Statistical analysis of network data lecture 4

Sofia Olhede

October 8, 2025

1. Centralities
2. Nonparametric Summaries Cont'd
3. The Stochastic Blockmodel
4. Other block model estimators

- As an alternative Boldi and Vigna have proposed to study the harmonic centrality instead defined as

Definition

Harmonic centrality We define the harmonic centrality of vertex i in a graph G with n nodes as

$$C_i^{(H)} = \sum_{j \neq i} \frac{1}{\text{dist}_G(i,j)}.$$

- This is integrally related to the average efficiency of the network defined by Latorra

Definition

Average efficiency We define the average efficiency of a graph G with n nodes as

$$E(G) = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{\text{dist}_G(i,j)}.$$

- Latorra defined a local version thereof:

Definition

local efficiency We define the local efficiency of a node i in graph G with n nodes with G_i as the neighbours of i, i.e. those which have edges in common with i as

$$E = \frac{1}{n} \sum_{i \in v(G)} E(G_i).$$

- The efficiency E is often normalised further.
- This measure naturally takes account of the fact that some nodes are in different connected components.

- Unfortunately there are more than two measures of centrality. Betweenness centrality was introduced by Anthonisse and Freeman:

Definition

Betweenness centrality We define the betweenness centrality of vertex i in a graph G with n nodes, with n_{jk}^i as the number of shortest paths from j to k that pass through i, and with n_{jk} as the number of shortest paths between j and k as

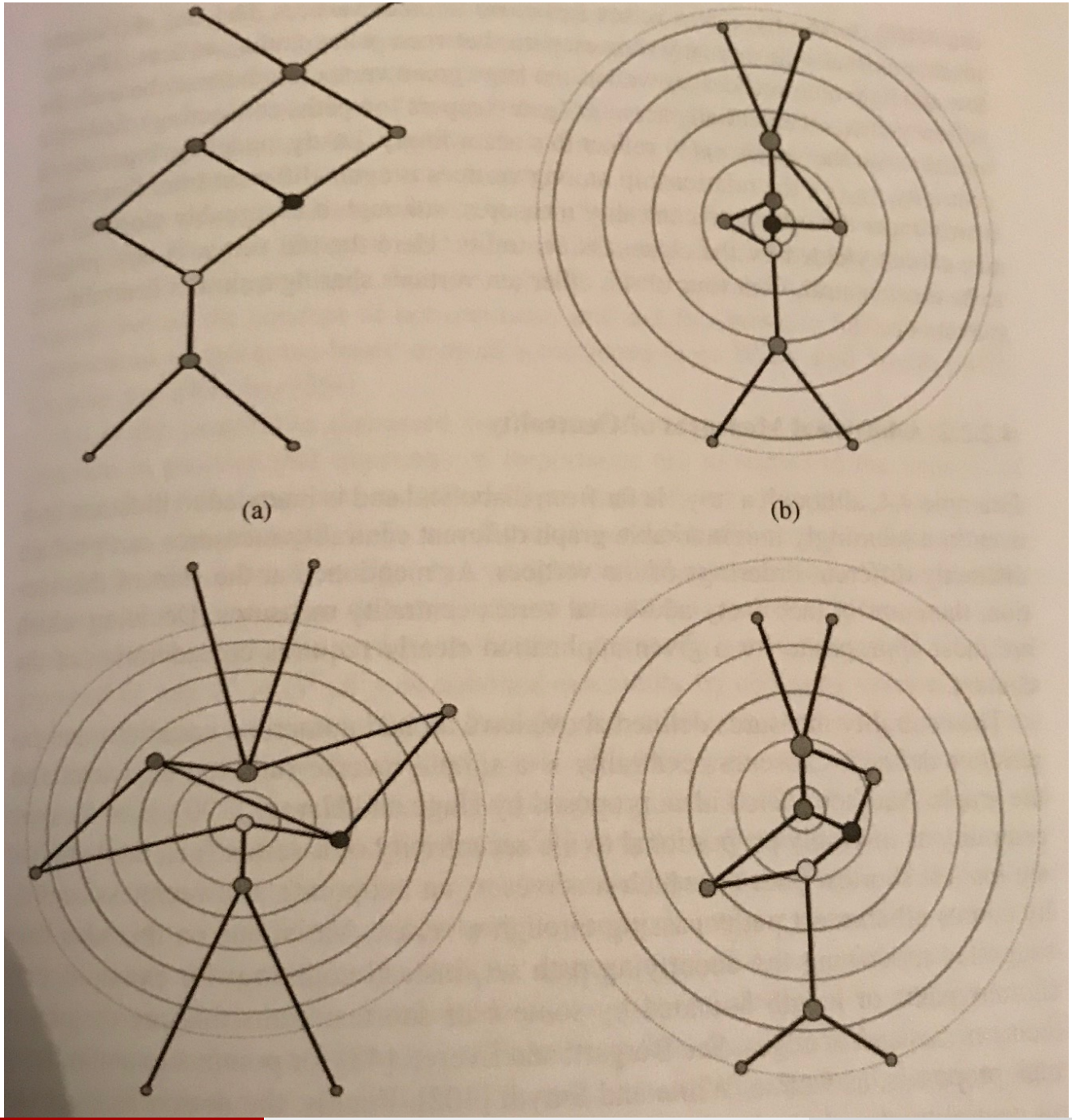
$$B_i = \sum_{1 \leq j < k \leq n} \frac{n_{jk}^i}{n_{jk}}.$$

- The Eigenvector centrality (also called eigencentrality) can also be used to measure the importance of a node i.

Definition

Eigenvector centrality. Let u be the eigenvector with eigenvalue λ of the largest eigenvalue (with positive entries due to the Perron–Frobenius theorem) of the adjacency matrix. We define the eigenvector centrality of vertex i in a graph G with n nodes, as

$$C_i = \frac{1}{\lambda} \sum_{ij \in E(G)} u_j.$$


(a)
(b)

- Additionally for the whole network we are interested in its degree of clustering. For a graph on the nodes $\{1, \ldots, n\}$ we let the number of paths with 3 nodes be

$$X_{P_3}(G) = \frac{1}{2} \sum_{1 \leq i,j,k \leq n} \mathsf{I}(ij, \ \ jk \in E).$$

Definition

Clustering coefficient We define the clustering coefficient of a graph G with n nodes as

$$\mathrm{CC}_G = \frac{X_{C_3}(G)}{X_{P_3}(G)}.$$

- For a sequence of graphs $\{G_n\}$ we can define a property of the sequence, namely to be highly clustered if

$$\lim_{n \to \infty} \inf \mathrm{CC}_{G_n} > 0.$$

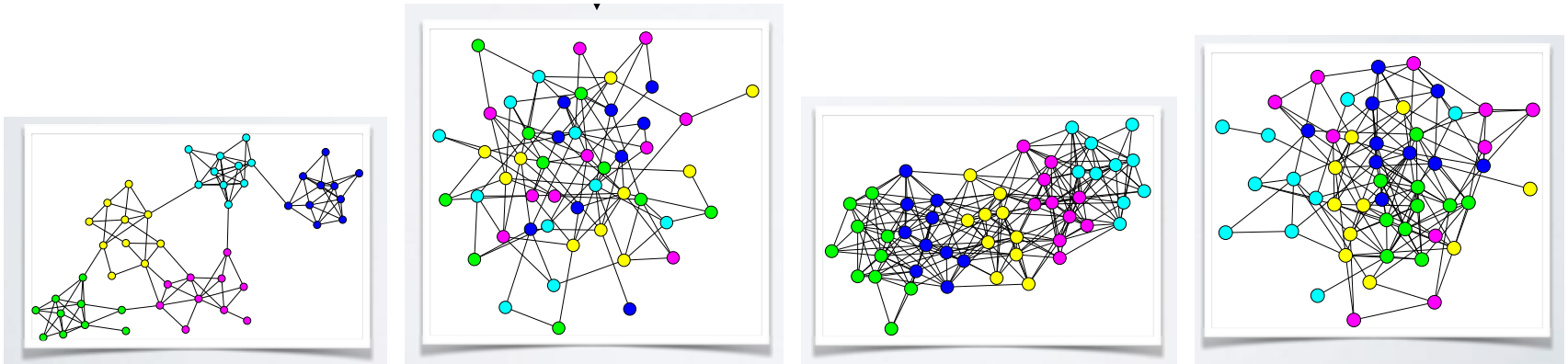
- With those non-parametric graph properties out of the way we may return to parametric properties of graphs.
- We already looked at estimating the growing–length parameter of the degree–based model.
- Let us return to the stochastic blockmodel of $\{z_i\}$ and $\{\theta_{ab}\}$ (The planted partition model).
- How can we estimate $\{z_i\}$ and $\{\theta_{ab}\}$?
- The most common methods are spectral clustering and an assessment via network modularity (the latter due to Newman).

- Recall that simple common network characteristics are clustering and blocking/grouping of nodes.
- By grouping nodes, and assuming the model is specified by the groups alone, we have gone from nodal permutations to group permutations. For each node i we define a random variable z_i that takes the value $\{1, \ldots, k\}$, where this variable is indicating the group membership of node i. We additionally define a connection probability matrix Θ which has entries θ_{ab} for $1 \leq a < b \leq k$. Then

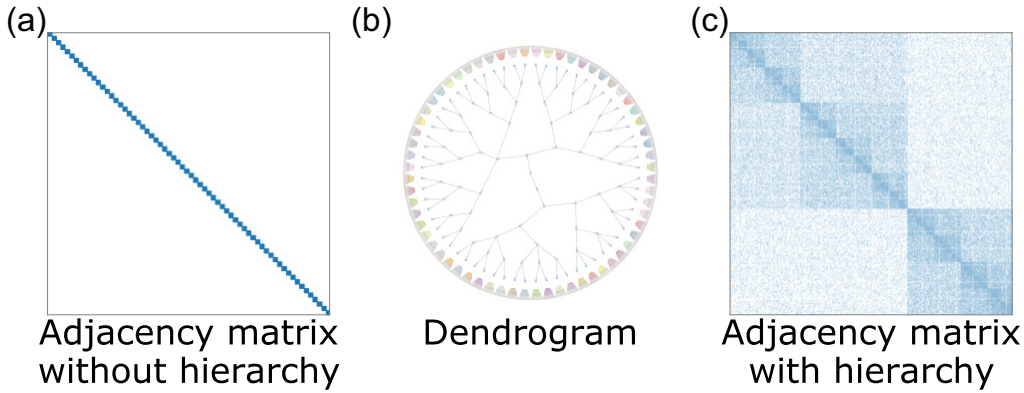
$$A_{ij}|z_i, z_j = \text{Bernoulli}(\theta_{z_i z_j}), \quad 1 \leq j < i \leq n. \tag{1}$$

  where each realization is independent. Furthermore $A_{ii} = 0$ for $1 \leq i \leq n$, and we complete the matrix by $A_{ji} = A_{ij}$ for $1 \leq j < i \leq n$. This is known as the stochastic block model.
- Block models are normally split up into types. The first is assortative stochastic blocks, that is the probability of connections within communities is higher than in between communities, e.g $\theta_{aa} \geq \theta_{ab}$ for $a, b \in \{1, \ldots, k\}$, and this shows edges inside groups. This captures "birds of a feather flock together".

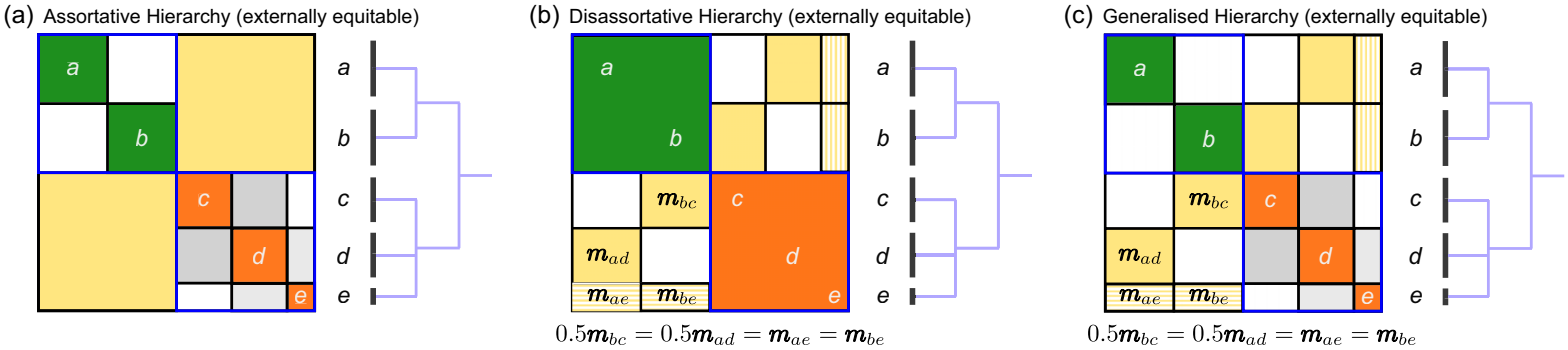
- Disassortative stochastic blocks, that is the probability of connections between communities is higher than within communities, e.g $\theta_{aa} \leq \theta_{ab}$ for $a, b \in \{1, \ldots, k\}$. People connect as they have different functions.
- Ordered stochastic blocks show hierarchical structure, and
- Core–periphery structure have a dense corse, but a sparse periphery.
- Note the distinction between a notional idea and a mathematical model can be loose.



Plots taken from Aaron Clauset.

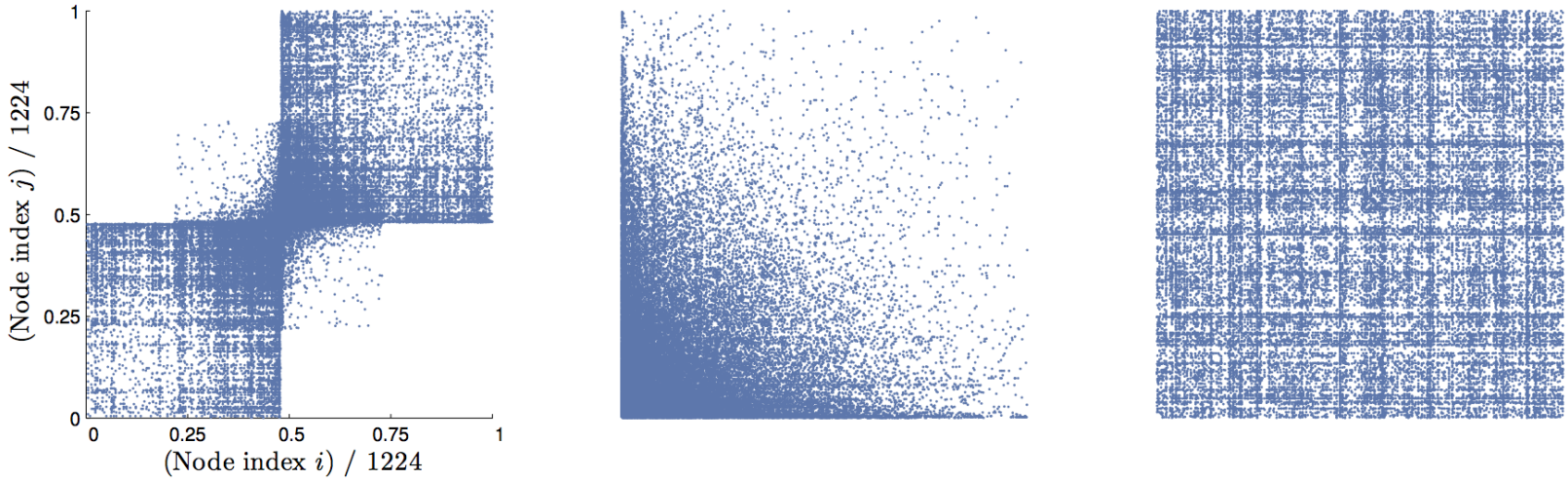


But how do we recover the latent structure? Picture from Peel et al.



But how do we recover the latent structure? Picture from Peel et al.

But how do we recover the latent blocks?



Remember..... With the right ordering everything looks easy... But we DO NOT have it.

- At the basic level the simplest stochastic block model are $\binom{n}{2}$ independent Bernoulli trials. We can formulate the likelihood dependent on link parameters $\{\theta_{ab}\}$ and a group label $z_i \in \{1, \ldots, k\}$ for $i = 1, \ldots, n$ when there are k groups. We then find

$$\Pr\{A\} = \prod_{i<j} \theta_{z_i z_j}^{A_{ij}} \{1 - \theta_{z_i z_j}\}^{1-A_{ij}}, \tag{2}$$

- The log-likelihood takes the form of

$$\ell(\theta, z) = \sum_{i<j} A_{ij} \log(\theta_{z_i z_j}) + (1 - A_{ij}) \log(1 - \theta_{z_i z_j}),$$

and that

$$\sum_i \mathsf{I}(z_i = b) = h_b,$$

so that with $h_{ab} = \sum_{i<j} \mathsf{I}(z_i = a)\mathsf{I}(z_j = b)$, $h_{ab} = h_a h_b$ for $a \neq b$

$$\ell(\theta, z) = \sum_{a \leq b} \{h_{ab} \bar{A}_{ab}(z) \log \theta_{ab} + (h_{ab} - h_{ab} \bar{A}_{ab}(z)) \log(1 - \theta_{ab})\}, \tag{3}$$

with $\bar{A}_{ab}(z) = h_{ab}^{-1} \sum_{i<j} A_{ij} \mathsf{I}(z_i = a)\mathsf{I}(z_j = b)$.

- As a first step for any fixed choice of z we can maximize the likelihood; we find that

$$\begin{aligned}
\frac{\partial \ell(\theta, z)}{\partial \theta_{ab}} &= h_{ab} \frac{\bar{A}_{ab}(z)}{\theta_{ab}} - h_{ab} \frac{1 - \bar{A}_{ab}(z)}{1 - \theta_{ab}} \\
\widehat{\theta}_{ab} &= \bar{A}_{ab}(z), \quad 1 \leq a \leq b \leq k \\
\frac{\partial^2 \ell(\theta, z)}{\partial \theta_{ab}^2} &= -h_{ab} \frac{\bar{A}_{ab}(z)}{\theta_{ab}^2} - h_{ab} \frac{1 - \bar{A}_{ab}(z)}{\{1 - \theta_{ab}\}^2} \\
&< 0.
\end{aligned}$$

- The profile likelihood (Cox & Reid and Bickel and Chen)then becomes

$$\ell(z) = \sum_{i<j} \{A_{ij} \log(\widehat{\theta}_{z_i z_j}) + (1 - A_{ij}) \log(1 - \widehat{\theta}_{z_i z_j})\}.$$

We can maximize this in z to obtain an estimate of the label vector z. This is unfortunately not computationally feasible.

- We first define
$$E = \frac{1}{2}\sum_{i,j} A_{ij} = \frac{1}{2}\sum_{l} d_l.$$
- Then we define the network modularity for label–vector z by
$$\widehat{Q}_G(z) = \sum_{i<j}\left\{A_{ij} - \frac{d_i d_j}{\sum_l d_l}\right\}\delta_{z_i z_j}.$$
- We maximise this quantity to arrive at a division of the nodes.
- Often this is done hierarchically; but it comes with no statistical guarantees.
- We have to decide when to stop; we stop when no further improvement in the modularity result.
- This treats the stochastic blockmodel parametrically; k is assumed to be known. There are other ways to cluster nodes; using the graph Laplacian is one of them.

- We now need a method to cluster the points $(A_{i\cdot})_i$ in $\mathbb{R}^k$.
- Clustering means dividing up a set of n points into $k \leq n$ groups. For each k we define a vector C_k of those group labels.
- This is not a well defined problem, see for example, C. Hennig, 'What are the true clusters?', Pattern Recognition Letters. 64 (2015), 53-62.
- To cluster we need a measure of similarity and a distance $d(x, y)$. These are obviously reciprocal concepts. Similarity measures how similar two vectors are like a covariance measure. Distance measures how far apart two points are or how dissimilar they are.
- What is a cluster?
  1. items that have very similar or the same properties;
  2. items whose distance is small, or their dissimilarity is small;
  3. have "contacts" with other items in the same cluster;
  4. are clearly different from the items in other clusters.