Statistical analysis of network data lecture 3

Sofia Olhede

March 9, 2021

1. Graph Statistics
2. Network Descriptors
3. Fitting network models
4. Non–Parametric summaries

- If G is an unlabelled graph and $v_1, \ldots v_k$ are sequences of vertices in G, then $G(v_1, \ldots, v_k)$ is obtained by randomly sampling the vertices uniformly. We allow $v_i = v_j$ but no self-loops.
- Let $G[k]$ for $k \geq 1$ be the random graph obtained by sampling $v_1, \ldots v_k$ uniformly at random among the vertices of G, with replacement.
- For $k \leq v(G)$, we further let $G[k]'$ be the random graph $G(v'_1, \ldots, v'_k)$ where we sample $v'_1, \ldots, v'_k$ uniformly at random without replacement.
- 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

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

- Note that both F and $G[k]$ are graphs on $[k]$, so the relation $F \subseteq G[k]$ is well-defined as containment of labelled graphs on the same vertex set.
- Although the relation $F \subseteq G[k]$ may depend on the labelling of F, the probability of $t(F; G)$ does not, by symmetry, so $t(F, G)$ is well defined for unlabelled F and G.
- $t_{\mathrm{inj}}(F, G) \equiv \Pr\{F \subseteq G[k]'\}$ and also $t_{\mathrm{ind}}(F, G) \equiv \Pr\{F = G[k]'\}$, for $|v(F)| \leq |v(G)|$, whilst if $|v(F)| > |v(G)|$ then $t(F; G) = 0$.
- Since the probability of a random sample $v_1, \ldots v_k$ of vertices in G contains some repeated vertex is $\leq k^2/(2|v(G)|)$ it follows

$$|t(F, G) - t_{\mathrm{inj}}(F, G)| \leq \frac{|v(F)|^2}{2|v(G)|}.$$

- In fact you can define a metric between two networks G_1 and G_2 by computing for an infinite sequence of subgraphs $\{F_j\}$

$$d(G_1, G_2) = \sum_i 2^{-i} |t(F_i, G_1) - t(F_i, G_2)|.$$

- This metric can be shown to be equivalent to the cut metric introduced by Lovasz and collaborators.
- What do people do in practice?
- The hard problem is picking motifs, or subgraphs F_i in practice. However for bioinformatics this is a standard way of comparing graphs. Often variances are computed empirically by using the bootstrap (making smaller networks).
- Various choices of comparison can be made. The connection probability of the ER model is estimated by the proportion of observed edges in the network.
- Also, we calculate the empirical distribution of the degrees in the network and used it as the distribution of the expected degrees $\mathbb{E}\, d_i$. This corresponds to the typical use of the Chung-Lu model.
- These maps are well defined for dense networks, where $|e(G)| = O(n^2)$. However most often we have sparse networks, e,g. we need to renormalize $t(F, G)$ appropriately.

- Denote the proportion of vertices with degree k in G (with the assumption $|v(G)| = n$) by $P_k^{(n)}$. Sometimes $\{P_k^{(n)}\}$ is referred to as the 'degree distribution' of G.
- Then we may write the empirical summary

$$P_k^{(n)} = \frac{1}{n} \sum_i \mathrm{I}\{d_i = k\}.$$

- This has expectation

$$\mathscr{P}_k^{(n)} = \frac{1}{n} \sum_i \Pr\{d_i = k\}.$$

- If we can assume that $\mathscr{P}_k^{(n)} \to p_k$ for some deterministic constants $\{p_k\}$ then we can characterize the degree distribution of G.
- If

$$\lim_{k \to \infty} \frac{\log p_k}{\log(1/k)} = \tau,$$

exists then a random graph G is said to be scale-free with exponent τ.

- 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

$$\mathbb{E}\,\widehat{\rho} = \frac{\sum_{i<j} \rho}{\binom{n}{2}} = \rho \tag{1}$$

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

- 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, and when a Gaussian limit is appropriate, following from the schedule in ρ as a function of n.
- What about 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.
- As the denominator converges in distribution to a non-zero constant, this leads us to keep the Gaussian distribution for the estimator.
- This is not a standard set-up as the number of parameters (dimensionality of π) increases with n.

- We can start therefore by understanding d_i better.
- Each network degree is a sum of $n-1$ $\text{Bernoulli}(\pi_i\pi_j)$ random variables.
- The probability that d_i takes the value k is the sum of all distinct ways in which k successes can occur in $n-1$ $\text{Bernoulli}(\pi_i\pi_j)$ trials.
- This yields a Poisson–Binomial distribution; a generalization of the Binomial distribution to summing Bernoullis with different success probabilities.
- Key to understanding this distribution are the moments of the distribution.

- We can note that if π is a deterministic vector of parameters, and let d_i be the degree vector of an n-node simple graph whose edges are independent Bernoulli($\pi_i\pi_j$) trials then

$$\mathbb{E}\{d_i\} = \pi_i(\|\pi\|_1 - \pi_i)$$
$$\mathbb{V}\text{ar}\{d_i\} = \pi_i(\|\pi\|_1 - \pi_i) - \pi_i^2 \sum_{j \neq i} \pi_j^2$$
$$\mathbb{C}\text{ov}\{d_i, d_j\} = \pi_i\pi_j(1 - \pi_i\pi_j).$$

A key characterizer of the distribution of d_i is its dispersion, e.g.

$$\frac{\mathbb{V}\text{ar}\{d_i\}}{\mathbb{E}\{d_i\}} = 1 - \pi_i \frac{\sum_{j \neq i} \pi_j^2}{(\|\pi\|_1 - \pi_i)}.$$

- Before describing more complex network statistics, let us note some more non-parametric statistics.
- Often one wishes to know how central or important some vertices are. Degrees are one way of determining how important a particular node i is.
- Define $\mathrm{dist}_G(u, v)$ as the graph distance between node u and v in G:

  $\mathrm{dist}_G(u, v) = \text{minimal number of edges in any path linking } u \text{ and}$

  Of no path exists then the distance is set to infinity.

- Then we define

Definition

Closeness centrality We define the closeness centrality of vertex i in a graph G with n nodes as

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

- Normally it is assumed that the closeness centrality plays an important role in understanding the functionality of a network, for example when information or disease is spreading on it.
- Sometimes to understand the full network C_i is only averaged over nodes in connected components.

- 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

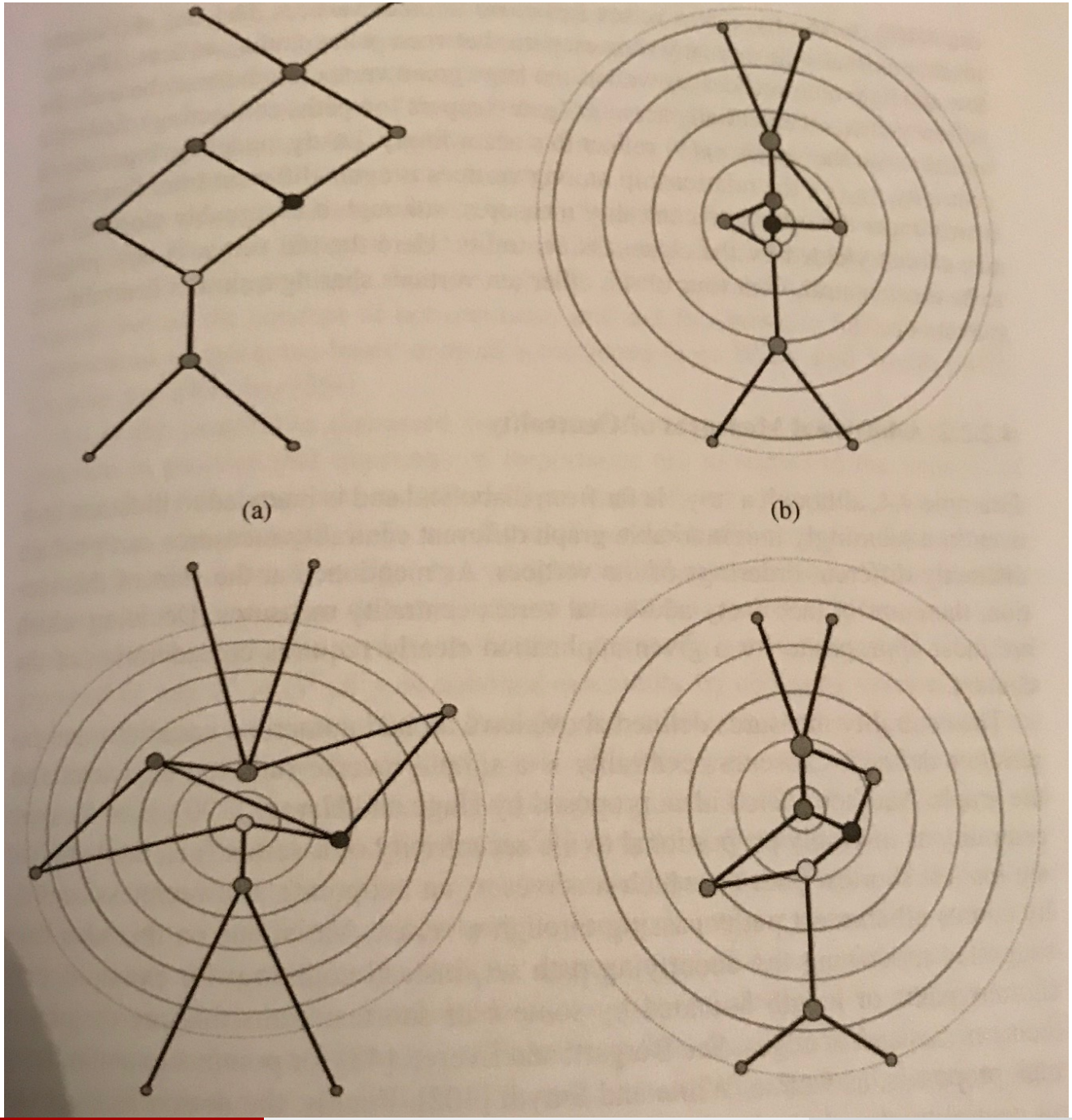
$$B_i = \sum_{1 \leq j < k \leq n} \frac{n_{jk}^i}{n_{jk}}.$$

- 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).