17/5/21

Objective: Given a network, made of nodes and observed edges, the aim is to find unknown groupings of nodes

↓

clusters, community, blak

↓

community detection method

We are going to look at two methods:

① Community detection via hierarchical clustering

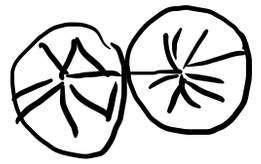
② Spectral analysis → K-means

COMMUNITY DETECTION VIA HIERARCHICAL CLUSTERING

We need to define a measure of quality of a cluster ⇝ Think of $W(\Gamma_i)$ in K-means

Let $(C_1, \dots, C_K) = C$ a partition of the nodes in a network.

Let $f_{ij}(C)$ denote the fraction of edges in the original network that connect nodes in cluster C_i with nodes in cluster C_j.



In particular, $f_{KK}(C)$ is the fraction of edges within cluster C_K.

MODULARITY

(K fixed)

$$M(C) = \sum_{j=1}^{K} \left(\underbrace{f_{jj}(C)}_{\substack{\text{percentage} \\ \text{of edges in} \\ C_j}} - \underbrace{f^*_{jj}(C)}_{\substack{\text{percentage of} \\ \text{edges under} \\ \text{random assignment (standardization)}}} \right)^2$$

Large values of modularity suggest non-trivial grouping structures

Hierarchical clustering can be based on modularity:

① AGGLOMERATIVE: Successive coarsening of partitions through the process of merging.
At each step, consider the merge that has the highest modularity

② DIVISIVE: Successive refinement of partitions through the process of splitting.
At each step, consider the split with the highest modularity.

Implementations would then return the optimum K, ie the partition with the highest modularity overall.

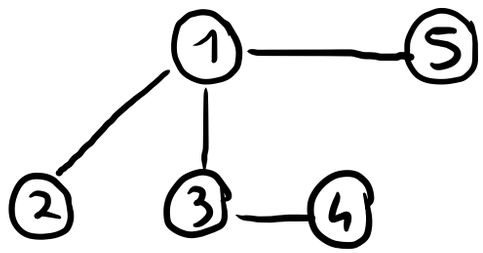
COMMUNITY DETECTION VIA SPECTRAL CLUSTERING

The connectivity of a graph (network) is related to the eigenanalysis of its associated **Laplacian matrix**

Def Given a network, one can define its **adjacency matrix**:

$$A = (a_{ij}) \quad \text{where} \quad a_{ij} = \begin{cases} 1 & \text{if node } i \text{ and } j \text{ are connected} \\ 0 & \text{otherwise} \end{cases}$$

Ex:



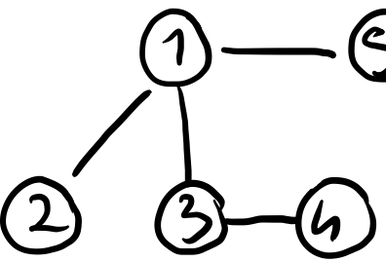
$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Def Laplacian matrix is defined as

$$L = D - A$$

where A is the adjacency matrix and D is a diagonal matrix with node degrees in the diagonal.

Ex



$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad D = \begin{bmatrix} 3 & & & & \\ & 1 & & & \\ & & 2 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}$$

So the Laplacian is

$$L = \begin{bmatrix} 3 & -1 & -1 & 0 & -1 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 2 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Row and column sums are zero since $d_i = \sum_{j=1}^{p} a_{ij}$

L has some nice mathematical properties:

L is $p \times p$, where p = number of nodes

(1) For every vector $\underline{v} \in \mathbb{R}^p$,

$$\underline{v}^t L \underline{v} = \frac{1}{2} \sum_{i,j=1}^{p} a_{ij} (v_i - v_j)^2$$

Proof:

$$\underline{v}^t L \underline{v} = \underline{v}^t (D - A) \underline{v} = \underline{v}^t D \underline{v} - \underline{v}^t A \underline{v} = \sum_{i=1}^{p} d_i v_i^2 - \sum_{i,j=1}^{p} a_{ij} v_i v_j$$

$d_i \rightarrow (D)_{ii}$

$$= \frac{1}{2} \left[\sum_{i=1}^{p} d_i v_i^2 + \sum_{j=1}^{p} d_j v_j^2 - 2 \sum_{i,j=1}^{p} a_{ij} v_i v_j \right]$$

$$= \frac{1}{2} \left[\sum_{i,j=1}^{p} a_{ij} v_i^2 - 2 \sum_{i,j=1}^{p} a_{ij} v_i v_j + \sum_{i,j=1}^{p} a_{ij} v_j^2 \right]$$

$d_i = \sum_{j=1}^{p} a_{ij}$

$$= \frac{1}{2} \sum_{i,j=1}^{p} a_{ij} (v_i - v_j)^2$$

■

② L is symmetric and positive semi-definite

$$l_{ij} = d_{ij} - a_{ij} = d_{ji} - a_{ji} = l_{ji}$$

symmetric

(D, A symmetric)

For any $\underline{v} \in \mathbb{R}^p$, $\underline{v}^t L \underline{v} = \frac{1}{2} \sum_{i,j} a_{ij} (v_i - v_j)^2 \geq 0$

positive semi-definite

($a_{ij} \geq 0$ ①; $(v_i - v_j)^2 \geq 0$)

③ L has p non-negative real-valued eigenvalues

$$0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_p$$

④ The smallest eigenvalue is zero and the corresponding eigenvector is the constant one vector.

Take $\underline{v} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$, then $L\underline{v} = \lambda \underline{v}$ with $\lambda = 0$

Indeed $L \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}_{\underline{v}} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \overbrace{0}^{\lambda} \cdot \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}_{\underline{v}}$

row sums are zero in L

These properties can be extended to the case of weighted graphs

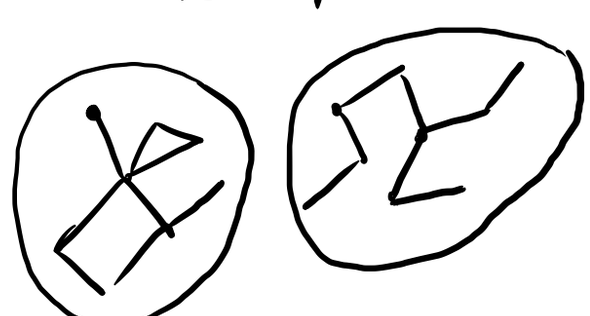
$d_i = \sum_{j=1}^{p} w_{ij}$ with $w_{ij} \geq 0$

$L = D - W$

LAPLACIAN AND CONNECTIVITY

Def: A connected component (or just a component) of a graph is a subgraph in which any two vertices are connected by a path (ie a sequence of connected nodes)

Prop The number of zero eigenvalues of the Laplacian matrix is equal to the number of connected components in the graph.



Two zero eigenvalues

Proof Let us consider the case of 2 connected components.

Then I can rearrange the adjacency matrix like this:

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \begin{matrix} \} \text{ number of nodes in component 1} \\ \} \text{ component 2} \end{matrix}$$

Then consider

$$\underline{v}_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \} \text{ \# nodes in component 1} \\ \} \text{ \# nodes in component 2} \end{matrix} \qquad \underline{v}_2 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

These two are eigenvectors with eigenvalue zero.

With K components:

$$A = \begin{pmatrix} A_1 & & 0 \\ & A_2 & \\ 0 & & \ddots \\ & & & A_K \end{pmatrix}$$

$$\underline{v}_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \left.\vphantom{\begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}\right\} \#\text{ nodes in component 1} \qquad \underline{v}_2 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \left.\vphantom{\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}\right\} \#\text{ nodes in component 2} \qquad \cdots \qquad \underline{v}_K = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \left.\vphantom{\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix}}\right\} \#\text{ nodes in component K}$$

$\underline{v}_1, \ldots, \underline{v}_K$ are eigenvectors with zero eigenvalues.

EXAMPLE ($K = 3$)

Consider the matrix of eigenvectors with zero eigenvalues:

$$\begin{matrix} \underline{v}_1 & \underline{v}_2 & \underline{v}_3 \\ \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & \\ 0 & 1 & \\ \vdots & \vdots & \\ & 1 & 0 \\ & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix} \end{matrix} \quad \text{equal rows}$$

In practice, networks will not have completely disconnected components.

SPECTRAL CLUSTERING works by applying K-means to the $p \times K$ matrix of eigenvectors corresponding to the smallest eigenvalues.