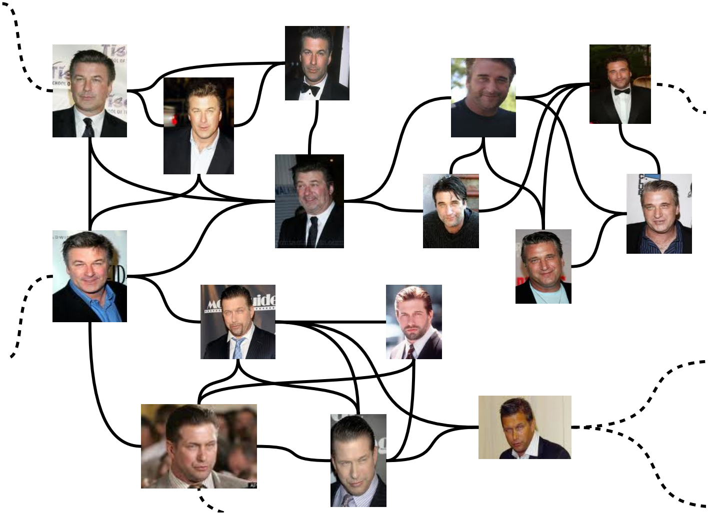
Scaleable correlation clustering algorithms

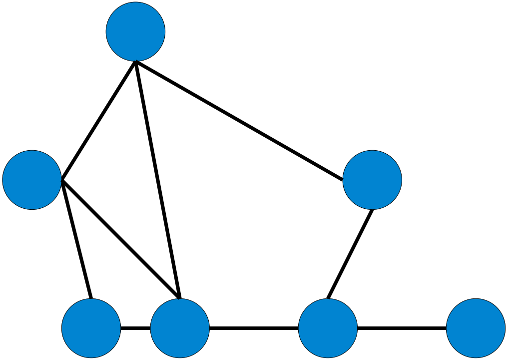
Edo Liberty

Joint work (in progress) with Nir Ailon.

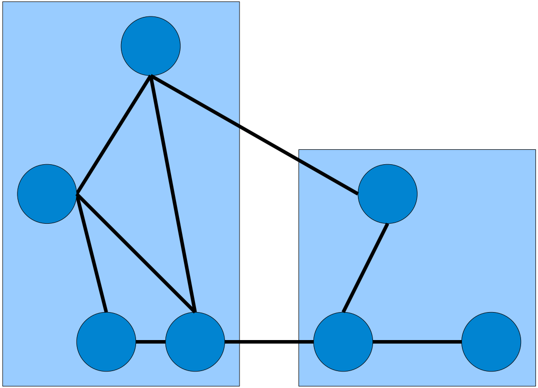
Correlation clustering



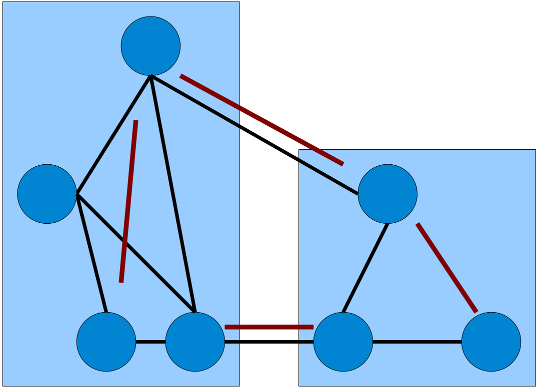
Input for correlation clustering



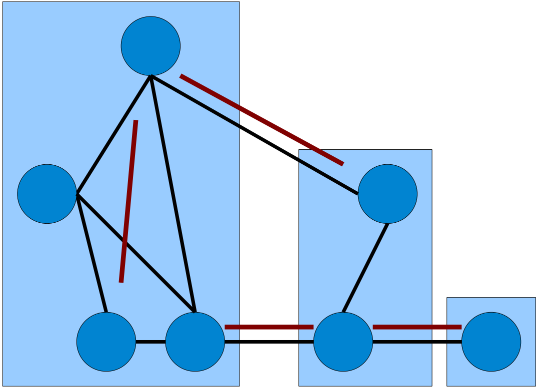
Output of correlation clustering



Cost of a correlation clustering solution



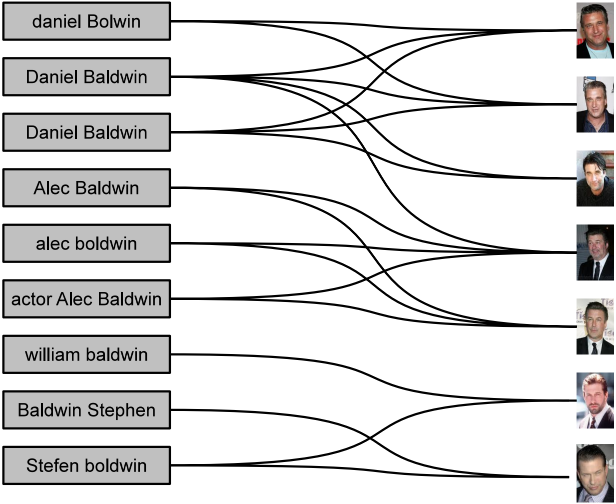
Cost of a correlation clustering solution



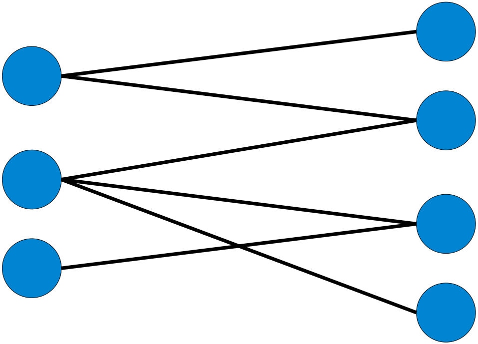
Correlation clustering results

	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	$O(n) + cost(OPT)$
	3	$log(n)$ message passing rounds*

Correlation bi-clustering

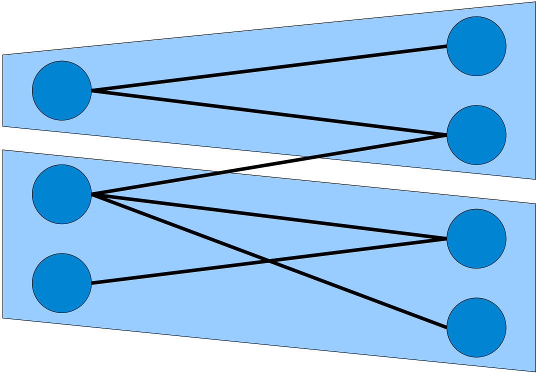


Input for correlation bi-clustering



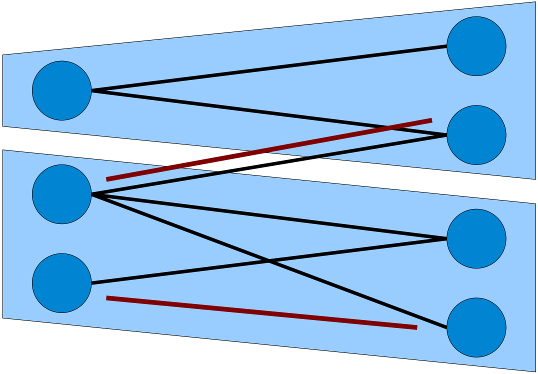
The input is an undirected unweighted bipartite graph.

Output of correlation bi-clustering



The output is a set of bi-clusters.

Cost of a correlation bi-clustering solution

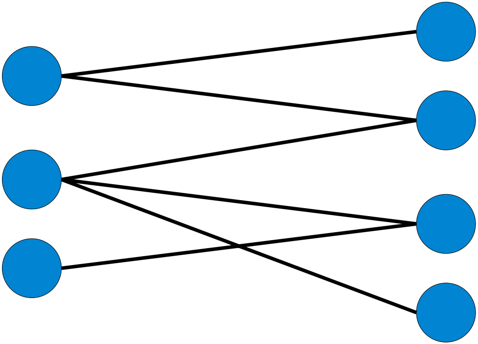


The cost is the number of erroneous edges.

Correlation bi-clustering results

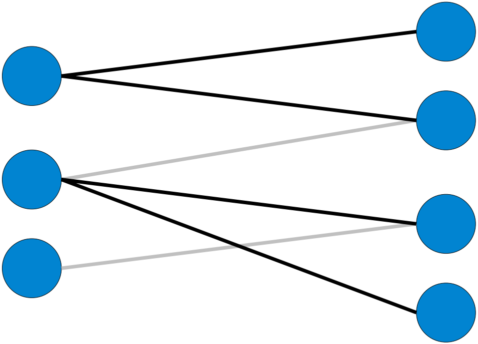
	approx const	running time
Demaine, Emanuel, Fiat, Immorlica	$O(\log(n))*$	LP
Charikar, Guruswami, Wirth	$O(\log(n))*$	LP
Guo, Huffner, Komusiewicz, Zhang	4	sequential $O(m)$
	4	3 message passing rounds $O(n + cost(OPT))$ communication

Quick-bi-cluster



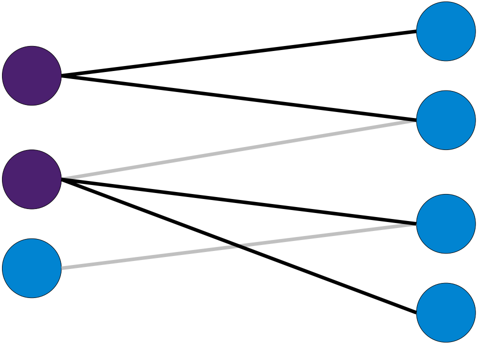
Permute both left and right side randomly.

Quick-bi-cluster



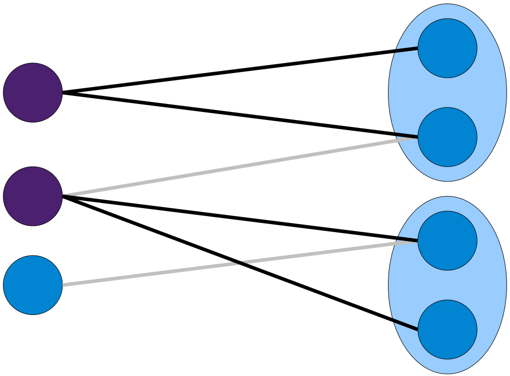
For every node on the right keep only the top edge.

Quick-bi-cluster



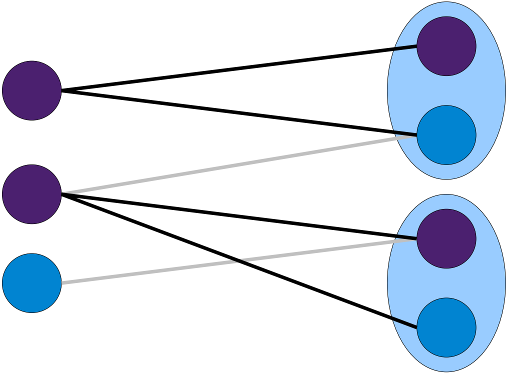
Mark connected nodes on the left as "left-centers".

Quick-bi-cluster



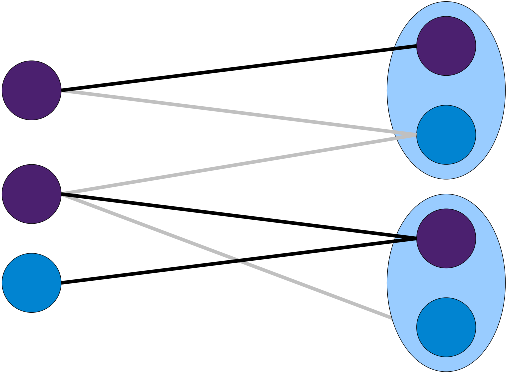
Cluster all nodes on the right by their left-center.

Quick-bi-cluster



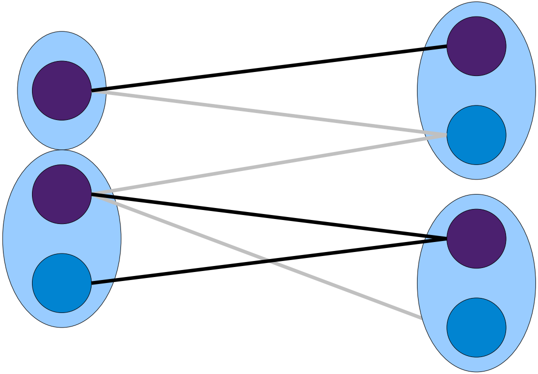
Mark top nodes in their right clusters and "right-centers".

Quick-bi-cluster



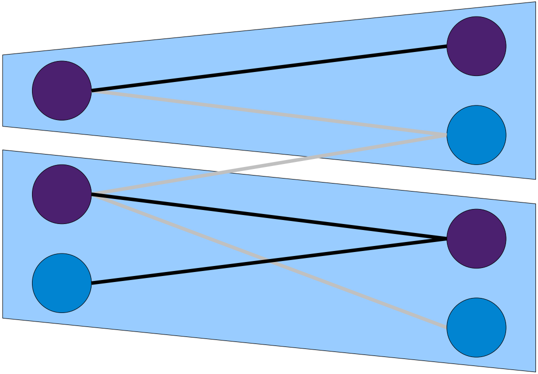
For every node on the left keep only the top edge to a right-center.

Quick-bi-cluster



Cluster nodes on the left by their right-center.

Quick-bi-cluster



Bi-cluster right-clusters and left-clusters by centers connections.

Quick-bi-cluster

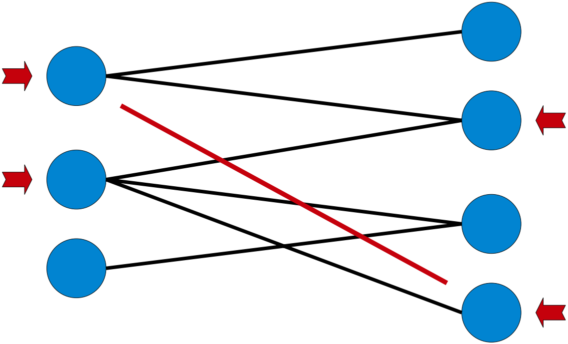
Lemma

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

$$E\left[cost(B)\right] \leq 4cost(OPT)$$

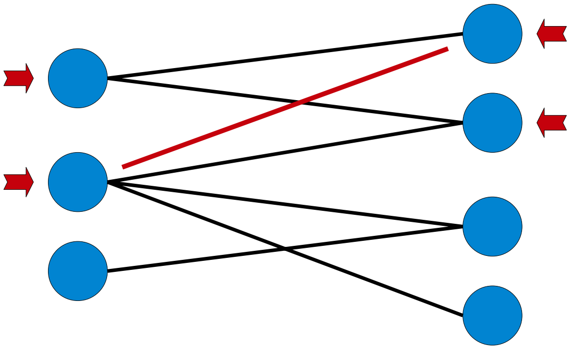
Let's prove this...

Bad squares, bad events, and erroneous edges



A bad square, is a set of four nodes (two on each side) between which there are exactly three edges.

Bad squares, bad events, and erroneous edges



A bad square, is a set of four nodes (two on each side) between which there are exactly three edges.

Bounding $cost(OPT)$

OPT must make at least one mistake in every bad square.

$$
\begin{aligned}
cost(OPT) \quad &\geq \quad \min \sum_e x_e \\
&\text{s.t.} \quad \forall \ s \ \sum_{e \subset s} x_e \geq 1
\end{aligned}
$$

$x_e \in [0, 1]$ indicates whether edge e is erroneous.

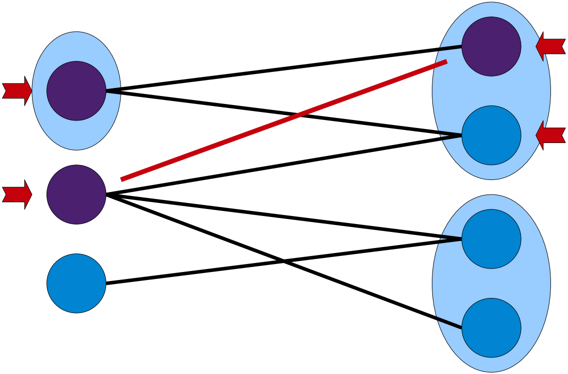
Bounding $cost(OPT)$

Defining dual of the problem and using that $PRIMAL \geq DUAL$ we have:

$$\begin{aligned} cost(OPT) \quad &\geq \quad max \sum_s \beta_s \\ \text{s.t.} \quad &\forall\, e \;\; \sum_{s \supset e} \beta_s \leq 1 \end{aligned}$$

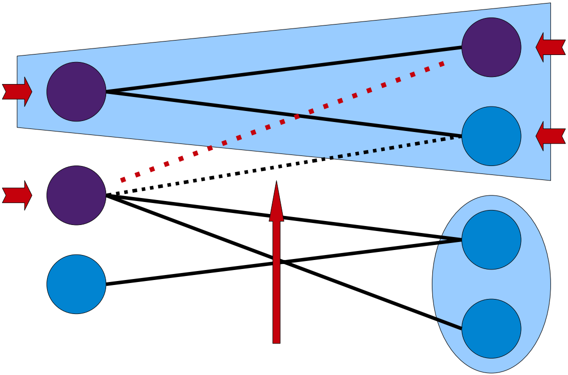
β_s is a fractional "blame" factor for each square s.

Bad squares, bad events, and erroneous edges



A bad event, happens to a bad square when two nodes that belong to it are chosen, one on each side.

Bad squares, bad events, and erroneous edges



This is a bad event because it generates an erroneous edge.

Bad squares, bad events, and erroneous edges

Property 1

There is a one to one mapping between bad events occurring and errors in the output clustering.

$$\mathbb{E}(cost(B)) = \mathbb{E}\sum_{e} z_e = \mathbb{E}\sum_{s} z_s = \sum_{s} p_s$$

- z_e denote the indicator variable that edge e is erroneous
- z_s denote the event that a bad event happened to bad square s.
- p_s is the probability that a bad event happens to bad square s.

Bad squares, bad events, and erroneous edges

Property 2

The probability of each edge in a bad event being erroneous is $1/4$.

We have that $p_e \leq 1$ for every edge, and also:

$$p_e = \sum_{s \supset e} \Pr(z_e | z_s = 1) \cdot p_s = \sum_{s \supset e} \frac{1}{4} p_s$$

$$\forall \; e \quad \sum_{s \supset e} \frac{1}{4} p_s \leq 1$$

Putting it all together

$$\mathbb{E}(cost(B)) = \sum_s p_s$$

$$\forall \ e \quad \sum_{s \supset e} \frac{1}{4} p_s \leq 1$$

$$cost(OPT) \geq max \sum_s \beta_s$$

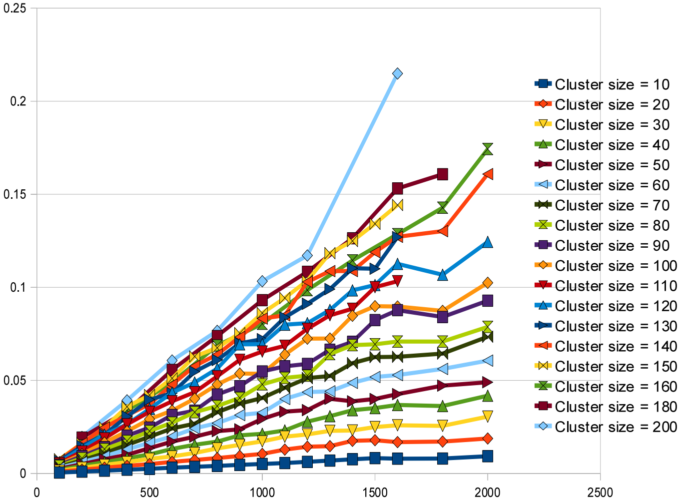
$$\text{s.t.} \quad \forall\, e \ \sum_{s \supset e} \beta_s \leq 1$$

We finally get...

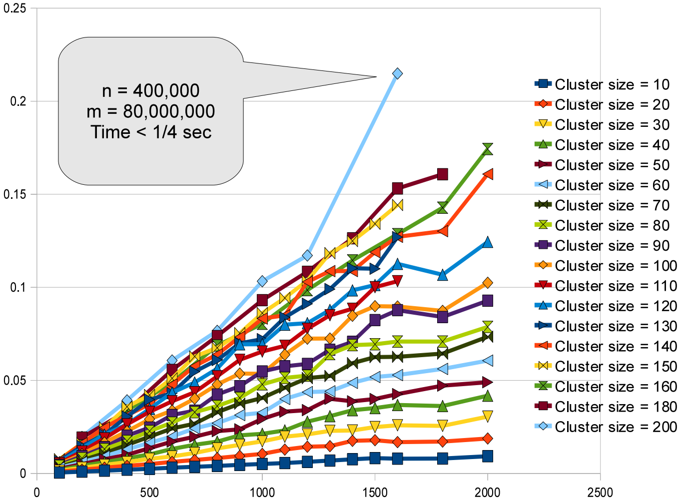
The values $\beta_s = \frac{1}{4} p_s$ give a feasible solution to DUAL and so

$$E\left[cost(B)\right] \leq 4cost(OPT)$$

Some experiments....



Some experiments....



Fin