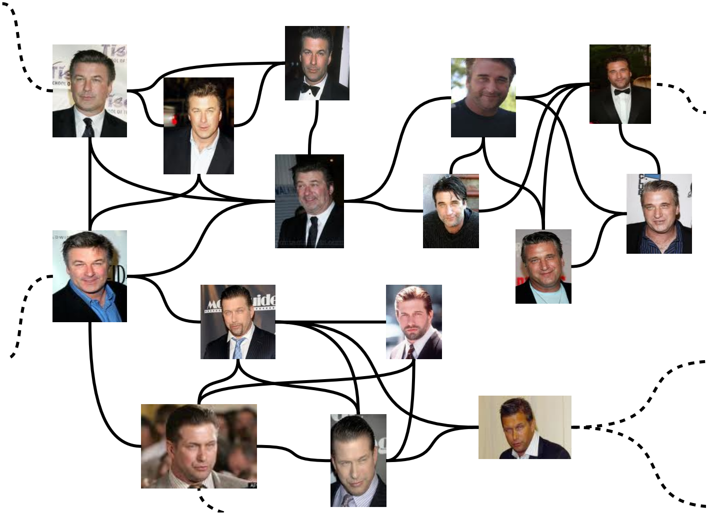
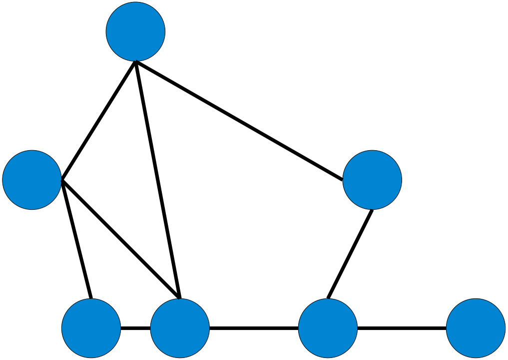
Improved Approximation Algorithms for Bipartite Correlation Clustering

Nir Ailon Noa Avigdor-Elgrabli Edo Liberty Anke van Zuylen

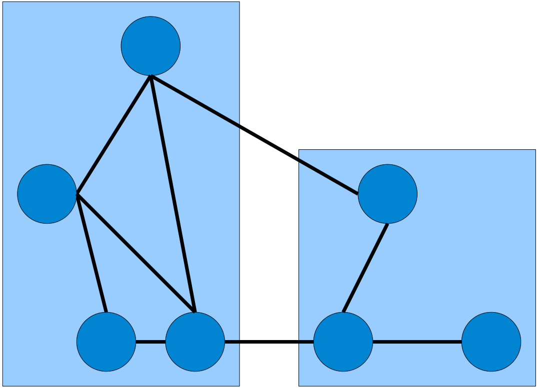
Correlation clustering



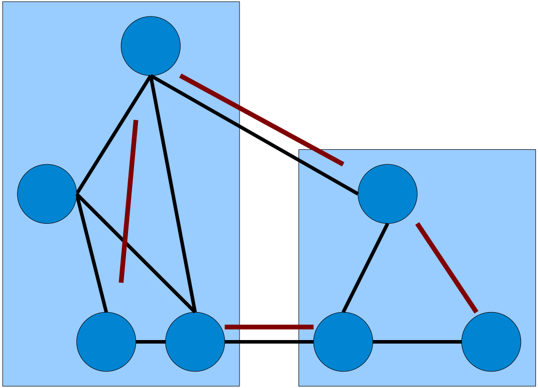
Input for correlation clustering



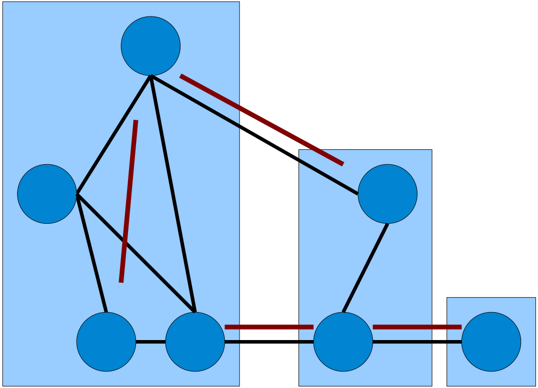
Output of correlation clustering



Cost of a correlation clustering solution



Cost of a correlation clustering solution

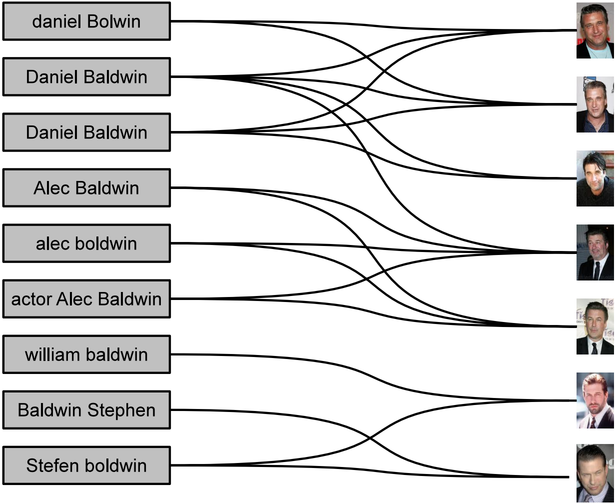


Correlation clustering results

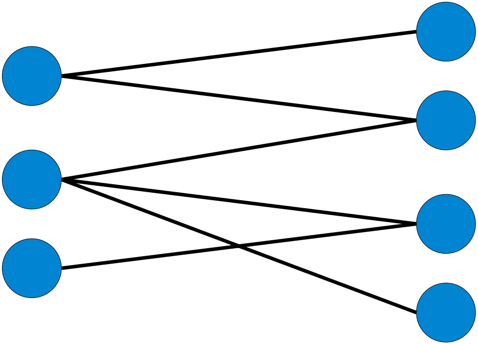
	approx const	running time
Bansal, Blum, Chawla	$\approx 20,000$	$\Omega(n^2)$
Demaine, Emanuel, Fiat, Immorlica	$4\log(n)$	LP
Charikar, Guruswami, Wirth	4	LP
Ailon, Charikar, Newman, Alantha	2.5	LP
Ailon, Charikar, Newman, Alantha	3	$O(m)$
Ailon, Liberty	< 3	$O(n) + cost(OPT)$

n and m are the number of nodes and edges in the graph.

Correlation bi-clustering

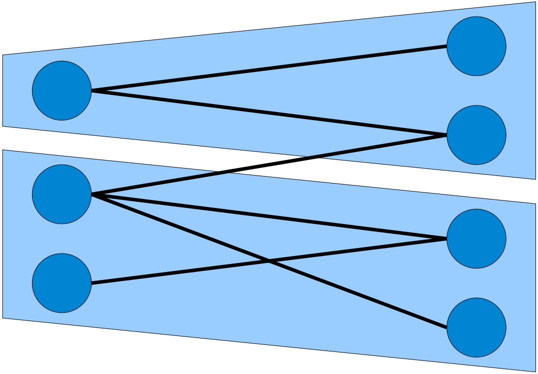


Input for correlation bi-clustering



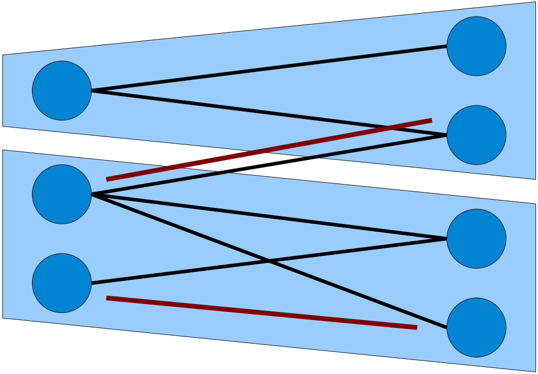
The input is an undirected unweighted bipartite graph.

Output of correlation bi-clustering



The output is a set of bi-clusters.

Cost of a correlation bi-clustering solution



The cost is the number of erroneous edges.

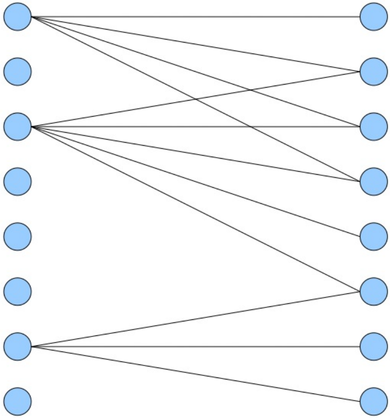
Correlation bi-clustering results

	approx const	running time
Demaine, Emanuel, Fiat, Immorlica	$O(\log(n))*$	LP
Charikar, Guruswami, Wirth	$O(\log(n))*$	LP
Noga Amit	12	LP
This work	4	LP (deterministic)
This work	4	$O(m)$ (randomized)

* The first two results hold for general weighted graph.

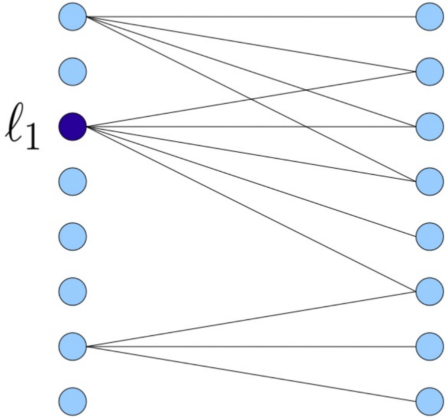
n and m are the number of nodes and edges in the graph.

PivotBiCluster



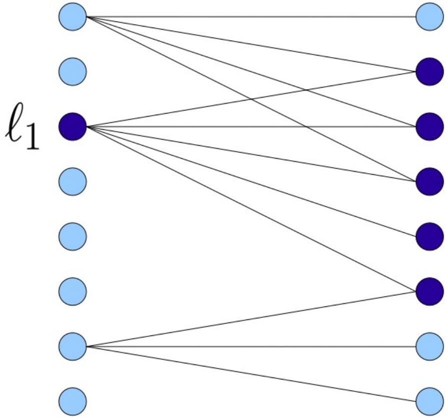
Consider the following graph

PivotBiCluster



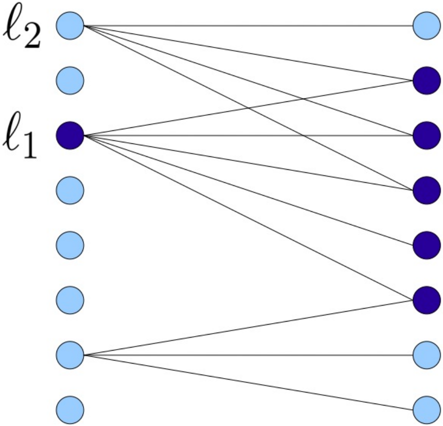
Choose ℓ_1 uniformly at random from the left side.

PivotBiCluster



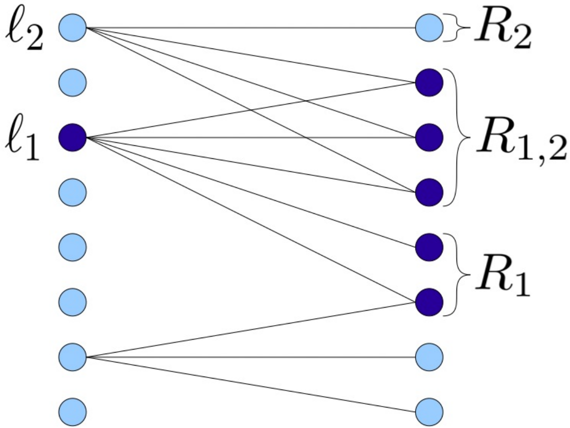
Add the neighborhood of ℓ_1 to the cluster

PivotBiCluster



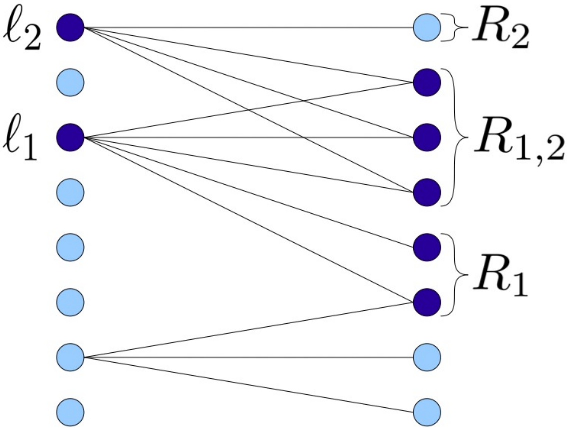
For each other node on the left (ℓ_2) do the following:

PivotBiCluster



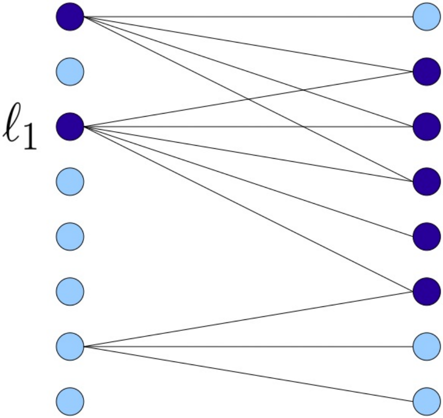
w.p. $\min(|R_{1,2}|/|R_2|, 1)$ add ℓ_2 to the cluster if $|R_{1,2}| \geq |R_1|$.

PivotBiCluster



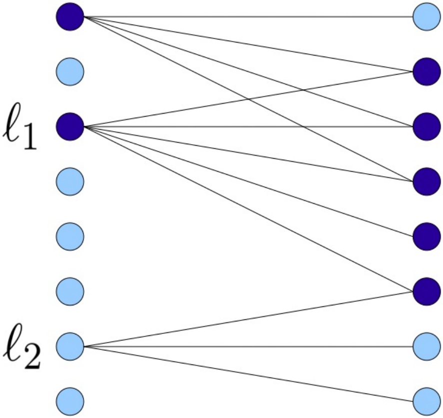
Here ℓ_2 joins the cluster because $R_{1,2} \geq R_1$.

PivotBiCluster



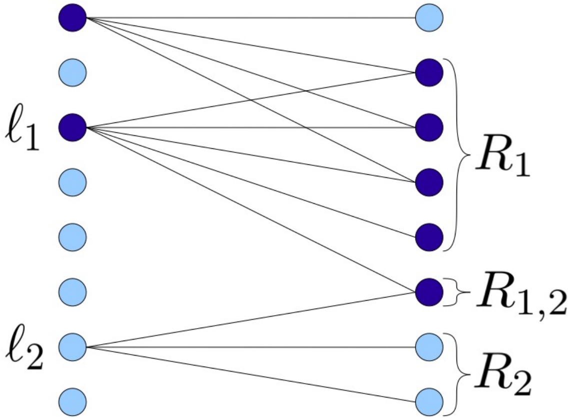
Let's consider another example

PivotBiCluster



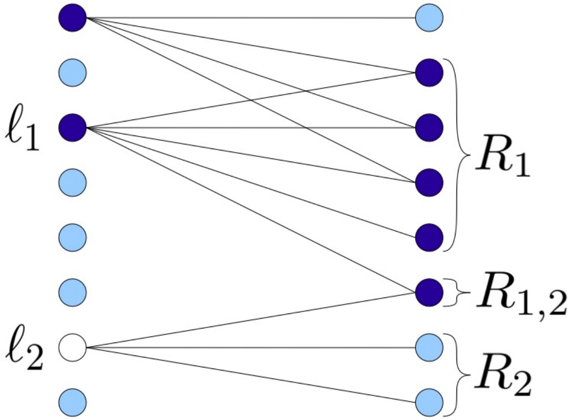
Let's consider another example

PivotBiCluster



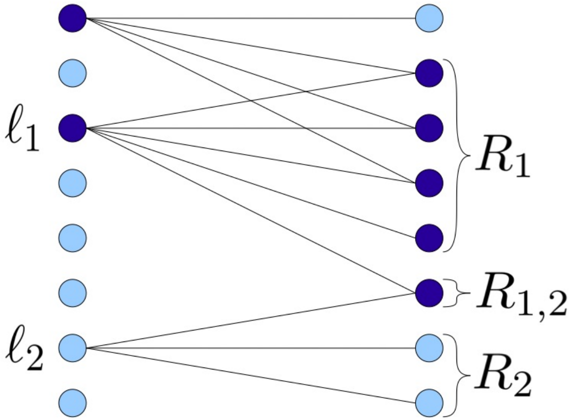
Since $|R_{1,2}|/|R_2| = 1/2$ with probability $1/2$ we decide what to do with ℓ_2

PivotBiCluster



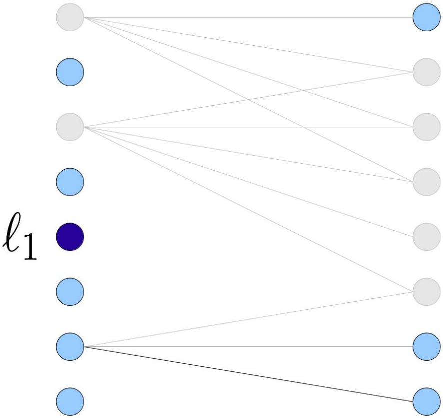
Since $|R_{1,2}| < |R_1|$ that decision should be to make ℓ_2 a singleton

PivotBiCluster



Otherwise (w.p. $1/2$) we decide nothing about ℓ_2 and continue.

PivotBiCluster



We remove the clustered nodes from the graph and repeat.

PivotBiCluster

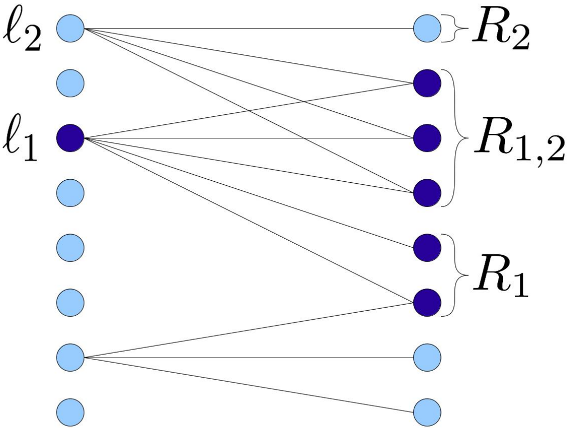
Lemma

Let OPT denote the best possible bi-clustring of G.
Let B be a random output of PivotBiCluster. Then:

$$E_{B \sim PivotBiCluster}\left[cost(B)\right] \leq 4cost(OPT)$$

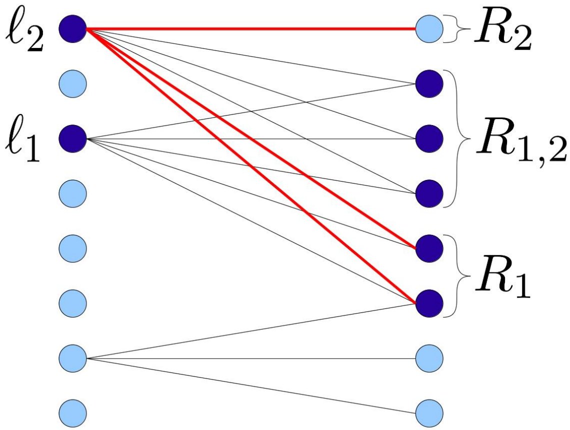
Let's see how to prove this...

Tuples, bad events, and violated pairs



A "bad event" (X_T) happens to tuple $T = (\ell_1, \ell_2, R_1, R_{1,2}, R_2)$.

Tuples, bad events, and violated pairs



We "blame" bad event X_T for the violated (red) pairs, $\mathbb{E}[cost(T)|X_T] = 3$.

Tuples, bad events, and violated pairs

Since every violated pair can be blamed on (or colored by) one bad event happening we have:

$$\mathbb{E}_{B \sim PivotBiCluster}\left[cost(B)\right] \leq \sum_{T} q_T \cdot \mathbb{E}[cost(T)|X_T]$$

where q_T denotes the probability that a bad event happened to tuple T.

Note: the number of tuples is exponential in the size of the graph.

Proof sketch

1. We have (previous slide)

$$ALG \leq \sum_T q_T \cdot \mathbb{E}[cost(T)|X_T]$$

2. Write the dual linear program

$$OPT \geq \sum_T \beta(T) \text{ s.t. constrains on } \beta(T)$$

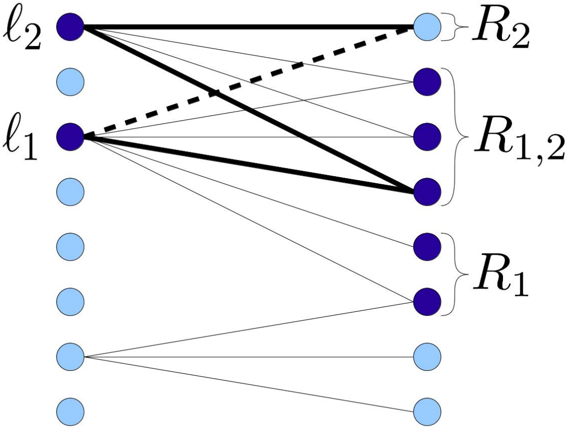
3. Set a feasible solution $\beta(T) \leftarrow q_T f(T)$.
4. Show that:

$$\mathbb{E}[cost(T)|X_T] + E[cost(\overline{T})|X_{\overline{T}}] \leq 4(f(T) + f(\overline{T}))$$

5. Which gives

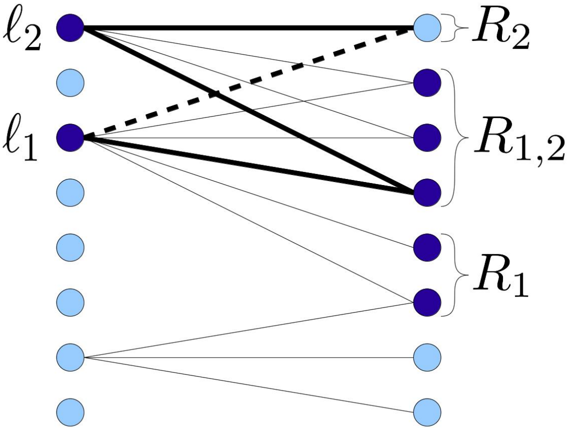
$$ALG \leq \sum_T q_T \cdot \mathbb{E}[cost(T)|X_T] \leq 4 \sum_T q_T f(T) \leq 4 \cdot OPT$$

The linear program



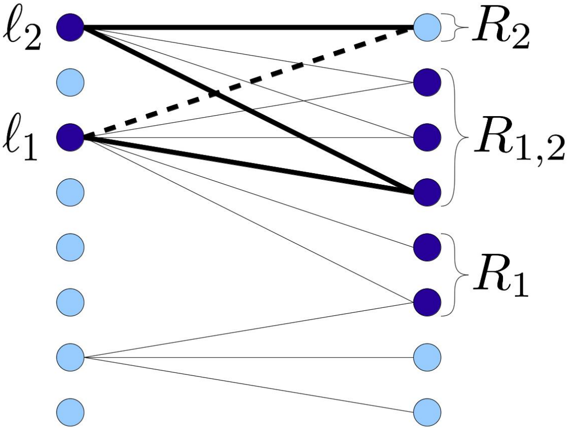
In a bad square, any clustering must err at least once.

The linear program



Let $x_{\ell,r}$ be equal 1 if the clustering errs on pair (ℓ, r) and 0 otherwise.

The linear program



For $r_2 \in R_2$ and $r_{1,2} \in R_{1,2}$ we have $x_{\ell_1,r_2} + x_{\ell_1,r_{1,2}} + x_{\ell_2,r_2} + x_{\ell_2,r_{1,2}} \geq 1$

The linear program

Since each tuple corresponds to $|R_2^T| \cdot |R_{1,2}^T|$ bad squares, we get the following constraint:

$$\forall\ T : \sum_{r_2 \in R_2^T, r_{1,2} \in R_{1,2}^T} \left(x_{\ell_1^T, r_2} + x_{\ell_1^T, r_{1,2}} + x_{\ell_2^T, r_2} + x_{\ell_2^T, r_{1,2}} \right)$$

$$= \sum_{r_2 \in R_2^T} |R_{1,2}^T| \cdot (x_{\ell_1^T, r_2} + x_{\ell_2^T, r_2}) + \sum_{r_{1,2} \in R_{1,2}^T} |R_2^T| \cdot (x_{\ell_1^T, r_{1,2}} + x_{\ell_2^T, r_{1,2}})$$

$$\geq |R_2^T| \cdot |R_{1,2}^T|$$

Minimizing the cost corresponds to a minimization over $\sum x_{\ell,r}$ and subject to $x_{\ell,r} \geq 0$.

The linear program

The dual of which is as follows:

$$max \sum_{T} \beta(T)$$

s.t. $\forall (\ell, r) \in E$:

$$\sum_{T:\, \ell_2^T = \ell, r \in R_2^T} \frac{\beta(T)}{|R_2^T|} + \sum_{T:\, \ell_1^T = \ell,\, r \in R_{1,2}^T} \frac{\beta(T)}{|R_{1,2}^T|} + \sum_{T:\, \ell_2^T = \ell,\, r \in R_{1,2}^T} \frac{\beta(T)}{|R_{1,2}^T|} \leq 1$$

and $\forall (\ell, r) \notin E$: $\displaystyle\sum_{T:\, \ell_1^T = \ell,\, r \in R_2^T} \frac{1}{|R_2^T|} \beta(T) \leq 1$

Feasibilty

Lemma

The solution

$$\beta(T) = q_T \cdot f(T)$$

is a feasible solution to the dual of the linear program when:

$$f(T) = \min\{|R^T_{1,2}|, |R^T_2|\} \min\left\{1, \frac{|R^T_{1,2}|}{\min\{|R^T_{1,2}|, |R^T_1|\} + \min\{|R^T_{1,2}|, |R^T_2|\}}\right\}$$

Lemma

For any tuple T,

$$\mathbb{E}[\mathrm{cost}(T)|X_T] + \mathbb{E}[\mathrm{cost}(\bar{T})|X_{\bar{T}}] \leq 4 \cdot \left(f(T) + f(\bar{T})\right).$$

We will not show these here, they are very technical...

Conclusion and discussion

	PivotBiCluster	LP based
Running time	$O(m)$	Solves LP on n^3 constraints
Takes advantage of bipartiteness	Yes	no
Approximation factor	4	4
Symmetric	**No**	Yes
Deterministic	**No**	Yes

- Can a combinatorial algorithm bit the 4 approximation factor?
- Maybe it should take advantage of the symmetry?
- Can this algorithm be derandomized?

Things I'll be happy discuss

(Some of which I don't know the answers for...)

- The rest of the proof
- Why we believe the analysis must use tuples and bad squares are not enough
- The running time of the algorithm (easily seen to be $O(m)$)
- An LP based 4-approximation deterministic algorithm
- Is there a tight bad example for which the algorithm achieves the approximation bound
- Is there a combinatorial **and** deterministic algorithm with this bound (or better)
- Is there an approximation hardness result? (for what factor)

Thank you