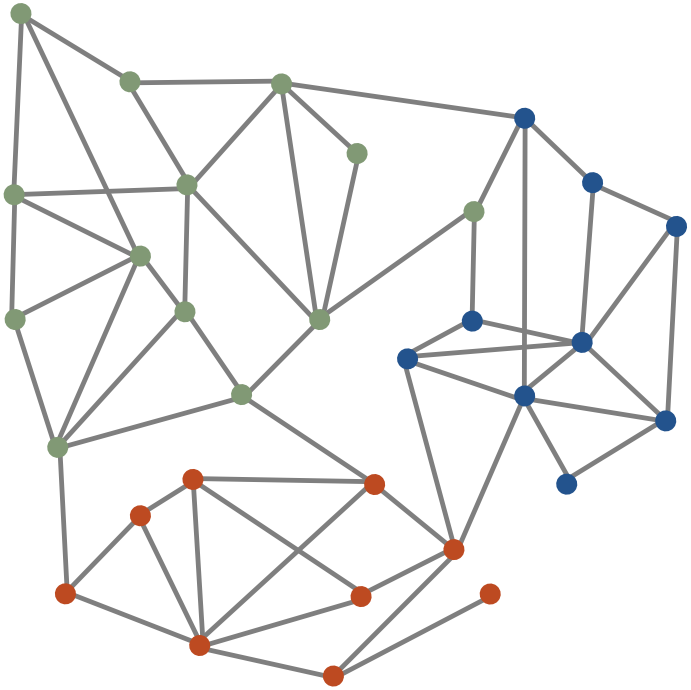
Partitioning Well-Clustered Graphs: Spectral Clustering Works!



Luca Zanetti

University of Bristol

Joint work with Richard Peng

He Sun

University of BRISTOL

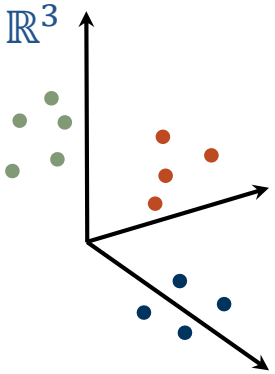
spectral clustering

Let L be the normalized Laplacian matrix of G, with eigenvalues $0 = \lambda_1 \leq \cdots \leq \lambda_n$ and the corresponding eigenvectors $f_1, \cdots, f_n$.

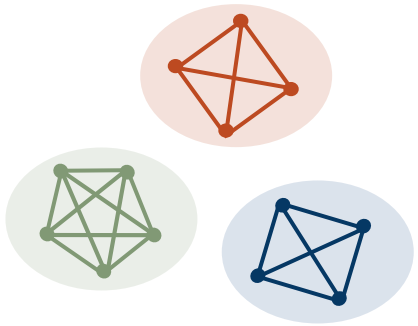
We want to partition G into k clusters.

Compute for every $u \in V[G]$

$$F(u) = \frac{1}{\text{NormalizationFactor}(u)} \cdot (f_1(u), \cdots, f_k(u))$$



Apply a k-means algorithm



Partition G into k clusters

spectral clustering

Let L be the normalized Laplacian matrix of G, with eigenvalues $0 = \lambda_1 \leq \cdots \leq \lambda_n$ and the corresponding eigenvectors $f_1, \cdots, f_n$.

We want to partition G into k clusters.

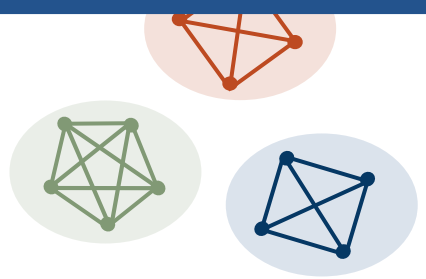
Compute for every $u \in V[G]$

$$F(u) = \frac{1}{\text{NormalizationFactor}(u)} \cdot (f_1(u), \cdots, f_k(u))$$

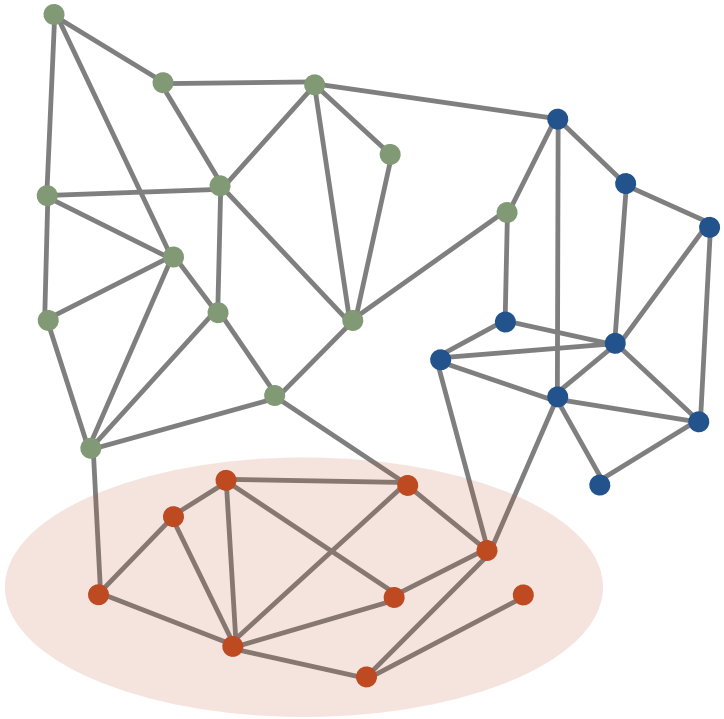
Apply a k-means algorithm

Can we analyze this framework theoretically?

Partition G into k clusters



clusterability assumption



$$\phi_G(S_1) = \frac{4}{31}$$

The conductance of set S is defined by

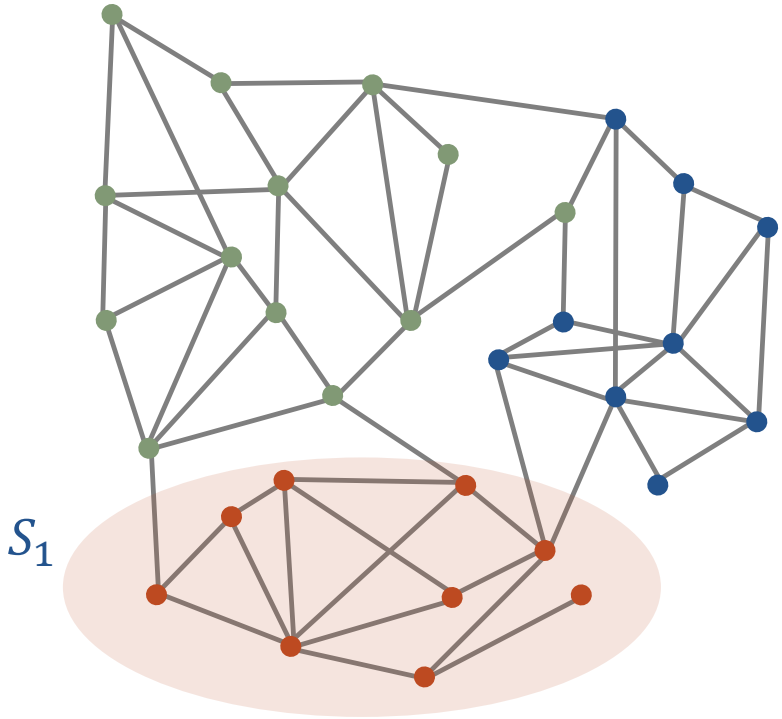
$$\phi_G(S) = \frac{|E(S, V \backslash S)|}{\text{vol}(S)}$$

The k-way expansion constant is defined by

$$\rho(k) = \min_{\text{partition } A_1, \cdots, A_k} \max_{1 \le i \le k} \phi_G(A_i)$$

We call $S_1, \cdots, S_k$ achieving $\rho(k)$ an optimal partition.

clusterability assumption



$$\phi_G(S_1) = \frac{4}{31}$$

The conductance of set S is defined by

$$\phi_G(S) = \frac{|E(S, V\backslash S)|}{\text{vol}(S)}$$

The k-way expansion constant is defined by

$$\rho(k) = \min_{\text{partition } A_1,\cdots,A_k} \max_{1\le i\le k} \phi_G(A_i)$$

We call $S_1, \cdots, S_k$ achieving $\rho(k)$ an optimal partition.

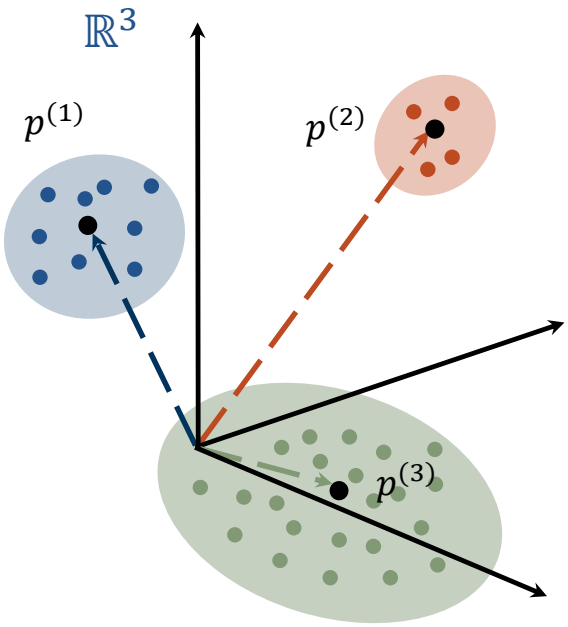
Assumption: G is a well-clustered graphs with k clusters, i.e. any $(k+1)$-partition contains a subset with conductance much bigger than $\rho(k)$.

Gap Assumption: Lower bound on

$$\Upsilon = \frac{\lambda_{k+1}}{\rho(k)}$$

University of BRISTOL

analysis of spectral clustering



Embedded points are concentrated around the approximate centers $\{p^{(i)}\}$

$$\sum_{i=1}^{k} \sum_{u \in S_i} d_u \left\| F(u) - p^{(i)} \right\|^2 \leq k^2 / \Upsilon$$

Distance between different clusters is inversely proportional to the volume of the ***smaller*** cluster.

$$\left\| p^{(i)} - p^{(j)} \right\|^2 \geq \frac{1}{1000k \cdot \min\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\}}$$

provable guarantees for spectral clustering

- Let $S_1, \cdots, S_k$ be an optimal partition.
- Let $A_1, \cdots, A_k$ be the output of spectral clustering.

Main Result: Let $\Upsilon = \Omega(k^3)$. For any $1 \leq i \leq k$, the following holds:

- Symmetric difference between A_i and S_i is bounded:

$$\text{vol}(A_i \,\Delta\, S_i) = O(k^3 \text{vol}(S_i)/\Upsilon).$$

- Conductance of each A_i is bounded:

$$\phi_G(A_i) = (\phi_G(S_i) + k^3/\Upsilon).$$

provable guarantees for spectral clustering

- Let $S_1, \cdots, S_k$ be an optimal partition.
- Let $A_1, \cdots, A_k$ be the output of spectral clustering.

Main Result: Let $\Upsilon = \Omega(k^3)$. For any $1 \leq i \leq k$, the following holds:

- Symmetric difference between A_i and S_i is bounded:

$$\text{vol}(A_i \,\Delta\, S_i) = O(k^3 \text{vol}(S_i)/\Upsilon).$$

- Conductance of each A_i is bounded:

$$\phi_G(A_i) = (\phi_G(S_i) + k^3/\Upsilon).$$

Result 2: There is an algorithm with comparable guarantees that runs in $O(m \cdot \text{polylog}\, n)$ time, i.e. independent of k.

provable guarantees for spectral clustering

- Let $S_1, \cdots, S_k$ be an optimal partition.
- Let $A_1, \cdots, A_k$ be the output of spectral clustering.

Main Result: Let $\Upsilon = \Omega(k^3)$. For any $1 \leq i \leq k$, the following holds:

- Symmetric difference between A_i and S_i is bounded:

$$\text{vol}(A_i \,\Delta\, S_i) = O(k^3 \text{vol}(S_i)/\Upsilon).$$

- Conductance of each A_i is bounded:

$$\phi_G(A_i) = (\phi_G(S_i) + k^3/\Upsilon).$$

Result 2: There is an algorithm with comparable guarantees that runs in $O(m \cdot \text{polylog}\, n)$ time, i.e. independent of k.

Thank you!

University of BRISTOL