Understanding Regularized Spectral Clustering via Graph Conductance

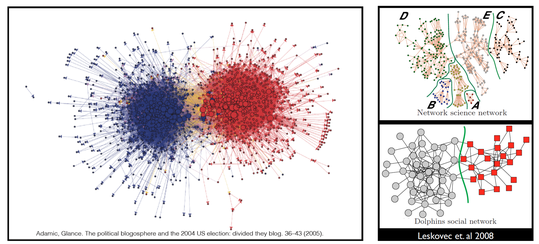
Yilin Zhang Karl Rohe

yilin.zhang@wisc.edu, karl.rohe@wisc.edu, University of Wisconsin-Madison

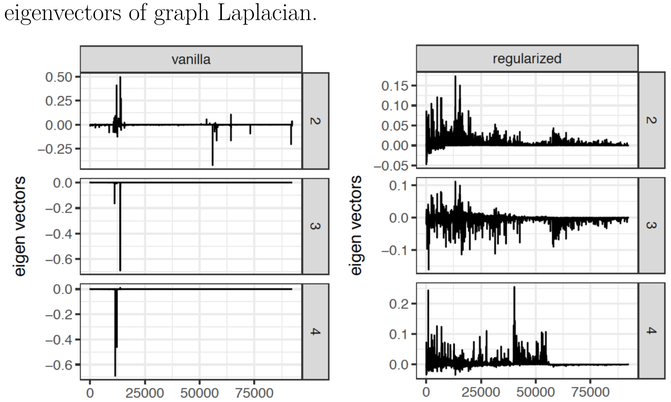
Video: `https://youtu.be/lOCoa3hYR4Y`

In practice, spectral clustering fails but regularization helps for community detection.

Community detection helps understand massive networks.



Spectral clustering is one popular approach. It partitions the network based on eigenvectors of graph Laplacian.



Spectral Clustering:

Adjacency Matrix $A \in \{0,1\}^{N\times N}$ with $A_{ij} = \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{o.w.} \end{cases}$.

Graph Laplacian $L = I - D^{-1/2}AD^{-1/2}$, where $D_{ii} = \sum_j A_{ij}$.

Regularization: Add a tiny edge between ever pair of nodes.

Adjacency Matrix $A_\tau = A + \frac{\tau}{N}J$, where J is an all-one matrix.

Graph Laplacian $L_\tau = I - D_\tau^{-1/2}A_\tau D_\tau^{-1/2}$, where $D_\tau = D + \tau I$.

Why this happens? graph conductance and noises

Spectral clustering likes sets with small graph conductance!

Graph conductance:

$$\text{conductance}(S) = \frac{cut(S)}{vol(S)} = \frac{\text{num of edges cut}}{\text{sum of node degrees}}.$$

Noises like g-dangling sets have small conductance. Many!

g-dangling sets: (one type of **noises**): In a graph $G = (V, E)$, a subset $S \subset V$ is g-dangling if and only if the following holds.

- S contains exactly g nodes.
- There are exactly $g-1$ edges within S and they do not form any cycles (i.e. the node induced subgraph from S is a tree).
- There is exactly one edge between nodes in S and nodes in S^c.

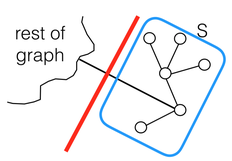


Figure 1: 6-dangling set.

Fact 1: g-dangling set has small conductance 1/(2g-1).

Theorem 1: *(many dangling sets)* Suppose an inhomogeneous random graph model such that for some $\epsilon > 0$, $p_{ij} > (1+\epsilon)/N$ for all nodes i, j. If that model contains a non-vanishing fraction of peripheral nodes $V_p \subset V$, such that $|V_p| > \eta N$ for some $\eta > 0$, then the expected number of distinct g-dangling sets in the sampled graph grows proportionally to N.

Theorem 2: *(many small eigenvalues)* If a graph contains Q g-dangling sets, and the rest of the graph has volume at least $4g^2$, then there are at least $Q/2$ eigenvalues that is smaller than $(g-1)^{-1}$. *(conceals true cluster even with large k and causes computational inefficiency)*

Spectral clustering is fooled by randomness!

Think regression: What do you say if the model perfectly interpolates the data (MSE = 0)?

Spectral clustering overfits to conductance.

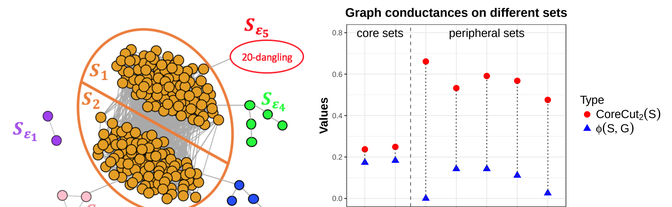
Don't be fooled by randomness!

Why regularization fixes this? Regularization changes the graph conductance.

Regularized spectral clustering likes sets with small CoreCut!

CoreCut: (graph conductance on the regularized graph)

$$\text{CoreCut}_\tau(S) = \frac{cut(S) + \frac{\tau}{N}|S||S^c|}{vol(S) + \tau|S|}.$$



Graph conductance likes peripheral sets.

Regularization increases conductance for peripheral sets significantly, but does not affect core sets that much.

CoreCut (conductance with regularization) prefers core sets.

Regularized SC ignores peripheral sets and focuses on the core!

Simulation