Statistical analysis of network data revision class

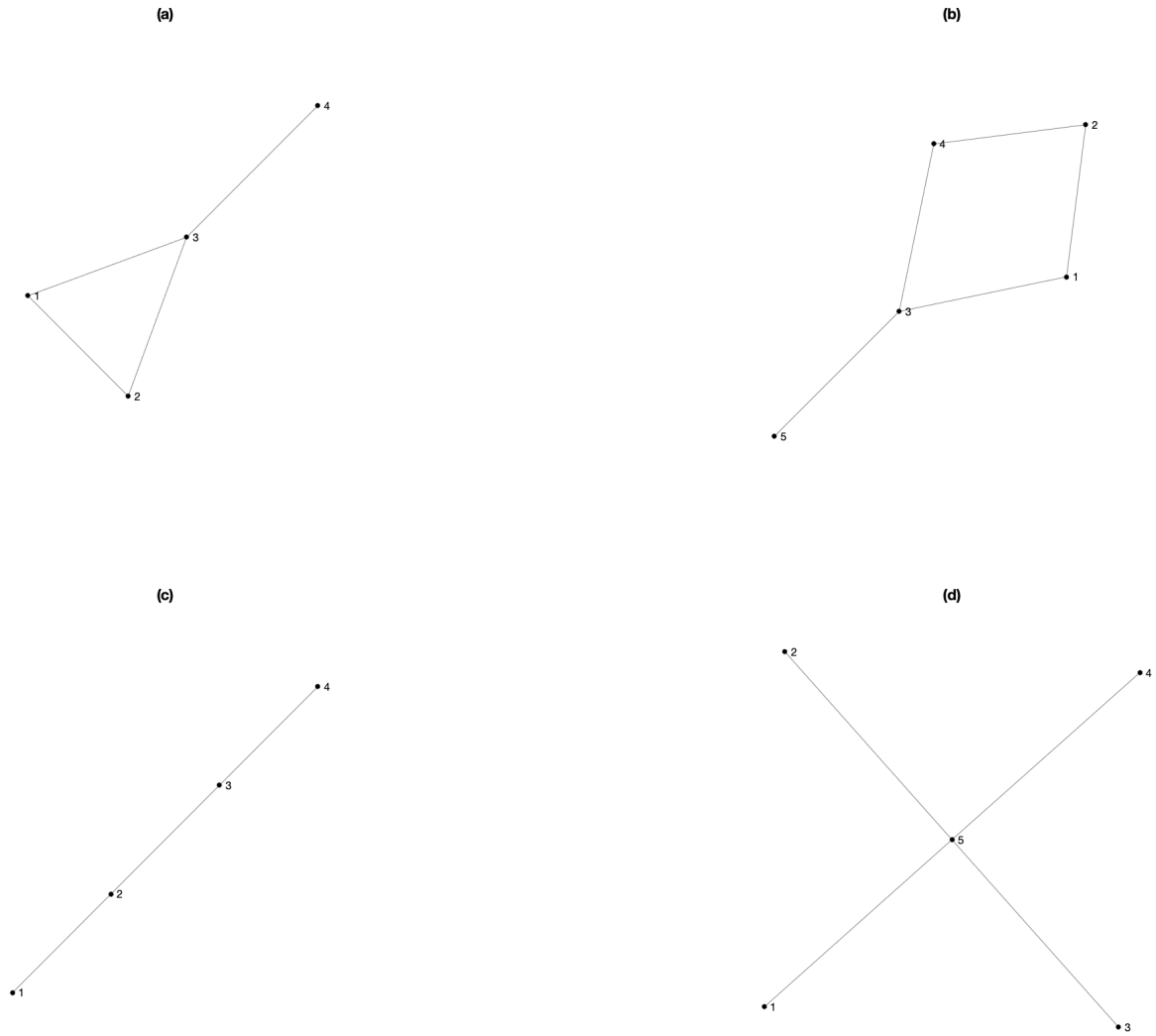
Sofia Olhede

December 18, 2024

Please note that this is a summary, not the complete course.

1. Graphs or network data structures
2. Network data models
3. Network Statistics

- This course concerns the stochastic properties of network data.
- A network represents interactions between entities (nodes or vertices), where the presence of an interaction is indicated by an edge.
- We write a network or a graph as G, usually represented by an adjacency matrix A. A_{ij} where if node i and node j are linked A_{ij} takes the value unity, otherwise it takes the value zero. A network can also be represented by a list of edges, an edge list, that just specifies the existing edges, e.g. $\{(1, 15), (1, 32), ...\}$.
- We usually write the collection of nodes as $v(G)$ and the collection of edges as $e(G)$. We normally assume that $|v(G)| = n$ if not specified otherwise. We write $G = (v(G), e(G))$.
- A graph $H = (v(H), e(H))$ is a subgraph of G if $v(H) \subseteq v(G)$ and $e(H) \subseteq e(G)$.



Labelled graphs or networks.

- Additionally we define the incidence matrix C that catalogues the relationships between nodes and edges. This C_{ij} is unity if vertex v_i and edge e_j are incident.
- A summary statistic defined from A the adjacency matrix is the degree vector. This takes the form

$$d_i = \sum_{j \neq i} A_{ij}.$$

- In this course we shall consider simple and undirected networks. A simple network is unweighted, e.g. the strength of a connection is not weighting the adjacency matrix. Instead all weights are either zero or unity. For example when looking at trade relationships between countries it makes sense to report the magnitude of trade either by weight or cost.
- Some special graphs have names and symbols. The complete graph on n nodes is written as K_n.
- A cycle on n nodes is written C_n. Every node in a cycle has degree 2.

- A d-regular graph is a graph where all nodes have the same degree d. Cycles are regular graphs.
- A connected graph with no cycles is a tree. A disjoint union of such graphs is a forest.
- A bipartite graph $G = (v(G), e(G))$ is one where the vertices can be split into $v_1(G)$ and $v_2(G)$, and where each edge has one endpoint in $v_1(G)$ and the other in $v_2(G)$.

- The simplest graph is an Erdős-Rényi network on n nodes with edge probability $0 < p < 1$. We write this as $\mathrm{ER}(n, p)$. The adjacency matrix is generated element by element as

$$A_{ij} = \mathrm{Bernoulli}(p), \quad 1 \leq j < i \leq n. \tag{1}$$

  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 $1 \leq j < i \leq n$.
- The simplest generalization introduces n parameters π_i and then generates edges independently by

$$A_{ij} = \mathrm{Bernoulli}(\min(\pi_i \pi_j, 1)), \quad 1 \leq j < i \leq n. \tag{2}$$

  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 Chung-Lu or the configuration model.

- An Inhomogeneous Random Graph (Soderberg). This generates edges independently by

$$A_{ij} = \text{Bernoulli}(p_{ij}), \quad 1 \leq j < i \leq n. \tag{3}$$

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 $1 \leq j < i \leq n$.

- Some graphs display clear group structure. 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

$$A_{ij}|z_i, z_j = \text{Bernoulli}(\theta_{z_i z_j}), \quad 1 \leq j < i \leq n. \tag{4}$$

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.

- Another generalization of the stochastic block model is the mixed membship model (Airoldi, Fienberg and Xing). Generate latent variable $\boldsymbol{\xi}_i$ for each node i from the Dirichlet distribution of dimension k with parameters $\boldsymbol{\alpha}$. Define $\boldsymbol{\Theta} = (\theta_{pq})$ and draw

$$A_{ij} \,|\, \boldsymbol{\xi}_i, \boldsymbol{\xi}_j \sim \mathrm{Ber}(\boldsymbol{\xi}_i^T \boldsymbol{\Theta} \boldsymbol{\xi}_j). \tag{5}$$

- The random dot product graph (RPDG) is a latent position model of

$$\mathbb{E}\{A_{ij} \,|\, \boldsymbol{\Xi}\} = \rho_n \cdot \boldsymbol{\xi}_i^T \boldsymbol{\xi}_j,$$

where the latent position of node i, namely $\boldsymbol{\xi}_i$ is generated by probability density function $f(\boldsymbol{\xi})$.

- Degree corrected stochastic block model. 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, and a latent uniform ξ_i. Define a connection probability matrix Θ which has entries θ_{ab} for $1 \leq a < b \leq k$. Then with 1-d function $g(x)$ we draw

$$A_{ij} | z_i,\, z_j, \xi_i, \xi_j = \mathrm{Bernoulli}(\theta_{z_i z_j} + g(\xi_i) g(\xi_j)), \quad 1 \leq j < i \leq n. \tag{6}$$

where each realization is independent.

- Most of these models fall in a more general framework of permutation invariance. Thus if we introduce a permutation Π that remaps all indices, the nature of the model should not change.
- Let Π be a permutation on the ordering, and let the repermuted adjacency matrix be A^{Π}.

Definition

Permutation-invariance of the distribution holds when $\Pr(A = a) = \Pr(A^{\Pi} = a)$ *for any permutation* Π *and any adjacency matrix* A*. That is, permuting the adjacency matrix does not change its distribution. Then we say that the distribution is permutation-invariant.*

Definition (Exchangeable arrays)

More generally, for any $k \geq 1$ *we can consider an array of* E*-valued r.v.s* $(X_e)_{e \in \mathbb{N}^{(k)}}$ *indexed by size-*k *subsets of* $\mathbb{N}$*, and say it is (jointly) exchangeable if* $(X_e)_e \stackrel{d}{=} (X_{\Pi(e)})_e \; \forall \Pi \in \mathrm{Sym}(\mathbb{N})$*, where if* $e = \{n_1, \ldots, n_k\}$ *then* $\Pi(e) := \{\Pi(n_1), \ldots, \Pi(n_k)\}$.

Theorem (Aldous Hoover)

An array $\mathbf{A}$ *is jointly exchangeable, iff it has the same distribution as*

$$A_{ij} = f(\alpha, \xi_i, \xi_j, \zeta_{ij}),\ 1 \leq i < j,$$

with $f : \mathbb{R}^4 \mapsto \mathbb{R}$ *and some iid random uniform variables* α, ξ_i *and* ζ_{ij}.

- A basic concept is to count the occurance of subgraphs in G. A subgraph count simply corresponds to

$$X_F(G) = \sum_{F' \subset G} \mathsf{I}(F' \equiv F).$$

  Here $\equiv$ means 'isomorphic' to, which means that F' can be mapped to F.
- It has been shown that $X_F(G)$ has an asymptotically Gaussian distribution as long as F is a strictly balanced graph.
- The functional $t(F, G)$ is defined for two graphs F and G as the proportion of all mappings $V(F) \to V(G)$ that are graph homomorphisms $F \to G$, i.e., map adjacent vertices to adjacent vertices.
- In probabilistic terms, $t(F, G)$ is the probability that a uniform random mapping $V(F) \to V(G)$ is a graph homomorphism. Assuming F is labelled and $k = |v(F)|$ then we define

$$t(F, G) \equiv \Pr\{F \subseteq G[k]\}.$$

- Already for the Erdős–Renyi model we have discussed estimating

$$\widehat{\rho} = \frac{\sum_{i<j} A_{ij}}{\binom{n}{2}}.$$

- We can calculate the moments of this estimator. We have from the independence of the trials

$$\mathbb{E}\,\widehat{\rho} = \frac{\sum_{i<j} \rho}{\binom{n}{2}} = \rho \tag{7}$$

$$\mathbb{V}\mathrm{ar}\,\widehat{\rho} = \frac{\sum_{i<j} \rho(1-\rho)}{\binom{n}{2}^2} = \frac{\rho(1-\rho)}{\binom{n}{2}}. \tag{8}$$

- Using standard Central Limit Theorems we can deduce that a function of $\sum_{i<j} A_{ij}$ becomes Gaussian if ρ is sufficiently large. We have

$$\frac{1}{\sqrt{\binom{n}{2}\rho(1-\rho)}}\left\{\sum_{i<j} A_{ij} - \binom{n}{2}\rho\right\} \xrightarrow{L} N(0,1).$$

Rucinski discusses under what conditions a Poisson limit follows.

- For the Chung–Lu model we can now estimate

$$\widehat{\pi}_i = \frac{d_i}{\sqrt{\|d\|_1}}, \quad i = 1, \ldots, n.$$

- We can with this model determine that d_i becomes Gaussian under suitable conditions.

- We can now look at non–parametric statistics in this setting.
- Often one wishes to determine how important a given vertex is.
- We define the graph distance between nodes u and v in graph G:

$$\text{dist}_G(u, v) = \text{minimal number of edges linking } u \text{ and } v.$$

If no path exists between u and v then the distance is set to infinity.

- Then we define closeness centrality of vertex i in graph G with n nodes as

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

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

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

- We define the betweenness centrality of vertex i in graph G with n nodes in terms of n_{jk}^i as th enumber of shortest paths from j to k that pass through i

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

- We denote by $X_{P_3}(G)$ and $X_{C_3}(G)$ as the number of paths with three nodes, and the number of cycles on three nodes respectively.
- We can then define the clustering coefficient as

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

- Network modularity. We define the network modularity to be

$$\hat{Q}_G = \sum_i \sum_{i<j} \left\{ A_{ij} - \frac{d_i d_j}{\sum_l d_l} \right\} \delta_{z_i z_j}.$$

- This measures how well we have picked z.
- Given z we can estimate the connection probability:

$$\widehat{\theta}_{ab}(z) = \frac{\sum_{i<j} A_{ij} \mathsf{I}(z_i = a)\mathsf{I}(z_j = b)}{\sum_{i<j} \mathsf{I}(z_i = a)\mathsf{I}(z_j = b)}.$$

- We need to learn z from matrix A.
- We shall use spectral clustering to this purpose.

- We calculate the (unnormalized graph) Laplacian from the adjacency matrix:

$$L = \mathrm{diag}\{d_1, \ldots, d_n\} - A.$$

- From this matrix we can define unnormalized spectral clustering.

- Input: Adjacency matrix $A \in \mathbb{R}^{n \times n}$, and a fixed number of clusters k.
- Compute the unnormalized Laplacian L.
- Compute the first k eigenvectors $v_1, \ldots, v_k$ of L.
- Let $V \in \mathbb{R}^{n \times k}$ be the matrix containing vectors $v_1, \ldots, v_k$ as columns.
- For $i = 1, \ldots, n$ let y_i correspond to the ith row of V.
- Cluster the points $(y_i)_i$ in $\mathbb{R}^k$ with the k-means algorithm into clusters $\widetilde{C}_1, \ldots, \widetilde{C}_k$.
- Outputs are clusters $C_1, \ldots, C_k$ with

$$C_i = \{j \,|\, y_j \in C_i\}.$$

Estimating interactions

- There are various modifications of k-means. This leads to the estimated labels $\widehat{z}_i$.

$$\widehat{\theta}_{ab}(z) = \frac{\sum_{i<j} A_{ij} \mathsf{I}(z_i = a)\mathsf{I}(z_j = b)}{\sum_{i<j} \mathsf{I}(z_i = a)\mathsf{I}(z_j = b)}.$$

- We can also define the combinatorial least squares problem:

$$\widehat{z} = \min_{z \in \mathcal{Z}_k} \sum_{i<j} \{A_{ij} - \widehat{\theta}_{ab}(z)\}^2 \mathsf{I}(z_i = a)\mathsf{I}(z_j = b).$$

- Least squares is normally computationally efficient; but the addition of z makes this problem intractable.

- For any two graphs G and G' we have already met the Hamming or edit distance

$$d_1(G, G') = \frac{\|e(G)\Delta e(G')\|}{n^2} = \frac{\|A(G) - A(G')\|_1}{n^2}.$$

  Sometimes the matrix norm is divided by $n(n-1)$ as in networks without self-loops the diagonal is set to zero.
- The Hamming distance treats as uniform all changes in the graph structure, whether they are addition in edges or deletions.
- The distance is also sensitive to the edge density of the graphs.

- To adapt to the sparsity we need to rescale the distance.
- We use Jaccard distance (Levandowsky and Winter (1971)):

$$d_J(G, G') = \frac{\|A(G) - A(G')\|_1}{\|A(G) + A(G')\|_*},$$

  where the latter is the nuclear norm.
- Details were provided for calculating expected counts with the blockmodels, and ERGMs were introduced.

Latent Space Models

- We can also assume (see e.g. Hoff (2002)) that conditionally on unobserved latent variables z_i (unobserved positions in a latent space) and x_{ij} we have

$$\Pr\{A|\mathbf{z}, \mathbf{X}, \boldsymbol{\theta}) = \prod_{i<j} \Pr\{a_{ij}|z_i, z_j, x_{ij}, \boldsymbol{\theta}).$$

- We can use the GLM parameterisation to do this.
- Latent class/stochastic block model fits in this framework, latent distance models and the latent eigenmodel.
- The latent space model can be rewritten as a random dot produce graph.

Network Sampling

- Relational or edge sampling. Relational sampling corresponds to sampling exactly relations or edges. This could be sampling phone calls. In many networks applications the relations are the primitive objects and the vertices are derivative from these.
- Hyperedge sampling. Sampling academic articles from a research repository involves more than actor in every relationship. Then every article represents an hyperedge.
- Path sampling. In the early days of network sciecne it was thought that one could ascertain network topology by analyzing the paths traversed when sending information from one part of the Internet to another.
- Snowball sampling.
- We discussed the Preferential attachment network generative mechanism.

Additional topics:

- Multilayer networks.
- Directed networks.
- Hypergraphs.
- Link prediction.
- Biclustering problems.
- Alternative forms of exchangeability.