Statistical analysis of network data lecture 13

Sofia Olhede

December 10, 2025

1. Percolation continued
2. Extensions of Exchangeability
3. Graphex
4. Edge Exchangeability
5. Geometry of Networks

Percolation

- Why teach percolation in a random graph course?
- There are many similarities between random graph theory and percolation.
- The main difference is that we have been treating discrete rather than continuous percolation.
- Here our underlying discrete structure is a fixed infinite lattice where bonds can appear only between a site and its lattice neighbours.
- In the case of a finite rectangular lattice each site can have two, three or four nearest neighbours; for a finite graph as you have seen the situation is more complex.

Percolation Probability

- We wish to determine the probability that a given site is contained in an infinite open cluster.
- Wlog we take the site to be the origin.
- As the number of bonds tends to infinity we wish to determine the probability that the origin is part of an infinite connected open set in $\mathbb{L}^d$.
- We need another definition:

Definition (Percolation probability)

The percolation probability is defined to be:

$$\theta(p) = \Pr_p\{\mathbf{0} \leftrightarrow \infty\} = \Pr_p\{|\mathcal{C}(\mathbf{0})| = \infty\}, \tag{1}$$

with $|\mathcal{C}|$ $\sharp$ *of sites in* $\mathcal{C}$, $\Pr_p\{|\mathcal{C}(\mathbf{0})| = \infty\} = \lim_{k\to\infty} \Pr_p\{|\mathcal{C}(\mathbf{0})| \geq k\}$.

Percolation Probability

- When $p = 0$ then every bond will be blocked in the lattice. We deduce $\theta(0) = 0$.
- When $p = 1$ then every bond will be open. Thus $\theta(1) = 1$.
- As p increases more bonds will be open, and the possibility of long open paths emerges.
- Writing $\mathcal{C}_p$ as the open cluster of $\mathbf{0}$ with percolation probability p, then we note $\mathcal{C}_p \subset \mathcal{C}_{p'}$ and $\theta(p) < \theta(p')$. We can deduce that $\theta(p)$ is a nondecreasing function of p.
- The general shape of $\theta(p)$ is sketched on the next page.
- There exists a "critical" probability $0 < p_c < 1$ in two dimensions or higher. Because $\theta(p)$ is non-decreasing we may define:

Definition (The critical probability)

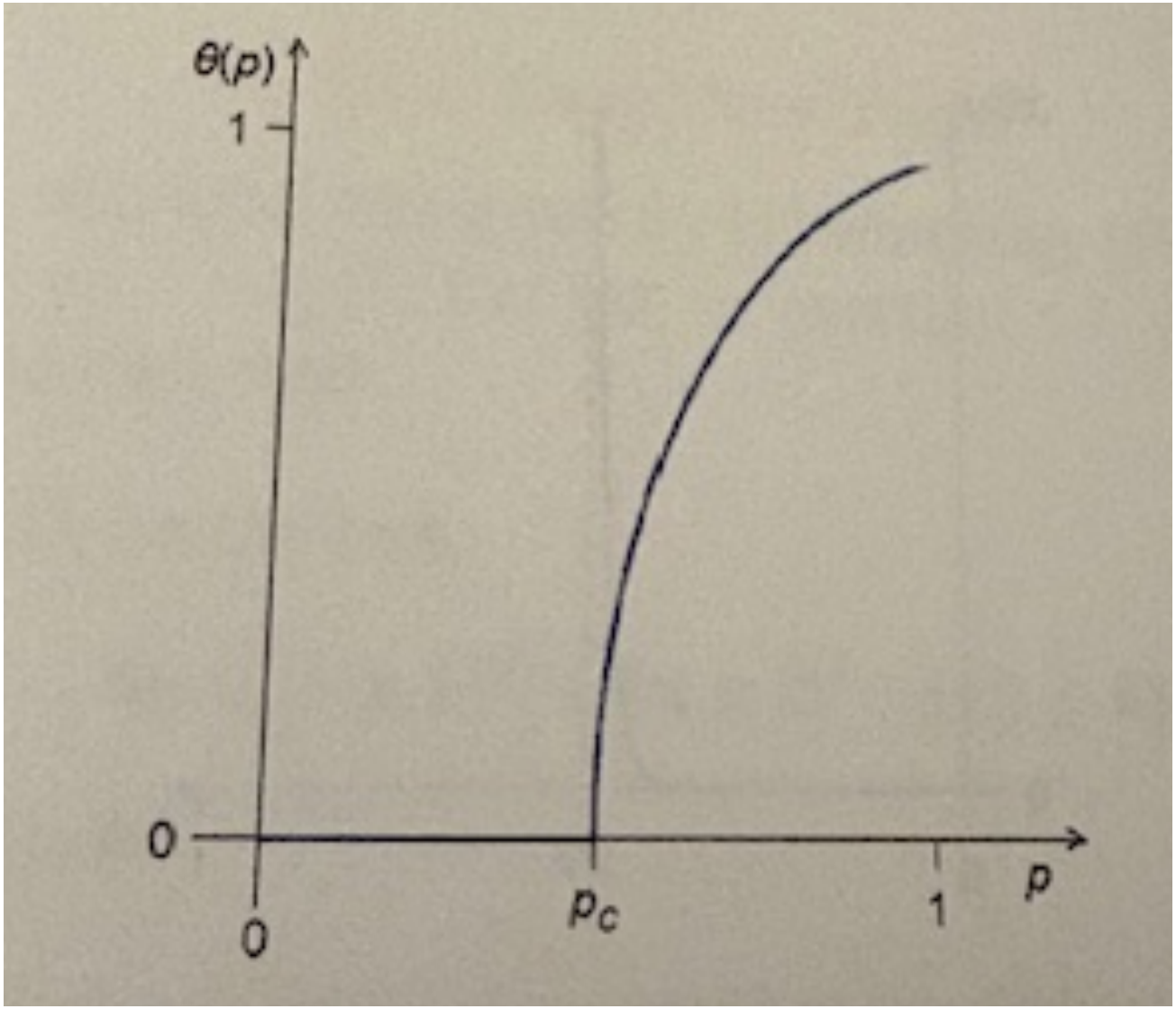
The critical probability is defined as:

$$p_c = \sup_p\{p : \quad \theta(p) = 0\} = \inf_p\{p : \quad \theta(p) > 0\}.$$

Percolation Probability

- The percolation model exhibits a phase transition with two phases:
- A subcritical phase $p < p_c$ where $\theta(p) = 0$ so that each open cluster $\mathcal{C}$ is, almost surely, finite.
- A supercritical phase $p > p_c$ where $\theta(p) > 0$ so that there is a nonzero probability that $\mathcal{C}$ is an infinite open cluster.
- A critical point at $p = p_c$.
- When $\theta(p) > 0$ it is customary to say "percolation occurs".
- When $\theta(p) = 0$ it is customary to say "percolation does not occur".

Percolation Probability



- We have noticed that permutation invariance is a natural probabilistic symmetry for networks.
- We can either consider finite exchangeability, or a finite sample from an infinite exchangeable array.
- Exchangeability has a number of drawback:

  (a) Sparsity: The simplest (vertex) exchangeable representation is not consistent with sparsity;
  (b) Internal Heterogeneity: Neither are simple (vertex) exchangeable models easily brought together with "power-law degrees"
  (c) External Heterogeneity: It is more challenging to bring exchangeability together with covariates and other information.

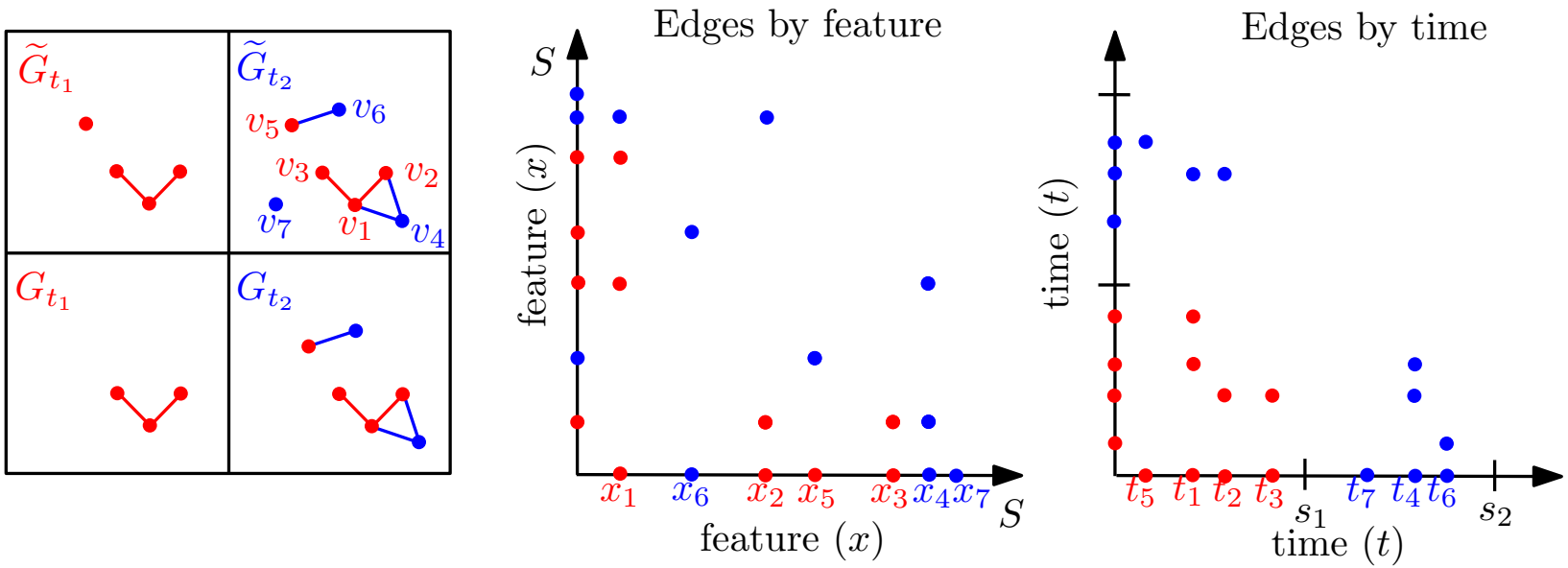
- What alternatives are there?

- The graphon Aldous–Hoover framework used a function $f(x, y), [0, 1]^2 \mapsto [0, 1]$ to parameterise the distrubtion of $\{A_{ij}\}_{i>j}$.
- As an alternative a network can be constructed from a point process.
- We define a point process $Y_t \subset [0, t] \times [0, t]$.
- We say Y_t is exchangeable if its distribution is unchanged by applying any measure preserving transformation.
- We observe A where A_{ij} is unity if $(\theta_i, \theta_j) \in Y$, where $\theta_1, \theta_2, \ldots$ are the arrival times of vertices in the point process.
- $\{Y_t\}$ being stationary does not mean X is exchangeable.
- The occurence of $(t, t') \in y$ means that there is an edge between vertices labelled t and t'.
- It is tempting to interpret t as the time a vertex enters the system; this is overinterpreting. But its connection to real data is easier to make with this interpretation.

- So then what? How do we get sparse and powerlaw?
- A graphex is a triplet (I, S, W). $I \in [0, \infty)$ (non-negative real number), $S: [0, \infty) \mapsto [0, \infty)$ is a measurable map, and $W: [0, \infty)^2 \mapsto [0, 1]^2$, a graphon function.
- We take realisations of a unit-rate point processes $\Xi = \{(\theta_i, \theta_j)\}$, $\Xi_i' = \{(\sigma_{ij}, \chi_{ij})\}$ and $\Xi'' = \{(\rho_j, \rho_j', \eta_j\}_j$. θ_i, σ_{ij} as potential vertex labels, while can be regarded as types of the corresponding labels.
- The point process determines the stochastic properties of θ_i.
- We then realize a graph A by a suitable combination of the three random arrays.
- The most important edges are present with probability/conditional expectation

$$\Pr\{W(\theta_i, \theta_j) \leq \zeta_{ij}\},$$

where $\{\zeta_{ij}\}$ are iid uniform random variables. Additional edges are present due to i) isolated 'stars' and ii) isolated 'edges'.



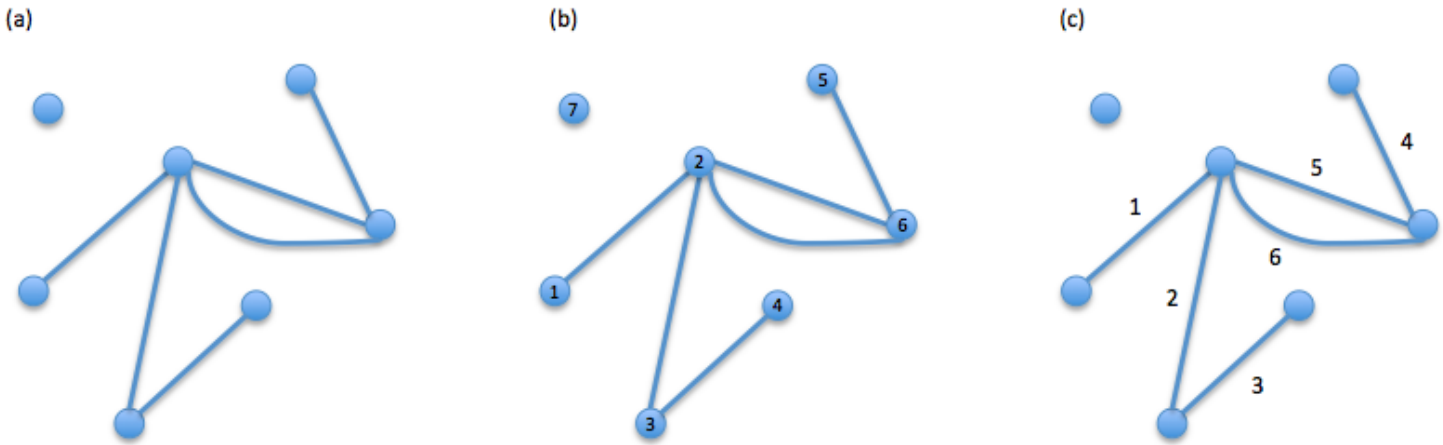
From Borgs, Chayes, Cohn and Holden 2018.

Edge Exchangeable RVs

- OK, but what if we start to observe edges rather that nodes?
- We assume we observe an edge list $E_n = \{E_i\}_{i\in[n]}$. Also assume we have permutations $\sigma : [n] \mapsto [n]$ and define $E_n^\sigma = \{E_{\sigma(i)}\}_{i\in[n]}$.
- We write a random edge labeled network $Y = \{Y_i\}_{i\in\mathbb{N}}$.
- We then have the following definition:

Definition (Edge exchangeability)

A randomly edge labeled network $\mathcal{Y} = \{\mathcal{Y}_i\}_{i\in\mathbb{N}}$ is edge exchangeable if $\mathcal{Y}^\sigma \stackrel{\mathcal{L}}{=} \mathcal{Y}$ for all permutations $\sigma : \mathbb{N} \to \mathbb{N}$.



From Crane and Dempsey 2018 (JASA).

- Edge exchangeable models have the same probability assigned to all edge labeled graphs that are equivalent up to a choice of relabeling.
- Any edge labelled network $\mathcal{Y} = \{\mathcal{Y}_i\}_{i\in\mathbb{N}}$ yields a compatible sequence of finite networks $\mathcal{Y} = \{\mathcal{Y}_i\}_{i\in\mathbb{N}}$ by taking $\mathcal{Y}_n = \mathcal{Y}|_n$ to be the restriction of subsampling $[n] \subset \mathbb{N}$. Any such sequence is infinitely edge exchangeable namely $\mathcal{Y}^\sigma \stackrel{\mathcal{L}}{=} \mathcal{Y}$ for all permutation σ $[n] \to [n]$ and $\mathcal{Y}_n|_{[m]} \stackrel{\mathcal{L}}{=} \mathcal{Y}_m$ for all $n \geq m \geq 1$.
- Even if all edges arrive in an exchangeable process, the vertices arrive in biased order weighted by the relative frequency of their occurrence in the network interactions.
- The sample of vertices, therefore, can be argued does not represent an exchangeable draw from the population of vertices.

Geometry of Networks

- We have studied networks in terms of bivariate interactions.
- How can we o/w study a network?
- We shall use d–dimensionak simplices.
- A d–dimensionak simplex is formed by a set of $d+1$ interacting nodes.
- This includes all the subsets of $\delta+1$ nodes (with $\delta < d$) that are called the δ–dimensional faces of the simplex.
- A simplicial complex of dimension d is formed by simplice sof dimension at most d.
- A simplex is a generalization of a triangle to an arbitrary dimensions.
- We write as $\mathcal{Q}_d(N)$ the set of all possible and distinct d–dimensional simplices in a d-dimensional simplicial complex of N nodes while we write as $\mathcal{S}_{d,\delta}$ the set of all δ-dimensional simplices that are in a d-dimensional simplicial complex.

- We shall describe the simplices using an adjacency tensor **A**.
- The adjacency tensor **A** has elements A_α that takes the values $A_\alpha = 0, 1$ which for any simplex $\alpha \in \mathcal{Q}_d(N)$ tells us if the simplex is present ($A_\alpha = 1$) or absent ($A_\alpha = 0$).
- Thus

$$A_\alpha = \begin{cases} 1 & \text{if} \quad \alpha \in \mathcal{S}_{d,\delta} \\ 0 & \text{o/w} \end{cases} .$$

- The generalized degree $k_{d,\delta}(\alpha)$ of a δ–dimensional face α of a d–dimensional simplicial complex corresponds to the number of d–dimensional simplices incident to the δ–face α.
- Thus the generalized degree $k_{d,\delta}(\alpha)$ can be defined as

$$k_{d,\delta}(\alpha) = \sum_{\alpha' \in \mathcal{Q}_d(N) \,|\, \alpha \subset \alpha'} A_{\alpha'}.$$

In this expression we see that the generalized degrees are not independent of each other.

- The generalized degree of a δ-face α is linked to a generalized degree of the δ' dimensional faces that are incident with $\delta' > \delta$.
- We can also use the combinatorial relationship

$$k_{d,\delta}(\alpha) = \frac{1}{\binom{d-\delta}{\delta'-\delta}} \sum_{\alpha' \in \mathcal{Q}_d(N) \,|\, \alpha \subset \alpha'} k_{d,\delta'}(\alpha').$$

- As every d–dimensional simplex belongs to $\binom{d+1}{\delta+1}$, δ–dimensional faces it follows that the generalized degrees satisfy in a simplicial complex with M d–dimensional simplices:

$$\sum_{\alpha \in \mathcal{S}_{d,\delta}} k_{d,\delta}(\alpha) = \binom{d+1}{\delta+1} M.$$

- If we focus on node r then we have

$$k_{d,0}(r) = \sum_{\alpha' \in \mathcal{Q}_d(N) \,|\, r \subset \alpha'} A_{\alpha'}.$$

Thus the generalized degree of the node indicates the number of d-dimensional simplices incident to r.

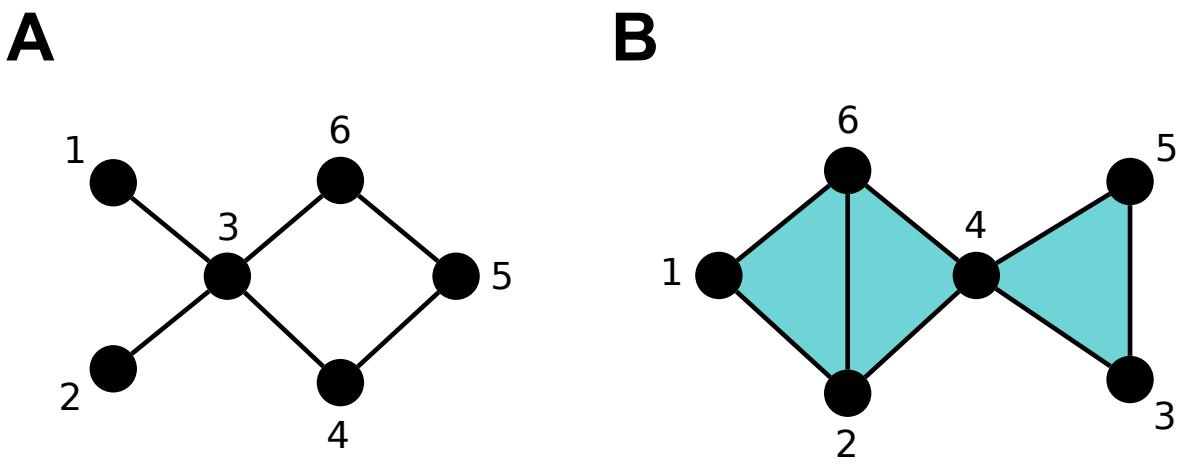
- Since the simplicial complexes under investigation are formed by d-dimensional simplices, the generalized degrees satisfy:

$$\sum_{r=1}^{n} k_{d,0}(r) = (d+1)M.$$

  We recall in this expression that M is the number of d–dimensional simplices in the simplicial complex.
- We shall consider networks such that the number of simplices M is the same order to the number of nodes n.

- If $d = 1$ then the 1-dimensional simplices are nodes and edges.
- The adjacency tensor is then simply the adjacency matrix where element A_{ij} is indicating if edge ij is in the network.
- Then the generalized degrees $k_{1,0}(r)$ is the normal degree of node r. Thus it follows

$$k_{1,0}(r) = \sum_{m=1}^{n} A_{rm}.$$



From Courtney and Bianconi 2016 (Phys Rev E).

- If $d = 2$ then we are summarizing interactions between 3 nodes out of the n.
- This simplicial complex is also determined by its adjacency tensor $\{A_{rmj}\}$.
- We have $A_{rmj} = 1$ if the nodes (r, m, j) are linked by a triangle.
- $A_{rmj} = 0$ if the nodes (r, m, j) are not linked by a triangle.
- There are two types of (generalized) degrees in this instance:

$$k_{2,0}(r) = \sum_{m<j} A_{rmj},$$

and also

$$k_{2,1}(r, m) = \sum_{j} A_{rmj}.$$

- The first type of generalized degree $k_{2,0}(r)$ denotes the triangles that are incident to r.

- The second type of generalized degree $k_{2,1}(r, m)$ denotes the number of triangles next to an edge rm.
- One can deduce that

$$k_{2,0}(r) = \sum_{m<j} A_{rmj} = \frac{1}{2} \sum_{m,j} A_{rmj} = \frac{1}{2} \sum_{m} k_{2,1}(r, m).$$

- Furthermore as each triangle is incident to exactly 3 nodes, we must also have

$$\sum_{r=1}^{n} k_{d,0}(r) = 3M.$$

Recall that we consider networks such that the number of simplices M is the same order to the number of nodes n.

- Bianconi and collaborators suggest a number of variants of the configuration model in this setting.
- For the case of $d = 1$ it then follows that the probability $\Pr\{G\}$:

$$\Pr\{G\} = \frac{1}{Z} e^{-\sum_{r \subset \alpha} \lambda_r}.$$

- The normalizing constant Z can be recovered from

$$\sum_{G} \Pr\{G\} = 1.$$