# Bipartite correlation clustering: Maximizing agreements

Megasthenis Asteris, Anastasios Kyrillidis, Dimitris Papailiopoulos, Alexandros Dimakis

The University of Texas at Austin
Electrical and Computer Engineering
*Cockrell School of Engineering*

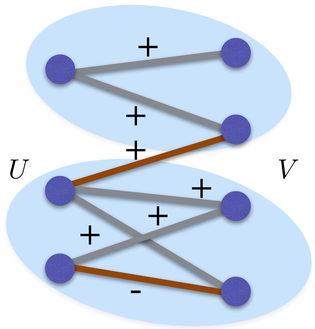
## [ What is bipartite correlation clustering (BCC)? ]

**Task:** given a bipartite graph $G = (U, V, E)$, partition objects into clusters, based on pairwise relationships, in order to maximize the number of ***agreements***.

Two vertices are similar: '+' edge / otherwise: '-' edge.

**Number of agreements:** # of '+' within resulting clusters
\+ # of '-' across resulting clusters.

**Example:**




Number of agreements: 5 '+'
Number of disagreements: 1 '+' and 1 '-'

# of output clusters: *not part of the input*

Might be useful to restrict number of output clusters:

**k-BCC**

Applications:

1. Recommendation systems
2. Gene expression data analysis
3. Graph partitioning

## [ Contributions ]

We develop a novel approximation algorithm for k-BCC with provable guarantees for the "maximizing agreements" problem.

In particular:

- **k-BCC:** given $G = (U, V, E)$, a number of clusters k and any accuracy parameter $\delta \in (0, 1)$, our algorithm computes a clustering of at most k clusters and within $(1 - \delta)$-factor from the optimal.
- **BCC:** We show that, in order to achieve a $(1 - \delta)$-approximation in the "maximizing agreements" objective, it suffices to use at most $O(\delta^{-1})$ clusters.
- **Proposed scheme:** We propose a new algorithm, based on constrained bilinear maximization techniques, that yields an EPTAS for the unconstrained BCC problem

## [ Related work ]

| Ref. | Min./Max. | Guarantee | Complexity | D/R | Setting |
|---|---|---|---|---|---|
| [13] | MaxAgree | (E)PTAS | $n/\delta \cdot k^{O\left(\delta^{-2}\log k \log(1/\delta)\right)}$ | R | $k$-CC |
| [13] | MinDisagree | PTAS | $n^{O(100^k/\delta^2)} \cdot \log n$ | R | $k$-CC |
| [15] | MinDisagree | PTAS | $n^{O(9^k/\delta^2)} \cdot \log n$ | R | $k$-CC |
| [11, 9] | MinDisagree | $O(\log n)$-OPT | LP | D | Inc. CC |
| [17] | MaxAgree | 0.766-OPT | SDP | D | Inc. CC |
| [5] | MinDisagree | 11-OPT | LP | D | BCC |
| [2] | MinDisagree | 4-OPT | LP | D | BCC |
| [2] | MinDisagree | 4-OPT | $\lvert E\rvert$ | R | BCC |
| Ours | MaxAgree | (E)PTAS | $2^{O(k/\delta^2\cdot\log\sqrt{k}/\delta)} \cdot (\delta^{-2}+k)\cdot n + \mathrm{T}_{\mathrm{SVD}}(\delta^{-2})$ | R | $k$-BCC |
| | MaxAgree | (E)PTAS | $2^{O(\delta^{-3}\cdot\log\delta^{-3})} \cdot O(\delta^{-2})\cdot n + \mathrm{T}_{\mathrm{SVD}}(\delta^{-2})$ | R | BCC |

Table 1: Summary of results on BCC and related problems. For each scheme, we indicate the problem setting, the objective (MaxAgree/MinDisagree), the guarantees ($c$-OPT implies a multiplicative factor approximation), and its computational complexity ($n$ denotes the total number of vertices, LP/SDP denotes the complexity of a linear/semidefinite program, and $\mathrm{T}_{\mathrm{SVD}}(r)$ the time required to compute a rank-$r$ truncated SVD of a $n \times n$ matrix). The D/R column indicates whether the scheme is deterministic or randomized.

## [ k-BCC as a bilinear maximization problem ]

**Setting:** For any instance $G = (U, V, E)$ of k-BCC problem, let

$\mathbf{B} \in \{\pm 1\}^{|U|\times|V|}$ : Bi-adjacency matrix

$\mathbf{X} \in \{0,1\}^{|U|\times k}$, $\mathbf{Y} \in \{0,1\}^{|V|\times k}$ : Cluster assignment matrices (i.e., $X_{ij} = 1$ iff vertex $i \in U$ is assigned to cluster $C_j$.)

**Lemma 3.2 :** For any clustering $\mathcal{C}$ of $U \cup V$ into k clusters, the number of agreements achieved by $\mathcal{C}$ is:

$$\textsc{Agree}(\mathcal{C}) = \textsc{Tr}\left(\mathbf{X}^\top \mathbf{B}\mathbf{Y}\right) + |E^-|$$

where $E^-$ denotes the set of negative edges in the graph.

Thus, computing a k-clustering that achieves the **maximum number of agreements,** builds down to a **constrained bilinear maximization problem:**

$$\textsc{MaxAgree}[k] = \max_{\mathbf{X}\in\mathcal{X},\mathbf{Y}\in\mathcal{Y}} \textsc{Tr}\left(\mathbf{X}^\top \mathbf{B}\mathbf{Y}\right) + |E^-|,$$

where

$$\mathcal{X} \triangleq \left\{\mathbf{X} \in \{0,1\}^{m\times k} : \|\mathbf{X}\|_{\infty,1} = 1\right\},$$
$$\mathcal{Y} \triangleq \left\{\mathbf{Y} \in \{0,1\}^{n\times k} : \|\mathbf{Y}\|_{\infty,1} = 1\right\}.$$

## [ Bilinear maximization framework ]

Consider the following maximization problem:

$$\max_{\mathbf{X}\in\mathcal{X},\mathbf{Y}\in\mathcal{Y}} \quad \textsc{Tr}\left(\mathbf{X}^\top \mathbf{A}\mathbf{Y}\right)$$

where $\mathbf{A}$ is a real $m \times n$ input matrix.

In general **NP-hard**; we approximate the above problem as:

$$\max_{\mathbf{X}\in\mathcal{X},\mathbf{Y}\in\mathcal{Y}} \quad \textsc{Tr}\left(\mathbf{X}^\top \widetilde{\mathbf{A}}\mathbf{Y}\right)$$

where $\widetilde{\mathbf{A}}$ is the rank-r approximation of $\mathbf{A}$.

## [ Algorithm for bilinear maximization ]

**Algorithm 1** BiLinearLowRankSolver

**input** : $m \times n$ real matrix $\widetilde{\mathbf{A}}$ of rank $r$, $\epsilon \in (0, 1)$
**output** $\widetilde{\mathbf{X}} \in \mathcal{X}$, $\widetilde{\mathbf{Y}} \in \mathcal{Y}$

$\mathcal{C} \leftarrow \{\}$ {Candidate solutions}
$\widetilde{\mathbf{U}}, \widetilde{\mathbf{\Sigma}}, \widetilde{\mathbf{V}} \leftarrow \texttt{SVD}(\widetilde{\mathbf{A}})$ { $\widetilde{\mathbf{\Sigma}} \in \mathbb{R}^{r\times r}$ }
**for each** $\mathbf{C} \in (\epsilon\text{-net of } \mathbb{B}_2^{r-1})^{\otimes k}$ **do**
  $\mathbf{L} \leftarrow \widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\mathbf{C}$ {$\mathbf{L} \in \mathbb{R}^{m\times k}$}
  $\mathbf{X} \leftarrow \mathrm{P}_{\mathcal{X}}(\mathbf{L})$
  $\mathbf{R} \leftarrow \mathbf{X}^\top \widetilde{\mathbf{A}}$ {$\mathbf{R} \in \mathbb{R}^{k\times n}$}
  $\mathbf{Y} \leftarrow \mathrm{P}_{\mathcal{Y}}(\mathbf{R})$
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{(\mathbf{X}, \mathbf{Y})\}$
**end for**
$(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}) \leftarrow \arg\max_{(\mathbf{X},\mathbf{Y})\in\mathcal{C}} \textsc{Tr}\left(\mathbf{X}^\top \widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^\top \mathbf{Y}\right)$

**Lemma 3.3.** *For any $\mathbf{A} \in \mathbb{R}^{m\times n}$, let*

$$(\mathbf{X}_\star, \mathbf{Y}_\star) \triangleq \arg\max_{\mathbf{X}\in\mathcal{X},\mathbf{Y}\in\mathcal{Y}} \textsc{Tr}\left(\mathbf{X}^\top \mathbf{A}\mathbf{Y}\right),$$

*where $\mathcal{X}$ and $\mathcal{Y}$ satisfy the conditions of Lemma 3.2. Let $\widetilde{\mathbf{A}}$ be a rank-$r$ approximation of $\mathbf{A}$, and $\widetilde{\mathbf{X}} \in \mathcal{X}$, $\widetilde{\mathbf{Y}} \in \mathcal{Y}$ be the output of Alg. 1 with input $\widetilde{\mathbf{A}}$ and accuracy $\epsilon$. Then,*

$$\textsc{Tr}\left(\mathbf{X}_\star^\top \mathbf{A}\mathbf{Y}_\star\right) - \textsc{Tr}\left(\widetilde{\mathbf{X}}^\top \mathbf{A}\widetilde{\mathbf{Y}}\right) \le 2 \cdot \left(\epsilon\sqrt{k} \cdot \|\widetilde{\mathbf{A}}\|_2 + \|\mathbf{A} - \widetilde{\mathbf{A}}\|_2\right) \cdot \mu_{\mathcal{X}} \cdot \mu_{\mathcal{Y}}.$$

Comments:

- Our approach has complexity **exponential** in the rank of $\widetilde{\mathbf{A}}$
- The quality of the output depends on the spectrum of $\mathbf{A}$ and the level of rank approximation r.

## [ Our k-BCC algorithm ]

**Algorithm 4** $k$-BCC/MaxAgree

**input** : Bi-adjacency matrix $\mathbf{B} \in \{\pm 1\}^{m,n}$,
Target number of $k$,
Accuracy $\delta \in (0, 1)$.
**output** Clustering $\widetilde{\mathcal{C}}^{(k)}$ of $U \cup V$ such that $\textsc{Agree}(\widetilde{\mathcal{C}}^{(k)}) \ge (1-\delta)\cdot\textsc{Agree}(\mathcal{C}_\star^{(k)})$.

1: Set up parameters:
$\epsilon \leftarrow 2^{-3}\cdot\delta\cdot k^{-1/2}$, $r \leftarrow 2^6\cdot\delta^{-2} - 1$.
2: Return output of Alg. 3 for input $(\mathbf{B}, k, r, \epsilon)$.

**Theorem 1.** ($k$-BCC.) *For any instance $(G = (U, V, E), k)$ of the k-BCC problem with bi-adjacency matrix $\mathbf{B}$ and for any desired accuracy parameter $\delta \in (0, 1)$, Algorithm 4 computes a clustering $\widetilde{\mathcal{C}}^{(k)}$ of $U \cup V$ into at most $k$ clusters, such that*

$$\textsc{Agree}(\widetilde{\mathcal{C}}^{(k)}) \ge (1-\delta)\cdot\textsc{Agree}(\mathcal{C}_\star^{(k)}),$$

*where $\mathcal{C}_\star^{(k)}$ is the optimal $k$-clustering, in time $2^{O(k/\delta^2\cdot\log\sqrt{k}/\delta)}\cdot(\delta^{-2}+k)\cdot(|U|+|V|) + \mathrm{T}_{\mathrm{SVD}}(\delta^{-2})$.*

## [ A (E)PTAS algorithm for BCC ]

**Algorithm 5** A PTAS for BCC/MaxAgree

**input** : Bi-adjacency matrix $\mathbf{B} \in \{\pm 1\}^{m,n}$,
Accuracy $\delta \in (0, 1)$.
**output** Clustering $\widetilde{\mathcal{C}}$ of $U \cup V$ (into at most $2^3\cdot\delta^{-1}$ clusters) such that $\textsc{Agree}(\widetilde{\mathcal{C}}) \ge (1-\delta)\cdot\textsc{Agree}(\mathcal{C}_\star)$.

1: Set up parameters:
$k \leftarrow 2^3\cdot\delta^{-1}$, $\epsilon \leftarrow 2^{-6}\cdot\delta^2$, $r \leftarrow 2^8\cdot\delta^{-2} - 1$.
2: Return output of Alg. 3 for input $(\mathbf{B}, k, r, \epsilon)$.

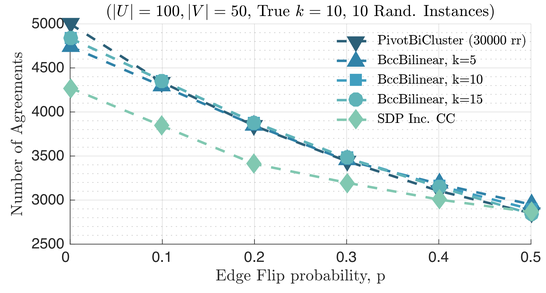
**Theorem 2.** (PTAS for BCC.) *For any instance $G = (U, V, E)$ of the BCC problem with bi-adjacency matrix $\mathbf{B}$ and for any desired accuracy parameter $\delta \in (0, 1)$, Algorithm 5 computes a clustering $\widetilde{\mathcal{C}}$ of $U \cup V$ into (at most) $2^3\cdot\delta^{-1}$ clusters, such that*

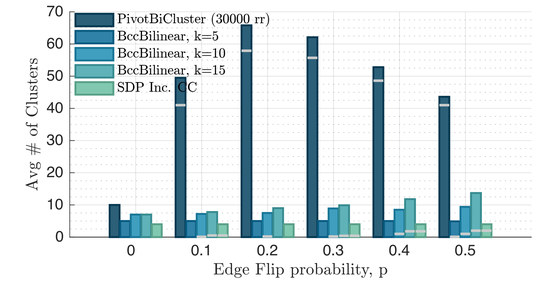
$$\textsc{Agree}(\widetilde{\mathcal{C}}) \ge (1-\delta)\cdot\textsc{Agree}(\mathcal{C}_\star),$$

## [ In practice… ]

Generate BCC instances as follows:

- Generate complete bipartite graphs with k=5 clusters
- Perturb label signs with probability p
- Generate 10 random instances and compute vertex clustering using PivotBiCluster (Ailon et al 2012), SDP approach (Swamy 2004) and ours.




| | $p = 0.0$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|---|---|---|---|---|---|---|
| PivotBiCluster [2] | 124 | 83 | 66 | 54 | 45 | 39 |
| BccBil. (k=5) | 69 | 70 | 70 | 71 | 71 | 71 |
| BccBil. (k=10) | 73 | 74 | 75 | 75 | 75 | 75 |
| BccBil. (k=15) | 78 | 79 | 79 | 80 | 80 | 80 |
| SDP Inc CC [17] | 198 | 329 | 257 | 258 | 266 | 314 |

Average runtimes/instance (in seconds).