Improved Clustering Algorithms for the Random Cluster Graph Model

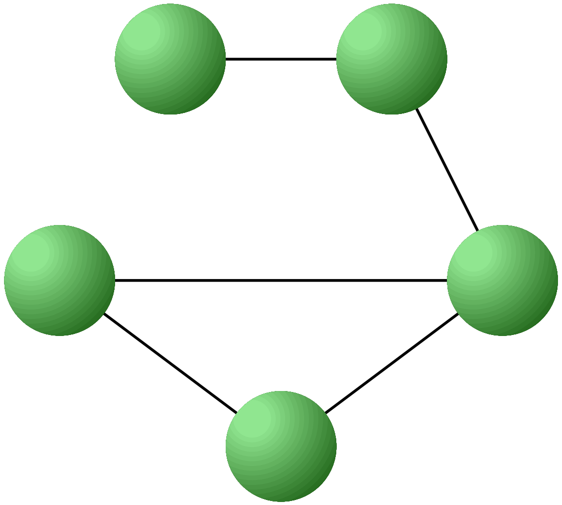
Ron Shamir   Dekel Tsur

Tel Aviv University

The Clustering Problem

Input: A graph G. (edges in G represent similarity between the vertices)

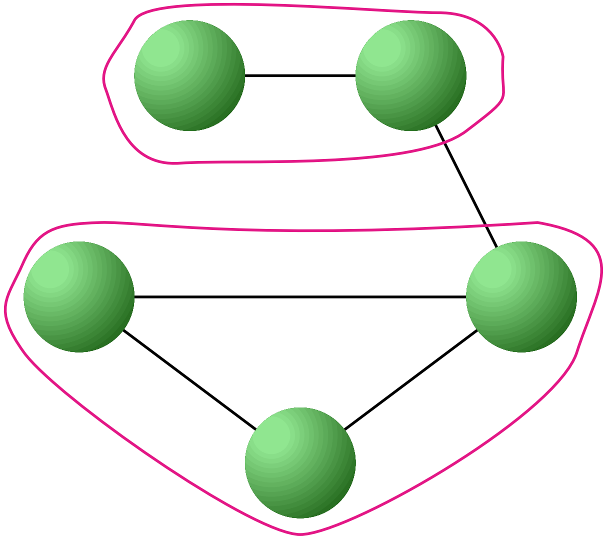
Output: A partition of the vertices of V into sets such that there are many edges between vertices from the same set, and few edges between vertices from different sets.



The Clustering Problem

Input: A graph G. (edges in G represent similarity between the vertices)

Output: A partition of the vertices of V into sets such that there are many edges between vertices from the same set, and few edges between vertices from different sets.



The Random Cluster Graph Model

A graph $G = (V, E)$ which is built by the following process:

1. V is partitioned into disjoint sets $V_1, \ldots, V_m$ (clusters).
2. Mates (= vertices from the same set) are connected by an edge with probability p.
3. Non-mates are connected by an edge with probability $r < p$.

The edges are independent.

The Clustering Problem

Input: A cluster graph G.

Output: The clusters $V_1, \ldots, V_m$.

$$n = |V|$$
$$k = \min_i |V_i|$$
$$\Delta = p - r$$

Previous Results

General case

Paper	Requirements		Complexity
	k	Δ	
Ben-Dor et al 99	$\Omega(n)$	$\Omega(1)$	

Equal sized clusters

	m	Δ
Dyer and Frieze 86	2	$\Omega(n^{-1/4}\log^{1/4} n)$
Boppana 87	2	$\Omega(n^{-1/2}\sqrt{\log n})$
Jerrum and Sorkin 93	2	$\Omega(n^{-1/6+\varepsilon})$
Condon and Karp 99	$O(1)$	$\Omega(n^{-1/2+\varepsilon})$

$$n = |V| \qquad k = \min_i |V_i| \qquad \Delta = p - r$$

Previous Results

General case

Paper	Requirements		Complexity
	k	Δ	
Ben-Dor et al 99	$\Omega(n)$	$\Omega(1)$	
This paper	$\Omega(\Delta^{-1}\sqrt{n}\max(\log n, \Delta^{-\varepsilon}))$		

Equal sized clusters

	m	Δ	
Dyer and Frieze 86	2	$\Omega(n^{-1/4}\log^{1/4} n)$	
Boppana 87	2	$\Omega(n^{-1/2}\sqrt{\log n})$	
Jerrum and Sorkin 93	2	$\Omega(n^{-1/6+\varepsilon})$	
Condon and Karp 99	$O(1)$	$\Omega(n^{-1/2+\varepsilon})$	
This paper		$\Omega(mn^{-1/2}\sqrt{\log n})$	

$$n = |V| \qquad k = \min_i |V_i| \qquad \Delta = p - r$$

Previous Results

General case

Paper	Requirements		Complexity
	k	Δ	
Ben-Dor et al 99	$\Omega(n)$	$\Omega(1)$	$n^2 \log^{O(1)} n$
This paper	$\Omega(\Delta^{-1}\sqrt{n}\max(\log n, \Delta^{-\varepsilon}))$		$O(mn^2/\log n)$

Equal sized clusters

	m	Δ	
Dyer and Frieze 86	2	$\Omega(n^{-1/4}\log^{1/4} n)$	$O(n^2)$
Boppana 87	2	$\Omega(n^{-1/2}\sqrt{\log n})$	$n^{O(1)}$
Jerrum and Sorkin 93	2	$\Omega(n^{-1/6+\varepsilon})$	$O(n^4)$
Condon and Karp 99	$O(1)$	$\Omega(n^{-1/2+\varepsilon})$	$O(n^2)$
This paper		$\Omega(mn^{-1/2}\sqrt{\log n})$	$O(mn^2 \log n)$

$$n = |V| \qquad k = \min_i |V_i| \qquad \Delta = p - r$$

Previous Results

General case

Paper	Requirements		Complexity
	k	Δ	
Ben-Dor et al 99	$\Omega(n)$	$\Omega(1)$	$n^2 \log^{O(1)} n$
This paper	$\Omega(\Delta^{-1}\sqrt{n}\max(\log n, \Delta^{-\varepsilon}))$		$O(n \log n)$

Equal sized clusters

	m	Δ
Dyer and Frieze 86	2	$\Omega(n^{-1/4}\log^{1/4} n)$
Boppana 87	2	$\Omega(n^{-1/2}\sqrt{\log n})$
Jerrum and Sorkin 93	2	$\Omega(n^{-1/6+\varepsilon})$
Condon and Karp 99	$O(1)$	$\Omega(n^{-1/2+\varepsilon})$
This paper		$\Omega(mn^{-1/2}\sqrt{\log n})$

$$n = |V| \qquad k = \min_i |V_i| \qquad \Delta = p - r$$

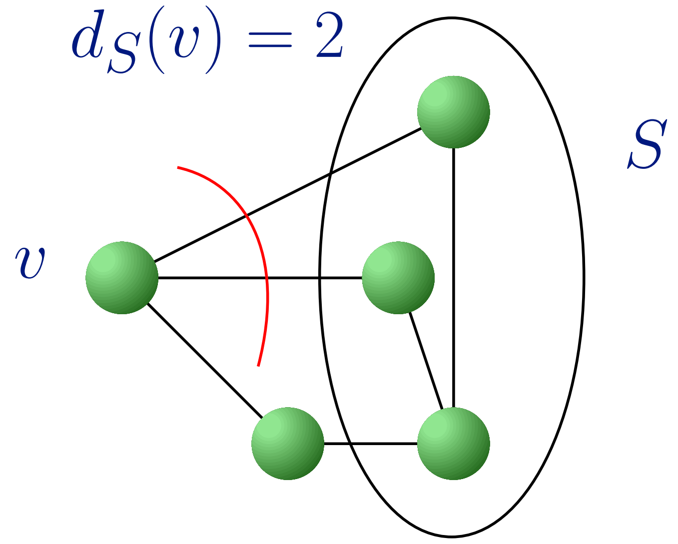
More Notation

For a graph $G = (V, E)$,

$$\text{w.h.p.} = \text{With probability } 1 - n^{-\Omega(1)}$$

$$N(v) = \text{The neighbors of } v$$

$$d_S(v) = |N(v) \cap S|$$



Top Level Description

A set $S \subseteq V$ is called a subcluster if $S \subseteq V_i$ for some cluster V_i.

Our algorithm:

While G is not empty:

Find seed:	Find a subcluster S of size $\Theta(\log n/\Delta^2)$.
Expand:	Find the whole cluster V_i which contains S, and remove it from G.

Expanding a subcluster S

Suppose that $S \subseteq V_i$ and $|S| = \Theta(\log n/\Delta^2)$.
Consider $d_S(v)$ for $v \in V - S$:

$$\mathrm{E}[d_S(v)] = \begin{cases} |S|p & \text{if } v \in V_i \\ |S|r & \text{otherwise} \end{cases}$$

Expanding a subcluster S

Suppose that $S \subseteq V_i$ and $|S| = \Theta(\log n/\Delta^2)$.

Consider $d_S(v)$ for $v \in V - S$:

$$\mathrm{E}[d_S(v)] = \begin{cases} |S|p & \text{if } v \in V_i \\ |S|r & \text{otherwise} \end{cases}$$

Using Chernoff-like bound, w.h.p.

$|d_S(v) - \mathrm{E}[d_S(v)]| < \frac{1}{2}D$, where $D = \Theta(\sqrt{|S| \log n})$

Expanding a subcluster S

Suppose that $S \subseteq V_i$ and $|S| = \Theta(\log n/\Delta^2)$.
Consider $d_S(v)$ for $v \in V - S$:

$$\mathrm{E}[d_S(v)] = \begin{cases} |S|p & \text{if } v \in V_i \\ |S|r & \text{otherwise} \end{cases}$$

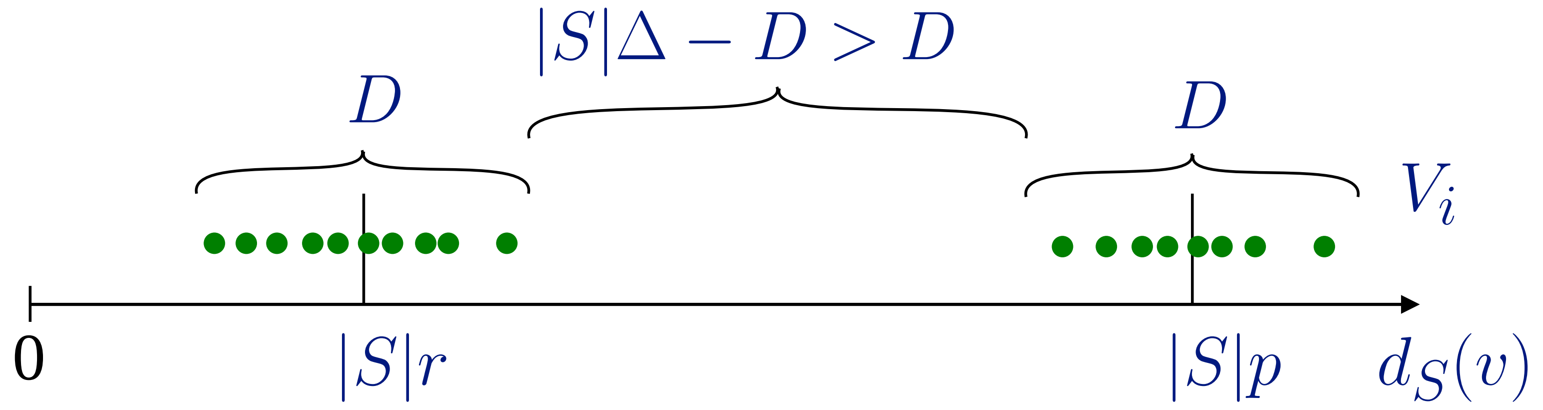
Using Chernoff-like bound, w.h.p.
$|d_S(v) - \mathrm{E}[d_S(v)]| < \frac{1}{2}D$, where $D = \Theta(\sqrt{|S| \log n})$

Expanding a subcluster S

Suppose that $S \subseteq V_i$ and $|S| = \Theta(\log n/\Delta^2)$.
Consider $d_S(v)$ for $v \in V - S$:

$$\mathrm{E}[d_S(v)] = \begin{cases} |S|p & \text{if } v \in V_i \\ |S|r & \text{otherwise} \end{cases}$$

Using Chernoff-like bound, w.h.p.
$|d_S(v) - \mathrm{E}[d_S(v)]| < \frac{1}{2}D$, where $D = \Theta(\sqrt{|S| \log n})$



Expanding a subcluster S

1. Order $V - S = \{v_1, \ldots, v_{n-|S|}\}$ such that $d_S(v_1) \geq d_S(v_2) \geq \cdots \geq d_S(v_{n-|S|})$.
2. Let $D = \Theta(\sqrt{|S| \log n})$.
3. If $\max_j \{d_S(v_j) - d_S(v_{j+1})\} < D$, then return V.
4. Otherwise, let j be the first index for which $d_S(v_j) - d_S(v_{j+1}) \geq D$.
   Return $S \cup \{u_1, \ldots, u_j\}$.

Finding a Subcluster — Imbalance

For two disjoint sets L, R of vertices of equal size, the L, R-imbalance of V_i (Jerrum and Sorkin 93) is

$$\mathrm{I}(V_i, L, R) = \frac{|V_i \cap L| - |V_i \cap R|}{|L|}.$$

The imbalance of L, R is

$$\max\{\mathrm{I}(V_1, L, R), \ldots, \mathrm{I}(V_m, L, R)\}.$$

The secondary imbalance of L, R is the second largest value.

Finding a Subcluster

1. Find L, R with large imbalance and small secondary imbalance.
2. Let $f(v) = d_L(v) - d_R(v)$, $D = \Theta(\sqrt{|L| \log n})$.
3. Randomly choose $\Theta(\frac{m^2 \log n}{\Delta^2})$ vertices from $V - (L \cup R)$ into a set S.
4. Order $S = \{v_1, \ldots, v_s\}$ such that $f(v_1) \geq \cdots \geq f(v_s)$.
5. If $\max_j\{f(v_j) - f(v_{j+1})\} < D$, then return. ($L, R$ are "bad")
6. Let j be the first index for which $f(v_j) - f(v_{j+1}) \geq D$. Return $\{v_1, \ldots, v_j\}$.

Correctness of the Algorithm

Denote $b_i = \mathrm{I}(V_i, L, R)$ and $l = |L|$.
Suppose that $b_1 \geq b_2 \geq \cdots \geq b_m$.

Lemma If $b_1 \geq \Omega(\frac{\sqrt{\log n}}{\Delta\sqrt{l}})$ and $b_2 \leq \frac{1}{2}b_1$ then w.h.p. the alg. returns a subcluster.

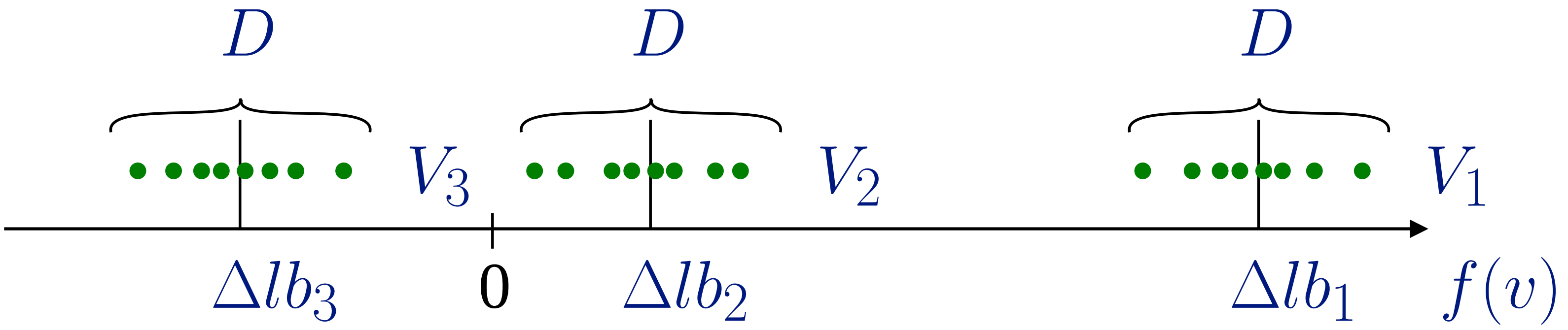
Proof For $v \in V_i$, $\mathrm{E}[f(v)] = \Delta l b_i$.

Correctness of the Algorithm

Denote $b_i = \mathrm{I}(V_i, L, R)$ and $l = |L|$.
Suppose that $b_1 \geq b_2 \geq \cdots \geq b_m$.

Lemma If $b_1 \geq \Omega(\frac{\sqrt{\log n}}{\Delta\sqrt{l}})$ and $b_2 \leq \frac{1}{2}b_1$ then w.h.p. the alg. returns a subcluster.

Proof For $v \in V_i$, $\mathrm{E}[f(v)] = \Delta l b_i$.
$|f(v) - \mathrm{E}[f(v)]| < \frac{1}{2}D$

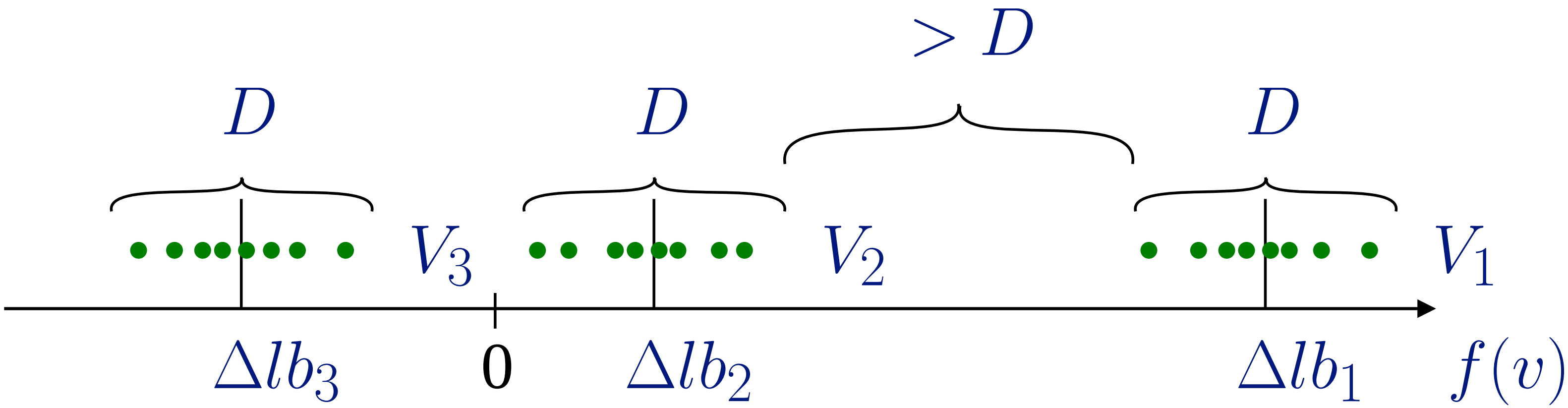


Correctness of the Algorithm

Denote $b_i = \mathrm{I}(V_i, L, R)$ and $l = |L|$.
Suppose that $b_1 \geq b_2 \geq \cdots \geq b_m$.

Lemma If $b_1 \geq \Omega(\frac{\sqrt{\log n}}{\Delta\sqrt{l}})$ and $b_2 \leq \frac{1}{2}b_1$ then w.h.p. the alg. returns a subcluster.

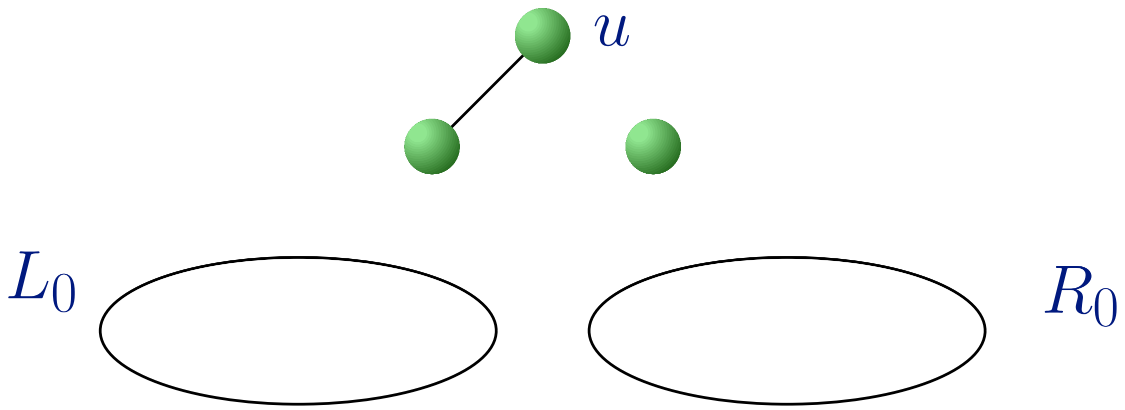
Proof For $v \in V_i$, $\mathrm{E}[f(v)] = \Delta l b_i$.
$|f(v) - \mathrm{E}[f(v)]| < \frac{1}{2}D$



$\square$

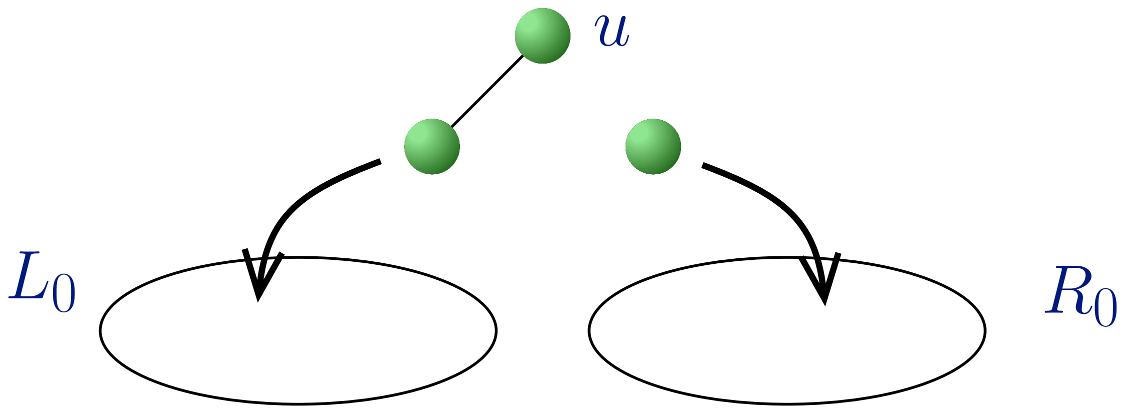
Finding the Sets L, R — Initialization

1. $L_0, R_0 \leftarrow \phi$. Let $l = \Theta(\frac{m^2}{\Delta^2})$.
2. Randomly select a vertex u and l pairs of vertices.
3. For each pair of vertices, if only one vertex is a neighbor of u, place that vertex in L_0 and the other vertex in R_0.



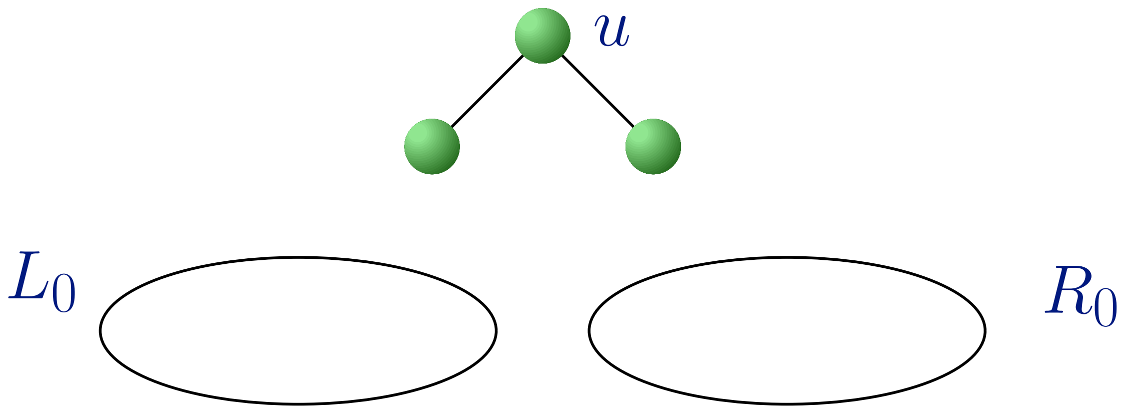
Finding the Sets L, R — Initialization

1. $L_0, R_0 \leftarrow \phi$. Let $l = \Theta(\frac{m^2}{\Delta^2})$.
2. Randomly select a vertex u and l pairs of vertices.
3. For each pair of vertices, if only one vertex is a neighbor of u, place that vertex in L_0 and the other vertex in R_0.



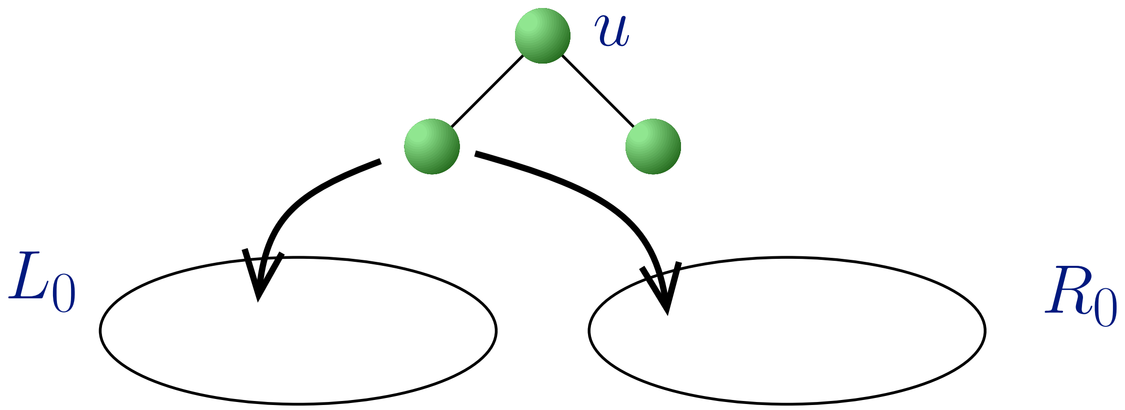
Finding the Sets L, R — Initialization

1. $L_0, R_0 \leftarrow \phi$. Let $l = \Theta(\frac{m^2}{\Delta^2})$.
2. Randomly select a vertex u and l pairs of vertices.
3. For each pair of vertices, if only one vertex is a neighbor of u, place that vertex in L_0 and the other vertex in R_0. Otherwise randomly place one vertex in L_0 and the other vertex in R_0.



Finding the Sets L, R — Initialization

1. $L_0, R_0 \leftarrow \phi$. Let $l = \Theta(\frac{m^2}{\Delta^2})$.
2. Randomly select a vertex u and l pairs of vertices.
3. For each pair of vertices, if only one vertex is a neighbor of u, place that vertex in L_0 and the other vertex in R_0. Otherwise randomly place one vertex in L_0 and the other vertex in R_0.



Analysis of the Initialization

Suppose that $u \in V_1$.
If $v \in V_1$ and $w \notin V_1$, then

$$\mathrm{P}\left[v \text{ is a neighbor of } u\right] = p > r = \mathrm{P}\left[w \text{ is a neighbor of } u\right]$$

$\Rightarrow$ Using Chernoff bounds and Hoeffding-Azuma's Inequality, w.h.p.,

$$\mathrm{I}(V_1, L_0, R_0) \approx (1 - \frac{1}{m})\frac{\Delta}{m}$$

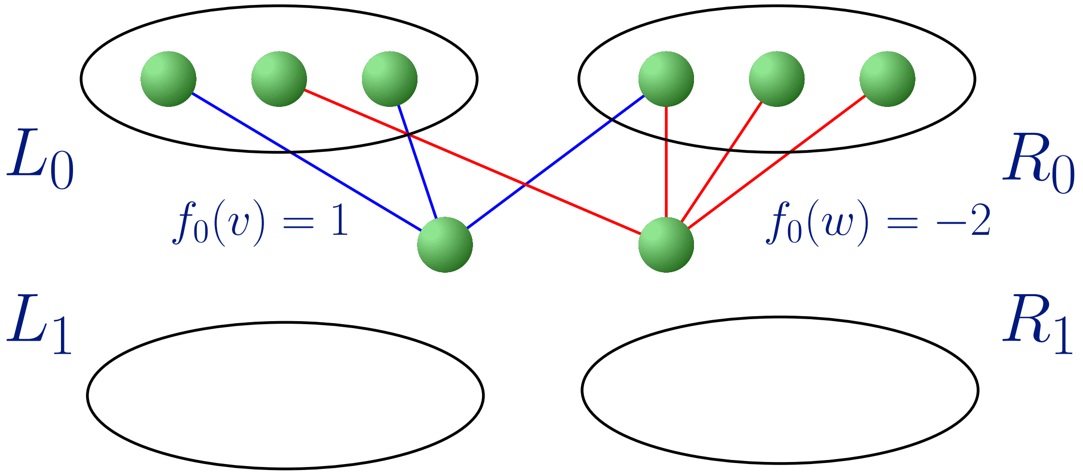
$$\mathrm{I}(V_i, L_0, R_0) \approx -\frac{1}{m} \cdot \frac{\Delta}{m} \qquad i > 1$$

Finding the Sets L, R — 1st Iteration

4. If L_0, R_0 are “good” (yielding a subcluster) stop.

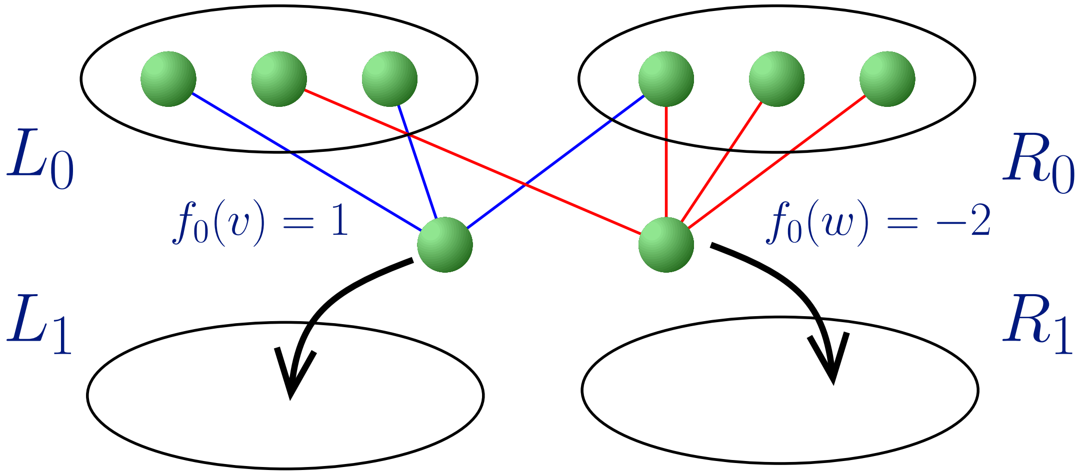
Finding the Sets L, R — 1st Iteration

4. If L_0, R_0 are "good" (yielding a subcluster) stop.
5. Let $f_0(v) = d_{L_0}(v) - d_{R_0}(v)$.
6. $L_1, R_1 \leftarrow \phi$. Randomly select l pairs of unchosen vertices.
7. For each pair v, w, if $f_0(v) \neq f_0(w)$ place the vertex with larger f_0-value in L_1 and the other vertex in R_1.



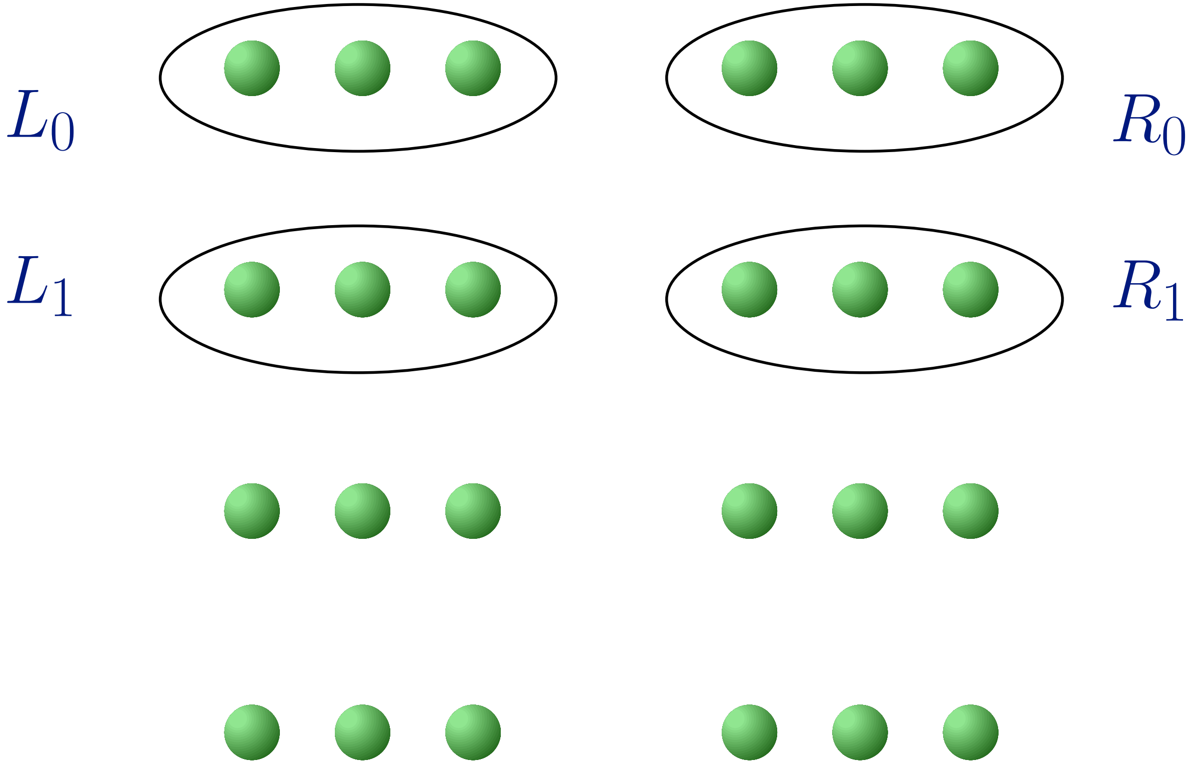
Finding the Sets L, R — 1st Iteration

4. If L_0, R_0 are "good" (yielding a subcluster) stop.
5. Let $f_0(v) = d_{L_0}(v) - d_{R_0}(v)$.
6. $L_1, R_1 \leftarrow \phi$. Randomly select l pairs of unchosen vertices.
7. For each pair v, w, if $f_0(v) \neq f_0(w)$ place the vertex with larger f_0-value in L_1 and the other vertex in R_1.



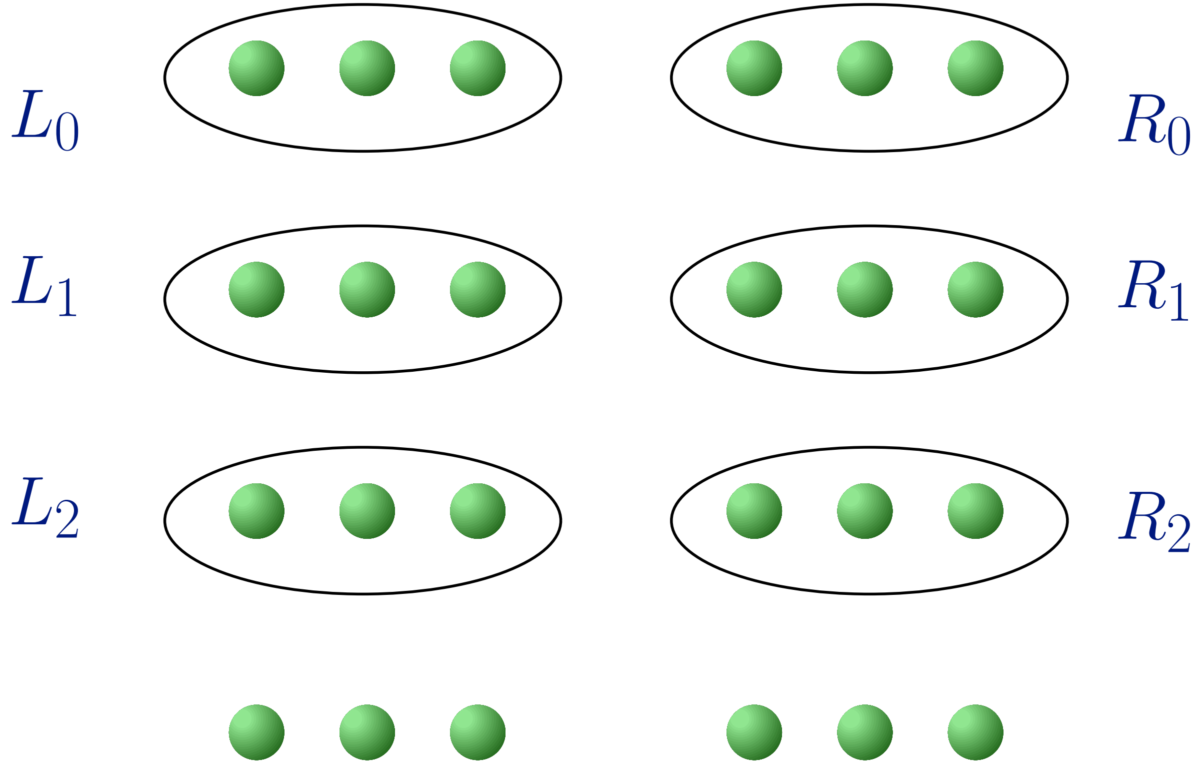
Finding the Sets L, R — Iterations

8. If L_1, R_1 are “good” stop.
9. Otherwise repeat this process (i.e. build L_2, R_2 from L_1, R_1, build L_3, R_3 from L_2, R_2 etc.) until a “good” pair is found.



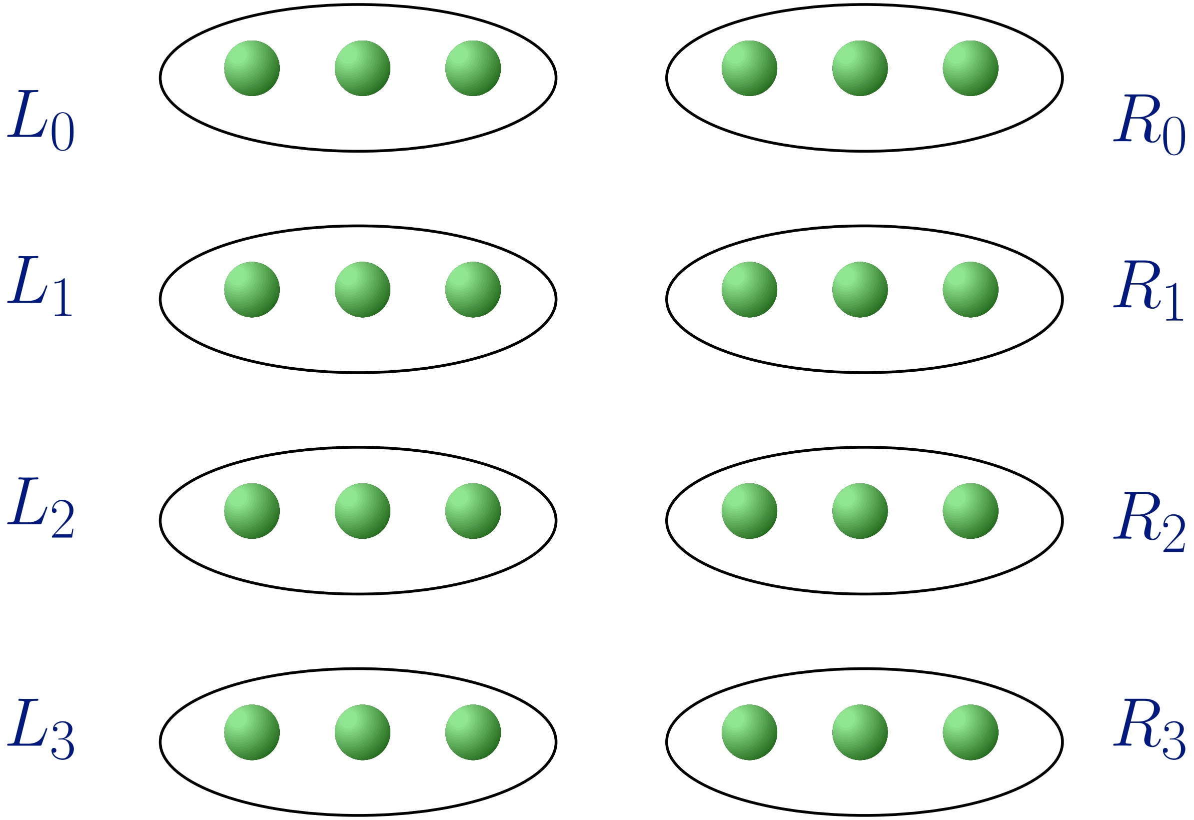
Finding the Sets L, R — Iterations

8. If L_1, R_1 are "good" stop.
9. Otherwise repeat this process (i.e. build L_2, R_2 from L_1, R_1, build L_3, R_3 from L_2, R_2 etc.) until a "good" pair is found.



Finding the Sets L, R — Iterations

8. If L_1, R_1 are "good" stop.
9. Otherwise repeat this process (i.e. build L_2, R_2 from L_1, R_1, build L_3, R_3 from L_2, R_2 etc.) until a "good" pair is found.



Analysis of the Iterations

Denote $b_i^t = \mathrm{I}(V_i, L_t, R_t)$.
Using Hoeffding-Azuma's Inequality and Esseen's Inequality we show that w.h.p.

1. The imbalance of V_1 grows exponentially:
   $b_1^t \geq 2b_1^{t-1}$ for all t.
2. The imbalance of other V_i-s is much smaller:
   $b_i^t = o(b_1^t)$ for all i, t.

$\Rightarrow$ After at most $\log n$ iterations we reach L_t, R_t with high imbalance.

Concluding Remarks

Main results:

- An algorithm for (almost) equal sized cluster (shown).
  The algorithm requires $k = \Omega(\Delta^{-1}\sqrt{n \log n})$.
- An algorithm for unequal sized cluster (not shown)
  The algorithm requires $k = \Omega(\Delta^{-1}\sqrt{n}\max(\log n, \Delta^{-\varepsilon}))$.