Supplementary Material
Monochromatic Bi-Clustering

Sharon Wulff SHARON.WULFF@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Switzerland

Ruth Urner RURNER@CS.UWATERLOO.CA
Shai Ben-David SHAI@CS.UWATERLOO.CA
Department of Computer Science, ETH Zurich, Switzerland

A. NP-Hardness Proofs

For the sake of concreteness, we focus on the case where $D = \{0, 1\}$. The proofs readily generalize to arbitrarily domain D.

A.1. NP-hardness of $K = L = 2$

We construct a reduction from MaxCut. A *cut* in a graph $G = (V, E)$ is a partition of the vertex set into V_1 and V_2. The size of the cut is the number of edges in E that connect vertices from V_1 to vertices from V_2. The decision version of MaxCut is defined as follows:

Input A graph $G = (V, E)$, an integer r.

Question Is there a cut of G of size at least r?

The size of a cut can also be defined as the total number of edges, $|E|$ minus the edges within V_1 and the edges within V_2. For a subset $U \subseteq V$ of vertices and a vertex $v \in V$ let $v[U]$ denote the number of neighbors of v in U. We can re-write the size of the cut as

$$s(V_1, V_2) := \frac{1}{2}\left(\sum_{v\in V} v[V] - \sum_{v\in V_1} v[V_1] - \sum_{v\in V_2} v[V_2]\right) \tag{3}$$

Thus maximizing the size of a cut is equivalent to minimizing the *cost* of the cut $c(V_1, V_2)$ defined as

$$c(V_1, V_2) = \frac{1}{2}\left(\sum_{v\in V_1} v[V_1] + \sum_{v\in V_2} v[V_2]\right) \tag{4}$$

The MaxCut question can be reformulated as: Is there a cut of G of cost at most r?

Given an instance $G = (V, E)$, as well as a cost r, we construct an instance $\boldsymbol{M}$, with a $2, 2$-MCBC cost $\frac{2r}{|\boldsymbol{M}|}$. The construction is shown in figure 3. We start by defining the "left half" of the matrix $\boldsymbol{M}$. For every vertex $v \in V$ we introduce $n = |V|$ rows $r_1^v \dots r_n^v$ and n columns $c_1^v \dots c_n^v$. We set the entries of $\boldsymbol{M}$ corresponding to rows and columns of the same vertex (the "diagonal blocks"), to 1, i.e. $\forall v,\ 1 \le i, j \le n : \boldsymbol{M}[r_i^v, c_j^v] = 1$. Let $V = \{v_1 \dots v_n\}$ be an ordering of the vertices of G, if vertices v_i and v_j of G are connected by an edge, we set the entry $\boldsymbol{M}[r_j^{v_i}, c_i^{v_j}]$ and $\boldsymbol{M}[r_i^{v_j}, c_j^{v_i}]$ to 0. All other entries of the left half are set to $\star$.

The "right side" of $\boldsymbol{M}$ is an $n^2 \times n^2$ matrix as well, where the diagonal $n \times n$ blocks are set to 0 and the rest to $\star$. More formally, we introduce another set of n columns $0_1^v, \dots 0_n^v$ for each vertex $v \in V$. We set $\forall v,\ 1 \le i, j \le n : \quad \boldsymbol{M}[r_i^v, 0_j^v] = 0$. The remaining right half of $\boldsymbol{M}$ is set to $\star$. We refer to these columns as the 0-columns of a vertex v, and use O^v for the set $\{0_1^v, \dots 0_n^v\}$. Similarly we refer to the set of columns $\{c_1^v \dots c_n^v\}$ as the 1-columns of v, and denote it as C^v (note that while the 0 columns contain only 0 and $\star$ entries, the 1-columns consist mostly of 1 and $\star$, but also contain a few 0 entries, corresponding to edges of the graph). Finally we use R^v to refer to the set of rows $\{r_1^v \dots r_n^v\}$ associated with vertex v.

The NP-hardness of $2, 2$-MCBC now follows directly from the NP-hardness of MaxCut and the following lemma.

Lemma A.1. *G has a cut of cost at most r if and only if $\boldsymbol{M}$ has a $2, 2$-bi-clustering of monochromatic cost at most $\frac{2r}{|\boldsymbol{M}|}$.*

Proof. We first show that a cut of cost at most r induces a solution of the $2, 2$-bi-clustering of cost $\frac{2r}{|\boldsymbol{M}|}$. Let V_1, V_2 be a cut of G of cost at most r. We define a partition $P_R = \{R_1, R_2\}$ of the rows and $P_C = \{C_1, C_2\}$ of the columns as follows: For all

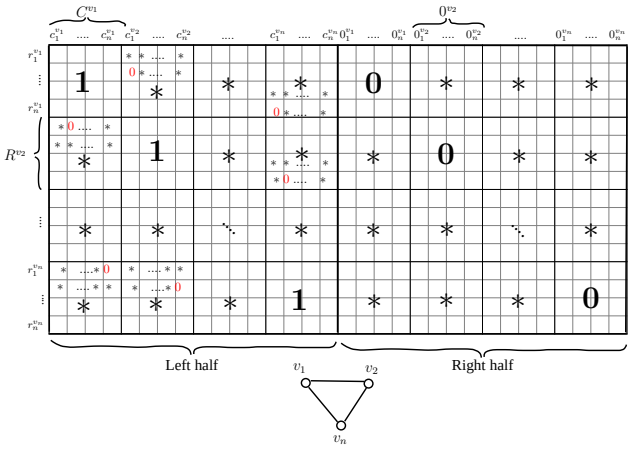

Figure 3. Construction of the reduction from MaxCut to $2,2$-MCBC. *Bottom*: The graph instance (only 3 out of n vertices are displayed)

$v \in V_1$ we put R^v in R_1, C^v in C_1 and O^v in C_2. For all $v \in V_2$ we do the opposite, put R^v in R_2, C^v in C_2 and O^v in C_1. The partition is depicted in figure 4. This results in a 2×2 bi-clustering with a majority of 1 in the diagonal blocks (upper left and lower right) and a majority of 0 is the off-diagonal ones (upper right and lower left). The off-diagonal blocks consist of only 0 and $\star$ entries, such that the monochromatic cost due to these blocks is 0. The diagonal blocks consist of: $n^2|V_t|$ 1-entries, and $\sum_{v \in V_t} v[V_t]$ 0-entries, for $t \in \{1,2\}$. The remaining entries are $\star$. Clearly there is a majority of 1 in these blocks, such that their total monochromatic cost amounts to exactly $2c(V_1, V_2) = 2r$ (see (4)). Normalizing by the size of the matrix yields the required cost.

Next we show that, if there is a bi-clustering of $\boldsymbol{M}$ with cost at most r, then there is cut of G of size $\frac{|\boldsymbol{M}|r}{2}$. Note that if there is any bi-clustering solution of cost at most r, then the optimal one has cost at most r, thus we consider a bi-clustering of optimal cost. For the sake of the proof we start by considering an optimal $n^2, 2$-bi-clustering, namely a bi-clustering where every row is a set of the row partition (thus, we are only interested in the resulting 2-partition of the columns). We will then argue, that this optimal $n^2, 2$-bi-clustering is actually a $2,2$-bi-clustering and therefore also the optimal $2,2$-bi-clustering.

We can assume that in an optimal solution identical columns are in the same cluster. Similarly, we can assume that two columns that are "inverse" of each other (one can be obtained from the other by replacing each 0 with a 1 and each 1 with a 0) are in different clusters. Thus, for any v, all columns in O^v are in one cluster. Consider a vertex v and its corresponding set of rows R^v. Without loss of generality let us assume that the 0-columns of v are in C_2. Each of the rows $r_2^v, \ldots, r_n^v$ contains n entries 1 and at most one entry 0 in the left half. The row r_1^v contains no 0. Equally the column c_1^v contains no 0 entries. Therefore, we can assume that c_1^v is in C_1 (it is the inverse of the columns in O^v). By way of contradiction, assume that not all columns of C^v are in C_1, say c_i^v is in C_2 for some $i \geq 2$. As at least c_1^v is in C_1, all columns in O^v are in C_2 and the rows $r_1^v, \ldots, r_n^v$ contain at most one 0 entry in the left half, we can assume that all these rows have a 1-block for C_1 and a 0-block for C_2. Having a column c_i^v in C_2 incurs a cost of at least n by its 1 entries, however moving it into C_1 can incur a cost of at most 1 as the column has at most one 0 entry and all its 1 entries do not contribute to the cost anymore. Thus, all columns in C^v are in C_1.

We showed that for every vertex, all its 0 columns are in one cluster and all its 1-columns are in the other cluster (which group is in which cluster may vary). Every row therefore has a block pattern $1,0$ or $0,1$, and the only cost that it incurs per row is that of a 0 entry of the left half, which ended up in the 1-block of the row. Now, grouping all the rows with pattern $1,0$ into R_1 and all the rows with pattern $0,1$ in R_2, leads to a $2,2$-bi-clustering of the same cost. As this cost is optimal for an $n,2$-bi-clustering, it is also optimal for $2,2$-bi-clustering (if there was a $2,2$-solution of lower cost, separating the rows into singleton sets for an $n,2$-bi-clustering would lead to a lower cost solution for this as well).

If we set V_1 to be the vertices whose 1-columns are in C_1 and V_2 the vertices whose 1-columns are in C_2 we obtain a cut. The only matrix entries that contribute to the cost of the bi-clustering are the 0 coming from edges within one of these sets. □

A.2. NP-hardness for larger K, L-MCBC

Proof outline: We prove the claim by reducing the $2,2$-MCBC problem to the K, L-MCBC problem. Given K and L as in the theorem (without loss of generality, we assume $K \leq 2^{L-1}$ here), and an input matrix $\boldsymbol{M}$ to the $2,2$-MCBC problem, we construct another matrix $\boldsymbol{N}$ such that an optimal K, L-MCBC partitioning of $\boldsymbol{N}$ will induce an optimal $2,2$-MCBC partitioning of $\boldsymbol{M}$.

Let m and n be the number of rows and columns of $\boldsymbol{M}$, respectively. The matrix $\boldsymbol{N}$ will consist of $(K-1) \times (L-1)$ blocks, each of size $m \times n$. The top left corner block of $\boldsymbol{N}$ will be the input matrix $\boldsymbol{M}$. All other blocks will be either all-zero matrices

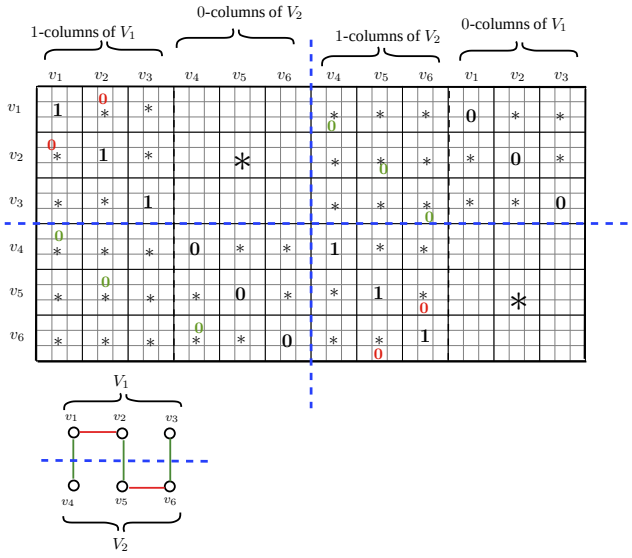

Figure 4. Optimal $2,2$-MCBC solution for the reduction matrix corresponding to the graph at the bottom. The zero entries highlighted in red correspond to edges that do not cross the cut, and incur a monochromatic cost respectively.

or all-one matrices. All the blocks in the top row of blocks, except the left-most block, will be all-zero matrices. The blocks indexed $(i,1)$, for $2 \le i \le 2^{L-2}-1$ (and $i \le K-1$) (blocks that reside below the top left corner block) will also be all-zero matrices. And the blocks indexed $(i,1)$, for $2^{L-2} \le i \le K-1$ (also residing below the top left corner block) will be all-one matrices. Finally, for every $2 \le i \le 2^{L-2}$ let v_i be vectors in $\{0,1\}^{L-2}$ such that for all $i \ne j$, $v_i \ne v_j$ and none is the all-zero vector (i.e. let the set of the v_i be all vectors in $\{0,1\}^{L-2}$ except the all-zero). Now let the (i,j) block, for $2 \le i \le 2^{L-2}$ (and $i \le K-1$) and $2 \le j \le L-1$, be a homogeneous matrix all of whose entries equal the $j-1$ entry of the vector v_i. Finally, set the entries of the $(2^{L-2}+i,j)$ block equal to those of the (i,j) block for all $1 \le i \le 2^{L-2}-1$ (and $i \le K-1$) and $2 \le j \le L-1$.

It is easy to see that the optimal K,L-MCBC partition of $\boldsymbol{N}$ induces an optimal $2,2$-MCBC partition over $\boldsymbol{M}$.

A.3. NP-hardness for matrices with arbitrary fraction of $\star$ entries

Given an $\epsilon > 0$, and K and L satisfying the condition of theorem 4.1. We show that $K+1, L+1$-MCBC is NP-hard restricted to input matrices containing at most an ϵ fraction of $\star$-entries by a reduction from general K,L-MCBC. Given a matrix $\boldsymbol{M}$, we construct matrix $\boldsymbol{N}$ as follows: We add $|M|\frac{1}{\epsilon}$ rows and columns to $\boldsymbol{M}$ such that the upper left block of $\boldsymbol{N}$ is identical to $\boldsymbol{M}$. We set the entries of the upper right and lower left blocks to 1 and the entries to the lower right block to 0. Now $\boldsymbol{N}$ has at most an ϵ-fraction of $\star$ entries. Further, it is easy to see that the optimal $K+1, L+1$-MCBC partition of $\boldsymbol{N}$ induces an optimal K,L-MCBC solution to $\boldsymbol{M}$.

B. Approximation Algorithm Proofs

B.1. Proof of theorem 5.2

To prove theorem 5.2 we first prove the following

Theorem B.1. *On input $\boldsymbol{M}$, $\boldsymbol{A}$, K, L, ϵ, δ, with probability at least 1 - δ the monochromatic approximation algorithm (given in Algorithm 1), lines $(4)-(11)$ outputs a partition P of $\boldsymbol{M}$ such that $Mon_{\boldsymbol{A}}(\boldsymbol{M},P) \le OPT_{\boldsymbol{A}} + 4\epsilon$ where $OPT_{\boldsymbol{A}}$ is the minimal monochromatic cost with respect to $\boldsymbol{A}$.*

Proof. We start by analyzing the partition of the columns by algorithm 1. Let $P_R^\star = \{R_1^\star, \dots R_K^\star\}$ denote the optimal partition of the rows of $\boldsymbol{M}$ (with respect to pattern $\boldsymbol{A}$). We say that a sample $R^S \subset R$ is *good*, if there exists a partition $P_R^S = \{R_1^S, \dots R_K^S\}$ of R^S such that, for all columns $j \in C$, except for at most $\epsilon|C|$ columns, the following holds:

$$\begin{aligned} &\forall\ 1 \le l \le L: \\ &\parallel Err(j,l|\boldsymbol{A},P_R^\star) - Err(j,l|\boldsymbol{A},P_R^S) \parallel \le \epsilon \end{aligned} \tag{5}$$

Namely, that in every column block l, the difference in the number of errors between the placement of the column j in the l'th block under the optimal partition of the rows, and the partition P_R^S of the sample of rows, is bounded by ϵ.

Let R^S be a good sample of the rows, and let $P_R^S = \{R_1^S, \dots R_K^S\}$ be the partition of the sample for which all but a fraction of ϵ of the columns in C satisfy 5. Consider a column $j \in C$,

1. If j satisfies 5, then placing j in a greedy manner with respect to P_R^S and the pattern $\boldsymbol{A}$, yields a cost increase compared to the optimal partition which is bounded by $t\epsilon \le m\epsilon$. The number of columns is n, and therefore the cost increase in this case compared to the optimal solution is bounded by $mn\epsilon$.

2. If j does not satisfy 5, we can still bound the number of entries of j disagreeing with the pattern, by m, the overall number of entries. Since we assumed that R^S is a good sample, the number of such columns is bounded by $n\epsilon$. This implies

a cost increase compared to the optimal solution, of $mn\epsilon$.

Altogether the number of errors incurred by the partition of the columns is bounded by $2mn\epsilon$. We can carry out the same analysis only for the partition of the rows, assuming we have a good sample of the columns. The overall increase compared to the optimal solution is then bounded by $4mn\epsilon$.

We defined the monochromatic pattern cost (2) as a fraction of the number of mistakes made by the partition divided by the size of the matrix (mn), we can therefore conclude that the algorithm yields a solution which is at most OPT+4ϵ.

Now it suffices to show the following.

Lemma B.2. *With probability at least 1 - δ over the random sampling, the rows and columns samples picked by the algorithm are good w.r.t the the optimal solution.*

Proof. As before we use R^S to denote a sample of the rows of size t, let $P_R^{\star S}$ denote the restriction of the optimal partition $P_R^\star$ to the sample R^S, that is

$$\forall\, 1 \le k \le K \quad R_k^{\star S} = R_k^\star \bigcap R^S$$

We can focus our analysis on this specific partition of the sample of rows since the approximation algorithm is going over all possible partitions, and is therefore guaranteed to consider this one.

Let $j \in C$ be a column, we define an indicator random variable ξ_i^l for each row index $i \in R^S$ and column block index $l \in [L]$ in the following way

$$\xi_i^l = \begin{cases} 1 & \text{if } \boldsymbol{M}[i,j] \neq \star \text{ and } \boldsymbol{M}[i,j] \neq \boldsymbol{A}[P_R^{\star S}(i), l] \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Where for a row index i, $P_R^{\star S}(i)$ denotes its row block assignment in the partition $P_R^{\star S}$. The variable ξ_i^l is simply the error associated with the i entry in the column j, if we place it in the column block l, given that the partition of the rows is given by $P_R^{\star S}$ and the target pattern is $\boldsymbol{A}$.

It is easy to see that the random variable corresponding to the sum over ξ_i^l, for all i in the sample R^S, is simply the error function defined in (5.1).

$$\frac{1}{t}\sum_{i \in R^S} \xi_i^l = Err(j, l | \boldsymbol{A}, P_R^{\star S}) \tag{7}$$

Using Chernoff additive bound we can guarantee that with a sample size of $t = \frac{1}{2\epsilon^2}\log\frac{L}{\delta\epsilon}$ equation (5) holds for a certain column j and column block l with probability at least $1 - \frac{\epsilon\delta}{L}$:

$$\begin{aligned} Pr(\parallel Err(j,l|\boldsymbol{A}, P_R^\star) - Err(j,l|\boldsymbol{A}, P_R^S) \parallel > \epsilon) \\ \le \exp(-2\epsilon^2 t) \end{aligned} \tag{8}$$

We apply markov inequality to get that for all columns $j \in C$ except for not more than $\epsilon|C|$, for a specific column block j equation (5) holds with probability at least $1 - \frac{\delta}{L}$. Finally we get a guarantee of (5) for all blocks l with probability at least $1 - \delta$.

The same analysis applies for the approximated partition of the rows determined by a sufficiently large sample of the columns. □

□

Proof of theorem 5.2

Proof. According to theorem B.1, the monochromatic approximation algorithm for a given pattern $\boldsymbol{A}$, computes a partition P such that $Mon_{\boldsymbol{A}}(\boldsymbol{M}, P) \le OPT_{\boldsymbol{A}}+4\epsilon$. This in turn translates into a bound on the agreement of $1 - Mon_{\boldsymbol{A}}(\boldsymbol{M}, P) \ge OPT_{\boldsymbol{A}} - 4\epsilon$. Since the algorithm goes over all possible patterns, and finally picks the pattern and partition with the lowest overall cost, the optimal pattern will be considered as well and thus the returned partition is guaranteed to have a cost $\le OPT + 4\epsilon$ or agreement $\ge OPT - 4\epsilon$. The run time increase due to the iteration over the patterns is exponential in K, L but is constant in $|\boldsymbol{M}|$. □

B.2. Proof Of Corollary 5.3

Proof. There is always a trivial solution to the monochromatic bi-clustering problem with an agreement score of at least $\frac{1}{2}$. This is simply assigning all of the rows and all of the columns to the same cluster. Note that the presence of missing entires implies that the agreement score of this trivial solution is even strictly larger than $\frac{1}{2}$ (see definition 1). Therefore an additive 4ϵ approximation translates into a relative $(1-\epsilon)OPT$ bound on the agreement of the solution, with a fixed increase of the sample size and therefore the running time.

$$\begin{aligned} 1 - Mon(\boldsymbol{M}, P) \ge \quad & OPT - 4\epsilon = OPT(1 - \tfrac{4\epsilon}{OPT}) \\ & \text{use } OPT \ge \tfrac{1}{2} \\ & \ge OPT(1 - 8\epsilon) \\ & \text{substitute } \epsilon' = 8\epsilon \end{aligned}$$

The corollary now follows from Theorem 5.2 □