = O(\log n)$  and  $K \geq 1$ , when  $n$  is sufficiently large,  $S_1 - \frac{S_2}{\lambda} \geq \frac{\epsilon N^2}{n^2}$ . This implies

$$\Pr\{\mathbb{E}(C')\} \leq \Pr\left\{S_1 - \frac{S_2}{\lambda} > \frac{\epsilon N^2}{n^2}\right\}$$

Note the expectation of both  $S_1$  and  $S_2$  are 0, therefore so is the linear combination  $S_1 - S_2/\lambda$ ; in the following we argue that the probability for  $S_1 - S_2/\lambda$  to deviate from its mean by  $\epsilon N^2/n^2$  is negligible when  $N$  is sufficiently large.

For any  $\gamma \in \mathcal{S}$ , let  $z_1(\gamma) = |E(C', \gamma) - E(U^N, \gamma)|$  be the number of former- $\gamma$ -edges contributing in  $C'$  but not  $U^N$  before rounding. Similarly, define  $z_2(\gamma) = |E(U^N) - E(C')|$  and  $z_3(\gamma) = |E(U^N) \cap E(C')|$ . We have

$$\begin{aligned} & \Pr\left\{S_1 - \frac{S_2}{\lambda} > \frac{\epsilon N^2}{n^2}\right\} \\ &= \Pr\left\{\sum_{\gamma \in \mathcal{S}} \left( \sum_{i=1}^{z_1(\gamma)} Y_i^{(\gamma)} + \frac{1}{\lambda} \sum_{j=1}^{z_2(\gamma)} (-Y_j^{(\gamma)}) + \frac{\lambda-1}{\lambda} \sum_{k=1}^{z_3(\gamma)} Y_k^{(\gamma)} \right) > \frac{\epsilon N^2}{n^2}\right\} \\ &\leq \Pr\left\{\sum_{\gamma \in \mathcal{S}} \left( \sum_{i=1}^{z_1(\gamma)} Y_i^{(\gamma)} + \sum_{j=1}^{z_2(\gamma)} (-Y_j^{(\gamma)}) + \sum_{k=1}^{z_3(\gamma)} Y_k^{(\gamma)} \right) > \frac{\epsilon N^2}{n^2}\right\} \quad (\lambda > 1) \\ &\leq \sum_{\gamma \in \mathcal{S}} \sum_{h \in \{1,2,3\}} \left( \Pr\left\{\sum_{i=1}^{z_h(\gamma)} (-1)^{(h-1)} Y_i^{(\gamma)} > \frac{\epsilon N^2}{3n^2|\mathcal{S}|}\right\} \right) \quad (\text{union bound}) \\ &\leq \sum_{\gamma \in \mathcal{S}} \sum_{h \in \{1,2,3\}} \left( \exp\left(-2z_h(\gamma) \left( \frac{\epsilon N^2}{3n^2|\mathcal{S}| \cdot z_h(\gamma) \cdot (\alpha + \beta)} \right)^2 \right) \right) \quad (\text{Hoeffding bound}) \\ &\leq 3|\mathcal{S}| \cdot \exp\left(-c_1 \cdot \frac{N^2/n}{n^8(\alpha + \beta)^2}\right) \quad (|\mathcal{S}| \leq n^2, z_h(\gamma) \leq N^2n) \end{aligned}$$

where  $c_1$  is some constant. Since we allow  $\alpha + \beta = O((Nn)^{(1/2-\delta)})$  and want  $(Nn)^{Nn} \cdot \Pr(\mathbb{E}(C'))$  to be negligible, it is now clear why we need  $N^2/n$  duplicates of  $E$  and thus the  $N$ -fold roll construction given in Definition 2. In contrast, had we adopted a naive construction with  $N$  isolated duplicates of  $G$ , there will be only  $N$  duplicates of  $E$ ; and it is readily verified that this is insufficient to prove that  $(Nn)^{Nn} \cdot \Pr(\mathbb{E}(C'))$  is negligible.

Now set  $N = n^{6/\delta}$ , we have

$$\Pr\{\mathbb{E}\} \leq (Nn)^{Nn} \cdot \Pr\{\mathbb{E}(C')\} \leq 3n^2 \cdot \exp\left((6/\delta + 1)n^{6/\delta+1} \ln n - c_2 \cdot n^{(6/\delta+2+2\delta)}\right)$$

for some constant  $c_2$ . Note this probability is bounded by  $\frac{1}{2}$  as the input size  $n$  is sufficiently large. Therefore we have obtained a polynomial time algorithm that approximates  $\mathcal{S}$ -MAXAGREE within a factor of  $\lambda + \epsilon$  with probability at least  $\frac{1}{2}$ .  $\square$

We give a similar result for  $\mathcal{S}$ -MINDISAGREE in Lemma 4, the proof follows essentially exactly the same construction and analysis as Lemma 3 so we only give a high level discussion without duplicating the proof.

**Lemma 4** *Given an  $\mathcal{S}$ -MINDISAGREE instance  $G = (V, E)$ , let  $G^N = (V^N, E^N)$  be the  $N$ -fold roll of  $G$  with  $\mathcal{S}$ -to- $\{-\alpha, 0, \beta\}$  randomized rounding. If*

1.  $\alpha + \beta = O((Nn)^{1/2-\delta})$ , where  $\delta \in (0, \frac{1}{2}]$ ; and
2. *there is a  $\lambda$ -approximation algorithm for  $\{-\alpha, 0, \beta\}$ -MINDISAGREE, where  $\lambda = O(\log n)$*

*then for any arbitrarily small number  $\epsilon > 0$  there exists a polynomial time algorithm that approximates  $\mathcal{S}$ -MINDISAGREE within a factor of  $(\lambda + \epsilon)$  with probability at least  $\frac{1}{2}$ .*

**Proof.** (Sketch) We define  $U^N$  and  $C'$  analogously as in that of Lemma 3. The weight of  $U^N$  before rounding is  $\frac{KN^2}{n}$ , and the weight of  $C'$  before rounding is greater than  $\frac{(\lambda+\epsilon)KN^2}{n}$ . Again since  $\epsilon$  is a fixed constant, there is a gap between  $\frac{(\lambda+\epsilon)KN^2}{n}$  and  $\frac{\lambda KN^2}{n}$ . For  $C'$  to be a  $\lambda$ -approximation after rounding, its new weight must necessarily be at most  $\lambda$  times of the new weight of  $U^N$ . Since the expectation of the new weight of  $U^N$  is  $\frac{KN^2}{n}$  and that of  $C'$  is greater than  $\frac{(\lambda+\epsilon)KN^2}{n}$ , again we need to rely on chance to close this gap of  $\frac{\epsilon KN^2}{n}$ . By applying a similar analysis as in Lemma 3 we can show that even  $(Nn)^{Nn}$  times of this probability, which upper bounds the probability of the bad event that a  $\lambda$ -approximation on  $G^N$  does not imply a  $(\lambda + \epsilon)$ -approximation on  $G$ , is negligible.  $\square$

Lemma 3 and 4 leads to the following theorem.

**Theorem 2** *If  $\mathcal{S}$ -MAXAGREE (resp.  $\mathcal{S}$ -MINDISAGREE) is  $\mathcal{NP}$ -hard to approximate within a factor of  $\lambda$  ( $\lambda = O(\log n)$ ) for any specific choice of  $\mathcal{S}$ , then for any choice of  $\mathcal{S}'$ , where  $\max_{\gamma \in \mathcal{S}'} |\gamma| = O(n^{1/2-\delta})$  for some  $\delta \in (0, \frac{1}{2}]$ , no polynomial time algorithm, possibly randomized, can approximate  $\mathcal{S}'$ -MAXAGREE (resp.  $\mathcal{S}'$ -MINDISAGREE) within a factor of  $\lambda + \epsilon$  with probability at least  $\frac{1}{2}$  unless  $\mathcal{NP} = \mathcal{RP}$ .*

**Proof.** This follows from Lemma 3 and 4 by setting  $\alpha = -\min \mathcal{S}$  and  $\beta = \max \mathcal{S}$ .  $\square$

In particular, invoking the result by Charikar *et al.* in Theorem 1 leads to the following improved inapproximability result.

**Theorem 3** *No polynomial time algorithm, possibly randomized, can approximate unweighted version of MAXAGREE in general graphs within a factor of  $80/79 - \epsilon$  unless  $\mathcal{NP} = \mathcal{RP}$ .*

**Acknowledgement.** The author would like to thank Tanmoy Chakraborty and the anonymous reviewers for their valuable comments and suggestions that helped to improve the presentation of the paper.

## References

- [1] N. Ailon, M. Charikar, A. Newman. Aggregating inconsistent information: Ranking and clustering. In *Proceedings of STOC'05*, 684C693, 2005.
- [2] N. Bansal, A. Blum, and S. Chawla. Correlation clustering, *Machine Learning*, 56:89-113, 2004.
- [3] A. Ben-Dor, R. Shamir, Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6:281C297, 1999.
- [4] M. Charikar, V. Guruswami, and A. Wirth. Clustering with Qualitative Information, *Journal of Computer and System Sciences*, 71:360-383, 2005.
- [5] E. Demaine, and N. Immorlica. Correlation clustering with partial information. In *Proceedings of APPROX'03*, 1-13, 2003.
- [6] D. Emanuel, and A. Fiat. Correlation Clustering - Minimizing Disagreements on Arbitrary Weighted Graphs. In *Proceedings of ESA'03*, 208-220, 2003.
- [7] I. Giotis, and V. Guruswami. Correlation clustering with a fixed number of clusters. *Theory of Computing*, 2:249-266, 2006.
- [8] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58:13-30, 1963.
- [9] M. Kearns, R. Schapire, L. Sellie. Toward efficient agnostic learning, *Machine Learning*, 17:115-142, 1994.
- [10] R. Shamir, R. Sharan, D. Tsur. Cluster graph modification problems. In *Proceedings of WG'02*, 379C390, 2002.
- [11] C. Swamy. Correlation Clustering: maximizing agreements via semidefinite programming. In *Proceedings of SODA'04*, 519-520, 2004.