Statistical analysis of network data lecture 1

Sofia Olhede

February 24, 2021

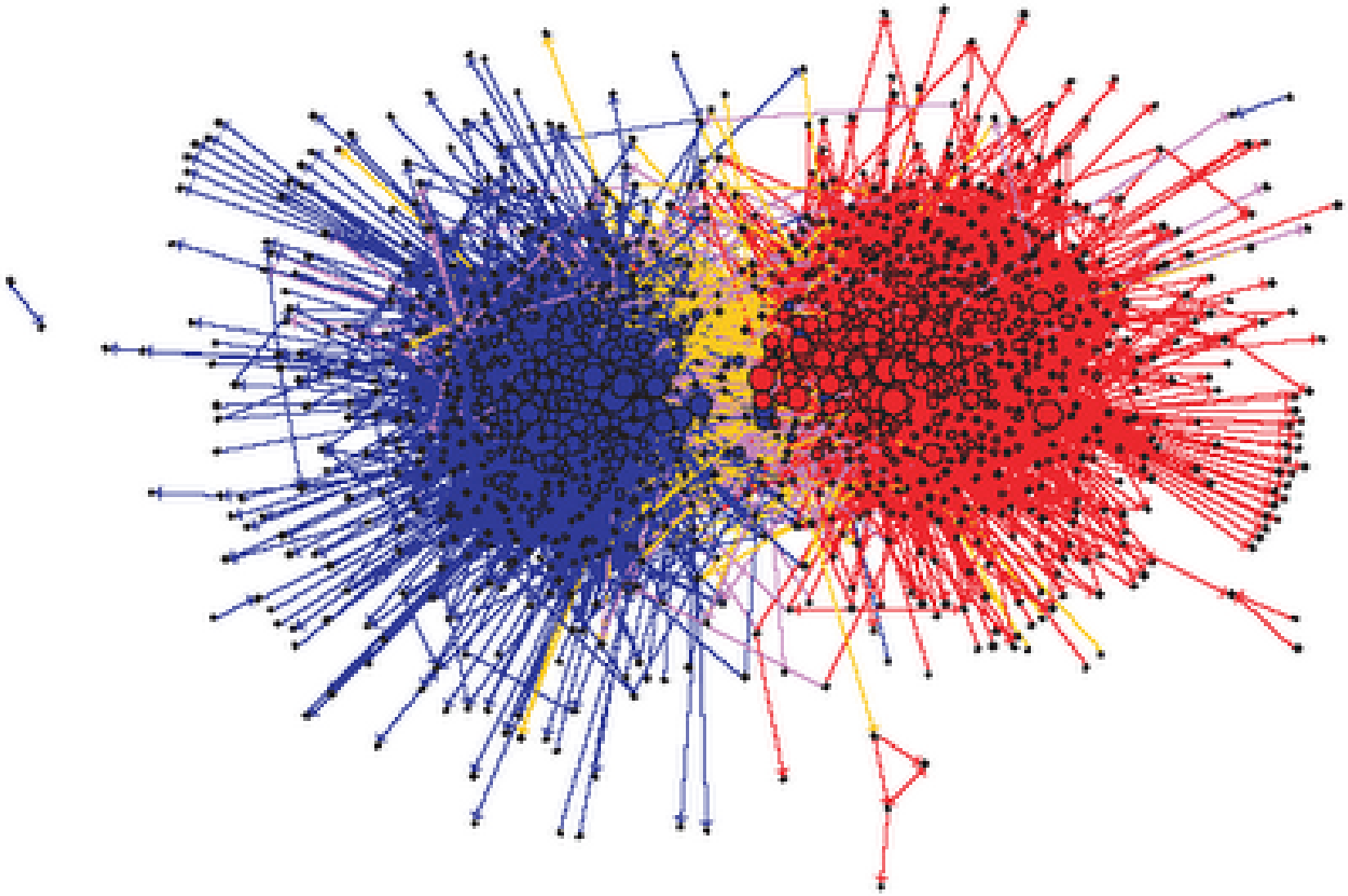
1. Practical Things
2. Graphs or network data structures
3. Network data models

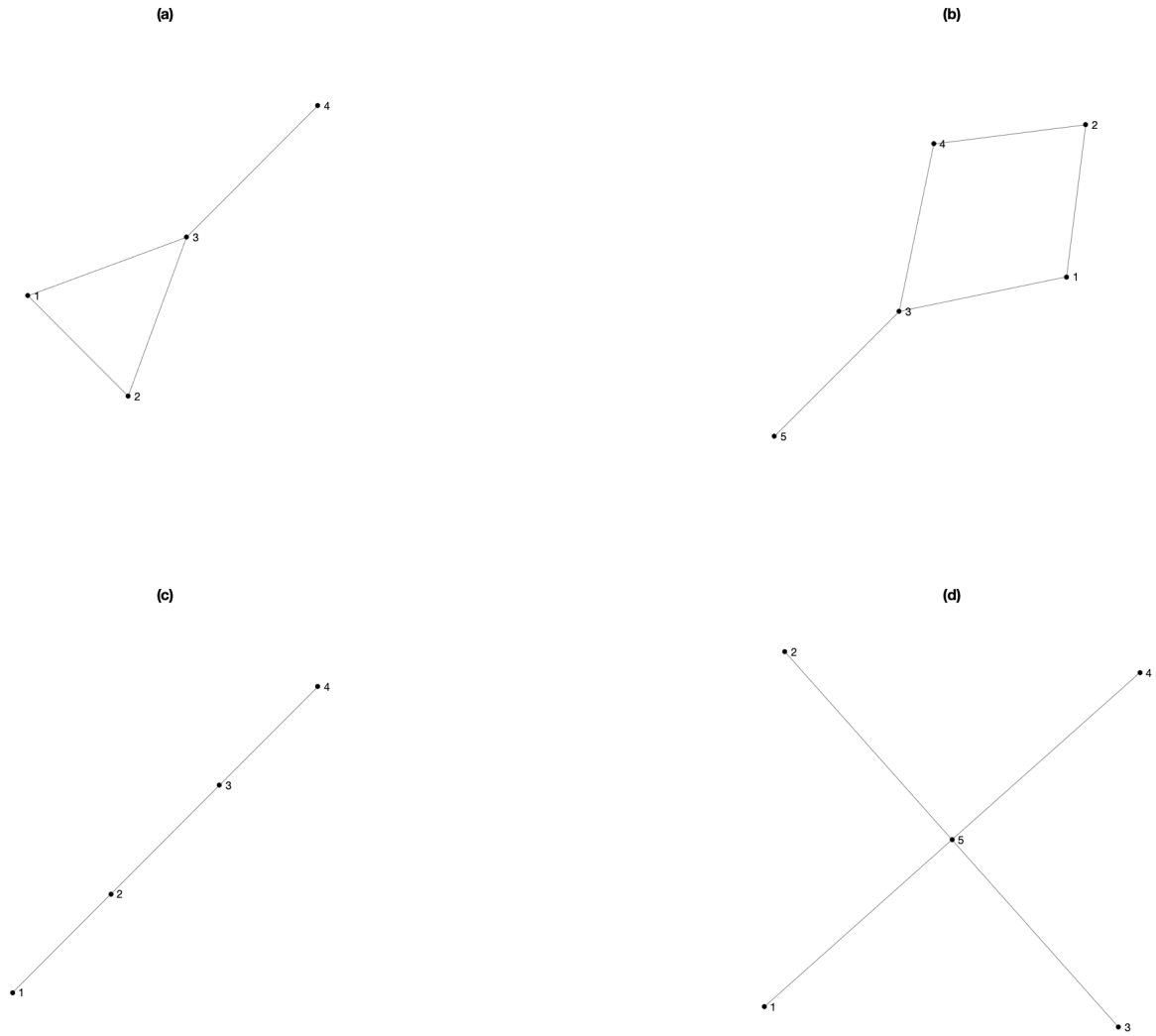
- This course will until further notice be given online on zoom; lectures will be uploaded on switchtube.
- Problem Sheets and lecture notes will be uploaded on Moodle.
- The structure of the course is as follows:
- The data structure of a network.
- Simple statistical models for network data.
- Probabilistic invariances; exchangeability.
- Statistics on networks, subgraph counts.
- Distributions of subgraph counts.
- Fitting network models.
- Nonparametric network summaries.

- Clustering networks, fitting blockmodels.
- Network metrics.
- Exponential random graphs.
- Latent space models.
- Network Sampling.
- Spreading on graphs.
- Directed networks.
- Hypergraphs.
- Link prediction.
- Biclustering.
- Alternative forms of exchangeability.

- Books on this topic:
- E.D. Kolaczyk: Statistical Analysis of Network Data. Springer, 2009
- E.D. Kolaczyk: Topics at the Frontier of Statistics and Network Analysis: (Re)Visiting The Foundations (SemStat Elements)
- R. van der Hofstad. Random Graphs and Complex Networks Volume One, 2016
- There is an exam in the summer for the course, no midterm.

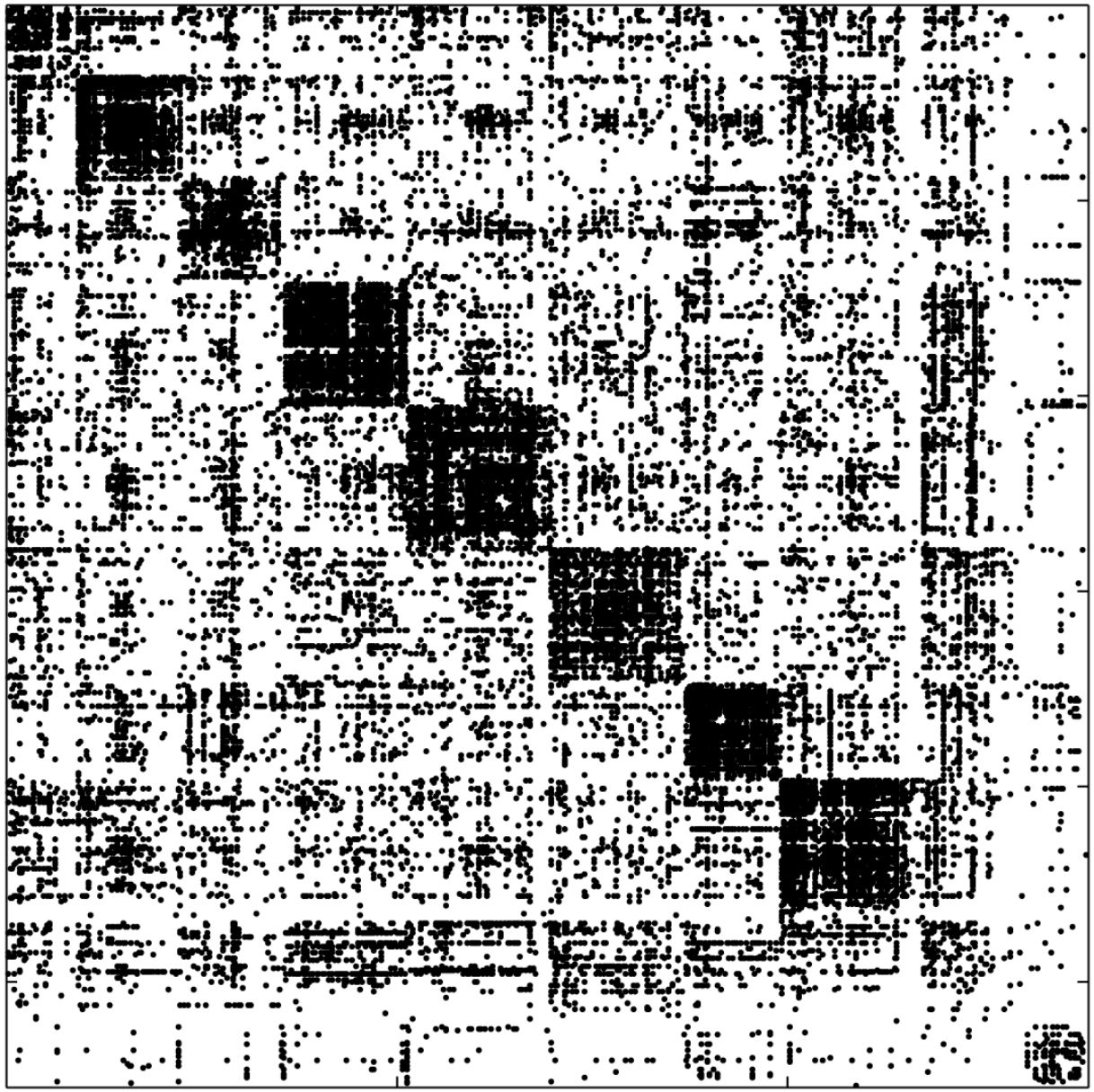
- This course concerns the stochastic properties of network data.
- A network represents interactions between entities (nodes or vertices), where the presence of an interaction is indicated by an edge.
- A network (or graph) G is a pair $G = (V, E)$ of sets so that $E \subseteq [V]^2$. We refer to the elements of V as the vertices (or nodes) of the graph, and E are the edges of G, written as $V(G)$ and $E(G)$.
- A vertex v is incident with an edge e if $v \in e$.
- Two vertices are adjacent or neighbours if connected by an edge.

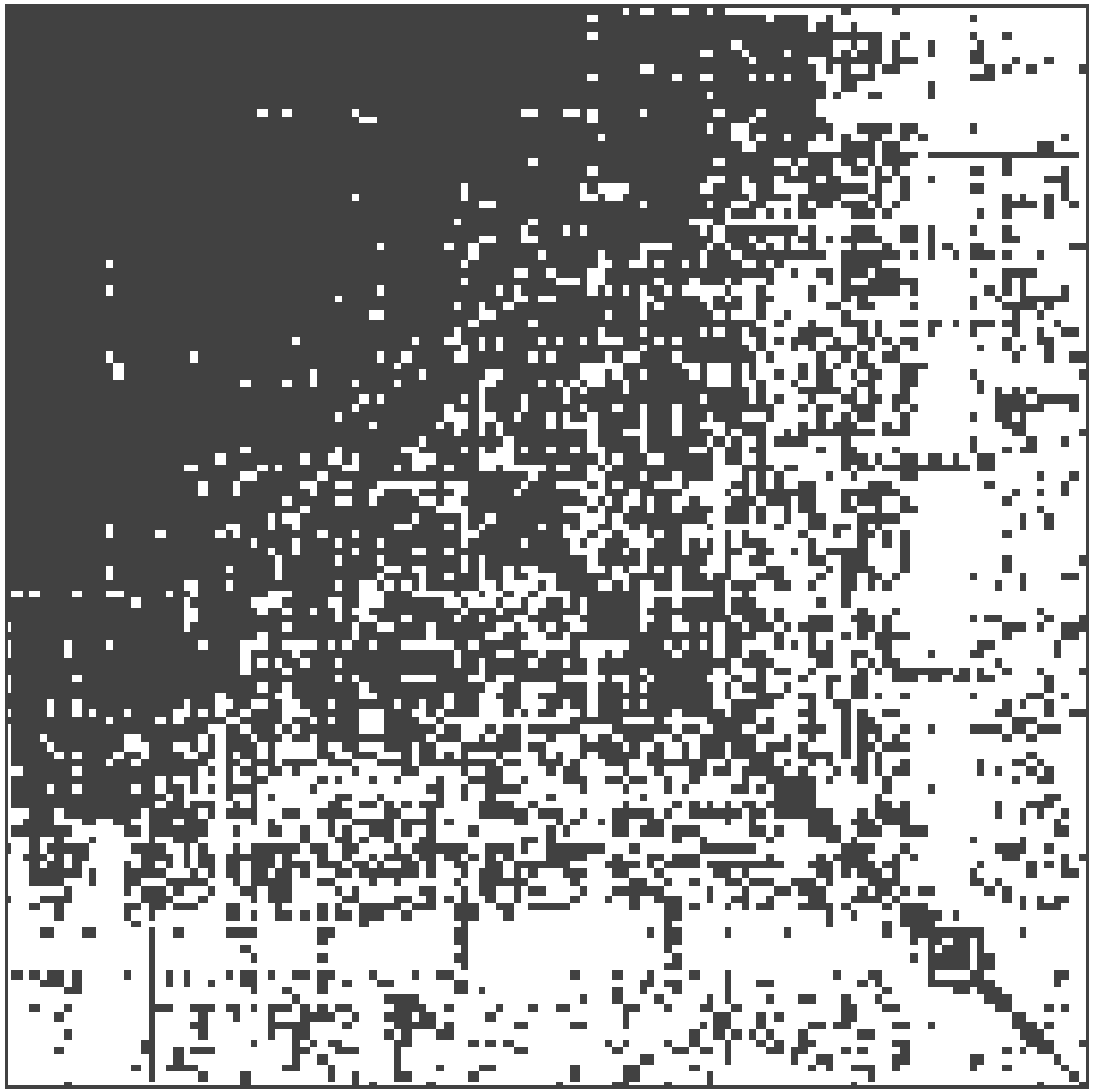


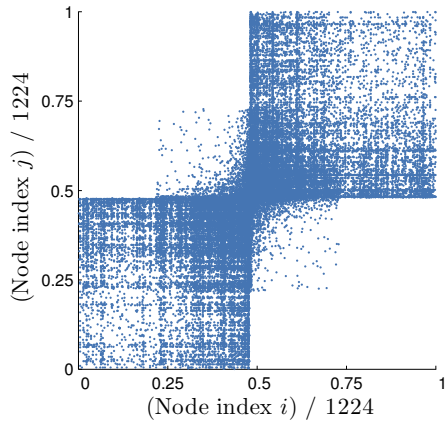


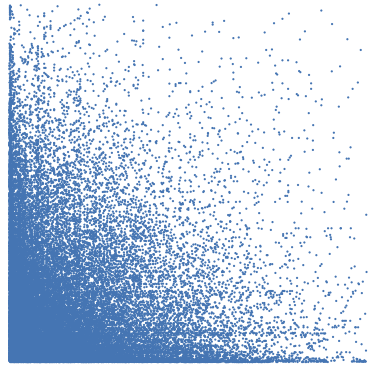
Labelled graphs or networks.

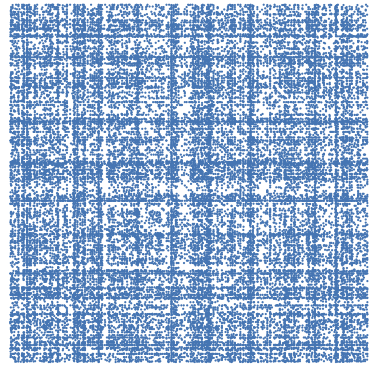
- We write a network or a graph as G, which is usually represented by an adjacency matrix A. The edge–variable A_{ij} where if node i and node j are linked A_{ij} takes the value unity, otherwise it takes the value zero.
- A network can also be represented by a list of edges, an edge list, that just specifies the existing edges, e.g. $\{(1, 15), (1, 32), ...\}$.
- We normally assume that $|V(G)| = n$ if not specified otherwise.
- A graph $H = (V(H), V(H))$ is a subgraph of G if $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$.



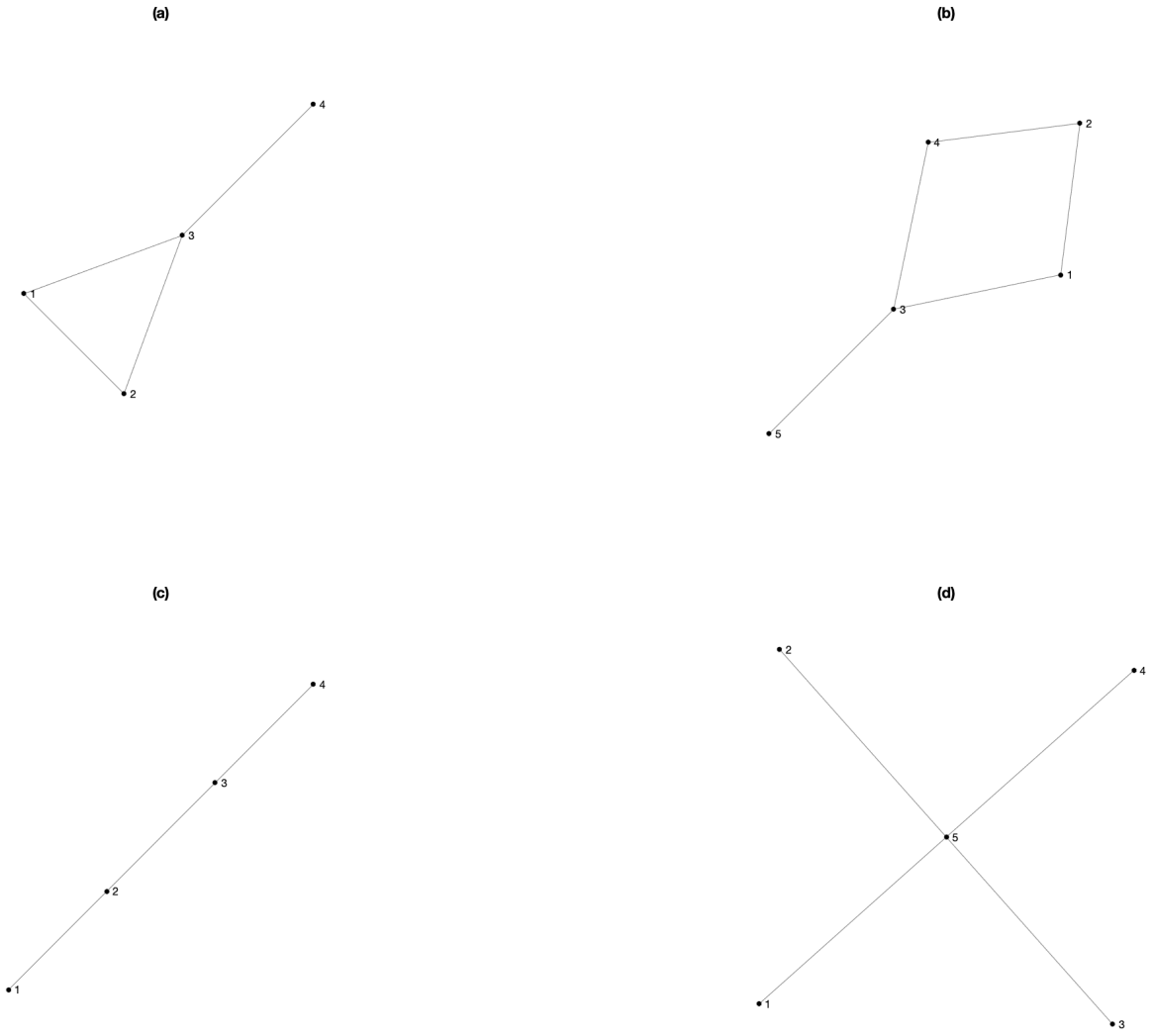



1
0.75
0.5
0.25
0
0
0.25
0.5
0.75
1
(Node index j) / 1224
(Node index i) / 1224





What is the adjacency matrix for these graphs? What is the difference of labelled and unlabelled graphs?



- 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. C is not a square matrix.
- 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 total number of edges in a network is twice the sum of degrees $2|E| = \sum_i d_i$.
- A path/walk on a graph G from vertex v_0 to v_1 is an alternating sequence $\{v_0, e_1, v_1, e_2, \ldots, v_l\}$.
- A graph is connected if it is not empty and any two vertices in the network are connected by a path/walk.
- if we equip edges with weights then we decorate the graph with auxiliary numerical values.

- The simplest graph model 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 $1 \leq j < i \leq n$.
- Erdős-Rényi networks can also be generated by rewiring, or randomly making pN connections. Either mechanism is equivalent.
- With this model the degree d_i is $\mathrm{Bin}(n-1, p)$. See exercise sheet 1.
- The first summary of a network is a degree distribution. The degree distribution is simply a histogram of the degrees, as if they were independent.
- This leads to degrees that have an equal distribution. In real networks the degrees are quite different. To encompass this heterogeneity we cannot make all edges have the same distribution.

- The simplest generalization introduces n parameters π_i and then generates edges independently by

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

  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.
- What do we chose for π_i? A common choice is to take $\pi_i = \theta i^{-\gamma}$.
- Some networks also have a strong latent spatial structure. A Random Geometric Graph is an example of such a model. This is generated by randomly placing n locations x_i in a specified spatial domain, specifying a radius $r > 0$ and connecting node i and j (e.g. setting $A_{ij} = 1$ if with some specified distance $d(x, y)$ if $d(x_i, x_j) < r$.

- 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. Namely, that for most of the enumerated model the value of i or j contain no information about the model structure. Thus if we introduce a permutation that remaps all indices, the nature of the model should not change.
- Permutation invariance is a stochastic invariance. What other examples have you met?
- Let Π be a permutation on the ordering so that $\Pi(\{1,\ldots,n\}) = \{\pi(1),\ldots,\pi(n)\}$, 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 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.*

- Furthermore this can be related to the underlying array.

Definition

Let E be a suitable space. A sequence of E-valued random variables $(X_n)_{n\in\mathbb{N}}$ is exchangeable if

$$(X_n)_n \stackrel{d}{=} (X_{\Pi(n)})_n \quad \forall \Pi \in \mathrm{Sym}(\mathbb{N}),$$

where $\mathrm{Sym}([n])$ is the group of all permutations of $[n]$ and $\mathrm{Sym}(\mathbb{N})$ is the group of all permutations of $\mathbb{N}$.

- Note that this is really an assertion about the measure which is the joint law of the r.v.s (X_n): it is invariant under the action of $\mathrm{Sym}(\mathbb{N})$ by the permutation of coordinates. When $E = \{0, 1\}$ these were studied by de Finetti in the 1930's; for more general E see results by Hewitt and Savage in the 1950's.