Recursive Clustering

Sunday, September 26, 2021 6:11 PM

YuRan

We can view the elements of a set as vertices of a graph whose edge weights represent similarities.

$G = (V, E)$ $(A)_{ij} = a_{ij} = sim(i, j) \geq 0$

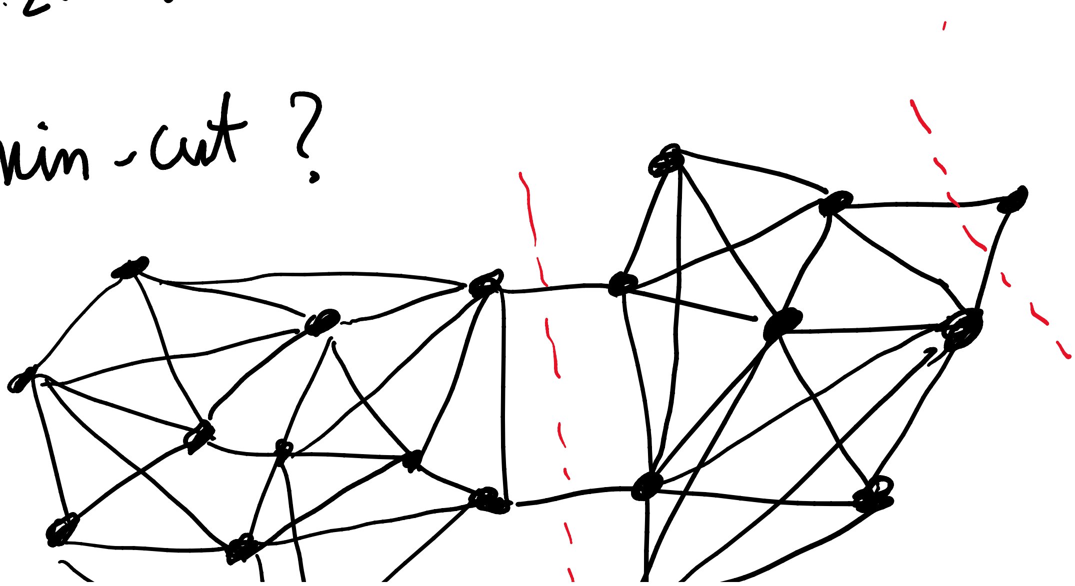
How to split into clusters?

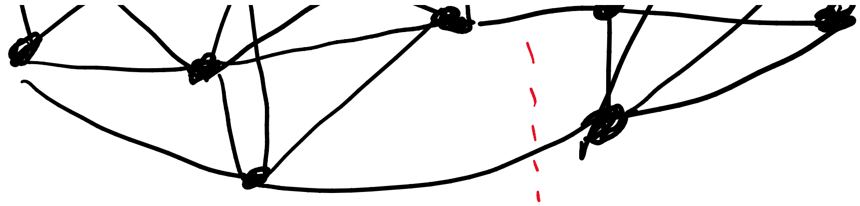
GOAL: want to keep similar elements in same cluster; and dissimilar elements in different clusters.

Q.1. How to cut?

Q.2. What is the quality of a cluster?

min-cut?





or is this better

$$a(S) = \sum_{i\in S,\, j\in V} a_{ij}$$

Conductance (expansion if unweighted)

$$\phi(S) = \frac{\sum_{i\in S\, j\notin S} a_{ij}}{\min\ a(S),\ a(V\setminus S)} \qquad \frac{\text{Similarity across } (S,\bar{S})}{\text{Similarity incident to } S \text{ or } \bar{S}.}$$

$S \subseteq V$

$$\phi(G) = \min_{S\subseteq V} \phi(S).$$

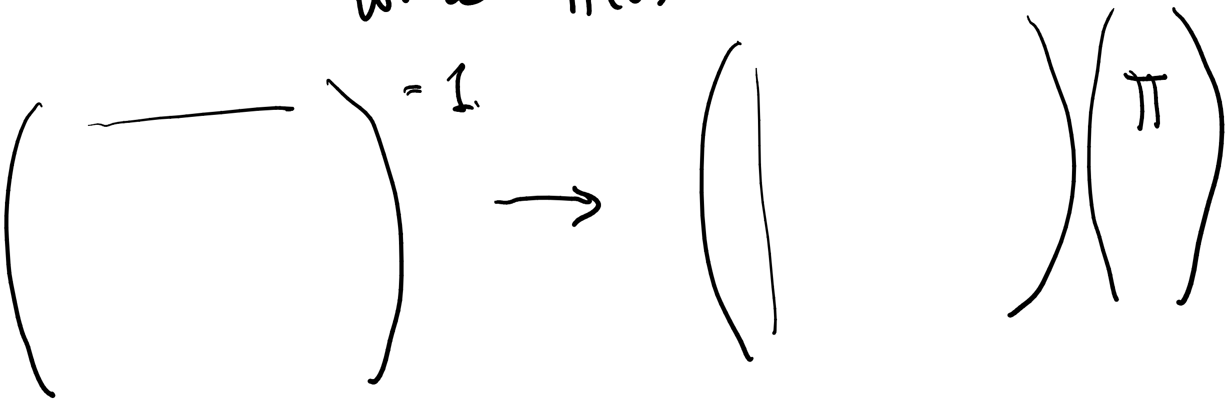
Since $a_{ij} \geq 0$, we can normalize rows to get B:

$$b_{ij} = \frac{a_{ij}}{\sum_j a_{ij}} = \frac{a_{ij}}{a(i)}$$

Then $B\mathbb{1} = \mathbb{1}$ and $B^T\pi = \pi$

where $\pi(i) \propto a(i)$

Where $\pi(i)$ ...



$$(B^T\pi)_i = \sum_j B_{ji}\pi_j = \sum_j \frac{a_{ij}}{\cancel{a(j)}} \cdot \frac{\cancel{a(j)}}{\|\pi\|_1}$$

$$= \frac{a(i)}{\|\pi\|_1} = \pi_i .$$

<u>Note:</u> $\pi_i b_{ij} = \pi_j b_{ji}$ (time reversible)

B^T represents the transition matrix of a time-reversible Markov chain.

<u>Thm</u> [Cheeger; JS] Let $B \geq 0$ $B \in \mathbb{R}^{N\times N}$

$B\mathbb{1} = \mathbb{1}$, $\pi_i b_{ij} = \pi_j b_{ji}$ $\pi_i > 0$.

Let v be the second (right) eigenvector of B with components $v_{i_1} \geq v_{i_2} \geq \ldots v_{i_N}$ and eigenvalue λ_2. ($\lambda_1 = 1$). Then,

$$\ldots \; 1 - \lambda_2 \geq \ldots \left(\min \phi(\{1,2,\ldots \ell\})\right)^2 .$$

$$2.\ \min_{S \subseteq V} \phi(S) \geq 1 - \lambda_2 \geq \frac{1}{2}\left(\min_{1 \leq \ell < N} \phi(\{1,2,\ldots \ell\})\right).$$

<u>Corollary:</u> $\exists$ polytime algorithm to find S

s.t. $\quad \phi(S) \leq \sqrt{2(1-\lambda_2)} \leq 2\sqrt{OPT}$.

<u>Algorithm:</u> 1) compute second eigenvector v

2) sort components $v_1 \geq v_2 \geq \ldots \geq v_N$

3) Output min conductance cut among $\{v_1 \ldots v_i\} / \{v_{i+1} \ldots v_N\}$.

<u>Pf</u> (of Thm). B is not symmetric in general.

Let D be a diagonal nonnegative matrix

s.t. $D^2 = \Pi$.

<u>claim:</u> $Q = D B D^{-1}$ is symmetric and has same eigenvalues as B.

eigenvectors are ...

Note $D^2 B = B^T D^2 \Rightarrow DBD^{-1} = D^{-1} B^T D$
symmetric.

Next $\forall v$ $Bv = \lambda v$

$DBD^{-1}(Dv) = \lambda (Dv)$. same eigenvalues.

We also know $\lambda_1 = 1$. and $\mathbb{1}$ is top eigenvector of B.

$(\pi^T D^{-1}) Q = \pi^T B D^{-1} = \pi^T D^{-1} \leftarrow$ e.v. of Q.

So $$\lambda_2 = \max_{x:\, \pi^T D^{-1} x = 0} \frac{x^T D B D^{-1} x}{x^T x}$$

$$1 - \lambda_2 = \min \frac{x^T (I - DBD^{-1}) x}{x^T x} = \frac{x^T D (I-B) D^{-1} x}{x^T x}$$

let $y = D^{-1} x$

$$= \min_{y:\, \pi^T y = 0} \frac{y^T D^2 (I-B) y}{y^T D^2 y}$$

Numerator $= y^T D^2 (I-B) y = -\sum_{i \neq j} \pi_i b_{ij} y_i y_j + \sum \pi_i (1-b_{ii}) y_i^2$

$$(\neq) \quad + \sum_i \pi_i (1-b_{ii}) y_i$$

$$= - \sum_{i \neq j} \pi_i b_{ij} y_i y_j + \sum_{i \neq j} \pi_i b_{ij} \frac{(y_i^2 + y_j^2)}{2}$$

$$= \sum_{i<j} \pi_i b_{ij} (y_i - y_j)^2 = \mathcal{E}(y,y) \text{ (say)}$$

$$1 - \lambda_2 = \frac{\mathcal{E}(y,y)}{\sum_i \pi_i y_i^2}$$

For the first inequality, let $(S, \bar{S})$ be min conductance cut. Define

$$w_i = \begin{cases} \sqrt{\dfrac{\pi(\bar{S})}{\pi(S)}} & i \in S \\ -\sqrt{\dfrac{\pi(S)}{\pi(\bar{S})}} & i \notin S. \end{cases}$$

Then $\pi^T w = 0$

and

$$1 - \lambda_2 \leq \frac{\mathcal{E}(y,y)}{\sum_i \pi_i y_i^2} = \frac{\sum_{i \in S, j \notin S} \pi_i b_{ij} \left(\sqrt{\frac{\pi(\bar{S})}{\pi(S)}} + \sqrt{\frac{\pi(S)}{\pi(\bar{S})}} \right)^2}{\sum_{i \in S} \pi_i \frac{\pi(\bar{S})}{\pi(S)} + \sum_{i \notin S} \pi_i \frac{\pi(S)}{\pi(\bar{S})}}$$

$$= \sum \pi_i b_{ij} \left(\pi(\bar{S}) + \pi(S) \right)^2$$

$$= \frac{\sum_{i\in S, j\notin S} \pi_i b_{ij} \left(\pi(\bar{S}) + \pi(S)\right)}{\pi(S)\cdot\pi(\bar{S})}.$$

$$\le 2 \frac{\sum_{i\in S, j\notin S} \pi_i b_{ij}}{\min\{\pi(S), \pi(\bar{S})\}} = 2\phi(S).$$

$v_1 \ge v_2 \ge \quad \cdots \quad \ge v_N$

Let r be s.t.

$$\pi_1 + \pi_2 + \cdots \pi_{r-1} \le \frac{1}{2} < \pi_1 + \cdots \pi_r$$

and $z_i = v_i - v_r$ so that

$$z_1 \ge z_2 \ge \cdots \ge z_r = 0 \ge z_{r+1} \ge \cdots \ge z_n$$

$$\frac{\mathcal{E}(v,v)}{\sum_i \pi_i v_i^2} = \frac{\mathcal{E}(z,z)}{-v_r^2 + \sum_i \pi_i z_i^2} \ge \frac{\mathcal{E}(z,z)}{\sum_i \pi_i z_i^2}$$

$$\sum_i \pi_i v_i^2 = \sum_i \pi_i (z_i + v_r)^2 = \sum \pi_i z_i^2 + \left(\sum_i \pi_i\right)\cdot v_r^2 + \sum_i 2\pi_i z_i v_r$$

$$= \sum \pi_i z_i^2 + v_r^2 - 2v_r^2$$

$$= \sum \pi_i z_i^2 + v_r^2 - 2 v_r$$

$$\pi^T z = \pi^T v - v_r^2 = -v_r^2 \qquad = \sum \pi_i z_i^2 - v_r^2$$

$$\frac{\sum_{i<j} \pi_i b_{ij} (z_i - z_j)^2}{\sum \pi_i z_i^2} \quad \frac{\left(\sum_{i<j} \pi_i b_{ij} (|z_i| + |z_j|)^2 \right)}{\left(\sum_{i<j} \pi_i b_{ij} (|z_i| + |z_j|)^2 \right)}$$

$$\text{Num} \geq \left(\sum \pi_i b_{ij} \; |z_i - z_j| \, (|z_i| + |z_j|) \right)^2$$

Claim $|z_i - z_j| (|z_i| + |z_j|) \geq \sum_{k=i}^{j-1} |z_{k+1}^2 - z_k^2|$

z_i, z_j have same sign

$LHS = |z_i^2 - z_j^2| = RHS.$

else $LHS = (|z_i| + |z_j|)^2 > z_i^2 + z_j^2 = RHS.$

$i \leq r \leq j$

$z_r = 0.$

So Numerator $\geq \left(\sum_{i<j} \pi_i b_{ij} \sum_{k=i}^{j-1} |z_{k+1}^2 - z_k^2| \right)^2$

Denominator $\leq \sum_i \pi_i z_i^2 \cdot 2 \sum_{i<j} \pi_i b_{ij} (z_i^2 + z_j^2)$

$\leq 2 \left(\sum \pi_i z_i^2 \right)^2$

$$\leq 2 \left(\sum_i \pi_i z_i^2\right)^2$$

$$1 - \lambda_2 \geq \frac{1}{2} \cdot \left(\frac{\sum_{i<j} \pi_i b_{ij} \sum_{k=i}^{j-1} |z_{k+1}^2 - z_k^2|}{\sum \pi_i z_i^2} \right)^2$$

Let $$\hat{\alpha} = \min_{1 \leq k < N} \frac{\sum_{i \leq k < j} \pi_i b_{ij}}{\min \pi(\{1 \ldots k\}), \pi(\{k+1, \ldots N\})}$$

Then

$$\sum_{i<j} \pi_i b_{ij} \sum_{k=i}^{j-1} |z_{k+1}^2 - z_k^2| = \sum_{k=1}^{N-1} |z_{k+1}^2 - z_k^2| \sum_{i \leq k < j} \pi_i b_{ij}$$

$$\geq \sum_{k=1}^{N-1} |z_{k+1}^2 - z_k^2| \, \hat{\alpha} \cdot \min \pi(\{1 \ldots k\}), \pi(\{k+1 \ldots N\})$$

$$= \hat{\alpha} \left(\sum_{k=1}^{r-1} (z_k^2 - z_{k+1}^2) \pi(S_k) + \sum_{k=r}^{N-1} (z_{k+1}^2 - z_k^2)(1 - \pi(S_k)) \right)$$

$$= \hat{\alpha} \left(\sum_{k=1}^{N-1} (z_k^2 - z_{k+1}^2) \pi(S_k) + z_N^2 - z_r^2 \right)$$

$$= \hat{\alpha} \left(\sum_{k=1}^{N-1} z_k^2 (\pi(S_k) - \pi(S_{k-1})) + z_N^2 \right)$$

$$= \hat{\alpha} \left(\sum_{k=1}^{N} \pi_k z_k^2 \right)$$

$$= \alpha \left(\sum_{k=1} \pi_k \tau_k \right)$$

$$\text{i.e. } 1 - \lambda_2 \geq \frac{1}{2} \cdot \hat{\alpha}^2 .$$

Multiway Cheeger

Recursive Partitioning.

- cut into two parts
- Recurse while cluster is not high enough quality.

α = conductance of cluster

ε = fraction of similarity between clusters.

Thm. $\exists\ (\alpha, \varepsilon)$ - bicriteria clustering, Recursive Spectral Partitioning finds a

$$\left(c \cdot \frac{\alpha^2}{\log^2\left(\frac{n}{\varepsilon}\right)}, c\sqrt{\varepsilon} \log \frac{n}{\varepsilon} \right) \text{- clustering.}$$

---

can be used with any

Can be used with any approximate conductance cut algorithm.

<u>Thm</u>. $\exists$ algorithm that finds S s.t.

$$\phi(S) \le c\sqrt{\log n} \cdot OPT.$$