Statistical analysis of network data lecture 2

Sofia Olhede

March 3, 2021

1. Exchangeability

2. Statistics for Networks

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 $\Pi(e) := \{\Pi(n_1), \ldots, \Pi(n_k)\}$.

- General arrays were studied by Hoover, Aldous, Fremlin and Talagrand and Kallenberg. Finite exchangeability simply puts e in a finite space. A finite $n \times m$ random matrix A is row-column exchangeable if for n-permutation σ and m-permutations π

$$\begin{aligned}&\Pr\{A_{11} \in N_{11},\ A_{12} \in N_{12}, \ldots, A_{nm} \in N_{nm}\} \\ &= \Pr\{A_{\sigma(1)\pi(1)} \in N_{11},\ A_{\sigma(1)\pi(2)} \in N_{12}, \ldots, A_{\sigma(n)\pi(m)} \in N_{nm}\},\end{aligned}$$

for all Borel sets $N_{11}, \ldots, N_{nm}$.

- For $r > n$ and $q > m$ the matrix A is called (r, q)-extendible if there are matrices T, Z and W with

$$\mathrm{T} = (A_{ij}), \quad i = 1, \ldots, n, \quad j = m+1, \ldots, q \tag{1}$$
$$\mathrm{Z} = (A_{ij}), \quad i = n+1, \ldots, r, \quad j = 1, \ldots, m \tag{2}$$
$$\mathrm{W} = (A_{ij}), \quad i = n+1, \ldots, r, \quad j = m+1, \ldots, q, \tag{3}$$

such that

$$\mathrm{A}^* = \begin{pmatrix} \mathrm{A} & \mathrm{T} \\ \mathrm{Z} & \mathrm{W} \end{pmatrix},$$

is row-column exchangeable. A matrix that is (r, q) extendible for all $r > n$ and $q > m$ is called infinitely extensible.

Theorem (Aldous Hoover)

An array 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}.

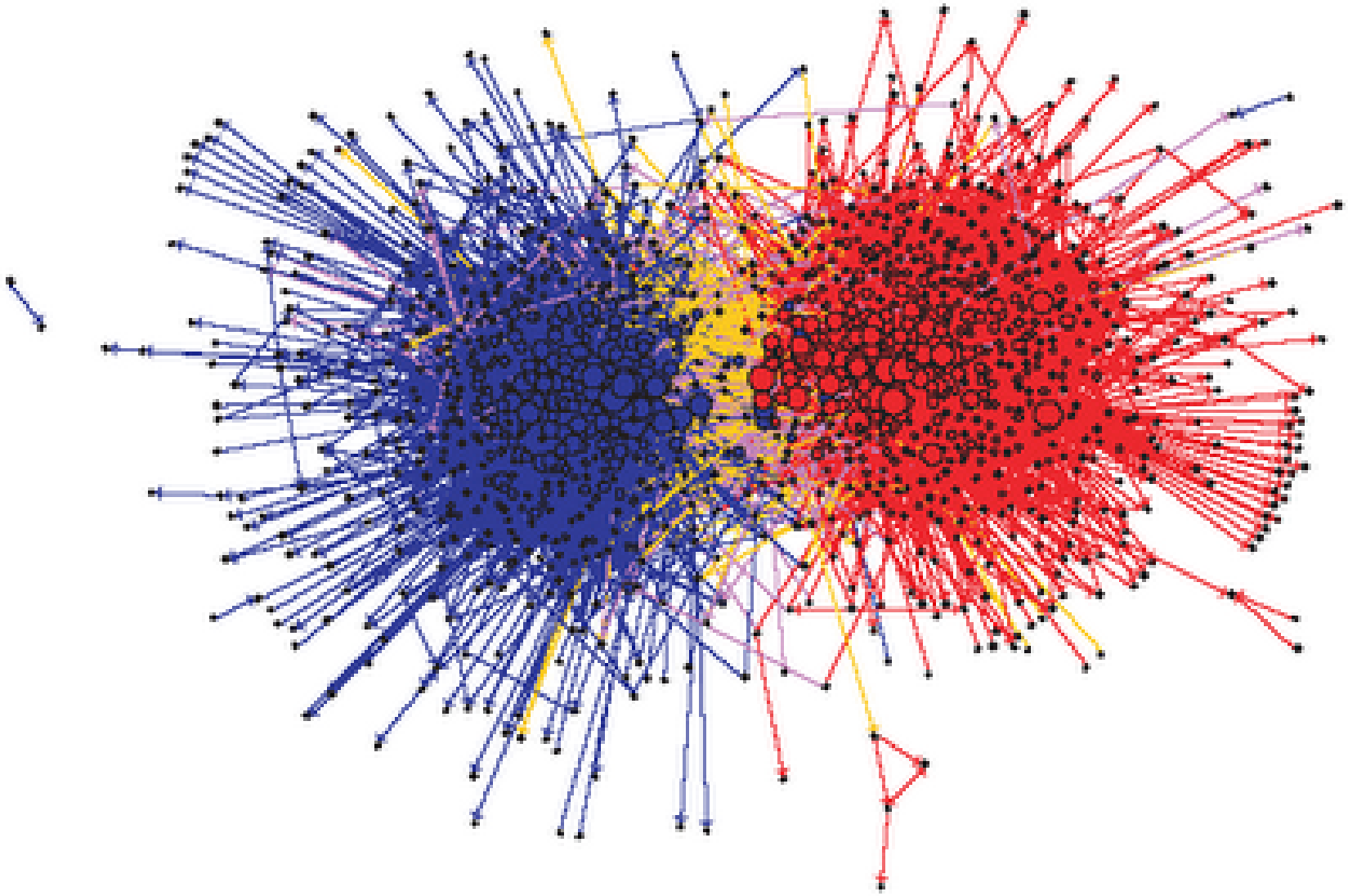
- To define statistical summaries for networks we need to understand what natural network summaries are.
- A basic concept is to count the occurance of subgraphs in G. A subgraph count simply corresponds to

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

  Here $\equiv$ means 'isomorphic' to, which means that F' can be mapped to F.
- There are more than one form of mapping; $\equiv$ can imply that there is a mapping between F' and F. We can constrain our mapping to be a graph homomorphism that maps adjacent vertices to adjacent vertices, and define an injective map if greater specificity is required.

- The problem with $X_F(G)$ is that it is not invariant to the number of nodes n, or the marginal probability of getting an edge $\rho = \Pr(Aij = 1) > 0$.
- For example if A is generated as an Erdős-Rényi graph with marginal edge probability ρ then with K_n the complete graph on n nodes then

$$\begin{aligned}
\mathbb{E}\{X_F(G)\} &= \mathbb{E}\left\{\sum_{F'\subset G} I\{F \equiv F'\}\right\} \\
&= \mathbb{E}\left\{\sum_{F'\subset K_n} I\{F \equiv F'\} I\{F' \subseteq G\}\right\} \\
&= \sum_{F'\subset K_n} I\{F \equiv F'\} \Pr\{F' \subseteq G\} \\
&= \sum_{F'\subset K_n} I\{F \equiv F'\} \Pr\{F \subseteq G\} \\
&= \Pr\{F \subseteq G\} \sum_{F'\subset K_n} I\{F \equiv F'\} = \rho^{|e(F)|} \sum_{F'\subset K_n} I\{F \equiv F'\}.
\end{aligned}$$



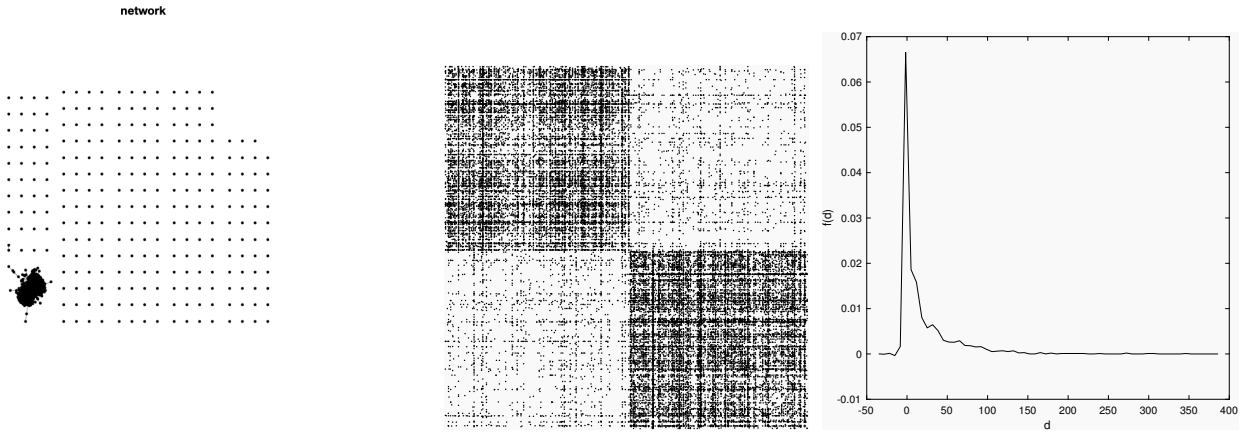


Figure 11: A spaghetti plot, an adjacency matrix spy plot, and a density estimate of degrees for the political blogs data.

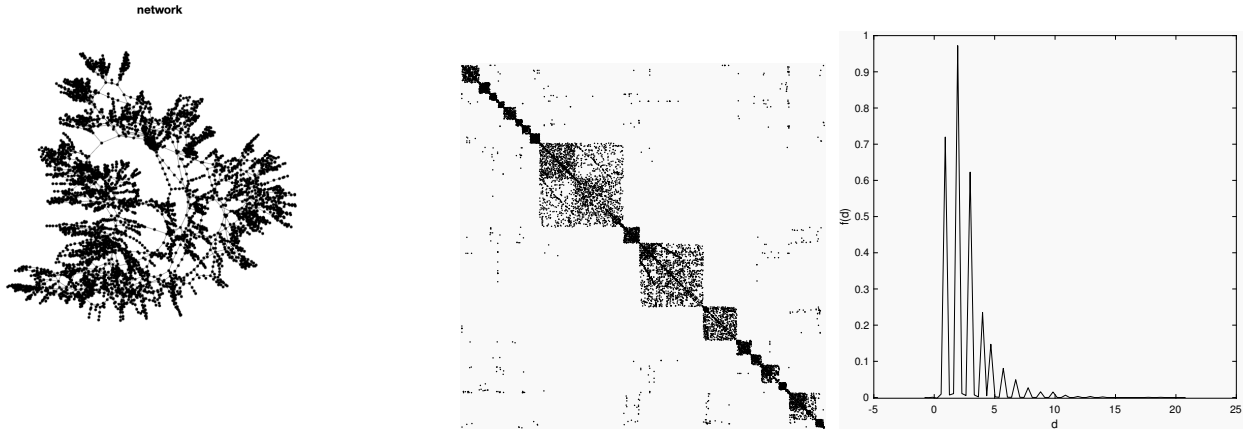


Figure 12: A spaghetti plot, an adjacency matrix spy plot, and a density estimate of degrees for the power grid network in the US.

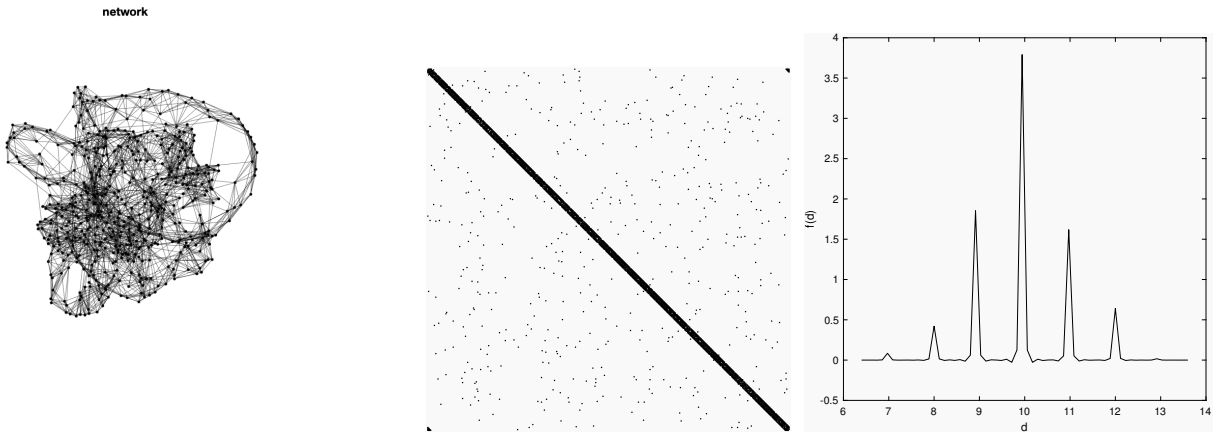


Figure 13: A spaghetti plot, an adjacency matrix spy plot, and a density estimate of degrees for a Watts-Strogatz network.

- Note that $\sum_{F' \subset K_n} I\{F \equiv F'\} = X_F(K_n)$.
- This tells us how $X_F(G)$ scales with ρ. In addition $X_F(K_n)$ is deterministic and takes the form of $(n)_{|v(F)|}/\mathrm{aut}(F)$: The symbol $(n)_k$ is interpreted as $(x)_y = x(x-1)\ldots(x-y+1)$. There is a lot of confusion regarding rising/falling factorials and Pochhammer symbols (see e.g. the Wikipedia page!)
- We write $\mathrm{aut}(G)$ for

$$|\{\Phi \in \mathrm{Sym}(v(G)) : pq \in e(G) \Leftrightarrow \Phi(p)\Phi(q) \in e(G)\}|$$

  the order of the auto–morphism group of G; i.e., the number of adjacency–preserving permutations of $v(G)$.
- It has been shown that $X_F(G)$ has an asymptotically Gaussian distribution as long as F is a strictly balanced graph.

- Define the average degree of a graph on n nodes with e edges as $\overline{d} = 2e/n$. Define the maximum average degree of a graph F as

$$m(F) = \max_{H \subset F} \{\overline{d}(H)\}.$$

Definition

Strictly balanced graph A graph H is balanced if $m(H) = d(H)$ and is strictly balanced if $F \subseteq H$ and $\overline{d}(F) = m(H)$ implies that $F = H$.

- Trees, cycles and complete graphs are strictly balanced. See Random Graphs by Bollobas for more details.

Theorem

Poisson Limit Suppose that H is a fixed strictly balanced graph with k vertices and $l \geq 2$ edges, and its automorphism group has a members. Let $c > 0$ be a constant and set $p = c \cdot n^{-k/l}$. For G generated as an Erdős-Rényi graph with success probability p denote by $X_F(G)$ the number of copies of F in G. Then $X_F(G)$ asymptotically becomes a Poisson random variable with mean $c^l/\mathrm{aut}(H) = a$.

- Note that the edge probability p has been specified so that $\mathbb{E}\, X_F(G) = (c \cdot n^{-k/l})^l (n)_k / \mathrm{aut}(H) \approx c^l / \mathrm{aut}(H)$ is an order one quantity.
- We defined the $X_F(G)$ the count of the number of copies of F in G. Note that regular (isomorphic) copies only count adjacency and not non-adjacency. A copy counting non- adjacency is an induced copy.
- Define G_n as the Stochastic blockmodel on Θ with h_c nodes in group c when counting a strictly balanced graph H in G_n. We note

$$\lambda = \mathbb{E}\, X_H(G_n) = \binom{n}{|v(H)|} \frac{(v(H))!}{\mathrm{aut}(H)} \mu(H),$$

where

$$\mu(H) = \sum_{c_1, \ldots c_{v(G)}}^{Q} h_{c_1} \ldots h_{c_{|v(G)|}} \prod_{uv} \theta_{c_u c_v}.$$

It has been shown by Coulson et al (2015) that the distribution of $X_H(G_n))$ becomes Poisson with mean λ.