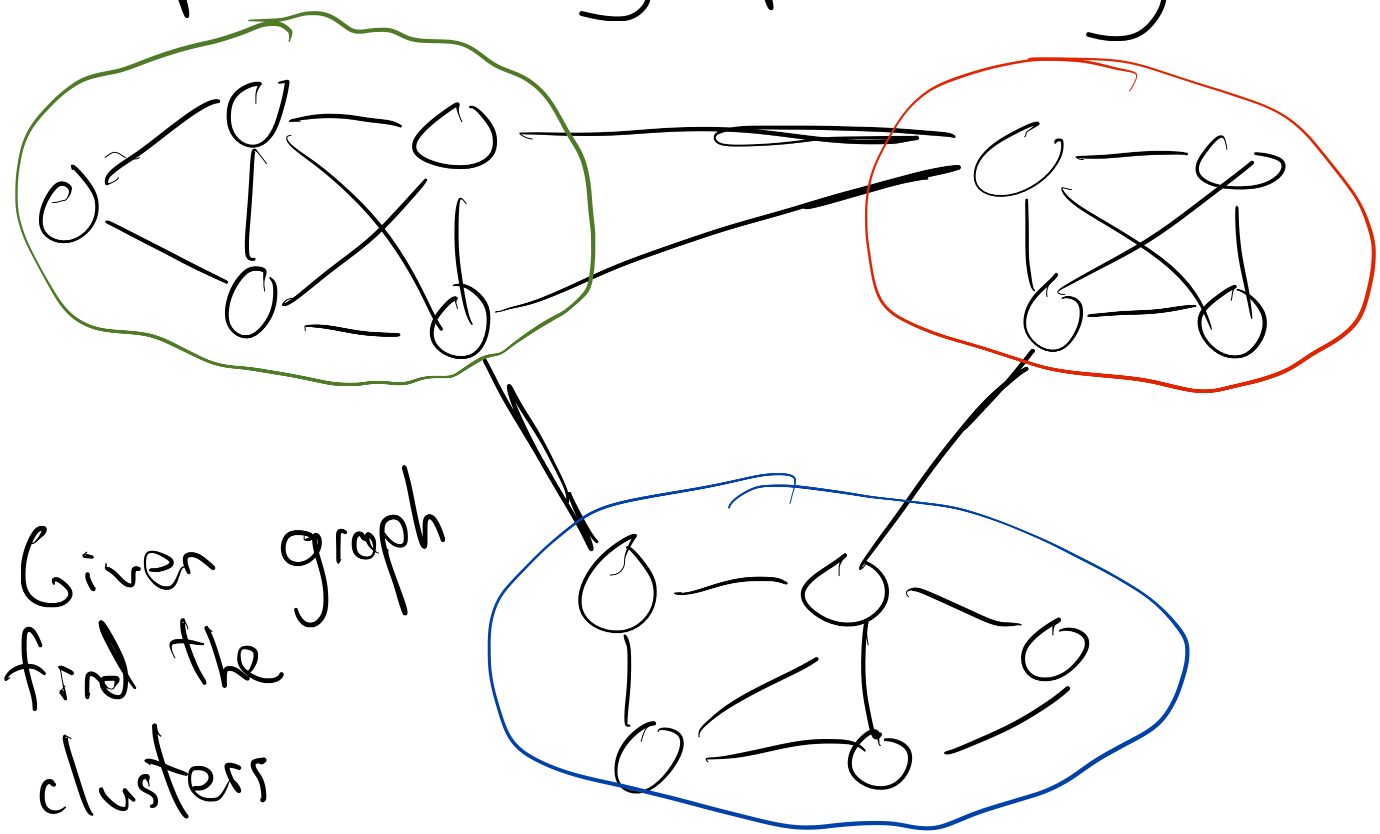
March 10, 2020

Graph clustering / partitioning / community detection



- communities in social networks
- modules in biological networks
- papers on a topic in citation network
- brain regions

Given graph find the clusters

mesoscopic

(no labels given)

Lots of methods! (see surveys)

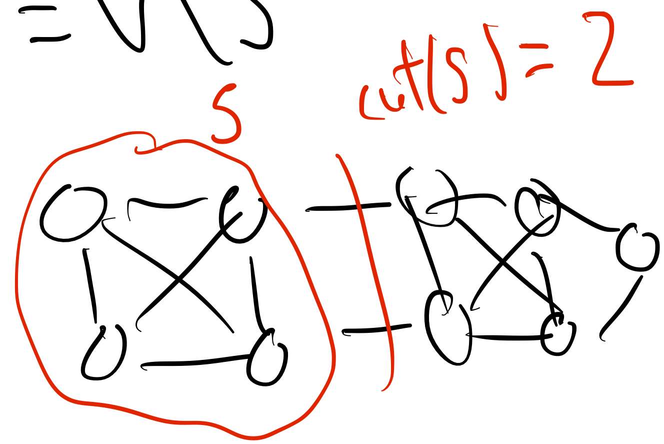
Our focus: spectral methods

$G = (V, E)$ undir. $|V| = n$ $|E| = m$

partition V into S and $\bar{S} = V \setminus S$

$\mathrm{cut}(S) = \#$ of edges leaving S

$$= \sum_{i \in S,\, j \in \bar{S}} A_{ij}$$



$$\mathrm{RatioCut}(S) = \frac{\mathrm{cut}(S)}{|S|} + \frac{\mathrm{cut}(\bar{S})}{|\bar{S}|}$$

Test vectors: $x_i = \frac{1}{\sqrt{n}} \begin{cases} \sqrt{|\bar{S}|/|S|} & i \in S \\ -\sqrt{|S|/|\bar{S}|} & i \in \bar{S} \end{cases}$ $(*)$

Claim: $x^T L x = \mathrm{RatioCut}(S)$ under $(*)$

$$x^T L x = \sum_{(i,j)\in E} (x_i - x_j)^2$$

$$= \cancel{\sum_{i\in S, j\in S} (x_i - x_j)^2} + \cancel{\sum_{i\in \bar{S}, j\in \bar{S}} (x_i - x_j)^2} + \sum_{i\in S, j\in \bar{S}} (x_i - x_j)^2$$

$$= \sum_{i\in S, j\in \bar{S}} \frac{1}{n}\left(\sqrt{|\bar{S}|/|S|} + \sqrt{|S|/|\bar{S}|}\right)^2$$

$$\frac{|S|}{|S|} + \frac{|\bar{S}|}{|\bar{S}|}$$

$$= \frac{1}{n}\sum |\bar{S}|/|S| + |S|/|\bar{S}| + 2$$

$$= \cancel{\frac{1}{n}} \sum \frac{|S| + |\bar{S}|}{|S|} + \frac{|\bar{S}| + |S|}{|\bar{S}|}$$

n

$$= \sum_{\substack{i\in S \\ j\in \bar{S}}} \frac{1}{|S|} + \frac{1}{|\bar{S}|} = \left(\frac{1}{|S|} + \frac{1}{|\bar{S}|}\right) \mathrm{cut}(S)$$

$$= \frac{\mathrm{cut}(S)}{|S|} + \frac{\mathrm{cut}(\bar{S})}{|\bar{S}|}$$

Claim: $1^T x = 0$ under ⊛

$$1^T x = \sum_{i \in V} x_i = \sum_{i \in S} x_i + \sum_{j \in \bar{S}} x_j$$

$$= \sum_{i \in S} \frac{1}{\sqrt{n}} \sqrt{|\bar{S}|/|S|} + \sum_{j \in \bar{S}} \frac{1}{\sqrt{n}} \left(-\sqrt{|S|/|\bar{S}|} \right)$$

$$= \frac{1}{\sqrt{n}} \left(\sqrt{|S||\bar{S}|} - \sqrt{|\bar{S}||S|} \right) = 0$$

Claim: $x^T x = 1 = \|x\|_2^2 = \sum_i x_i^2$ under ⊛

$$x^T x = \frac{1}{n} \left(\sum_{i \in S} |\bar{S}|/|S| + \sum_{j \in \bar{S}} |S|/|\bar{S}| \right)$$

$$= \frac{1}{n} \left(|\bar{S}| + |S| \right)$$

$$= 1$$

$$\min_S \; x^T L x = \mathrm{RatioCut}(S) = \sum_{(i,j)\in S} (x_i - x_j)^2$$

$$\text{s.t.} \quad \cancel{x_i = \frac{1}{\sqrt{n}} \begin{cases} \sqrt{|\bar{S}|/|S|} & i \in S \\ -\sqrt{|S|/|\bar{S}|} & i \in \bar{S} \end{cases}} \quad \text{relax} \qquad 1^T x = 0 \quad x^T x = 1$$

Claim: $L x_* = \lambda_2 x_* \quad (x_*^T x_* = 1)$

Proof: (assume graph is connected) $\qquad V^T V = I = V V^T$

$$L = V \Lambda V^T \qquad x^T L x = x^T V \Lambda V^T x$$

$$Vy = VV^T x \qquad y^T \Lambda y = \sum_{i=1}^n y_i^2 \lambda_i \qquad 0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$$

$$v_1 = \frac{1}{\sqrt{n}} 1 \qquad 0 = v_1^T x = v_1^T V y = e_1^T y = y_1$$

$$y_2 = 1 \qquad y_1 = y_3 = \dots = y_n = 0$$

$$x = Vy = Ve_2 = v_2 \qquad v_2^T v_2 = 1$$

Still need S! round: $x_i > 0 \iff i \in S$

$$L x_* = \lambda_2 x_* \qquad x_*^T L x_* = \lambda_2 x_*^T x_* = \lambda_2$$

For any S, we had test vector x_S

$$\lambda_2 \le \min_S \text{RatioCut}(S)$$

$$L = D - A \qquad D \approx cI$$

$$\lambda(cI - A) = c - \lambda_i(A)$$

second largest for A

$\approx$ second smallest L

Normalized cut

$$v(S) = \sum_{i \in S} d_i \qquad d_i = \text{degree of node } i$$

$= \#$ edge end points in S

$$NCUT(S) = \frac{cut(S)}{v(S)} + \frac{cut(\bar{S})}{v(\bar{S})}$$

Probabilistic interp.

$$\pi_i = \frac{d_i}{2m}$$

① Draw node from stat. dist. π

② Take one RW step

z_i = node after step i $\quad z_1, z_2$

$$\Pr(Z_1 \in S) = \sum_{i\in S} \pi_i = \sum_{i\in S} \frac{d_i}{2m} = \frac{1}{2m} \underbrace{\sum_{i\in S} d_i}_{v(S)}$$

$$\Pr(Z_1 \in S,\ Z_2 \in \bar{S})$$

$$= \sum_{i\in S} \pi_i \Pr(i \to j,\ j \in \bar{S})$$

$$= \sum_{i\in S,\ j\in \bar{S}} \pi_i P_{ji} = \sum \pi_i \frac{1}{d_i} A_{ij} = \sum_{\substack{i\in S\\ j\in \bar{S}}} \frac{1}{2m} A_{ij}$$

$$= \frac{1}{2m}\, cut(S)$$

$$\Pr(Z_2 \in \bar{S} \mid Z_1 \in S)$$

$$= \Pr(Z_1 \in S,\ Z_2 \in \bar{S}) \Pr(Z_1 \in S) = \frac{\frac{1}{2m}\, cut(S)}{\frac{1}{2m}\, v(S)}$$

$$NCut(S) = \Pr(Z_2 \in \bar{S} \mid Z_1 \in S) + \Pr(Z_2 \in S \mid Z_1 \in \bar{S})$$

Test vector: $x_i = \frac{1}{\sqrt{v(V)}} \begin{cases} \sqrt{v(\bar{S})/v(S)} & i \in S \\ -\sqrt{v(S)/v(\bar{S})} & i \in \bar{S} \end{cases}$ $\circledast$

$v(V) = 2m$

Claim: $x^T L x = NCUT(S)$

$$\sum_{\substack{i \in S \\ j \in \bar{S}}} (x_i - x_j)^2 = \frac{1}{v(V)} \sum_{\substack{i \in S \\ j \in \bar{S}}} \left(\sqrt{v(\bar{S})/v(S)} + \sqrt{v(S)/v(\bar{S})} \right)^2$$

$$= \frac{1}{v(V)} \sum_{\substack{i \in S \\ j \in \bar{S}}} v(\bar{S})/v(S) + v(S)/v(\bar{S}) + 2$$

$$2 = \frac{v(S)}{v(S)} + \frac{v(\bar{S})}{v(\bar{S})}$$

$\vdots$

$= NCUT(S)$

Claim: $\mathbb{1}^T D x = 0$ under $\circledast$

Claim: $x^T D x = 1$ under $\circledast$

$$\min_S \; x^T L x = \mathrm{Ncut}(S)$$

$$\text{s.t.} \quad \cancel{x_i = \frac{1}{v(V)} \begin{cases} \sqrt{v(\bar{S})/v(S)} & i \in S \\ -\sqrt{v(S)/v(\bar{S})} & i \in \bar{S} \end{cases}} \quad \text{relax}$$

$$\mathbb{1}^T D x = 0 \qquad x^T D x = 1$$

$$z = D^{1/2} x$$

$$\min_z \; z^T D^{-1/2} L D^{-1/2} z \quad = \text{normalized Laplacian } N$$

$$\text{s.t.} \quad \mathbb{1}^T D^{1/2} z = 0 \qquad z^T z = 1$$

Claim: $D^{1/2}\mathbb{1}$ is evec for $\lambda_1(N) = 0$

$$N D^{1/2}\mathbb{1} = D^{-1/2} L D^{-1/2} D^{1/2} \mathbb{1}$$

$$= D^{-1/2} L \mathbb{1} = 0$$

$$Ly = 0 \iff N D^{1/2} y = 0$$

$$N z_* = \lambda_2 z_*$$

$$D^{-1/2} z_* = x_*$$

$$\lambda_2(N) \leq NCUT(S)$$