Random graphs and network statistics

Version 0.91

Lasse Leskelä
Aalto University

March 27, 2019

Abstract

These lectures notes contain material that was used in Fall 2018 during the course *MS-E1603 Random graphs and network statistics* at Aalto University. Many chapters are still under construction, and the notes are subject to updates in the future. The current version also contains several missing references which will be added later.

Contents

Chapter 1

Random and nonrandom graphs

1.1 Introduction

1.1.1 Graph statistics

Graphs are mathematical structures consisting of nodes (vertices, actors, points) and links (edges, dyads, arcs, ties, bonds, lines) between them. In addition, the nodes and the links of a graph are often associated with some attributes (weights, labels, features, categories). Commonly observed graph data include various types of measurements related to biological networks, information networks, financial networks, and social networks. A basic statistical task related to graphs is the following:

- Describe relevant features of a large graph in a compact form using a small set of human-readable numbers, tables, or figures. This type of descriptive statistical tasks are very common in today's society, and might often require a considerable amount of raw computing power, a sophisticated computational algorithm, or advanced visualization techniques.

Instead of merely describing data, statistics can be used for much more. Statistical inference refers to making predictions and decisions based on partially observed and noisy data. Typical statistical inference tasks associated with graphs are:

- Learning global graph features from a partially observed graph. For example, estimate the relative proportion of nodes having an exceptionally large number of neighbors, or predict the frequency of large cliques. This type of techniques can be applied for example to detect anomalies in a noisy data set, or to predict how likely a disease or a rumor is to spread in a large population of animals or in an online social messaging platform.

- Learning node attributes from an observed graph. Based on an observed graph structure (which node pairs are linked and which are not), try to infer the unknown attribute of each node. When a collection of nodes having a common attribute value is identified as a community, this task corresponds to learning the hidden community structure from the graph.

- Learning a graph from observed node attributes. Based on observed measurements of node attribute values, try to infer the unknown graph structure. This type of techniques are widely used in medical studies for example in discovering hidden interactions between different biochemical compounds.

- Learning node attributes from an observed graph and partially observed attributes. This type of problems are typical in phylogenetics, where nodes correspond to individuals or groups of living organisms and node attributes to heritable traits. The goal is to infer attributes of unobserved nodes based on observations of the end nodes of in evolutionary tree graph.

To make learning possible, the unknown variables must somehow be related to the observed variables. Such relations can be analyzed by using a *statistical graph model*. In case there are no attributes (or they are ignored), then a relevant statistical model is a probability distribution $x \mapsto p_\theta(x)$ on the space of all graphs under study, where θ contains the parameters characterizing the model. In contexts with node attributes, a relevant model is a joint probability distribution $(x, z) \mapsto p_\theta(x, z)$ where x refers to a graph structure and z to a list of node attributes.

Because the set of all graphs on a finite node set is finite, computing probabilities and expectations related to a statistical graph model can in principle done by counting sums over finite sets. However, such brute force computing is only feasible for very small graphs, because for example, the set of all undirected graphs on 10 nodes contains $2^{\binom{10}{2}} \geq 3 \times 10^{13}$ elements. This is why we need to work with good statistical graph models and mathematical approximations to handle big data sets.

1.2 Graphs

For most parts, the terminology and notations follow those in [Die17, vdH17]. In these notes, a *graph* is a pair $G = (V, E)$ where V is a finite set of elements called *nodes* (vertices, points), and E is a set of unordered node pairs called *links* (edges, lines, arcs, bonds). Hence there are no loops, no parallel links, nor directed links. The node set of graph G is denoted by $V(G)$ and the link set by $E(G)$. Nodes i and j such that $\{i, j\} \in E(G)$ are called *adjacent*. The *adjacency matrix* of graph G is the square matrix with entries

$$G_{ij} \;=\; \begin{cases} 1, & \text{if } \{i,j\} \in E(G), \\ 0, & \text{else.} \end{cases}$$

Every adjacency matrix is symmetric and has zero diagonal, and the collection such matrices is in one-to-one correspondence with the set of all graphs on node set V. For this reason we will employ the same symbol G both for the graph $G = (V, E)$ and its adjacency matrix $G = (G_{ij})$. See Figure 1.1 for an example.

$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

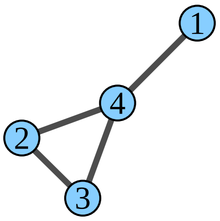

Figure 1.1: The adjacency matrix and a visualization of a graph with node set $V = \{1, 2, 3, 4\}$ and link set $E = \{\{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}\}$.

The *degree* of node i, denoted $\deg_G(i)$, is the number of nodes adjacent to i. The degrees are obtained as rows sums of the adjacency matrix according to

$$\deg_G(i) \;=\; \sum_{j=1}^{n} G_{ij}.$$

In Figure 1.1, node 1 has degree one and the other nodes degree two.

1.3 Random graphs

A *random graph* is a random variable G whose realizations are graphs. Usually the node set is assumed to be nonrandom, in which case a random graph on node set V is a random variable whose realizations belong to the finite set $\mathcal{G}(V)$ of all graphs on node set V. In this case the probability distribution of the random graph G is characterized by the probabilities

$$\mathbb{P}(G = g), \quad g \in \mathcal{G}(V).$$

Here $\mathbb{P}$ refers to a *probability measure* defined on some measurable space on which the random variables under study are defined. The underlying measurable space is usually hidden from the notation, and the reader is expected to recognize from the context which graphs (numbers, vectors, ...) are random and which are nonrandom. A random graph on node set $[n] = \{1, 2, \dots, n\}$ can be identified as a collection of $n(n-1)/2$ binary random variables $G_{ij} \in \{0, 1\}$ indexed by $1 \le i < j \le n$, which constitute the upper diagonal of the adjacency matrix.

Example 1.1 (Bernoulli random graph)**.** Let (p_{ij}) be a symmetric n-by-n matrix with $[0, 1]$-valued entries and zero diagonal, and let G be a random graph on node set $[n] = \{1, \dots, n\}$ in which every node pair $\{i, j\}$ is linked with probability p_{ij}, independently of all other node pairs. The probability distribution of G can be written as

$$\mathbb{P}(G = g) \;=\; \prod_{1 \le i < j \le n} (1 - p_{ij})^{1 - g_{ij}} p_{ij}^{g_{ij}}, \qquad g \in \mathcal{G}_n, \tag{1.1}$$

where $\mathcal{G}_n$ denotes the set of all graphs g on node set $[n]$. A random graph with probability distribution (1.1) is called a *Bernoulli random graph* with rate matrix (p_{ij}). The link indicator G_{ij} of a node pair $\{i, j\}$ is a binary random variable which follows a *Bernoulli distribution* with rate parameter p_{ij}, so that

$$\mathbb{P}(G_{ij} = x) \;=\; \begin{cases} 1 - p_{ij}, & x = 0, \\ p_{ij}, & x = 1. \end{cases}$$

A Bernoulli random graph is called *homogeneous* when all nonzero entries of the rate matrix have the same value, and inhomogeneous otherwise.

Example 1.2 (Uniform random graphs)**.** Let G be a random graph distributed according to

$$\mathbb{P}(G = g) \;=\; \frac{1}{|\mathcal{A}|}, \qquad g \in \mathcal{A},$$

where $\mathcal{A}$ is a subset of the set $\mathcal{G}_n$ of all graphs on node set $[n] = \{1, 2, \dots, n\}$. The following two cases have received the most attention in the literature:

(i) A *uniform random graph with m links* is obtained by setting

$$\mathcal{A} \;=\; \Big\{ g \in \mathcal{G}_n : |E(g)| = m \Big\}.$$

For this graph model, all realizations with the same number of links are equally likely.

(ii) A *uniform random graph with degree list $(d_1, \dots, d_n)$* is obtained by setting

$$\mathcal{A} \;=\; \Big\{ g \in \mathcal{G}_n : \deg_g(i) = d_i \text{ for all } i = 1, \dots, n \Big\},$$

where $d_1, \dots, d_n$ are nonnegative integers such that $\mathcal{A}$ is nonempty. When all entries of the degree list are equal, this model is called a *random regular graph*.

The homogeneous Bernoulli graph and the uniform random graph with a given link count are two of the most studied random graphs. Both are often called *Erdős–Rényi random graphs* after Pál Erdős and Alfred Rényi who in 1959 published a famous article [ER59] on the connectivity of the latter model. Edgar Gilbert published a similar connectivity analysis [Gil59] in the same year, but for some reason his name has not become as well known in this context.

1.4 Latent position graph models

Let (p_{ij}) be a symmetric n-by-n matrix with entries in $[0, 1]$. An *inhomogeneous Bernoulli graph* G on node set $[n]$ with link rate matrix (p_{ij}) is the one where each unordered node pair $\{i, j\}$ is linked with probability p_{ij}, independently of other node pairs. The probability distribution of the graph is then given by

$$\mathbb{P}(G = g) \;=\; \prod_{1 \le i < j \le n} (1 - p_{ij})^{1 - g_{ij}} p_{ij}^{g_{ij}}, \quad g \in \mathcal{G}_n.$$

This is an extremely flexible model, but in practice the number of model parameters $n(n-1)/2$ is way too large to do feasible inference. A feasible approach is to consider a model where each node i is assigned an *attribute* (type, mark, label, weight) z_i which is an element in a set $\mathcal{S}$, and assume that

$$p_{ij} = \rho K(z_i, z_j)$$

for some symmetric function $K : \mathcal{S} \times \mathcal{S} \to [0,1]$ called a *kernel* and some constant $\rho \in [0,1]$ which controls the overall link density. Then we get a random graph model on node set $[n]$ parametrized by the attribute list $z = (z_1, \dots, z_n)$, the kernel K, and the constant ρ. Such a model is called a *latent position graph* in the statistics literature. This model generalizes the Erdős–Rényi model to an inhomogeneous setting and is still rich enough to contain lots of different models as special cases, summarized in Table 1.1.

Nickname	Label space	Kernel
Erdős–Rényi graph	$\{1\}$	Constant
Stochastic block model	$\{1, \dots, m\}$	Symmetric $K : [m]^2 \to [0,1]$
Rank-1 inhomogeneous graph	$[0, \infty)$	$K(z_i, z_j) = \phi(z_i z_j)$
Multiplicative attribute graph	$\prod_r \{1, \dots, m_r\}$	$K(z_i, z_j) = \prod_r K_r(z_{ir}, z_{jr})$
Random intersection graph	$\{0,1\}^m$	$K(z_i, z_j) = \min\{\langle z_i, z_j \rangle, 1\}$
Random dot product graph	$\mathbb{R}^d$	$K(z_i, z_j) = \phi(\langle z_i, z_j \rangle)$
Random geometric graph	$\mathbb{R}^d$	$K(z_i, z_j) = \phi(\|z_i - z_j\|)$
Graphon	$[0,1]$	Symmetric $K : [0,1]^2 \to [0,1]$

Table 1.1: Latent position graph models. Here $\phi : [0, \infty) \to [0,1]$ truncates positive real numbers into probability values in $[0,1]$.

1.5 Notes and further reading

Statistical estimation of latent position graph models are discussed in [AFT+18].

Chapter 2

Connectivity

2.1 Connectivity probability

A graph is *connected* if between any two nodes there exists a path connecting them. In this section we will investigate the probability that G is connected, where G is a homogeneous Bernoulli graph on node set $[n] = \{1, \dots, n\}$ where every node pair is linked with probability p, independently of other node pairs. In principle, we can write this probability using the density (**??**) as

$$\mathbb{P}(G \text{ connected}) \;=\; \sum_{g \in \mathcal{G}_n^{\text{con}}} \prod_{1 \le x < y \le n} (1-p)^{1-g_{xy}} p^{g_{xy}}$$

where $\mathcal{G}_n^{\text{con}}$ denotes the set of all connected graphs on node set $[n]$. Although the above formula provides an exact answer to our question, as such it is not very useful in practice because the number of summands on the right is astronomical even for small graphs (for $n = 10$ the sum contains 34 496 488 594 816 terms[1]). Moreover, the above formula does not reveal much insight on whether or not the probability is small or large for a given value of p.

2.2 Number of isolated nodes

A node in a graph is *isolated* if it has no neighbors. The following result shows that the value $\frac{\log n}{n}$ is a sharp threshold for the existence of isolated nodes in large Bernoulli graphs. Here ω_n is thought to be a function having an arbitrarily slow growth to infinity, for example, $\omega_n = \log \log n$. The proof technique based on applying Markov's and Chebyshev's inequalities analyze

[1] This is the integer sequence A001187 in the On-Line Encyclopedia of Integer Sequences http://oeis.org/A001187.

the probability of a random nonnegative integer being nonzero is called as the *second moment method*.

Theorem 2.1. *The probability that a homogeneous Bernoulli graph G_n on node set $[n]$ with link probability p_n contains isolated nodes satisfies*

$$\mathbb{P}(\exists \textit{ isolated nodes}) \ \to \ \begin{cases} 0, & \textit{if } p_n \ge \frac{\log n+\omega_n}{n} \textit{ for some } \omega_n \to \infty, \\ 1, & \textit{if } p_n \le \frac{\log n-\omega_n}{n} \textit{ for some } \omega_n \to \infty. \end{cases}$$

Proof. Denote by A_i the event that node x is isolated, and let $Y_n = \sum_{i=1}^n 1(A_i)$ be the number of isolated nodes in G_n. Because $\mathbb{P}(A_i) = (1-p_n)^{n-1}$ for all x, we see that the expectation of Y_n is

$$\mathbb{E}Y_n \ = \ \sum_{i=1}^n \mathbb{E}1(A_i) \ = \ \sum_{i=1}^n \mathbb{P}(A_i) \ = \ n(1-p_n)^{n-1}. \tag{2.1}$$

To analyze how this expectation behaves for large n, let us first analyze the function $f(t) = \log(1-t)$ on the interval $(-1,1)$. The derivatives of f are given by $f^{(k)}(t) = -(k-1)!(1-t)^{-k}$ and hence we obtain a Taylor series representation

$$\log(1-t) \ = \ -\sum_{k=1}^\infty \frac{1}{k} t^k.$$

The above formula implies that

$$\log(1-t) \ \le \ -t \qquad \text{for all } 0 \le t < 1. \tag{2.2}$$

We may also note that for $0 \le t \le \frac{1}{2}$,

$$0 \ \le \ \sum_{k=2}^\infty \frac{1}{k} t^k \ \le \ \frac{1}{2} \sum_{k=2}^\infty t^k \ = \ \frac{1}{2} \frac{t^2}{1-t} \ \le \ t^2,$$

which yields a a lower bound

$$\log(1-t) \ \ge \ -t - t^2 \quad \text{for all } 0 \le t \le \frac{1}{2}. \tag{2.3}$$

(i) Assume that $p_n \ge \frac{\log n+\omega_n}{n}$ for some $\omega_n \to \infty$. Then with the help of (2.1) and (2.2) we see that

$$\mathbb{E}Y_n \ = \ e^{\log n + (n-1)\log(1-p_n)} \ \le \ e^{\log n - (n-1)p_n} \ = \ e^{p_n} e^{\log n - np_n} \ \le \ e^{p_n} e^{-\omega_n}.$$

Because $p_n \le 1$ and $\omega_n \to \infty$, we may conclude by Markov's inequality that

$$\mathbb{P}(\exists \text{ isolated nodes}) \ = \ \mathbb{P}(Y_n \ge 1) \ \le \ \mathbb{E}Y_n \to 0,$$

which confirms the first part of the claim.

(ii) Assume next that $p_n \le \frac{\log n - \omega_n}{n}$ for some $\omega_n \to \infty$. Then $p_n \le \frac{1}{2}$ for all large enough n, and with the help of (2.1) and (2.3),

$$\mathbb{E}Y_n \;=\; e^{\log n + (n-1)\log(1-p_n)} \;\ge\; e^{\log n - (n-1)(p_n + p_n^2)} \;\ge\; e^{\log n - np_n - np_n^2}.$$

Now $np_n \le \log n - \omega_n$ and $np_n^2 = n^{-1}(np_n)^2 \le n^{-1}(\log n)^2 \le 1$ for all large values of n, so that

$$\mathbb{E}Y_n \;\ge\; e^{\omega_n - 1} \to \infty \quad \text{as } n \to \infty.$$

Hence we conclude that the expected number of isolated nodes converges to infinity, but this is not yet sufficient to conclude that $\mathbb{P}(Y_n = 0) \to 0$ (finding a counterexample is a good exercise). One way to arrive at the desired conclusion is to show that the the probability mass of Y_n is sufficiently well concentrated around its mean, and Chebyshev's inequality then is the easiest way to do this. In our current case where $\mathbb{E}Y_n \ge 0$ we may apply Chebyshev's inequality to conclude that

$$\mathbb{P}(Y_n = 0) \;\le\; \mathbb{P}(Y_n \le 0) \;\le\; \mathbb{P}(|Y_n - \mathbb{E}Y_n| \ge \mathbb{E}Y_n) \;\le\; \frac{\mathrm{Var}(Y_n)}{(\mathbb{E}Y_n)^2}. \tag{2.4}$$

To analyze the variance above, note that

$$\mathbb{E}(Y_n^2) \;=\; \mathbb{E}\sum_{i=1}^n \sum_{j=1}^n 1(A_i)1(A_j) \;=\; \sum_{i=1}^n \sum_{j=1}^n \mathbb{P}(A_i, A_j).$$

Hence by recalling that $\mathbb{E}Y_n = \sum_{i=1}^n \mathbb{P}(A_i)$, it follows that

$$\mathrm{Var}(Y_n) \;=\; \mathbb{E}(Y_n^2) - (\mathbb{E}Y_n)^2 \;=\; \sum_{i=1}^n \sum_{j=1}^n \Big(\mathbb{P}(A_i, A_j) - \mathbb{P}(A_i)\mathbb{P}(A_j)\Big).$$

Observe next that for $x \ne y$, the conditional probability that y is isolated given that x is isolated equals $(1-p_n)^{n-2}$, so that

$$\mathbb{P}(A_i, A_j) \;=\; \mathbb{P}(A_i)\mathbb{P}(A_j \,|\, A_i) \;=\; (1-p_n)^{n-1}(1-p)^{n-2} \;=\; (1-p_n)^{2n-3}.$$

By recalling that $\mathbb{P}(A_i) = (1-p_n)^{n-1}$ for all i, we find that

$$\begin{aligned}
\mathrm{Var}(Y_n) \;&=\; n\Big(\mathbb{P}(A_1) - \mathbb{P}(A_1)^2\Big) + n(n-1)\Big(\mathbb{P}(A_1, A_2) - \mathbb{P}(A_1)^2\Big) \\
&\le\; n\mathbb{P}(A_1) + n^2\Big(\mathbb{P}(A_1, A_2) - \mathbb{P}(A_1)^2\Big) \\
&=\; n(1-p_n)^{n-1} + n^2\Big((1-p_n)^{2n-3} - (1-p_n)^{2n-2}\Big) \\
&=\; n(1-p_n)^{n-1} + n^2(1-p_n)^{2n-2}\frac{p_n}{1-p_n} \\
&=\; \mathbb{E}Y_n + (\mathbb{E}Y_n)^2\frac{p_n}{1-p_n}.
\end{aligned}$$

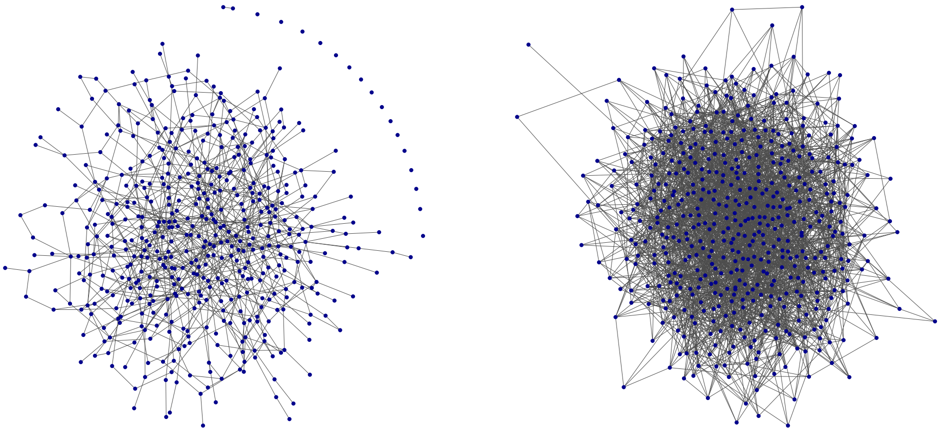

Figure 2.1: Simulated realizations of homogeneous Bernoulli graphs with $n = 500$ nodes and link rate $p = 0.5\frac{\log n}{n}$ (left) and $p = 1.5\frac{\log n}{n}$ (right).

Because $p_n \to 0$ and $\mathbb{E}Y_n \to \infty$, we conclude using (2.4) that

$$\mathbb{P}(Y_n = 0) \;\le\; \frac{\mathrm{Var}(Y_n)}{(\mathbb{E}Y_n)^2} \;\le\; \frac{1}{\mathbb{E}Y_n} + \frac{p_n}{1-p_n} \;\to\; 0,$$

and hence the probability that G_n contains isolated nodes tends to one. □

2.3 Connectivity threshold

The following striking result shows that for a large graph (with n large) there is a sharp transition from a disconnected graph to a connected one as the link probability p_n crosses the critical value $\frac{\log n}{n}$. This type of result was first discovered by Erdős and Rényi in 1959 for the uniform random graph with a given link count [ER59]. The sharp threshold is illustrated in Figure 2.1.

Theorem 2.2. *For a sequence of homogeneous Bernoulli graphs G_n on node set $[n]$ with link rate p_n,*

$$\mathbb{P}(G_n \text{ is connected}) \;\to\; \begin{cases} 0, & \text{if } p_n \le \frac{\log n - \omega_n}{n} \text{ for some } \omega_n \to \infty, \\ 1, & \text{if } p_n \ge \frac{\log n + \omega_n}{n} \text{ for some } \omega_n \to \infty. \end{cases}$$

Remark 2.3. The proof of Theorem 2.2 below shows that

$$\mathbb{P}(G_n \text{ is disconnected}) \;\le\; 3e^4 e^{\log n - np}$$

whenever $np \geq 9e \vee \log n$. This upper bound can be applied[2] in the case with $p_n \geq (1+\epsilon)\frac{\log n}{n}$ for some $\epsilon > 0$ to conclude that

$$\mathbb{P}(G_n \text{ is disconnected}) \;=\; O(n^{-\epsilon})$$

or

$$\mathbb{P}(G_n \text{ is disconnected}) \;=\; O(e^{-\delta np})$$

for $\delta = \frac{\epsilon}{1+\epsilon}$.

The proof is based on counting arguments related to the components of the graph. A set of nodes C in a graph G is called *isolated* the graph G contains no links between C and its complement, and a *component* of G is a connected isolated node set. The components of a graph form a partition of its node set, and a connected graph only has one single component containing all its nodes. An important fact related to connectivity is that any connected component in G contains a spanning tree as a subgraph, see Figure 2.2.

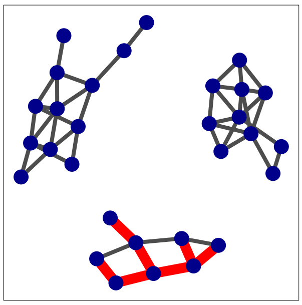

Figure 2.2: A spanning tree (red) of the smallest component of a disconnected graph having three components.

Proof. If $p_n \leq \frac{\log n - \omega_n}{n}$ for some $\omega_n \to \infty$, then by Theorem 2.1 it follows that

$$\mathbb{P}(G_n \text{ is connected}) \;\leq\; \mathbb{P}(G_n \text{ contains no isolated nodes}) \;\to\; 0.$$

Assume next that $p_n \geq \frac{\log n + \omega_n}{n}$ for some $\omega_n \to \infty$. In this case Theorem 2.1 states that, with high probability there are no isolated nodes, but of course this does not yet imply that G_n is connected. Instead, we also need to verify that the graph does not contain larger sets of isolated nodes. Denote[3] by

[2] exercise

[3] For clarity, we omit the subscript n from $p = p_n$ and some other variables where the dependence on n is clear from the context.

$Y^{(k)} = Y_n^{(k)}$ be the number of components of size k in graph G_n. If the graph is disconnected, then $Y^{(k)} > 0$ for some $1 \le k \le n/2$. Therefore, Markov's inequality implies that

$$\mathbb{P}(G_n \text{ disconnected}) \;\le\; \mathbb{P}\Big(\sum_{k \le n/2} Y^{(k)} \ge 1 \Big) \;=\; \sum_{k \le n/2} \mathbb{E} Y^{(k)}. \tag{2.5}$$

Hence for proving that the graph is connected with high probability, it suffices to show that the sum on right of (2.5) converges to zero as $n \to \infty$.

Note that $Y^{(1)}$ equals the number isolated nodes, and we already saw in the proof of Theorem 2.1 that

$$\mathbb{E} Y^{(1)} \;=\; n(1-p)^{n-1} \;\le\; ne^{-(n-1)p} \;=\; e^{p-\omega_n} \;\le\; e^{1-\omega_n} \;\to\; 0. \tag{2.6}$$

The random variable $Y^{(2)}$ equals the number isolated links, and

$$\mathbb{E} Y^{(2)} \;=\; \binom{n}{2} p(1-p)^{2(n-2)},$$

because any particular node pair is linked with probability p, and the node pair is isolated from all other nodes with probability $(1-p)^{2(n-2)}$. By applying the inequalities $\binom{n}{2} \le n^2$ and $1 - t \le e^{-t}$, it follows that

$$\mathbb{E} Y^{(2)} \;\le\; n^2 pe^{-2(n-2)p} \;=\; pe^{4p} e^{-2(np - \log n)} \;\le\; e^{4 - 2\omega_n} \;\to\; 0. \tag{2.7}$$

We could continue this way to verify that $\mathbb{E} Y^{(k)} \to 0$ for larger values of k as well. However, this approach has the problem that the range of summands in (2.5) grows as n grows. To overcome this, we will next derive a generic upper bound of $Y^{(k)}$.

We get an upper bound by noting that any connected set of nodes in a graph contains a spanning tree (see Figure 2.2), and that any component C is isolated from the rest of the nodes in the graph. Therefore,

$$\begin{aligned} Y^{(k)} \;&=\; \sum_C 1(C \text{ is a connected component}) \\ &\le\; \sum_C \sum_T 1(T \text{ is a subgraph of } G)\, 1(C \text{ is isolated}), \end{aligned}$$

where C ranges over all node sets of size k, and T ranges over all trees on C. By taking expectations and noting that the events $\{T \text{ is a subgraph of } G\}$ and $\{C \text{ is isolated}\}$ are independent, we find that

$$\mathbb{E} Y^{(k)} \;\le\; \sum_C \sum_T \mathbb{P}(T \text{ is a subgraph of } G)\, \mathbb{P}(C \text{ is isolated}). \tag{2.8}$$

By noting that any tree T with k nodes has precisely $k-1$ links[4], and that between any k-element set C and its complement there are precisely $k(n-k)$ node pairs, we see that

$$\mathbb{P}(T \text{ is a subgraph of } G) \;=\; p^{k-1}$$

and

$$\mathbb{P}(C \text{ is isolated}) \;=\; (1-p)^{k(n-k)}.$$

Moreover, by Cayley's theorem [vdH17, Theorem 3.17], the number of trees on a set of k nodes equals k^{k-2}. Hence the inequality (2.8) can be written as

$$\mathbb{E}Y^{(k)} \;\le\; \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)}. \tag{2.9}$$

The binomial coefficient above can be bounded by applying a Stirling lower bound $k! \ge k^k e^{-k}$ according to

$$\binom{n}{k} \;=\; \frac{n(n-1)\cdots(n-k+1)}{k!} \;\le\; \frac{n^k}{k!} \;\le\; \frac{n^k}{k^k} e^k.$$

For the probability that a set of k nodes is isolated we get an upper bound using $1-t \le e^{-t}$ in the form

$$(1-p)^{k(n-k)} \;\le\; e^{-k(n-k)p} \;\le\; e^{-knp/2}, \quad k \le n/2.$$

By plugging in these bounds into (2.9) we obtain

$$\mathbb{E}Y^{(k)} \;\le\; \frac{n^k}{k^2} e^k p^{k-1} e^{-knp/2} \;=\; \frac{1}{pk^2} \left(enp\, e^{-np/2} \right)^k, \quad k \le n/2.$$

Now $e^t \ge \frac{1}{2}t^2$ implies that $a_n := enp\, e^{-np/2} \le 8e(np)^{-1}$, and hence $a_n \le \frac{8}{9}$ whenever $np \ge 9e$. Therefore, we find that

$$\sum_{3 \le k \le n/2} \mathbb{E}Y^{(k)} \;\le\; \sum_{3 \le k \le n/2} \frac{1}{pk^2} a_n^k \;\le\; \frac{1}{9p} \sum_{k=3}^{\infty} a_n^k \;=\; \frac{1}{9p} \frac{a_n^3}{1-a_n} \;\le\; \frac{1}{p} a_n^3.$$

Also, $e^s \ge s + \frac{1}{2}s^2 + \frac{1}{6}s^3 = 4s + (s-3)(s+6) \ge 4s$ for $s \ge 3$ implies that $\log np \le \frac{1}{4} np$ for all $np \ge e^3$, and hence

$$\sum_{3 \le k \le n/2} \mathbb{E}Y^{(k)} \;\le\; \frac{1}{p} a_n^3 \;=\; e^{3 + \log n + 2\log(np) - \frac{3}{2}np} \;\le\; e^{3 + \log n - np} \;\le\; e^{3 - \omega_n}$$

[4] This is a good exercise to verify independently.

whenever $np \geq 9e$. By combining this inequality with the bounds (2.5)–(2.7), we conclude that

$$\begin{aligned} \mathbb{P}(G_n \text{ disconnected}) &\leq \mathbb{E}Y^{(1)} + \mathbb{E}Y^{(2)} + \sum_{3 \leq k \leq n/2} \mathbb{E}Y^{(k)} \\ &\leq e^{1-(np-\log n)} + e^{4-2(np-\log n)} + e^{3-(np-\log n)} \\ &\leq e^4 e^{-2(np-\log n)} + 2e^3 e^{-(np-\log n)}. \end{aligned}$$

whenever $np \geq 9e$. The right side tends to zero because $np - \log n \geq \omega_n \to \infty$. □

Chapter 3

Coupling and stochastic ordering

3.1 Stochastic coupling method

Stochastic coupling is a powerful general method for analyzing probability distributions of random variables defined via nonlinear interactions related to graphs and more general random structures. Among other things, this method can be used to show that:

- The cumulative distribution functions two random variables are ordered.
- Two probability distributions are close to each other with respect to a suitable metric.
- A stochastic process mixes rapidly and approaches its statistical equilibrium.

In the sequel, the probability distribution of a random variable X is denoted by $\mathrm{Law}(X)$. In general, a *coupling* of random variables X_1 and X_2 is a pair of random variables

$$(\hat{X}_1, \hat{X}_2)$$

defined on a common probability space, such that $\mathrm{Law}(\hat{X}_1) = \mathrm{Law}(X_1)$ and $\mathrm{Law}(\hat{X}_2) = \mathrm{Law}(X_2)$. The requirement of a common probability space implies that $\hat{X}_1$ and $\hat{X}_2$ are usually dependent. For any two random variables there exists an unlimited number of couplings, and the key idea is to

define a coupling with a suitable dependence structure which can reveal useful relationships between the probability distributions of X_1 and X_2. Algorithmically, this corresponds to constructing a simulator which outputs random variables that are distributed according to $\mathrm{Law}(X_1)$ and $\mathrm{Law}(X_2)$, which are suitably dependent when generated using a common seed as input. The general ideas underlying this abstract principle are perhaps best illustrated by the following example.

Example 3.1. Consider random integers X_1 and X_2 such that $\mathrm{Law}(X_i) = \mathrm{Bin}(n_i, p)$ for some positive integers $n_1 \le n_2$ and some $p \in [0,1]$. Because the binomial distribution $\mathrm{Bin}(n_i, p)$ represents the number of successes in n_i independent trials with success probability p, intuition suggests that

$$\mathbb{P}(X_1 \ge k) \ \le \ \mathbb{P}(X_2 \ge k) \tag{3.1}$$

for all k. But how can we verify this claim mathematically? Note that (3.1) is equivalent to

$$\sum_{\ell \ge k} \binom{n_1}{\ell} (1-p)^{n_1-\ell} p^{\ell} \ \le \ \sum_{\ell \ge k} \binom{n_2}{\ell} (1-p)^{n_2-\ell} p^{\ell},$$

so that in principle it should be possible to prove (3.1) using combinatorial techniques. The stochastic coupling method provides an alternative, perhaps more intuitive way. It aims at producing replicas of X_1 and X_2 using a carefully designed simulation algorithm which produces suitably correlated samples when run using a common seed as input. In this example we may select the seed to be a sequence $(B_1, B_2, \dots)$ of independent $\mathrm{Ber}(p)$-distributed random variables, and we define our algorithm as the map $(B_1, B_2, \dots) \mapsto (N_1, N_2)$ where

$$N_1 \ = \ \sum_{i=1}^{n_1} B_i \quad \text{and} \quad N_2 \ = \ \sum_{i=1}^{n_2} B_i. \tag{3.2}$$

To verify that the above algorithm produces proper replicas of X_1 and X_2, note that $N_1 = k$ if and only if precisely k out of the n_1 random variables $(B_1, \dots, B_{n_1})$ equal one, so that

$$\mathbb{P}(N_1 = k) \ = \ \binom{n_1}{k} (1-p)^{n_1-k} p^k$$

for all k. Analogously, a similar statement holds for N_2, and we may conclude that

$$\mathrm{Law}(N_1) = \mathrm{Law}(X_1) \quad \text{and} \quad \mathrm{Law}(N_2) = \mathrm{Law}(X_2) \tag{3.3}$$

as desired. Hence (N_1, N_2) constitutes a coupling of X_1 and X_2. Among the unlimited choices of such couplings, this coupling based on algorithm (3.2) has by design one special feature, namely $N_1 \le N_2$ holds for *every possible realization* of the seed, and hence

$$\mathbb{P}(N_1 \le N_2) \;=\; 1. \tag{3.4}$$

As a consequence of (3.3) and (3.4), it follows that

$$\begin{aligned} \mathbb{P}(X_1 \ge k) \;&=\; \mathbb{P}(N_1 \ge k) \\ &=\; \mathbb{P}(N_1 \ge k,\; N_1 \le N_2) \\ &\le\; \mathbb{P}(N_2 \ge k) \\ &=\; \mathbb{P}(X_2 \ge k). \end{aligned}$$

This concludes a pure probabilistic proof of (3.1) where we did not need to work out any formulas for binomial coefficients, or invoke any other combinatorial principles.

Comment on coupling of distributions

Exercise: Show using the coupling method that $\mathrm{Ber}(p) \le_{\mathrm{st}} \mathrm{Ber}(q)$ for $p \le q$. Extend this to $\mathrm{Bin}(n,p) \le_{\mathrm{st}} \mathrm{Bin}(n,q)$ for $p \le q$.

3.2 Coupling and stochastic ordering

3.2.1 Real line

This subsection could be skipped.

Example 3.1 provided an ordered coupling of two binomially distributed random variables. As it turns out, this is an example of a strong stochastic ordering between two random variables. In general, for real-valued random variables we say that X_1 is less than X_2 in the *strong stochastic order*, denoted

$$X_1 \le_{\mathrm{st}} X_2,$$

if $\mathbb{E} f(X_1) \le \mathbb{E} f(X_2)$ for all increasing[1] functions $f : \mathbb{R} \to \mathbb{R}_+$.

Theorem 3.2. *The following are equivalent for any real-valued random variables with cumulative distribution functions $F_i(t) = \mathbb{P}(X_i \le t)$, $i = 1, 2$:*

(i) $X_1 \le_{\mathrm{st}} X_2$

[1] Here a function is called increasing if $x \le y \implies f(x) \le f(y)$. Any increasing function is Borel-measurable.

(ii) $1 - F_1(t) \le 1 - F_2(t)$ *for all* t.

(iii) There exists a coupling $(\hat X_1, \hat X_2)$ *of* X_1 *and* X_2 *for which* $\hat X_1 \le \hat X_2$ *with probability one.*

Proof. (i) $\implies$ (ii). Assume that $X_1 \le_{\rm st} X_2$, and observe that the function f_t defined by $f_t(x) = 1(x > t)$ is increasing for any t. Hence it follows that

$$\mathbb{P}(X_1 > t) \ = \ \mathbb{E} f_t(X_1) \ \le \ \mathbb{E} f_t(X_2) \ = \ \mathbb{P}(X_2 > t).$$

(ii) $\implies$ (iii). We construct a coupling using a generic method used to generate samples from a given distribution, using a uniformly distributed random number $U \in (0,1)$ as input. Assume for simplicity[2] that the cumulative distribution functions $F_1(t) = \mathbb{P}(X_1 \le t)$ and $F_2(t) \le \mathbb{P}(X_2 \le t)$ are invertible, and define a pair of random variables by

$$(\hat X_1, \hat X_2) \ = \ (F_1^{-1}(U), F_2^{-1}(U)). \tag{3.5}$$

Then

$$\mathbb{P}(\hat X_i \le t) \ = \ \mathbb{P}(F_i^{-1}(U) \le t) \ = \ \mathbb{P}(U \le F_i(t)) \ = \ F_i(t) \ = \ \mathbb{P}(X_i \le t)$$

implies[3] that $\operatorname{Law}(\hat X_i) = \operatorname{Law}(X_i)$ for $i = 1, 2$, and hence (3.5) is a coupling of X_1 and X_2. Moreover, assumption (ii) implies that

$$1 - F_1(t) \ \le \ 1 - F_2(t)$$

for all t. Applying this to $t = F_1^{-1}(u)$ shows that $F_2(F_1^{-1}(u)) \le u$, and the fact that the inverse of a cumulative distribution function is increasing implies $F_1^{-1}(u) \le F_2^{-1}(u)$. This fact together with definition (3.5) implies that $\hat X_1 \le \hat X_2$ with probability one.

(iii) $\implies$ (i). Assume that $(\hat X_1, \hat X_2)$ is a coupling of X_1 and X_2 such that $\mathbb{P}(\hat X_1 \le \hat X_2) = 1$. Then for any increasing function $f : \mathbb{R} \to \mathbb{R}_+$,

$$\begin{aligned} \mathbb{E} f(X_1) \ = \ \mathbb{E} f(\hat X_1) \ &= \ \mathbb{E} f(\hat X_1) 1(\hat X_1 \le \hat X_2) \\ &\le \ \mathbb{E} f(\hat X_2) 1(\hat X_1 \le \hat X_2) \ \le \ \mathbb{E} f(\hat X_2) \ = \ \mathbb{E} f(X_2). \end{aligned}$$

□

[2] The general case may proved by using the right-continuous generalized inverse $Q_i(u) = \inf\{t \in \mathbb{R} : F_i(t) > u\}$ in place of F_i^{-1}.

[3] The fact that the cumulative distribution of X fully determines the whole distribution $\operatorname{Law}(X)$ is proved in MS-E1600.

3.2.2 Partially ordered sets

A *partial order* on a set S is a binary relation $\leq$ which is:

(i) reflexive: $x \leq x$ for all $x \in S$,

(ii) transitive: $x \leq y$ and $y \leq z \implies x \leq z$, and

(iii) antisymmetric: $x \leq y$ and $y \leq x \implies x = y$.

Elements x and y are called *comparable* if $x \leq y$ or $y \leq x$. If all pairs of elements are comparable, the partial order is called a *total order*. Any set with at least two elements has several partial orders, but usually the context makes it clear which partial order is meant. A real function f defined on a partially ordered set S is called *increasing* if $x \leq y \implies f(x) \leq f(y)$. A set $U \subset S$ is called *upper* when $x \in U$ and $x \leq y$ imply $y \in U$.

Example 3.3 (Coordinatewise order)**.** The *coordinatewise order* on $\mathbb{R}^n$ is a partial order defined by denoting $(x_1, \dots, x_n) \leq (y_1, \dots, y_n)$ if $x_i \leq y_i$ for all $i = 1, \dots, n$. For $n = 1$ this reduces to the natural total order of the real numbers. In higher dimensions the coordinatewise order is not total because for example the unit vectors of different coordinate axes are not comparable. The coordinatewise order on the space $\mathbb{R}^{m \times n}$ of all m-by-n real matrices is defined similarly by denoting $(x_{ij}) \leq (y_{ij})$ if $x_{ij} \leq y_{ij}$ for all i and j.

Example 3.4 (Subset order)**.** The relation $A \subset B$ is a partial order on the collection of all subsets of a set S, called the *subset order*. Note that for any set A we can associate an indicator function $1_A : S \to \{0, 1\}$ such that

$$1_A(x) \;=\; \begin{cases} 1, & x \in A, \\ 0, & \text{else.} \end{cases}$$

For a finite set $S = \{s_1, \dots, s_n\}$, such indicator functions can be viewed as vectors $1_A = (1_A(s_1), \dots, 1_A(s_n)) \in \{0, 1\}^n$, and these vectors are in one-to-one correspondence with the subsets of S. Moreover, $A \subset B$ if and only if $1_A \leq 1_B$ in the coordinatewise order on $\{0, 1\}^n$.

The following remarkable result, sometimes known as Strassen's coupling theorem, extends Theorem 3.2 to general partial orders. For random variables X_1 and X_2 defined on a partially ordered space S, we denote

$$X_1 \leq_{\mathrm{st}} X_2$$

and say that X_1 is less than X_2 in the *strong stochastic order* on $(S, \leq)$, if $\mathbb{E} f(X_1) \leq \mathbb{E} f(X_2)$ for all increasing $f : S \to \mathbb{R}_+$.

Theorem 3.5. *The following are equivalent for random variables with values in a partially ordered space*[4] $(S, \leq)$*:*

(i) $X_1 \leq_{\mathrm{st}} X_2$.

(ii) $\mathbb{P}(X_1 \in A) \leq \mathbb{P}(X_2 \in A)$ *for all upper sets* $A \subset S$.

(iii) X_1 *and* X_2 *admit a coupling* $(\hat{X}_1, \hat{X}_2)$ *for which* $\hat{X}_1 \leq \hat{X}_2$ *with probability one.*

Proof. (i) $\Longrightarrow$ (ii). Assume that $X_1 \leq_{\mathrm{st}} X_2$. Because the indicator 1_A of any upper set $A \subset S$ is increasing, it then follows that

$$\mathbb{P}(X_1 \in A) \ = \ \mathbb{E}1_A(X_1) \ \leq \ \mathbb{E}1_A(X_2) \ = \ \mathbb{P}(X_2 \in A).$$

(ii) $\Longrightarrow$ (i). Let $f : S \to \mathbb{R}_+$ be increasing. Then the level set $A = \{x : f(x) > t\}$ is upper for all real numbers t, and it follows that

$$\mathbb{P}(f(X_1) > t) \ = \ \mathbb{P}(X_1 \in A) \ \leq \ \mathbb{P}(X_2 \in A) \ = \ \mathbb{P}(f(X_2) > t).$$

By integrating both sides of the above inequality we find that

$$\mathbb{E}f(X_1) \ = \ \int_0^\infty \mathbb{P}(f(X_1) > t)\, dt \ \leq \ \int_0^\infty \mathbb{P}(f(X_2) > t)\, dt \ = \ \mathbb{E}f(X_2).$$

(iii) $\Longrightarrow$ (i). The argument used in proving (iii) $\Longrightarrow$ (i) of Theorem 3.2 extends directly to this setting

(i) $\Longrightarrow$ (iii). This part requires techniques of convex duality and is omitted here, see the original article of Volker Strassen [Str65]. When S is finite, this can be proved by applying the max-flow min-cut theorem [Pre74]. □

3.2.3 Random graphs

For graphs $g, h \in \mathcal{G}_n$ on node set $[n]$ we denote $g \leq h$ if g is a subgraph of h, or equivalently, if the corresponding adjacency matrices satisfy $g_{ij} \leq h_{ij}$ for all i, j. This means that every node pair that is linked in g is also linked in h. This relation is a partial order on $\mathcal{G}_n$ called the *subgraph order*. A real function ϕ on $\mathcal{G}_n$ is increasing if the value of $\phi(g)$ increases or remains constant when new links are added to g. Upper sets with respect to the subgraph order are those $A \subset \mathcal{G}_n$ for which the property

$$g \text{ belongs to } A$$

[4] If S is uncountably infinite, then we need to assume that S is a Polish topological space, the set $\{(x, y) : x \leq y\}$ is a closed subset of $S \times S$, and that the upper sets and increasing functions are measurable.

remains valid whenever new links are added to g. Such a property is called[5] *monotone graph property*. Table 3.1 illustrates this.

Upper set A	Monotone property
$\{g \in \mathcal{G}_n : g_{ij} = 1\}$	i is linked to j in g
$\{g \in \mathcal{G}_n : \deg_g(i) \geq 3\}$	i has at least 3 neighbors in g
$\{g \in \mathcal{G}_n : \sum_{1 \leq i<j<k \leq n} g_{ij} g_{ik} g_{jk} > 0\}$	g contains a triangle
$\{g \in \mathcal{G}_n : g \text{ is connected}\}$	g is connected

Table 3.1: Upper sets of graphs.

The *stochastic order of random graphs* on $(\mathcal{G}_n, \leq)$ is defined by denoting $G \leq_{\mathrm{st}} H$ if $\mathbb{E}\phi(G) \leq \mathbb{E}\phi(H)$ for all increasing functions $\phi : \mathcal{G}_n \to \mathbb{R}_+$. When we apply Theorem 3.5 for the subgraph order, we find that the following are equivalent:

(i) $G \leq_{\mathrm{st}} H$.

(ii) $\mathbb{P}(G \text{ has property } \mathcal{A}) \leq \mathbb{P}(H \text{ has property } \mathcal{A})$ for any monotone graph property $\mathcal{A}$.

(iii) G and H admit a coupling $(\hat{G}, \hat{H})$ for which $\hat{G}$ is a subgraph of $\hat{H}$ with probability one.

Here the rate matrices must have zero diagonal.

Theorem 3.6. *Let G and H be Bernoulli random graphs on node set $[n]$ with rate matrices (p_{ij}) and (q_{ij}). Then $G \leq_{\mathrm{st}} H$ if and only if $p_{ij} \leq q_{ij}$ for all node pairs $\{i, j\}$.*

Proof. (i) Assume that $G \leq_{\mathrm{st}} H$. Fix a node pair $\{i, j\}$. Let $A = \{g \in \mathcal{G}_n : g_{ij} = 1\}$ be set of graphs on $[n]$ in which i and j are linked. Then A is an upper set and its indicator 1_A is increasing. Hence

$$p_{ij} \;=\; \mathbb{P}(G_{ij} = 1) \;=\; \mathbb{E}1_A(G) \;\leq\; \mathbb{E}1_A(H) \;=\; \mathbb{P}(H_{ij} = 1) \;=\; q_{ij}.$$

(ii) As input to our coupling construction, we will choose the graph G itself and another Bernoulli random graph D with rate matrix (d_{ij}), where

$$d_{ij} \;=\; 1 - \frac{1 - q_{ij}}{1 - p_{ij}} \qquad \text{if } p_{ij} < 1,$$

[5] Some authors require that a graph property is invariant with respect to graph isomorphisms.

and $d_{ij} = 0$ otherwise. Moreover, we will assume that D is sampled independently of G. Using the pair (G, D) as input, we define a new pair of random graphs $(\hat{G}, \hat{H})$ by setting

$$\hat{G}_{ij} = G_{ij} \quad \text{and} \quad \hat{H}_{ij} = \max\{G_{ij}, D_{ij}\}.$$

The above construction guarantees that $\hat{G} \le \hat{H}$ for all realizations of the random graphs. Hence it remains to verify that $(\hat{G}, \hat{H})$ is a coupling of G and H. Obviously,

$$\mathrm{Law}(\hat{G}) = \mathrm{Law}(G).$$

To study the distribution of $\hat{H}$, note first that a node pair $\{i, j\}$ is linked in $\hat{H}$ if and only if it is linked in G or D, or both. Therefore, the probability that $\{i, j\}$ is not linked in $\hat{H}$ equals

$$\mathbb{P}(\hat{H}_{ij} = 0) \;=\; (1 - p_{ij})(1 - d_{ij}) \;=\; 1 - q_{ij} \;=\; \mathbb{P}(H_{ij} = 0).$$

This implies that

$$\mathrm{Law}(\hat{H}_{ij}) \;=\; \mathrm{Law}(H_{ij})$$

for all i, j. Because the random variables $\hat{H}_{ij}$ are independent, and so are the random variables H_{ij}, this implies that

$$\mathrm{Law}(\hat{H}) \;=\; \mathrm{Law}(H).$$

Hence the pair $(\hat{G}, \hat{H})$ gives the desired coupling. □

Theorem 3.7. *For a sequence of Bernoulli graphs G_n on node set $[n]$ with rate matrices $p_{ij}^{(n)}$, and for any $\omega_n \to \infty$,*

$$\mathbb{P}(G_n \textit{ is connected}) \;\to\; \begin{cases} 0, & \textit{if } \max_{1 \le i < j \le n} p_{ij}^{(n)} \le \frac{\log n - \omega_n}{n}, \\ 1, & \textit{if } \min_{1 \le i < j \le n} p_{ij}^{(n)} \ge \frac{\log n + \omega_n}{n}. \end{cases}$$

Proof. Denote

$$p_n = \min_{1 \le i < j \le n} p_{ij}^{(n)} \quad \text{and} \quad q_n = \max_{1 \le i < j \le n} p_{ij}^{(n)}.$$

Assume first that $q_n \le \frac{\log n - \omega_n}{n}$, and let H_n be a homogeneous Bernoulli graph on node set $[n]$ with link rate q_n. Then by $G_n \le_{\mathrm{st}} H_n$ by Theorem 3.6. Because being connected is a monotone graph property (the set of connected graphs on $\mathcal{G}_n$ is an upper set), it follows by Theorem 3.5 that

$$\mathbb{P}(G_n \text{ is connected}) \;\le\; \mathbb{P}(H_n \text{ is connected}).$$

Moreover, by Theorem 2.2, the probability on the right converges to zero. Hence the first claim follows.

Assume next that $p_n \ge \frac{\log n + \omega_n}{n}$. Then an analogous argument can be used to prove the second statement. This is left as an exercise. □

Exercise. Stochastic block model with n nodes and m communities. Assume that a particular community k contains n^ϵ nodes, and that $Q_{k,k} = n^\alpha$. For which values of ϵ and α does it hold that all nodes in community k are connected to each other via paths inside community k?

Chapter 4

Giant components

4.1 Branching processes

4.1.1 Description

A *branching process* is a stochastic model for a population where each individual lives for one time unit and upon death produces a random number of children, independently of other individuals. The model is parametrized by an *offspring distribution* p on the nonnegative integers so that $p(x)$ equals the probability that an individual produces x children. The process started with one individual is a random sequence defined recursively by $Z_0 = 1$ and

$$Z_t \; = \; \sum_{i=1}^{Z_{t-1}} X_{t,i}, \quad t = 1, 2, \dots, \tag{4.1}$$

where $X_{t,i}$, $t, i \geq 1$, are independent p-distributed random integers. The random variable Z_t describes the size of the population after t time units, and $X_{t,i}$ is the number of children produced by the i-th individual in generation t.

The recursive structure (4.1) of the branching process implies that (Z_t) is a Markov chain for which state 0 is an absorbing: If $Z_t = 0$ for some t, then also $Z_{t+1} = Z_{t+2} = \cdots = 0$. The event that the process hits zero is called an extinction, and the *extinction probability* is defined by

$$\eta \; = \; \mathbb{P}(Z_t = 0 \text{ for some } t \geq 0). \tag{4.2}$$

Fundamental questions related to this model are:

- What is the extinction probability; is it less than one?
- Can the population grow infinitely large?

- Can the population remain bounded without eventually going extinct?

Because the model is entirely determined by the offspring distribution p, the shape and characteristics of p must contain answers to all of the above questions. This model is often called a *Galton–Watson process* after Francis Galton and Henry William Watson who developed this model in 1875 [WG75] to analyze the survival of surnames in England.

4.1.2 Population size distribution

The probability distribution of the population size Z_t at time t is not easy to compute directly due to the highly nonlinear structure of the recursive definition (4.1). A feasible way to get a handle on the population size process is via generating functions.

The *generating function* of a probability distribution p on the nonnegative integers $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$ is defined by

$$G_p(s) \;=\; \sum_{x=0}^{\infty} s^x p_x \tag{4.3}$$

for those real numbers s at which the series on right converges. The generating function of a nonnegative random integer X is defined by the same formula with $p_x = \mathbb{P}(X = x)$, so that

$$G_X(s) \;=\; \mathbb{E}s^X.$$

Because $\sum_x p_x \;=\; 1$, the series on right side of (4.3) converges absolutely for $s \in (-1, 1)$, and Abel's theorem for power series implies that $G_p(s)$ is infinitely differentiable on the interval $(-1, 1)$. Especially, it follows that the probability distribution p can be uniquely recovered from the derivatives of its generating function via the formula

$$p_x \;=\; \frac{G_p^{(x)}(0)}{x!}, \qquad x = 0, 1, \dots$$

Generating functions are good tools for analyzing sums of independent random numbers, because $G_{X+Y}(s) = G_X(s)G_Y(s)$ whenever X and Y are independent. This property extends to multiple independent summands, and also to the form

$$G_{\sum_{i=1}^N X_i}(s) \;=\; G_N(G_{X_1}(s)), \tag{4.4}$$

when $N, X_1, X_2, \dots$ are independent, and summands X_i are identically distributed. An application of the above formula allows to derive a formula for the generating function of Z_t in a branching process, stated next.

Theorem 4.1. *The generating function of the population size Z_t at time instant $t \geq 1$ is given by the t-fold composition of G_p with itself, so that*

$$G_{Z_t}(s) \;=\; \underbrace{G_p \circ \cdots \circ G_p}_{t}(s), \qquad |s| \leq 1.$$

Proof. The claim is true for $t = 1$ because $Z_1 = X_{1,1}$ is distributed according to p. Assume next that the claim is true for some time index $t-1$. Then by definition

$$Z_t \;=\; \sum_{i=1}^{Z_{t-1}} X_{t,i},$$

where the random integers $Z_{t-1}, X_{t,1}, X_{t,2}, \ldots$ are all mutually independent, and the summands $X_{t,i}$ are all distributed according to p. Hence by applying (4.4) we find that $G_{Z_t}(s) = G_{Z_{t-1}}(G_p(s))$. The claim now follows by induction. □

4.1.3 Extinction probability

The following result tells how the extinction probability of the branching process defined in (4.2) can be determined from the offspring distribution p.

Theorem 4.2. *The extinction probability η is the smallest fixed point of G_p on the interval* $[0, 1]$.

Proof. (i) Let us first verify that η is a fixed point of G_p. The monotone continuity of probability measures together with the fact that $\cup_{s=1}^t \{Z_s = 0\} = \{Z_t = 0\}$ shows that

$$\eta \;=\; \mathbb{P}\left(\bigcup_{t=1}^{\infty} \{Z_t = 0\} \right) \;=\; \lim_{t\to\infty} \mathbb{P}\left(\bigcup_{s=1}^{t} \{Z_s = 0\} \right) \;=\; \lim_{t\to\infty} \eta_t$$

where $\eta_t = \mathbb{P}(Z_t = 0)$ is the probability that the population has become extinct by time t. Observe next that we can write $\mathbb{P}(Z_t = 0) = G_{Z_t}(0)$, and by Theorem 4.1,

$$G_{Z_t}(0) \;=\; G_p(G_{Z_{t-1}}(0)),$$

so that

$$\eta_t \;=\; G_p(\eta_{t-1}) \tag{4.5}$$

for all $t \geq 1$. Because η ja η_t are probabilities, they belong to the interval $[0, 1]$. Moreover, the function G_p, being a convergent power series, is continuous on $[0, 1]$. Therefore,

$$\eta \;=\; \lim_{t\to\infty} \eta_t \;=\; \lim_{t\to\infty} G_p(\eta_{t-1}) \;=\; G_p(\lim_{t\to\infty} \eta_{t-1}) \;=\; G_p(\eta)$$

shows that η is a fixed point of G_p.

(ii) Assume next that $r \in [0,1]$ is an arbitrary fixed point of G_p. We will show that $\eta \le r$. Observe first that because G_p is increasing on $[0,1]$, and Z_1 is distributed according to p,

$$\eta_1 \;=\; \mathbb{P}(Z_1 = 0) \;=\; G_p(0) \;\le\; G_p(r) \;=\; r.$$

Hence $\eta_1 \le r$. Next, by applying (4.5) and the monotonicity of G_p, we see that

$$\eta_2 \;=\; G_p(\eta_1) \;\le\; G_p(r) \;=\; r.$$

Hence also $\eta_2 \le r$. By repeating this we conclude that $\eta_t \le r$ for all $t \ge 1$, and therefore

$$\eta \;=\; \lim_{t\to\infty} \eta_t \;\le\; r.$$

□

4.1.4 Sure extinction

The population is sure to go extinct when $\eta = 1$. The following result gives a simple criterion for sure extinction in terms of the expectation of the offspring distribution p.

Theorem 4.3. *For any branching process with offspring distribution p having mean $\mu = \sum_{x\ge 0} x p_x \in [0,\infty]$, the extinction probability η satisfies*

(i) $\eta = 0$ *for* $p_0 = 0$.

(ii) $\eta \in (0,1)$ *for* $p_0 > 0$ *and* $\mu > 1$

(iii) $\eta = 1$ *for* $p_0 > 0$ *and* $\mu \le 1$.

Proof. (i) This is left as an exercise to verify.

(ii) Observe that the power series $G_p(t) = \sum_{k\ge 0} t^k p_k$ converges at $t = 1$ and hence G_p can be differentiated term by term infinitely many times on $(-1,1)$. Especially, for any $t \in (0,1)$, Shorten this

$$G_p'(t) \;=\; \sum_{k=1}^{\infty} k t^{k-1} p_k.$$

By Lebesgue's monotone convergence theorem, this implies

$$\lim_{t\uparrow 1} G_p'(t) \;=\; \sum_{k=1}^{\infty} k p_k \;=\; \mu. \tag{4.6}$$

When $\mu > 1$, the convergence in (4.6) implies that there exists $t_0 \in (0,1)$ such that $G_p'(t) \geq 1 + \frac{\mu-1}{2}$ for all $t \in (t_0, 1)$. Hence

$$1 - G_p(t) \;=\; G_p(1) - G_p(t) \;=\; \int_t^1 G_p'(u)\,du \;\geq\; \left(1 + \frac{\mu-1}{2}\right)(1-t) \;>\; 1-t$$

implies $G_p(t) < t$ for all $t \in (t_0, 1)$. Because $G_p(0) = p_0 > 0$, we conclude that $G_p(t) - t$ is strictly positive for $t = 0$ and strictly negative for $t \in (t_0, 1)$. Because $t \mapsto G_p(t) - t$ is continuous, we conclude that G_p has a fixed point in $(0, t_0)$, and hence the smallest fixed point of G_p in $[0,1]$ belongs to $(0, t_0) \subset (0,1)$. Hence $\eta \in (0,1)$ by Theorem 4.2.

(iii) Assume now that $p_0 > 0$ and $\mu \leq 1$. Note that $G_p(0) = p_0 > 0$ and $G_p(1) = 1$. Therefore zero is not a fixed point of G_p but one is. Hence by Theorem 4.2 it is sufficient to show that G_p has no fixed points in $(0,1)$. We consider two cases separately.

(a) If $p_\ell = 0$ for all $\ell \geq 2$, $G_p(t) = p_0 + p_1 t > p_0 t + p_1 t = t$ for all $t \in (0,1)$ shows that there are no fixed point in the interval $(0,1)$.

(b) If $p_\ell > 0$ for some $\ell \geq 2$, then

$$G_p''(u) \;=\; \sum_{k=2}^{\infty} k(k-1)u^{k-2}p_k \;\geq\; ct^{\ell-2} \tag{4.7}$$

for all $0 < t \leq u < 1$, where $c = \ell(\ell-1)p_\ell > 0$. Now fix some $0 < t < t_1 < 1$, and note that

$$\begin{aligned} G_p(t) \;&=\; G_p(t_1) - \int_t^{t_1} G_p'(s)\,ds \\ &=\; G_p(t_1) - G_p'(t_1)(t_1 - t) + \int_t^{t_1} (G_p'(t_1) - G_p'(s))\,ds \\ &=\; G_p(t_1) - G_p'(t_1)(t_1 - t) + \int_t^{t_1}\int_s^{t_1} G_p''(u)\,du\,ds. \end{aligned}$$

Hence by (4.7) it follows that

$$G_p(t) \;\geq\; G_p(t_1) - G_p'(t_1)(t_1 - t) + \frac{c}{2}(t_1 - t)^2$$

for all $0 < t < t_1 < 1$. By letting $t_1 \uparrow 1$ and applying (4.6), it follows that

$$G_p(t) \;\geq\; 1 - \mu(1-t) + \frac{c}{2}(1-t)^2 \;\geq\; 1 - (1-t) + \frac{c}{2}(1-t)^2$$

for all $0 < t < 1$. Because $c > 0$, we conclude that $G_p(t) > t$ for all $t \in (0,1)$, so there are no fixed points in $(0,1)$. □

4.2 Components of sparse Bernoulli graphs

4.2.1 Emergence of a giant

For a random or nonrandom graph G we denote $x \overset{G}{\leftrightsquigarrow} y$ if G contains a path from node x to node y, and we denote the *component* x by

$$C_G(x) \ = \ \{y : x \overset{G}{\leftrightsquigarrow} y\}.$$

Understanding the sizes of components in a graph is important in statistical models of social sciences because the component $C_G(x)$ gives an upper limit on how far a virus or rumor initiated at x can spread.

The following is the main result about the component structure of sparse Bernoulli graphs. Below λ represents the average degree of the graph. When $\lambda > 1$, the theorem implies that there exists a component which contains a positive fraction of all nodes in the graph, see Figure 4.1. Such a component is called a *giant component*. When $\lambda < 1$, all components are small, of size at most a constant in $\log n$. The proof the theorem is long, and is presented in detail in [vdH17, Section 4]. Some ideas related to the proof are presented in the subsections below.

Theorem 4.4. *Let G_n be a homogeneous Bernoulli graph with n nodes and with link probability $p_n = \lambda n^{-1}$.*

(i) If $\lambda > 1$, then the relative size of the largest component satisfies

$$n^{-1} \max_{1 \le x \le n} |C_{G_n}(x)| \ \xrightarrow{\mathbb{P}} \ 1 - \eta$$

as $n \to \infty$, where η is the extinction probability of a branching process with Poi(λ)*-distributed offspring distribution.*

(ii) If $\lambda < 1$, then there exists a constant $a > 0$ such that

$$\max_{1 \le x \le n} |C_{G_n}(x)| \le a \log n \quad \textit{with high probability.}$$

4.2.2 Component exploration process

The component of node x may be discovered recursively by the following graph exploration algorithm which keeps track of two node sets: the set of *explored* nodes V_t and the set of *discovered* nodes W_t after t steps. The evolution of the algorithm is a set-valued sequence $(V_t, W_t)_{t \ge 0}$ defined as follows:

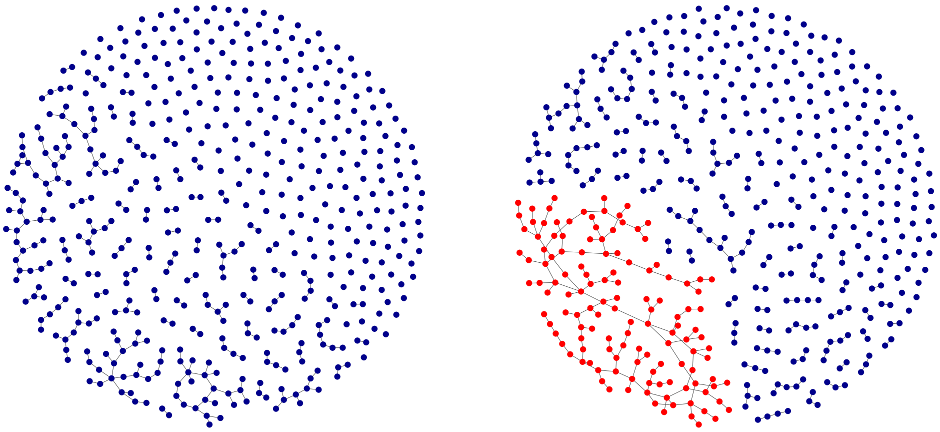

Figure 4.1: Simulated realizations of homogeneous Bernoulli graphs with $n = 500$ nodes and link rate $p = 0.9 \times n^{-1}$ (left) and $p = 1.1 \times n^{-1}$ (right). The giant component of the latter graph is displayed in red.

Step 0: Initialize $(V_0, W_0) = (\emptyset, \{x\})$, so that all nodes are initially unexplored, and node x is declared to be discovered.

Step t: If the set of discovered and yet unexplored nodes $W_{t-1} \setminus V_{t-1}$ is nonempty, select node $x_t = \min(W_{t-1} \setminus V_{t-1})$ to be explored, denote by $\Delta W_t = N_G(x_t) \setminus W_{t-1}$ the set of yet undiscovered nodes adjacent to x_t, and update the sets of explored and discovered nodes according to

$$(V_t, W_t) \;=\; (V_{t-1} \cup \{x_t\},\; W_{t-1} \cup \Delta W_t),$$

Otherwise do nothing and set $(V_t, W_t) = (V_{t-1}, W_{t-1})$.

To get an intuitive picture on how this algorithm works, the reader is recommended to run it by pen and paper for some small disconnected graph, for example a union of two 5-cycles.

The above definitions imply that V_t and W_t are increasing sequences with $V_t \subset W_t$ for all t. The exploration halts at the time when the set of explored nodes V_t reaches the set of discovered nodes W_t. This time instant is called the *halting time* and denoted by

$$T \;=\; \min\{t \geq 0 : S_t = 0\}, \tag{4.8}$$

where $S_t = |W_t \setminus V_t|$ denotes the size of the *exploration queue*, that is the number of discovered nodes waiting to be explored.

A crucial property of the exploration is that for any $t \geq 0$, the set of discovered nodes W_t equals the set of nodes reachable from x by a path of length at most t. Hence at the halting time the algorithm has explored and discovered all nodes in the component of x, so that $V_T = W_T = C_G(x)$. We also see that the number of explored nodes grows by one at every time step $t = 1, \dots, T$, and we conclude that

$$|C_G(x)| \;=\; |V_T| \;=\; T. \tag{4.9}$$

This key formula allows us to reduce the analysis of the exploration process (on a very high-dimensional state space) to a one-dimensional process.

Remark. In the definition of the exploration process, we can select the node x_t from $W_{t-1} \setminus V_{t-1}$ in an arbitrary fashion. This allows to do a breadth-first search or depth-first search if we wish.

Exercise 4.5. Denote by $\mathcal{C}_G = \{C_G(x) : x \in V(G)\}$ the set of components in a graph G with n nodes. Show that

(a) $x \overset{G}{\leftrightsquigarrow} y$ is an equivalence relation on $V(G)$.

(b) The size of $\mathcal{C}_G$ equals n if and only if all nodes are isolated.

(c) The size of $\mathcal{C}_G$ equals 1 if and only if G is connected.

(d) There exists a node x such that $|C_G(x)| \geq n/2$ if G is disconnected.

4.2.3 Heuristic description of the exploration queue in a large sparse graph

Consider now a homogeneous Bernoulli graph on node set $[n]$ with link probability $p_n = \lambda n^{-1}$, and let us inspect how the exploration behaves. We start with some node x and $S_0 = 1$. In the first exploration step exploration queue becomes

$$S_1 \;=\; S_0 - 1 + X_1,$$

where X_1 is distributed according to $\mathrm{Bin}(n-1, p_n) \approx \mathrm{Poi}(\lambda)$. The exploration halts after the first step with probability

$$\mathbb{P}(X_1 = 0) \;=\; (1 - p_n)^{n-1} \;\approx\; e^{-\lambda}.$$

Given that the exploration did not halt and $S_1 = s_1$ for some $s_1 > 0$, then during the second step the exploration queue becomes

$$S_2 \;=\; S_1 - 1 + X_2,$$

where now X_2 is distributed according to $\mathrm{Bin}(n-2-s_1, p_n)$ which is still approximately $\mathrm{Poi}(\lambda)$, and X_2 is independent of the past given the current state. This why during the initial phases of the exploration, the exploration queue is close in distribution to a *random walk* (S'_t) defined by $S'_0 = 1$ and

$$S'_{t+1} \;=\; S'_t - 1 + X'_{t+1},$$

where $X'_1, X'_2, \dots$ are *independent* $\mathrm{Poi}(\lambda)$-distributed random integers. Denote by

$$T' \;=\; \min\{t \geq 0 : S'_t = 0\}$$

the first hitting time into zero of the random walk (with $T' = \infty$ meaning that the random walk never hits zero). The one can show that (draw a picture, see [vdH17, Chap 4]) T' has the same distribution as the total progeny of a branching process (Z_t) with offspring distribution $\mathrm{Poi}(\lambda)$, especially

$$\mathbb{P}(T' < \infty) \;=\; \eta_\lambda$$

where the extinction probability η_λ is the smallest fixed point of the generating function $G_{\mathrm{Poi}(\lambda)}(s) = e^{\lambda(s-1)}$. This is why we expect that

$$\mathbb{P}(|C_G(x)| \text{ is large}) \;\approx\; \mathbb{P}(T' = \infty) \;=\; 1 - \eta_\lambda.$$

More precisely, it can be shown that for a suitably chosen $a > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|C_G(x)| > a \log n) \;=\; \mathbb{P}(T' = \infty) \;=\; 1 - \eta_\lambda.$$

When the limiting average degree of the graph $\lambda > 1$, we have $\eta_\lambda \in (0,1)$, so that large components may exist. When $\lambda < 1$, it follows that with high probability, all components are of size at most $a \log n$.

Proving these things rigorously is done in [vdH17, Chap 4], by a careful coupling of the three processes:

1. Graph exploration queue S_t

2. Random walk S'_t with Poisson (resp. Bin) step sizes. (If we declare 0 as absorbing state for this random walk, then (S'_t) has the same distribution as the exploration queue of the branching tree; the exploration process can also be defined for potentially infinite random graphs.)

3. Branching process Z_t with Poisson offspring.

Chapter 5

Coupling and stochastic similarity

Assume that we have observed two large graphs such that one is a realization of an ER graph with link probability p, and the other is a realization of an ER graph with link probability q. Can we identify which one is which? If the distributions of the data sources are sufficiently different from each other, we should be able to find a decision rule to detect this. Otherwise not. Here we will analyze this type of questions in detail.

Todo: Converge in distribution iff $d_{\mathrm{tv}} \to 0$

Todo: tv vs. Wasserstein

5.1 Total variation distance

Recall from Section 3.1 that a coupling of random variables X_1 and X_2 is a bivariate random variable $(\hat{X}_1, \hat{X}_2)$ such that $\mathrm{Law}(\hat{X}_1) = \mathrm{Law}(X_1)$ and $\mathrm{Law}(\hat{X}_2) = \mathrm{Law}(X_2)$. Besides stochastic ordering, the stochastic coupling method is also well suited for analyzing how close to each other two probability distributions are. For simplicity, we will restrict here to probability distributions on finite and countably infinite spaces.

A *probability density* on a finite or countably infinite set S is a function $x \mapsto p(x)$ on S such that $p(x) \geq 0$ for all x and $\sum_{x \in S} p(x) = 1$. We write $p(A) = \sum_{x \in A} p(x)$ and note that in this way we may associate each probability density $x \mapsto p(x)$ a unique probability measure $A \mapsto p(A)$. With this convention we write $\mathrm{Law}(X) = p$ when $\mathbb{P}(X = x) = p(x)$ for all x. Table 5.1 lists the most important discrete probability densities.

A *coupling* of discrete probability densities $p_1(x)$ and $p_2(y)$ is a probability

Abbreviation	Name	S	$p(x)$
$\mathrm{Uni}([n])$	Uniform	$\{1,\dots,n\}$	$\frac{1}{n}$
$\mathrm{Ber}(p)$	Bernoulli	$\{0,1\}$	$(1-p)^{1-x}p^x$
$\mathrm{Bin}(n,p)$	Binomial	$\{0,1,\dots,n\}$	$\binom{n}{x}(1-p)^{n-x}p^x$
$\mathrm{Poi}(\lambda)$	Poisson	$\{0,1,2,\dots\}$	$e^{-\lambda}\frac{\lambda^x}{x!}$

Table 5.1: Discrete probability densities.

density $(x,y) \mapsto \hat{p}(x,y)$ such that

$$\begin{aligned} \sum_y \hat{p}(x,y) &= p_1(x) \quad \text{for all } x, \\ \sum_x \hat{p}(x,y) &= p_2(y) \quad \text{for all } y. \end{aligned}$$

Then for discrete random variables we see that $(\hat{X}_1, \hat{X}_2)$ is a coupling of X_1 and X_2 if and only if the density $\hat{p} = \mathrm{Law}(\hat{X}_1, \hat{X}_2)$ is a coupling of the densities $p_1 = \mathrm{Law}(X_1)$ and $p_2 = \mathrm{Law}(X_2)$.

To study how similar two probability distributions, we need a concept of distance. Two natural metrics are the following. The *total variation distance* between probability measures p and q on S is defined[1] by

$$d_{\mathrm{tv}}(p,q) = \max_{A \subset S} |p(A) - q(A)|, \tag{5.1}$$

and the *L_1-distance* between probability densities p and q on a countable set S by

$$\|p-q\|_1 = \sum_{x \in S} |p(x) - q(x)|.$$

The result below summarizes the key facts about total variation distance. Inequality (5.2) below shows that the total variation distance between two probability distributions can be bounded by finding a coupling $(\hat{X}, \hat{Y})$ for which $\hat{X} = \hat{Y}$ occurs with large probability. The proof of the theorem is deferred to Section 5.4.

Theorem 5.1. *Let X and Y be random variables on S with probability densities p and q. Then*

(i) $d_{\mathrm{tv}}(p,q) = \frac{1}{2}\|p-q\|_1$.

[1] It can be shown that the maximum on the right side of (5.1) is always well defined, and it makes sense to write 'max' instead of 'sup'.

(ii) For any coupling $(\hat{X},\hat{Y})$ of X and Y,

$$d_{\mathrm{tv}}(p,q) \;\le\; \mathbb{P}(\hat{X}\neq\hat{Y}). \tag{5.2}$$

(iii) There exists a coupling for which the above inequality holds as equality.

The following example illustrates the total variation distance of coin flips.

Example 5.2. The total variation distance between Bernoulli distributions $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$ can be computed as half the L_1-distance between the corresponding densities f_p and f_q according to

$$\begin{aligned} d_{\mathrm{tv}}(\mathrm{Ber}(p),\mathrm{Ber}(q)) &= \frac{1}{2}\sum_{x=0}^{1}|f_p(x)-f_q(x)| \\ &= \frac{1}{2}\left(|(1-p)-(1-q)|+|p-q|\right) \\ &= |p-q|. \end{aligned}$$

If $p\le q$, then the probability density

$$\hat{f}(x,y) \;=\; \begin{cases} 1-q, & (x,y)=(0,0), \\ q-p, & (x,y)=(0,1), \\ 0, & (x,y)=(1,0), \\ p, & (x,y)=(1,1). \end{cases}$$

is a coupling of the densities f_p and f_q which is minimal in the sense that if $(\hat{X},\hat{Y})$ is $\hat{f}$-distributed, then

$$\mathbb{P}(\hat{X}\neq\hat{Y}) \;=\; \sum_{(x,y):x\neq y}\hat{f}(x,y) \;=\; \hat{f}(0,1)+\hat{f}(1,0) \;=\; q-p \;=\; d_{\mathrm{tv}}(\mathrm{Ber}(p),\mathrm{Ber}(q)).$$

Let us illustrate the application of the above theorem by computing the total variation distance between a Bernoulli distribution and a Poisson distribution.

Example 5.3. Compute the total variation distance between probability distribution $\mathrm{Ber}(p)$ and $\mathrm{Poi}(p)$, where $p\in[0,1]$.

We extend the probability density $f=\mathrm{Ber}(p)$, originally defined on $\{0,1\}$, to the set of all nonnegative integers by setting $f(x)=0$ for $x\ge 2$. Then by noting that $1-p \le e^{-p}$, we see that the L_1-distance between f and $g = \mathrm{Poi}(p)$ equals

$$\begin{aligned}
\|f-g\|_1 &= \sum_{x=0}^{\infty} |f(x)-g(x)| \\
&= |f(0)-g(0)| + |f(1)-g(1)| + \sum_{x=2}^{\infty} g(x) \\
&= |f(0)-g(0)| + |f(1)-g(1)| + 1 - g(0) - g(1) \\
&= |1-p-e^{-p}| + |p-pe^{-p}| + 1 - e^{-p} - pe^{-p} \\
&= e^{-p} - (1-p) + p(1-e^{-p}) + 1 - e^{-p} - pe^{-p} \\
&= 2p(1-e^{-p}).
\end{aligned}$$

Then by applying Theorem 5.1 we conclude that

$$d_{\mathrm{tv}}(\mathrm{Ber}(p), \mathrm{Poi}(p)) \;=\; p(1-e^{-p}). \tag{5.3}$$

Lemma 5.4. *The total variation distance between two Poisson distributions with means $\lambda_1, \lambda_2 \ge 0$ is bounded by*

$$d_{\mathrm{tv}}(\mathrm{Poi}(\lambda_1), \mathrm{Poi}(\lambda_2)) \;\le\; 1 - e^{-|\lambda_1-\lambda_2|}.$$

Proof. Exercise. (Hint: The sum of independent Poisson-distributed random integers is again Poisson distributed.) □

5.2 Total variation and hypothesis testing

Statistical hypothesis testing concerns the inference of an unknown parameter θ of a statistical model P_θ based on an observed data sample x. In the simplest case it is a priori known that there are only two options: a random data source output a random variable X distributed either according to P_{θ_1} or P_{θ_2}, and the problem is solved using a *decision rule* $x \mapsto \psi(x) \in \{\theta_1, \theta_2\}$ indicating the decision maker's guess of the unknown parameter.

Bear in mind that here the decision maker knows the distributions P_{θ_1} and P_{θ_2} precisely, which is usually in practice a rather strong assumption. Nevertheless, in certain cases there could be strong reasons to expect that, for example, both distributions are approximately Poisson, just with an unknown mean.

Let X be a random variable describing the output of the data source before observing the data. The probability of making an error equals

$$\begin{cases} P_{\theta_1}(\psi(X) = \theta_2), & \text{if the true parameter equals } \theta_1, \\ P_{\theta_2}(\psi(X) = \theta_1), & \text{if the true parameter equals } \theta_2, \end{cases}$$

and the average error probability of the decision rule ψ equals

$$p_e(\psi) \;=\; \frac{1}{2}\Big(P_{\theta_1}(\psi(X)=\theta_2) + P_{\theta_2}(\psi(X)=\theta_1)\Big).$$

In a setting where the decision maker needs to solve the hypothesis testing problem in a large collection of cases where θ_1 and θ_2 are equally prevalent, then the best decision rule in the long run is a function ψ which minimizes the average error probability $p_e(\psi)$.

How well can the decision maker do when the best decision rule is in use? If the distributions P_{θ_1} and P_{θ_2} produce very similar samples, then it is hard to do well. The following result, sometimes attributed to Le Cam, shows that the total variation distance precisely characterizes the smallest possible error rate.

Theorem 5.5. *The average error probability of the best possible decision rule equals*

$$\min_\psi p_e(\psi) \;=\; \frac{1}{2}\left(1 - d_{\rm tv}(P_{\theta_1}, P_{\theta_2})\right).$$

Proof. For an upper bound, let first show that no decision rule can do better than what is stated. Fix any decision rule (measurable function) $x \mapsto \psi(x) \in \{\theta_1, \theta_2\}$. Denote $B = \{x : \psi(x) = \theta_1\}$. Then

$$\begin{aligned}
p_e(\psi) &= \frac{1}{2}\left(P_{\theta_1}(B^c) + P_{\theta_2}(B)\right) \\
&= \frac{1}{2}\left(1 - (P_{\theta_1}(B) - P_{\theta_2}(B))\right) \\
&\ge \frac{1}{2}\left(1 - \sup_A |P_{\theta_1}(A) + P_{\theta_2}(A)|\right) \\
&= \frac{1}{2}\left(1 - d_{\rm tv}(P_{\theta_1}, P_{\theta_2})\right).
\end{aligned}$$

For a lower bound, we show that there exists an optimal decision rule. Namely, fix a set A for which $d_{\rm tv}(P_{\theta_1}, P_{\theta_2}) = P_{\theta_1}(A) - P_{\theta_2}(A)$ (such a set always exists, see Section 5.4). Then define a decision rule as

$$\psi(x) \;=\; \begin{cases} \theta_1, & x \in A, \\ \theta_2, & x \in A^c. \end{cases}$$

Then

$$\begin{aligned}
d_{\rm tv}(P_{\theta_1}, P_{\theta_2}) &= P_{\theta_1}(A) - P_{\theta_2}(A) \\
&= P_{\theta_1}(\psi(X)=\theta_1) - P_{\theta_2}(\psi(X)=\theta_1) \\
&= 1 - P_{\theta_1}(\psi(X)=\theta_2) - P_{\theta_2}(\psi(X)=\theta_1) \\
&= 1 - 2p_e(\psi),
\end{aligned}$$

so that $p_e(\psi) = \frac{1}{2}(1 - d_{\mathrm{tv}}(P_{\theta_1}, P_{\theta_2}))$. □

Indeed, it can be shown that when P_{θ_1} and P_{θ_2} have densities f_{θ_1} and f_{θ_2}, then the set $A = \{x : f_{\theta_1}(x) \geq f_{\theta_2}(x)\}$ is an optimal set. When both densities are strictly positive, an optimal decision rule can be written as

$$\psi(x) \ = \ \begin{cases} \theta_1, & \frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} \geq 1, \\ \theta_2, & \text{else}. \end{cases}$$

This is known as the *likelihood ratio test*, and the average error rate of this test is the best possible $\frac{1}{2}(1 - d_{\mathrm{tv}}(P_{\theta_1}, P_{\theta_2}))$. In the worst case where $P_{\theta_1} = P_{\theta_2}$ the error rate is $\frac{1}{2}$, the same as obtained by blind guessing. In the best case where the two measures have disjoint support, the error rate is zero.

5.3 Stochastic similarity of Bernoulli graphs

5.3.1 Hellinger distance

As a preliminary, we will discuss the following distance between probability measures. Let f and g be probability distributions on a finite or countable infinite space S. The *Hellinger distance*[2] between f and g is defined by

$$d_{\mathrm{H}}(f,g) \ = \ \sqrt{\frac{1}{2}\sum_{x\in S}\left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2}.$$

This distance is closely connected to the total variation distance via the inequalities [vdH17, Exercise 6.32]

$$d_{\mathrm{H}}(f,g)^2 \ \leq \ d_{\mathrm{tv}}(f,g) \ \leq \ 2^{1/2} d_{\mathrm{H}}(f,g). \tag{5.4}$$

Also, the Hellinger distance satisfies the formula

$$d_{\mathrm{H}}(f,g)^2 \ = \ 1 - \sum_{x\in S} f(x)^{1/2} g(x)^{1/2}.$$

An important feature of the Hellinger distance is that it behaves well with product distributions. Assume that $S = S_1 \times \cdots \times S_m$ for some finite or countably infinite spaces S_i, and

$$f(x) \ = \ \prod_i f_i(x_i), \quad g(x) \ = \ \prod_i g_i(x_i).$$

for some probability distributions f_i, g_i on S_i. Then (exercise)

$$1 - d_{\mathrm{H}}(f,g)^2 \ = \ \prod_{i=1}^m \left(1 - d_{\mathrm{H}}(f_i, g_i)^2\right). \tag{5.5}$$

[2] Named after the German mathematician Ernst Hellinger (1883–1950).

5.3.2 Asymptotic equivalence

The goal of this section is to show that Bernoulli random graphs $G(p)$ and $G(q)$ are statistically similar when the entries of the link probability matrices $p = (p_{ij})$ and $q = (q_{ij})$ are close to each other. We measure the difference using the total variation distance[3]

$$d_{\mathrm{tv}}(\mathrm{Law}(G(p)), \mathrm{Law}(G(q))) \;=\; \frac{1}{2} \sum_{g \in \mathcal{G}_n} |\mathbb{P}(G(p) = g) - \mathbb{P}(G(q) = g)|.$$

Theorem 5.6. *Assume that $p_{ij} \le 1 - \epsilon$ for all i, j. Then the total variation distance between the distributions on $G(p)$ and $G(q)$ is bounded by*

$$d_{\mathrm{tv}}(G(p), G(q))^2 \;\le\; (1 + \epsilon^{-1}) \sum_{1 \le i < j \le n} \frac{(p_{ij} - q_{ij})^2}{p_{ij}}.$$

Proof. Observe that the probability distribution of $G(p)$ factorizes according to

$$\mathbb{P}(G(p) = g) \;=\; \prod_{1 \le i < j \le n} f_{p_{ij}}(g_{ij}),$$

where $f_{p_{ij}}(x) = (1 - p_{ij})^{1-x} p_{ij}^x$ denotes the Bernoulli density with mean p_{ij}, and a similar factorization is valid also for $G(q)$. Therefore, by (5.5),

$$1 - d_{\mathrm{H}}(G(p), G(q))^2 \;=\; \prod_{1 \le i < j \le n} \left(1 - d_{\mathrm{H}}(f_{p_{ij}}, f_{q_{ij}})^2\right).$$

Because $\prod_k (1 - t_k) \ge 1 - \sum_k t_k$ for any $t_k \in [0, 1]$ (you may interpret this inequality as an union bound of some events of probability t_k), it follows that

$$1 - d_{\mathrm{H}}(G(p), G(q))^2 \;\ge\; 1 - \sum_{1 \le i < j \le n} d_{\mathrm{H}}(f_{p_{ij}}, f_{q_{ij}})^2,$$

so that

$$d_{\mathrm{H}}(G(p), G(q))^2 \;\le\; \sum_{1 \le i < j \le n} d_{\mathrm{H}}(f_{p_{ij}}, f_{q_{ij}})^2. \tag{5.6}$$

Observe next that

$$(\sqrt{a} - \sqrt{b})^2 \;=\; \frac{(a - b)^2}{(\sqrt{a} + \sqrt{b})^2} \;\le\; \frac{(a - b)^2}{a + b},$$

[3] For convenience, we sometimes use the shorthand $d_{\mathrm{tv}}(X, Y) = d_{\mathrm{tv}}(\mathrm{Law}(X), \mathrm{Law}(Y))$.

and that $1 - p_{ij} \geq \epsilon \geq \epsilon p_{ij}$ when $p_{ij} \leq 1 - \epsilon$. By applying these inequalities, we find that

$$\begin{aligned} d_{\mathrm{H}}(f_{p_{ij}}, f_{q_{ij}})^2 &= \frac{1}{2}(\sqrt{p_{ij}} - \sqrt{q_{ij}})^2 + \frac{1}{2}(\sqrt{1-p_{ij}} - \sqrt{1-q_{ij}})^2 \\ &\leq \frac{1}{2}\frac{(p_{ij}-q_{ij})^2}{p_{ij}+q_{ij}} + \frac{1}{2}\frac{(p_{ij}-q_{ij})^2}{(1-p_{ij})+(1-q_{ij})} \\ &\leq \frac{1}{2}\frac{(p_{ij}-q_{ij})^2}{p_{ij}} + \frac{1}{2}\frac{(p_{ij}-q_{ij})^2}{1-p_{ij}} \\ &\leq \frac{1}{2}(1+\epsilon^{-1})\frac{(p_{ij}-q_{ij})^2}{p_{ij}}. \end{aligned}$$

Now by (5.6) it follows that

$$d_{\mathrm{H}}(G(p), G(q))^2 \leq \frac{1}{2}(1+\epsilon^{-1}) \sum_{1 \leq i < j \leq n} \frac{(p_{ij}-q_{ij})^2}{p_{ij}},$$

so that by (5.4),

$$d_{\mathrm{tv}}(G(p), G(q))^2 \leq 2 d_{\mathrm{H}}(G(p), G(q))^2 \leq (1+\epsilon^{-1}) \sum_{1 \leq i < j \leq n} \frac{(p_{ij}-q_{ij})^2}{p_{ij}}.$$

Hence the claim follows. □

5.3.3 Application: 2-community SBM

Consider a SBM G which is a Bernoulli random graph with $p_{ij} = \rho_n K_{z_i, z_j}$, where $z_1, \dots, z_n \in \{1, \dots, m\}$ and K is a symmetric m-by-m matrix with nonnegative entries. The probability that a uniformly randomly chosen node pair $\{i,j\}$ is linked in a SBM equals

$$\begin{aligned}
\binom{n}{2}^{-1} \sum_{1\le i<j\le n} \rho_n K_{z_i,z_j} &= \binom{n}{2}^{-1} \frac{1}{2} \sum_i \sum_{j\ne i} \rho_n K_{z_i,z_j} \\
&= \rho_n \binom{n}{2}^{-1} \frac{1}{2} \left(\sum_i \sum_j K_{z_i,z_j} - \sum_i K_{z_i,z_i} \right) \\
&= \rho_n \binom{n}{2}^{-1} \frac{1}{2} \left(n^2 \sum_s \sum_t \mu_s K_{s,t} \mu_t - n \sum_s \mu_s K_{s,s} \right) \\
&= \rho_n \frac{n}{n-1} \left(\sum_s \sum_t \mu_s K_{s,t} \mu_t - n^{-1} \sum_s \mu_s K_{s,s} \right) \\
&= (1+O(n^{-1}))\rho_n \sum_s \sum_t \mu_s K_{s,t} \mu_t \\
&= (1+O(n^{-1}))\rho_n (\mu^T K \mu).
\end{aligned}$$

We compare a Bernoulli graph with link rates p_{ij} to an ER graph with constant link rate $p = \rho_n(\mu^T K \mu)$. Then Theorem 5.6 gives an upper bound (we omit the $1+\epsilon^{-1}$ constant)

$$\begin{aligned}
d_{\mathrm{tv}}(G, ER(p))^2 \le \sum_{1\le i<j\le n} \frac{(p-p_{ij})^2}{p} &\le \sum_{1\le i,j\le n} \frac{(p-p_{ij})^2}{p} \\
&= n^2 \sum_s \sum_t \mu_s \mu_t \frac{(\rho_n(\mu^T K\mu) - \rho_n K_{s,t})^2}{\rho_n(\mu^T K \mu)} \\
&= n^2 \rho_n \frac{\sum_s \sum_t \mu_s \mu_t (K_{s,t} - \mu^T K \mu)^2}{\mu^T K \mu}
\end{aligned}$$

Note that

$$\mu^T K \mu = \sum_s \sum_t K_{s,t} \mu_s \mu_t = \mathbb{E} K_{S,T}$$

is the expected link rate between communities S and T chosen uniformly at random and independently of each other according to community frequencies (μ_s), and

$$\sum_s \sum_t \mu_s \mu_t (K_{s,t} - \mu^T K \mu)^2 = \mathrm{Var}(K_{S,T}).$$

Hence the upper bound can be written as

$$d_{\mathrm{tv}}(G, ER(p))^2 \le n^2 \rho_n \frac{\mathrm{Var}(K_{S,T})}{\mathbb{E} K_{S,T}}.$$

In the special case with $m = 2$ communities, and $\mu_1 = \mu_2 = \frac{1}{2}$, and with $K_{11} = K_{22} = a$ and $K_{12} = K_{21} = b$, it follows that $\mathbb{E}K_{S,T} = \frac{1}{2}(a+b)$. Also

$$\mathbb{E}K_{S,T}^2 \;=\; \frac{1}{2}(a^2+b^2),$$

so that

$$\mathrm{Var}(K_{S,T}) \;=\; \frac{1}{2}(a^2+b^2) - \left(\frac{1}{2}(a+b)\right)^2 \;=\; \frac{1}{4}(a-b)^2.$$

And hence the bound becomes

$$d_{\mathrm{tv}}(G, ER(p))^2 \;\leq\; n^2\rho_n \frac{\frac{1}{4}(a-b)^2}{\frac{1}{2}(a^2+b^2)} \;=\; n^2\rho_n \frac{(a-b)^2}{2(a+b)}.$$

For example, for $\rho_n = n^{-1}$, we see that $|a-b| \ll n^{-1/2}$ is sufficient to conclude that $d_{\mathrm{tv}}(G, ER(p))$ is near zero. Mossel et al. [MNS15] have a more detailed analysis for the 2-community case where they find that

$$\frac{(a-b)^2}{2(a+b)} < 1$$

is a critical condition, under which the SBM cannot be distinguished from an ER graph with the same mean degree. The above inequality is known as the Kesten–Stigum condition.

5.4 Proof of Theorem 5.1

Proof. As preliminaries, denote $a \wedge b = \min\{a,b\}$ and $a_+ = \max\{a,0\}$, and set $I_0 = \sum_x (p_x \wedge q_x)$ and

$$\begin{aligned} I_p &= \sum_x (p_x - p_x \wedge q_x) &= \sum_x (p_x - q_x)_+, \\ I_q &= \sum_x (q_x - p_x \wedge q_x) &= \sum_x (q_x - p_x)_+. \end{aligned}$$

Then one can show (see Figure 5.1) that $I_0 + I_p = 1$, $I_0 + I_q = 1$, and $I_p + I_q = \|p-q\|_1$, from which we find that

$$1 - I_0 \;=\; I_p \;=\; I_q \;=\; \frac{1}{2}\|p-q\|_1.$$

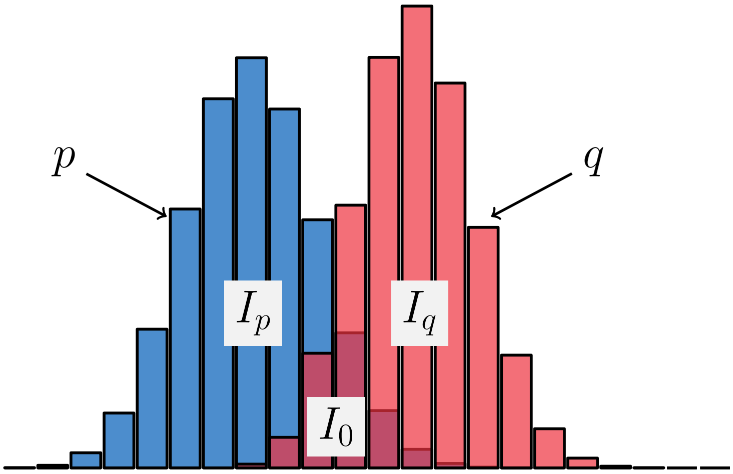

Figure 5.1: Partition of the region below densities p and q into three parts.

(i) To prove the first statement, denote $A_p = \{x : p_x > q_x\}$ and $A_q = \{x : q_x > p_x\}$. Because

$$\sum_{x \in A_p} (p_x - q_x) \;=\; \sum_{x \in S} (p_x - p_x \wedge q_x) \;=\; I_p \;=\; \frac{1}{2}\|p - q\|_1,$$

it follows that

$$p(A) - q(A) \;=\; \sum_{x \in A} (p_x - q_x) \;\le\; \sum_{x \in A \cap A_p} (p_x - q_x) \;\le\; \sum_{x \in A_p} (p_x - q_x) \;\le\; \frac{1}{2}\|p - q\|_1$$

for all $A \subset S$. By symmetry, the same upper bound holds for $q(A) - p(A)$, and we conclude that

$$d_{\mathrm{tv}}(p, q) \;=\; \sup_{A \subset S} |p(A) - q(A)| \;\le\; \frac{1}{2}\|p - q\|_1.$$

For the opposite direction, note that

$$\frac{1}{2}\|p - q\|_1 \;=\; \frac{1}{2}\left(\sum_{x \in A_p} (p_x - q_x) + \sum_{x \in A_q} (q_x - p_x) \right) \;=\; \frac{1}{2}\,(I_p + I_q) \;=\; I_p,$$

and

$$I_p \;=\; \sum_{x \in A_p} (p_x - q_x) \;=\; p(A_p) - q(A_p) \;\le\; |p(A_p) - q(A_p)| \;\le\; d_{\mathrm{tv}}(p, q).$$

Hence also $\frac{1}{2}\|p-q\|_1 \le d_{\mathrm{tv}}(p,q)$ and we are done with the first claim.

(ii) Fix a coupling $(\hat{X}, \hat{Y})$ of X and Y, and note that for all x,

$$\hat{X} = \hat{Y} \ \le\ \mathbb{P}(\hat{X} = x) \ =\ p_x$$

and

$$\mathbb{P}(\hat{X} = \hat{Y} = x) \ \le\ \mathbb{P}(\hat{Y} = x) \ =\ q_x.$$

This implies that

$$\mathbb{P}(\hat{X} = \hat{Y}) \ =\ \sum_{x \in S} \mathbb{P}(\hat{X} = \hat{Y} = x) \ \le\ \sum_{x \in S} (p_x \wedge q_x) \ =\ I_0 \ =\ 1 - \frac{1}{2}\|p-q\|_1.$$

Hence the second claim follows by applying (i).

(iii) We will construct a coupling for which the inequality holds as equality. Assume that $p \ne q$ (otherwise we can construct a trivial coupling, can you figure out which one?), and let $c = d_{\mathrm{tv}}(p,q) > 0$. Define a function $\ell : S \times S \to \mathbb{R}$ by

$$\ell_{x,y} \ =\ 1(x=y)(p_x \wedge q_x) + c^{-1}(p_x - p_x \wedge q_x)(p_y - p_y \wedge q_y).$$

We will show that this is what we want. Observe first that because $c = I_q$,

$$\begin{aligned}
\sum_y \ell_{x,y} \ &=\ p_x \wedge q_x + c^{-1}(p_x - p_x \wedge q_x) \sum_y (p_y - p_y \wedge q_y) \\
&=\ p_x \wedge q_x + c^{-1}(p_x - p_x \wedge q_x) I_q \\
&=\ p_x
\end{aligned}$$

for all x, so that

$$\sum_x \sum_y \ell_{x,y} \ =\ \sum_x p_x \ =\ 1.$$

Because $\ell \ge 0$, this shows that ℓ is a probability density on $S \times S$. By symmetry, the above computation also shows that $\sum_x \ell_{x,y} = p_y$ for all y, and we conclude that ℓ is a coupling of p and q. Now let $(\hat{X}, \hat{Y})$ be a random variable in $S \times S$ with (joint) probability distribution ℓ. Then $(\hat{X}, \hat{Y})$ is a coupling of X and Y, and

$$\mathbb{P}(\hat{X} = \hat{Y}) \ =\ \sum_x \mathbb{P}(\hat{X} = \hat{Y} = x) \ =\ \sum_x \ell_{x,x} \ =\ \sum_x (p_x \wedge q_x) \ =\ I_0.$$

Because $I_0 = 1 - d_{\mathrm{tv}}(p,q)$, it follows that (5.2) holds as equality. □

5.5 Remarks

Theorem 5.6 is due to Svante Janson [Jan10]. Mossel, Neeman, and Sly have sharp analysis of SBMs with two communities [MNS15].

Chapter 6

Degree distributions

6.1 Average degrees in stochastic block models

In this section we will study a stochastic block model with n nodes and m communities, having density parameter ρ and community link matrix K. This is a Bernoulli random graph G on node set $[n]$ where the link probabilities are of the form

$$p_{ij} \;=\; \rho K_{z_i,z_j},$$

where $\rho > 0$ is scalar, $z = (z_1, \dots, z_n)$ is a list of node attributes with values in $[m]$, and K is a symmetric nonnegative m-by-m matrix. Here we denote by

$$\mu_s \;=\; \frac{1}{n}\sum_{i=1}^{n} 1(z_i = s)$$

the relative frequency of nodes in community s. The vector $(\mu_s)_{s=1}^m$ is a probability distribution on $[m]$ called the *empirical community distribution*. The following result show that the in a large stochastic block model where all community frequencies are bounded away from zero, the mean degree is described by the community frequencies.

Theorem 6.1. *Assume that* $\min_{1\le s\le m} \mu_s \ge \epsilon$. *Then the expected degree of any node* i *community* s *equals*

$$\mathbb{E}\deg_G(i) \;\sim\; n\rho_n \sum_{t=1}^{m} K_{s,t}\mu_t,$$

and the expected average degree equals

$$\mathbb{E}\deg_G(U) \;\sim\; n\rho_n \sum_{s=1}^m \sum_{t=1}^m \mu_s K_{s,t}\mu_t.$$

Proof. The degree of node i may be written as $\deg_G(i) = \sum_{j\neq i} G_{ij}$ where G_{ij} are independent $\mathrm{Ber}(p_{ij})$-distributed random variables. By denoting the community of node i by $s = z_i$, we see that the expected degree of node i equals

$$\mathbb{E}\deg_G(i) \;=\; \sum_{j\neq i} p_{ij} \;=\; \sum_{j\neq i} \rho_n K_{z_i,z_j} \;=\; \rho_n \sum_{j=1}^n K_{z_i,z_j} - \rho_n K_{z_i,z_i}$$

Observe next that the number of nodes in community t equals $n\mu_t$, and therefore

$$\sum_{j=1}^n K_{z_i,z_j} \;=\; \sum_{t=1}^m K_{z_i,t} n\mu_t,$$

so that

$$\mathbb{E}\deg_G(i) \;=\; n\rho_n \sum_{t=1}^m K_{s,t}\mu_t - \rho_n K_{s,s}.$$

Because $1 \le \epsilon^{-1}\mu_s$, the last term on the right is bounded by

$$K_{s,s} \;\le\; \epsilon^{-1} K_{s,s}\mu_s \;\le\; \epsilon^{-1} \sum_{t=1}^m K_{s,t}\mu_t, \tag{6.1}$$

and hence we obtain

$$\mathbb{E}\deg_G(i) \;=\; (1 + O(\epsilon^{-1}n^{-1}) n\rho_n \sum_{t=1}^m K_{s,t}\mu_t.$$

For the expected average degree, by a similar argument we find that

$$\mathbb{E}\left(n^{-1} \sum_{i=1}^n \deg_G(i) \right) \;=\; n\rho_n \sum_{s=1}^m \sum_{t=1}^m \mu_s K_{s,t}\mu_t - \rho_n \sum_{s=1}^m \mu_s K_{s,s}.$$

By multiplying both sides of (6.1) by μ_s and summing over s, we find that

$$\sum_s \mu_s K_{s,s} \;\le\; \epsilon^{-1} \sum_s \sum_t \mu_s K_{s,s}\mu_t,$$

so that we obtain

$$\mathbb{E}\deg_G(U) \;=\; (1 + O(\epsilon^{-1}n^{-1}))n\rho_n \sum_{s=1}^m \sum_{t=1}^m \mu_s K_{s,t}\mu_t.$$

□

Exercise 6.2. Verify that under same the assumptions of Theorem 6.1, for a node in community s, the mean number of neighbors in community t is approximately $n\rho_n K_{s,t}\mu_t$. See Table 6.1.

ρ	Average degree	Regime
$\rho \ll n^{-1}$	$d_{\text{ave}} \ll 1$	Very sparse
$\rho \approx cn^{-1}$	$d_{\text{ave}} \approx c$	Sparse with bounded degree
$n^{-1} \ll \rho \ll 1$	$1 \ll d_{\text{ave}} \ll n$	Sparse with diverging degree
$\rho \approx c$	$d_{\text{ave}} \approx cn$	Dense

Table 6.1: Different regimes of large graph models.

6.2 Poisson approximation

The following result, sometimes called *Le Cam's inequality* after a famous statistician Lucien Le Cam, illustrates how to apply the stochastic coupling method to get an upper bound on the distance between a sum of independent $\{0,1\}$-valued random variables and a Poisson distribution.

Theorem 6.3. *Let A_i be independent $\{0,1\}$-valued random variables such that $\mathbb{E}A_i = a_i$ and $\sum_i a_i < \infty$. Then*

$$d_{\text{tv}}\Big(\text{Law}(\sum_i A_i),\, \text{Poi}(\sum_i a_i)\Big) \;\le\; \sum_i a_i^2.$$

Proof. By applying (5.3) and Theorem 5.1, we see that for every i there exists a coupling $(\hat{A}_i, \hat{B}_i)$ of X_i and a $\text{Poi}(a_i)$-distributed random integer B_i, so that

$$\mathbb{P}(\hat{A}_i \ne \hat{B}_i) \;\le\; a_i(1 - e^{-a_i}). \tag{6.2}$$

By a standard technique of probability theory, it is possible to construct all of the bivariate random variables $(\hat{A}_i, \hat{B}_i)$, $i \in I$, on a common probability space and in such a way that these bivariate random variables are mutually independent (nevertheless, $\hat{A}_i$ and $\hat{B}_i$ are dependent for each i). Then define $\hat{A} = \sum_i \hat{A}_i$ and $\hat{B} = \sum_i \hat{B}_i$. Then $\text{Law}(\hat{A}) = \text{Law}(\sum_i A_i)$. Moreover, because the sum of independent Poisson-distributed random integers is Poisson-distributed, it follows that $(\hat{A}, \hat{B})$ is a coupling of $\sum_i A_i$ and a $\text{Poi}(\sum_i a_i)$-distributed random integer B. By applying (6.2) and the union bound, this coupling satisfies

$$\mathbb{P}(\hat{A} \ne \hat{B}) \;=\; \mathbb{P}(\cup_{i\in I}\{\hat{A}_i \ne \hat{B}_i\}) \;\le\; \sum_{i\in I} \mathbb{P}(\hat{A}_i \ne \hat{B}_i) \;\le\; \sum_{i\in I} a_i(1 - e^{-a_i}).$$

By applying Theorem 5.1, it now follows that

$$d_{\mathrm{tv}}(\sum_i A_i, \mathrm{Poi}(\sum_i a_i)) \;\le\; \mathbb{P}(\hat{A} \neq \hat{B}) \;\le\; \sum_{i \in I} a_i(1 - e^{-a_i}).$$

This implies the claim after noting that $1 - e^{-a_i} \le a_i$. □

Exercise 6.4. For a sequence of probability distributions we denote $\mu_n \xrightarrow{tv} \mu$ when $d_{\mathrm{tv}}(\mu_n, \mu) \to 0$.

(a) Apply Le Cam's inequality to show that when $p_n \ll n^{-1/2}$,

$$d_{\mathrm{tv}}\Big(\mathrm{Bin}(n, p_n),\, \mathrm{Poi}(np_n)\Big) \;\to\; 0.$$

(b) As a consequence, derive Poisson's law of small numbers:

$$\mathrm{Bin}\left(n, \frac{\lambda}{n}\right) \xrightarrow{tv} \mathrm{Poi}(\lambda).$$

6.3 Model degree distributions in sparse SBMs

6.3.1 Degree of a node in a given community

Theorem 6.5. *In a stochastic block model with density $\rho_n \ll n^{-1/2}$, the degree distribution of any node i in community s is approximately Poisson according to*

$$d_{\mathrm{tv}}\big(\mathrm{Law}(\deg_G(i)),\, \mathrm{Poi}(n\rho_n\lambda_s)\big) \;\to\; 0,$$

where $\lambda_s = \sum_t K_{s,t}\mu_t$.

Proof. The degree of node i may be written as $\deg_G(i) = \sum_{j \neq i} G_{ij}$ where G_{ij} are independent $\mathrm{Ber}(p_{ij})$-distributed random variables. By denoting the community of node i by $s = z_i$, we see that the expected degree of node i equals $\lambda'_i = \sum_{j \neq i} p_{ij}$, which can be also written as

$$\lambda'_i \;=\; \sum_j \rho_n K_{z_i,z_j} - \rho_n K_{s,s} \;=\; n\rho_n \sum_t K_{s,t}\mu_t - \rho_n K_{s,s} \;=\; n\rho_n\lambda_s - \rho_n K_{s,s}.$$

By Le Cam's inequality (Theorem 6.3), it follows that

$$d_{\mathrm{tv}}\Big(\mathrm{Law}(\deg_G(i)),\, \mathrm{Poi}(\lambda'_i)\Big) \;\le\; \sum_{j \neq i} p_{ij}^2 \;=\; \rho_n^2 \sum_{j \neq i} (K_{z_i,z_j})^2 \;\le\; n\rho_n^2 \|K\|_\infty^2,$$

where $\|K\|_\infty = \max_{s,t} |K_{s,t}|$.

Now by Lemma 5.4 and the inequality $1-t \le e^{-t}$, it follows that it follows that

$$d_{\rm tv}\Big(\operatorname{Poi}(\lambda_i'), \operatorname{Poi}(n\rho_n\lambda_s)\Big) \ \le\ 1-e^{-|n\rho_n\lambda_s-\lambda_i'|} \ \le\ |n\rho_n\lambda_s-\lambda_i'| \ =\ \rho_n K_{s,s}.$$

In light of the above observations, the claim follows by the triangle inequality for the total variation distance:

$$\begin{aligned} & d_{\rm tv}\Big(\operatorname{Law}(\deg_G(i)), \operatorname{Poi}(n\rho_n\lambda_s)\Big) \\ &\le\ d_{\rm tv}\Big(\operatorname{Law}(\deg_G(i)), \operatorname{Poi}(\lambda_i')\Big) + d_{\rm tv}\Big(\operatorname{Poi}(\lambda_i'), \operatorname{Poi}(n\rho_n\lambda_s)\Big) \\ &\le\ n\rho_n^2\|K\|_\infty^2 + \rho_n K_{s,s}. \end{aligned}$$

The right side above tends to zero when $\rho_n \ll n^{-1/2}$. □

6.3.2 Overall degree distribution (Degree of a typical node)

By a *typical node* of a graph G we mean a node U sampled uniformly at random from the node set of the graph. When the graph is random, the degree of a typical node $\deg_G(U)$ involves two sources of randomness: the randomness associated with the graph G, and the randomness associated with the sampling of U.

A *mixed Poisson distribution* with mixing distribution f is the probability distribution $\operatorname{MPoi}(f)$ on the nonnegative integers with probability density

$$\mathbb{E} e^{-\Lambda}\frac{\Lambda^x}{x!}, \quad x = 0,1,\dots,$$

where Λ is a random variable distributed according to f, a probability distribution on $\mathbb{R}_+$. Samples from $\operatorname{MPoi}(f)$ can be generated by first sampling a random variable Λ from f, and conditionally on $\Lambda = \lambda$, sampling from a Poisson distribution with mean λ. We might denote this by $\int \operatorname{Poi}(s)f(s)\,ds$.

Theorem 6.6. *In a stochastic block model with density $\rho_n \ll n^{-1/2}$, the degree distribution of a typical node is approximately mixed Poisson according to*

$$d_{\rm tv}\Big(\operatorname{Law}(\deg_G(U)), \sum_{s=1}^m \operatorname{Poi}(n\rho_n\lambda_s)\mu_s\Big) \ \to\ 0,$$

where $\lambda_s = \sum_t K_{s,t}\mu_t$. Especially, for $\rho_n = n^{-1}$, the expected relative frequency of nodes of degree x satisfies

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n 1(\deg_G(i)=x)\right) \ =\ \mathbb{P}(\deg_G(U)=x) \ \to\ \sum_{s=1}^m e^{-\lambda_s}\frac{\lambda_s^x}{x!}\mu_s.$$

Proof. Denote by $f^{(n)}(x) = \mathbb{P}(\deg_G(U) = x)$ the typical node degree distribution. By conditioning on U we find that

$$\mathbb{P}(\deg_G(U) = x) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(\deg_G(i) = x),$$

and then by conditioning on the community of node i, we find that

$$f^{(n)}(x) \;=\; \sum_{s=1}^{m} f_s^{(n)}(x)\mu_s,$$

where

$$f_s^{(n)}(x) \;=\; \mathbb{P}(\deg_G(i) = x), \quad z_i = s,$$

the degree distribution of nodes in community s. By Theorem 6.5 (or actually its proof),

$$d_{\mathrm{tv}}\Big(f_s^{(n)}, g_s^{(n)}\Big) \;\le\; n\rho_n^2 \|K\|_\infty^2 + \rho_n K_{s,s}.$$

where $g_s^{(n)} = \mathrm{Poi}(n\rho_n\lambda_s)$. Then (use the L_1-distance representation of the total variation distance and triangle inequalities for the norm),

$$\begin{aligned} d_{\mathrm{tv}}\Big(f^{(n)}, g^{(n)}\Big) \;&=\; d_{\mathrm{tv}}\Big(\sum_s f_s^{(n)}\mu_s, \sum_s g^{(n)}\mu_s\Big) \\ &\le\; \sum_s \mu_s d_{\mathrm{tv}}\Big(f_s^{(n)}, g^{(n)}\Big). \end{aligned}$$

□

6.4 Joint degree distribution

Many results related to large random graph rely on the fact that several local characteristics of the graph are approximately independent for large n. In statistics it is important to quantify how close certain observables are to being fully independent. Here we discuss the case of degrees.

Let $G(p)$ be a Bernoulli random graph on node set $[n]$ where each unordered node pair $\{i, j\}$ is connected by a link with probability p_{ij}, independently of other node pairs. Denote by

$$\mathrm{Law}(D_i : i \in I)$$

the joint distribution of the degrees $D_i = \deg_G(i)$ for a set of nodes $I \subset [n]$. The degrees D_i are not independent, but the dependence is not strong in large sparse random graphs. We may quantify this by measuring how much the joint degree distribution deviates from the product distribution

$$\prod_{i\in I} \mathrm{Law}(D_i)$$

which represents the joint distribution of *independently* sampled random integers from the distributions $\mathrm{Law}(D_i)$.

A collection of random variables $(X_i :\in I)$ whose joint distribution depends on a scale parameter n, is called *asymptotically independent* if

$$d_{\mathrm{tv}}\left(\mathrm{Law}(X_i : i \in I),\ \prod_{i\in I}\mathrm{Law}(X_i)\right) \ \to\ 0 \quad \text{as } n \to \infty.$$

Theorem 6.7. *The joint degree distribution of an arbitrary set of nodes I in a Bernoulli random graph with link probabilities p_{ij} satisfies*

$$d_{\mathrm{tv}}\left(\mathrm{Law}(D_i : i \in I),\ \prod_{i\in I}\mathrm{Law}(D_i)\right) \ \le\ 4\sum_{i,j\in I : i<j}(1-p_{ij})p_{ij}.$$

As an immediate application of the above theory, we obtain the following result for sparse SBMs.

Proposition 6.8. *For a sparse stochastic block model with density parameter $\rho_n \ll 1$ and community link matrix K, the degrees of any set of $n_0 \ll \rho_n^{-1/2}$ nodes are asymptotically independent.*

Proof. For any node set I of size n_0,

$$4\sum_{i,j\in I:i<j}(1-p_{ij})p_{ij} \ \le\ 4\sum_{i,j\in I:i<j} p_{ij} \ \le\ 4\rho_n\|K\|_\infty n_0^2$$

The right side tends to zero when $\rho_n n_0^2 \to 0$ (and the community link matrix K does not depend on the scale parameter, which we implicitly assume here throughout). □

Proof of Theorem 6.7. The proof is based on a coupling argument described in [vdH17, Theorem 6.7(b)]. After relabeling the node set if necessary, we may and will assume that $I = \{1, 2, \dots, m\}$. Let G the adjacency matrix of the random graph, and let $\hat{G}$ be an *independent copy* of G. That is, we sample $\hat{G}$ from the same distribution as G, independently. Then we define

$$\tilde{D}_i \ =\ \sum_{j:j<i}\hat{G}_{ij} + \sum_{j:j>i} G_{ij}. \tag{6.3}$$

The random integers $\tilde{D}_i$ are *not* degrees of G nor $\hat{G}$. Nevertheless, we see that $\mathrm{Law}(\tilde{D}_i) = \mathrm{Law}(D_i)$ because all random variables on the right side above are independent. Note that

$$\begin{aligned} \tilde{D}_1 &= G_{12} + G_{13} + G_{14} + \cdots \\ \tilde{D}_2 &= \hat{G}_{12} + G_{23} + G_{24} + \cdots \\ \tilde{D}_3 &= \hat{G}_{13} + \hat{G}_{23} + G_{34} + \cdots \end{aligned}$$

and so on. Because all terms in the three above sums are independent, it follows that $\tilde{D}_1, \tilde{D}_2, \tilde{D}_3$ are independent. In fact, one may verify by induction that $\tilde{D}_1, \dots, \tilde{D}_n$ are all mutually independent. Hence so is the sublist $\tilde{D}_I = (\hat{D}_i : i \in I)$, and the distribution of the list $\tilde{D}_I$ equals $\prod_{i \in I} \mathrm{Law}(D_i)$. Now the pair $(D_I, \tilde{D}_I)$ constitutes a coupling of $\mathrm{Law}(D_I)$ and $\prod_{i \in I} \mathrm{Law}(D_i)$. Hence

$$\begin{aligned} d_{\mathrm{tv}}\left(\mathrm{Law}(D_i : i \in I), \prod_{i \in I} \mathrm{Law}(D_i)\right) &\le \mathbb{P}(D_I \ne \tilde{D}_I) \\ &= \mathbb{P}\Big(\bigcup_{i \in I} \{D_i \ne \tilde{D}_i\}\Big) \\ &\le \sum_{i \in I} \mathbb{P}(D_i \ne \tilde{D}_i). \end{aligned}$$

From (6.3) we see that $\tilde{D}_i - D_i = \sum_{j : j < i} (\hat{G}_{ij} - G_{ij})$. Hence $D_i = \tilde{D}_i$ unless $G_{ij} \ne \hat{G}_{ij}$ for one or more indices $j < i$. Therefore

$$\mathbb{P}(D_i \ne \tilde{D}_i) \le \mathbb{P}(\cup_{j : j < i} \{G_{ij} \ne \hat{G}_{ij}\}) \le \sum_{j : j < i} \mathbb{P}(G_{ij} \ne \hat{G}_{ij}).$$

Because

$$\begin{aligned} \mathbb{P}(G_{ij} \ne \hat{G}_{ij}) &= \mathbb{P}(G_{ij} = 0, \hat{G}_{ij} = 1) + \mathbb{P}(G_{ij} = 1, \hat{G}_{ij} = 0) \\ &= 2(1 - p_{ij}) p_{ij}, \end{aligned}$$

we conclude that

$$d_{\mathrm{tv}}\left(\mathrm{Law}(D_i : i \in I), \prod_{i \in I} \mathrm{Law}(D_i)\right) \le 2 \sum_{i \in I} \sum_{j \in I : j \ne i} (1 - p_{ij}) p_{ij},$$

and the claim follows. □

Exercise 6.9. If D_1 and D_2 are asymptotically independent, show that $\mathrm{cov}(\phi(D_1, \phi(D_2)) \to 0$ for any bounded function ϕ.

6.5 Empirical degree distributions

6.5.1 Empirical distributions of large data sets

To obtain a tractable sparse graph model, we need to impose some regularity assumptions on the behavior of node attributes. We will denote the *empirical distribution* of the list $x^{(n)}$ by

$$\mu_n(B) \;=\; \frac{1}{n}\sum_{i=1}^{n} 1(x_i^{(n)} \in B)$$

returns the relative frequency of node attributes with values in $B \subset \mathbb{R}$. Alternatively, μ_n is the probability distribution of random variable X_n obtained by picking an element of the list uniformly at random. We assume that for large graphs, the distribution of attributes can be approximated by a limiting probability distribution μ on $(0, \infty)$. More precisely, we assume that $\mu_n \to \mu$ *weakly*, that is,

example

$$\mathbb{E}\phi(X_n) \to \mathbb{E}\phi(X)$$

for any continuous and bounded function $\phi : (0, \infty) \to \mathbb{R}$ and random variables X_n distributed according to μ_n and X distributed according to μ. We also say that $\mu_n \to \mu$ *weakly with k-th moments* if in addition $\mathbb{E}X_n^k \to \mathbb{E}X^k$ and $\mathbb{E}X_n^k$, $\mathbb{E}X^k$ are finite[1]. For a thorough treatment of the aspects of weak convergence of probability measures, see for example [Kal02, Section 4]. The main fact is that when the limiting distribution has a continuous cumulative distribution F, then $\mu_n \to \mu$ weakly if and only if $F_n(t) \to F(t)$ for all t, where $F_n(t) = \frac{1}{n}\sum_{i=1}^n 1(x_i^{(n)} \le t)$ is the *empirical cumulative distribution* of the list $x^{(n)}$.

Example 6.10 (Random attribute lists)**.** A fundamental example is the following setting. Assume that $X_1, X_2, \dots$ are independent random numbers sampled from a probability distribution μ which has a finite k-th moment. Then the empirical distribution μ_n of the list $X^{(n)} = (X_1, \dots, X_n)$ is a *random* probability distribution. As consequence of the strong law of large numbers and the Glivenko–Cantelli theorem it follows that with probability one, $\mu_n \to \mu$ weakly with k-th moments.

Here, as elsewhere, we denote $f_n \ll g_n$ or $f_n = o(g_n)$ when $f_n/g_n \to 0$.

Lemma 6.11. *Assume the empirical distribution of $x^{(n)}$ converges weakly and with first moments to a probability distribution μ. Then $\max_{i\in[n]} x_i^{(n)} \ll n$.*

[1] This corresponds to convergence of probability measure in the Wasserstein-k metric.

Proof. Let X_n be a μ_n-distributed random number for each n. Then by Lemma A.5, the sequence (X_n) is uniformly integrable, and for any $\epsilon > 0$, it follows that

$$\begin{aligned} n\mathbb{P}(X_n > \epsilon n) &= \epsilon^{-1}\mathbb{E}\epsilon n 1(X_n > \epsilon n) \leq \epsilon^{-1}\mathbb{E}X_n 1(X_n > \epsilon n) \\ &\leq \epsilon^{-1}\sup_m \mathbb{E}X_m 1(X_m > \epsilon n) \to 0. \end{aligned}$$

But this means that

$$\sum_{i=1}^n 1(x_i^{(n)} > \epsilon n) = n\frac{1}{n}\sum_{i=1}^n 1(x_i^{(n)} > \epsilon n) = n\,\mathbb{P}(X_n > \epsilon n) \to 0.$$

Because the left side above is integer-valued, we conclude that exists n_0 such that $\sum_{i=1}^n 1(x_i^{(n)} > \epsilon n) = 0$ for all $n > n_0$. This implies that $n^{-1}x_i^{(n)} \leq \epsilon$ for all $i \in [n]$, or equivalently, $n^{-1}\max_{i\in[n]} x_i^{(n)} \leq \epsilon$ for all $n > n_0$, and the claim follows.

□

6.6 Product-form kernels (not part of Fall 2018 course)

Recall from Section 1.4 the definition of inhomogeneous Bernoulli graphs and latent position graphs. Many real-world data sets have highly varying degree distributions, where most nodes have a relatively small degree and a few hub nodes have an extremely high degree. Such data sets can be modeled as large inhomogeneous random graphs where the attribute space is $\mathcal{S} = [0,\infty)$ and the attributes are considered weights. A natural idea is the multiplicative model where the probability that i and j establish a link is an increasing function of the product $x_i x_j$ of their weights. Because probabilities are at most one, the product is truncated using

$$\phi(z_i, z_j) = \phi_0(cz_i z_j) \tag{6.4}$$

where $\phi_0 : [0,\infty) \to [0,1]$ is a monotone function such that $\phi_0(0) = 0$ and $\lim_{t\to\infty}\phi_0(t) = 1$, and where the constant c controls the overall link density of the model. Three alternative versions of this model are obtained by the following three truncations described in Table 6.2.

Sparse large graphs with a bounded average degree are obtained using rank-1 models when the normalizing constant in (6.4) is of the order n^{-1}. This can be obtained by setting $c = n^{-1}$ or $c = \|z\|^{-1}$ where $\|z\|_1 = |z_1| + \cdots + |z_n|$, under appropriate scaling assumptions on the labels z_i. This models are analyzed in [vdH17, Chap 6] and [vdH18, Chap 2].

Nickname	Truncation function ϕ_0
Beta model, generalized random graph	$\phi_0(t) = \frac{t}{1+t}$
Chung–Lu model	$\phi_0(t) = \min\{t, 1\}$
Norros–Reittu model	$\phi_0(t) = 1 - e^{-t}$

Table 6.2: Rank-1 graph models. The beta model is usually parametrized using $\beta_i = \log z_i$, in which case $\phi(z_i, z_j) = \frac{ce^{\beta_i+\beta_j}}{1+ce^{\beta_i+\beta_j}}$.

Theorem 6.12 (Expected link count, product-form kernel)**.** *Assume that $\mu_n \to \mu$ weakly and with first moments, where the limiting distribution has a finite nonzero mean. Then the number of links $e(G_n) = |E(G_n)|$ in the random graph $G_n = G_n(x^{(n)})$ converges according to*

$$\frac{1}{n} e(G_n) \xrightarrow{\mathbb{P}} \frac{1}{2}\mathbb{E}(X),$$

where X denotes a random number distributed according to μ.

Ok in wider generality, see [BJR07, Lemma 8.1]

BJR assume that the kernel is of the form $K(z_i, z_j) = \min\{n^{-1}\kappa(z_i, z_j), 1\}$ where $\kappa : S^2 \to \mathbb{R}_+$ is symmetric, the attribute space is (separable?) metric. The call things a vertex space when the empirical attribute distribution μ_n converges to a limit μ weakly in probability. They require that the kernel is continuous, and integrable with respect to the product of the limiting attribute distribution. A key condition is that the kernel is "graphical" in that the expected average degree[2] scales as

$$\mathbb{E}\frac{2e(G_n)}{n} \to \iint \kappa(z, w)\mu(dz)\mu(dw).$$

The analysis in BJR is first done in the SBM case. The setting is as follows. Node attributes take values in a finite set $[m]$. They can be nonrandom or random. It is assumed that the empirical attribute distribution converges weakly to a limiting distribution μ on $[m]$. The kernel is of the form $K(z_i, z_j) = \min\{n^{-1}\kappa(z_i, z_j), 1\}$ where the normalized kernel κ is automatically continuous. Now the kernel is also bounded. The one can show that the mean average degree is approximated by

$$\frac{2}{n}\mathbb{E}e(G_n) \to \sum_{s=1}^m \sum_{t=1}^m \kappa_{s,t}\mu_s\mu_t = \mu^T \kappa \mu.$$

Later we write in (8.4) that $K^{(n)}(u, v) = \rho_n K(u, v) \wedge 1$

[2] Recall that $\sum_i \deg_G(i) = 2e(G)$, so that the average degree equals $\frac{2e(G)}{n}$.

Proof. (i) We will first show that the expected average degree satisfies

$$\mathbb{E}\frac{2e(G_n)}{n} \;\to\; \mathbb{E}(X). \tag{6.5}$$

For this, observe first that (we omit the superscript n below)

$$\mathbb{E}e(G_n) \;=\; \frac{1}{2}\sum_{(i,j)\in[n]^2_{\neq}} \frac{x_i x_j}{\|x\|_1 + x_i x_j} \;\le\; \frac{1}{2}\sum_{(i,j)\in[n]^2} \frac{x_i x_j}{\|x\|_1} \;=\; \frac{\|x\|_1}{2}.$$

Hence

$$\mathbb{E}\frac{2e(G_n)}{n} \;\le\; n^{-1}\|x\|_1 \;=\; \mathbb{E}X_n, \tag{6.6}$$

where X_n is a random variable distributed according to μ_n.

We will next derive a matching lower bound. We do this by truncating the attributes by some sequence ω_n which grows to infinity at a suitable rate when $n \to \infty$. The analysis below shows a suitable rate[3]. Now we define the truncated attributes as $\tilde{x}_i^{(n)} = x_i^{(n)} \wedge \omega_n$ where $a \wedge b = \min\{a, b\}$. Again, we omit the superscript n for convenience. Because $t \mapsto t/(\|x\|_1 + t)$ is increasing, it follows that

$$\mathbb{E}e(G_n) \;\ge\; \frac{1}{2}\sum_{(i,j)\in[n]^2_{\neq}} \frac{\tilde{x}_i\tilde{x}_j}{\|x\|_1 + \tilde{x}_i\tilde{x}_j}.$$

so that

$$\begin{aligned}
\frac{\|\tilde{x}\|_1^2}{\|x\|_1} - 2\mathbb{E}e(G_n) \;&\le\; \frac{\|\tilde{x}\|_1^2}{\|x\|_1} - \sum_{(i,j)\in[n]^2_{\neq}} \frac{\tilde{x}_i\tilde{x}_j}{\|x\|_1 + \tilde{x}_i\tilde{x}_j} \\
&= \sum_{i\in[n]} \frac{\tilde{x}_i^2}{\|x\|_1 + \tilde{x}_i^2} + \sum_{(i,j)\in[n]^2} \tilde{x}_i\tilde{x}_j \left(\frac{1}{\|x\|_1} - \frac{1}{\|x\|_1 + \tilde{x}_i\tilde{x}_j} \right) \\
&= \sum_{i\in[n]} \frac{\tilde{x}_i^2}{\|x\|_1 + \tilde{x}_i^2} + \sum_{i\in[n]}\sum_{j\in[n]} \frac{\tilde{x}_i^2\tilde{x}_j^2}{\|x\|_1(\|x\|_1 + \tilde{x}_i\tilde{x}_j)} \\
&\le \sum_{i\in[n]} \frac{\tilde{x}_i^2}{\|x\|_1} + \sum_{i\in[n]}\sum_{j\in[n]} \frac{\tilde{x}_i^2\tilde{x}_j^2}{\|x\|_1^2} \\
&= \sum_{i\in[n]} \frac{\tilde{x}_i^2}{\|x\|_1} \left(1 + \sum_{i\in[n]} \frac{\tilde{x}_i^2}{\|x\|_1} \right).
\end{aligned}$$

[3] In real analysis we learn to choose a small $\epsilon > 0$. Here we learn to choose a sequence ω_n which grows to infinity slowly enough. This plays the role of a "large number" which is not too large.

By $\tilde{x}_i^2 \le \omega_n x_i$ we see that

$$\sum_{i\in[n]} \frac{\tilde{x}_i^2}{\|x\|_1} \;\le\; \sum_{i\in[n]} \frac{\omega_n x_i}{\|x\|_1} \;=\; \omega_n,$$

and hence we conclude that

$$\frac{\|\tilde{x}\|_1^2}{\|x\|_1} - 2\mathbb{E}e(G_n) \;\le\; \omega_n(1+\omega_n),$$

and

$$\mathbb{E}\frac{2e(G_n)}{n} \;\ge\; \frac{n^{-2}\|\tilde{x}\|_1^2}{n^{-1}\|x\|_1} - \frac{\omega_n(1+\omega_n)}{n} \;=\; \frac{\big(\mathbb{E}(X_n\wedge\omega_n)\big)^2}{\mathbb{E}X_n} - \frac{\omega_n(1+\omega_n)}{n}$$

By combining this with (6.6), we conclude that

$$\left(\frac{\mathbb{E}(X_n\wedge\omega_n)}{\mathbb{E}X_n}\right)^2 \mathbb{E}X_n - \frac{\omega_n(1+\omega_n)}{n} \;\le\; \mathbb{E}\frac{2e(G_n)}{n} \;\le\; \mathbb{E}X_n. \tag{6.7}$$

Now when ω_n is any sequence converging to infinity, one can show by applying Lebesgue's dominated and monotone convergence theorems that

$$\mathbb{E}(X_n\wedge\omega_n) \to \mathbb{E}X.$$

Moreover, $\frac{\omega_n(1+\omega_n)}{n} \to 0$ when we choose ω_n to grow slowly enough so that $\omega_n \ll n^{1/2}$. Hence for example by choosing $\omega_n = n^{0.4}$ it follows that both the lower and the upper bound in (6.7) converge to $\mathbb{E}X$, and we obtain (6.5).

(ii) To finish the proof, we use the second moment method by finding an upper bound for the variance of the link count. The variance is indeed quite easy to analyze, because the link indicators of the graph are independent and Ber(p_{ij})-distributed. Therefore

$$\mathrm{Var}(e(G_n)) \;=\; \sum_{(i,j)\in[n]^2_<} p_{ij}(1-p_{ij}) \;\le\; \sum_{(i,j)\in[n]^2_<} p_{ij} \;=\; \mathbb{E}e(G_n).$$

Now $\mathbb{E}e(G_n) = n(n^{-1}\mathbb{E}X_n) \sim 2n\mathbb{E}X \to \infty$ by (i), so that

$$\frac{\mathrm{Var}(e(G_n))}{(\mathbb{E}e(G_n))^2} \to 0,$$

and the claim follows by Chebyshev's inequality. □

Theorem 6.13 (Model degree distribution, product-form kernel)**.** *Consider an inhomogeneous Bernoulli graph with link probabilities $p_{ij} = \frac{x_i x_j}{\|x\|_1 + x_i x_j}$. Assume the empirical distribution of $x^{(n)}$ converges weakly and with first moments to a limiting distribution μ. Then the distribution of $\deg_G(U)$ in $G = G(x^{(n)})$ converges weakly to the μ-mixed Poisson distribution.*

Proof. Assume first that the empirical distribution of $x^{(n)}$ converges weakly and with second moments. Fix some n and let U be a random variable uniformly distributed in $[n]$. For each i there exists a coupling $(\hat{D}_i, \hat{Y}_i)$ of $\mathrm{Law}(\deg_G(i))$ and $\mathrm{Poi}(x_i)$ such that $\mathbb{P}(\hat{D}_i \neq \hat{Y}_i) \le \dots$. Now let $\tilde{U}, (\tilde{D}_1, \tilde{Y}_1), \dots, (\tilde{D}_n, \tilde{Y}_n)$ be independent random variables with the right distributions. Then $(\tilde{D}_{\tilde{U}}, \tilde{Y}_{\tilde{U}})$ form a coupling of $\mathrm{Law}(\deg_G(U))$ and a μ_n-mixed Poisson distribution such that

$$\mathbb{P}(\tilde{D}_{\tilde{U}} \neq \tilde{Y}_{\tilde{U}}) \; = \; \frac{1}{n} \sum_{i=1}^n \mathbb{P}(\tilde{D}_i \neq \tilde{Y}_i) \; \le \; \frac{1}{n} \sum_{i=1}^n \frac{x_i^2}{\|x\|_1} \left(1 + 2 \sum_{j=1}^n \frac{x_j^2}{\|x\|_1} \right)$$

The right side above can also be written as

$$\frac{1}{n} \sum_{i=1}^n \frac{x_i^2}{\|x\|_1} \left(1 + 2 \sum_{j=1}^n \frac{x_j^2}{\|x\|_1} \right) \; = \; \frac{\mathbb{E} X_n^2}{n \mathbb{E} X_n} \left(1 + 2 \frac{\mathbb{E} X_n^2}{\mathbb{E} X_n} \right).$$

Hence

$$d_{\mathrm{tv}}(\mathrm{Law}(\deg_G(U), \mathrm{MPoi}(\mu_n))) \; \le \; \frac{\mathbb{E} X_n^2}{n \mathbb{E} X_n} \left(1 + 2 \frac{\mathbb{E} X_n^2}{\mathbb{E} X_n} \right) \; \to \; 0.$$

Because $\mu_n \to \mu$ weakly, it can be shown (for example by computing the generating functions) that $\mathrm{MPoi}(\mu_n) \to \mathrm{MPoi}(\mu)$ weakly. Then also $\mathrm{Law}(\deg_G(U)) \to \mathrm{MPoi}(\mu)$ weakly.

If we only have convergence with the first moments, we can apply a truncation argument, see [vdH17, Proof of Corollary 6.9]. □

Theorem 6.13 can be strengthened into a statement of the empirical degree distribution

$$f_{G_n}(k) \; = \; \frac{1}{n} \sum_{i=1}^n 1(\deg_{G_n}(i) = k).$$

Here $\mathrm{MPoi}(\mu)$ denotes the μ-mixed Poisson distribution.

Theorem 6.14 (Empirical degree distribution)**.** *Consider a GRG. Assume the empirical distribution of $x^{(n)}$ converges weakly and with first moments to a limiting distribution μ. For all $\epsilon > 0$,*

$$\mathbb{P}\Big(d_{\mathrm{tv}}\big(f_{G_n}, \mathrm{MPoi}(\mu)\big) \le \epsilon \Big) \to 1.$$

Proof. See [vdH17, Theorem 6.10]. □

Theorem 6.15 (Asymptotic equivalence of sparse Chung–Lu and GRG)**.** *For each $n \ge 1$, let $x^{(n)} = (x_1^{(n)}, \dots, x_n^{(n)})$ be a list of weights with empirical distribution μ_n. Assume that μ_n converges weakly and with 2nd moments to some probability distribution μ. Then the Chung–Lu model and the generalized random graph model are asymptotically equivalent as $n \to \infty$.*

Proof. Fix some n, and denote the GRG probability matrix p and the Chung–Lu probability q by

$$p_{ij} \;=\; \frac{x_i x_j}{\|x\|_1 + x_i x_j} \quad \text{and} \quad q_{ij} \;=\; \min\left\{ \frac{x_i x_j}{\|x\|_1}, \, 1 \right\}.$$

Then by the inequality $1 - 1/(1+t) \le t$,

$$q_{ij} - p_{ij} \;\le\; \frac{x_i x_j}{\|x\|_1} - \frac{x_i x_j}{\|x\|_1 + x_i x_j} \;=\; \frac{x_i x_j}{\|x\|_1} \left(1 - \frac{1}{1 + \frac{x_i x_j}{\|x\|_1}} \right) \;\le\; \left(\frac{x_i x_j}{\|x\|_1} \right)^2.$$

Moreover, by denoting $\|x\|_\infty = \max_{i \in [n]} x_i$,

$$p_{ij} \;\ge\; \frac{x_i x_j}{\|x\|_1 + \|x\|_\infty^2},$$

so that

$$\sum_{(i,j) \in [n]^2_<} \frac{(q_{ij} - p_{ij})^2}{p_{ij}} \;\le\; \frac{\|x\|_1 + \|x\|_\infty^2}{\|x\|_1^4} \sum_{(i,j) \in [n]^2_<} (x_i x_j)^3.$$

Because

$$\sum_{(i,j) \in [n]^2_<} (x_i x_j)^3 \;\le\; \sum_{(i,j) \in [n]^2} (x_i x_j)^3 \;=\; \Big(\sum_{i \in [n]} x_i^3 \Big)^2 \;=\; \|x\|_3^6,$$

we conclude that

$$\sum_{(i,j) \in [n]^2_<} \frac{(q_{ij} - p_{ij})^2}{p_{ij}} \;\le\; \frac{\|x\|_1 + \|x\|_\infty^2}{\|x\|_1^4} \|x\|_3^6.$$

Now a key observation is that we may write $n^{-1} \|x\|_1 = \mathbb{E} X_n$ where X_n is a μ_n-distributed random integer. Because $\mu_n \to \mu$ weakly and with 1st moments, it follows that $\mathbb{E} X_n \to \mathbb{E} X$ where X is a μ-distributed random integer with a finite mean. Using this one can show (exercise) that $n^{-1/2} \|x\|_\infty \to 0$. Then also

$$\|x\|_3^3 \;=\; \sum_{i \in [n]} x_i^3 \;\le\; \|x\|_\infty \sum_{i \in [n]} x_i^2 \;=\; n \|x\|_\infty \mathbb{E} X_n^2,$$

and the right side above is bounded by

$$\frac{\|x\|_1 + \|x\|_\infty^2}{\|x\|_1^4} \|x\|_3^6 \;\le\; \frac{\|x\|_1 + \|x\|_\infty^2}{\|x\|_1^4} n^2 \|x\|_\infty^2 (\mathbb{E}X_n^2)^2 \;=\; \frac{n\mathbb{E}X_n + \|x\|_\infty^2}{n^4 \mathbb{E}X_n} n^2 \|x\|_\infty^2 (\mathbb{E}X_n^2)^2$$

The right side tends to zero as $n \to \infty$, so the claim follows by Theorem 5.6.

□

Asymptotic independence of GRG degrees

Consider an inhomogeneous Bernoulli graph with link probabilities $p_{ij} = \frac{x_i x_j}{\|x\|_1 + x_i x_j}$ defined by the weight lists $x^{(n)}$ such that the empirical distribution converges weakly and with 1st moments to a limiting distribution μ. Consider a node set $I^{(n)} \subset [n]$ of size $|I^{(n)}| = o(n^{1/2})$ and assume that the average weight of the nodes in $I^{(n)}$ satisfies

$$\frac{1}{|I^{(n)}|} \sum_{i \in I^{(n)}} x_i^{(n)}$$

remains bounded as $n \to \infty$. For GRG,

$$\sum_{(i,j) \in I^2_<} p_{ij} \;\le\; \sum_{(i,j) \in I^2} \frac{x_i x_j}{\|x\|_1} \;=\; \frac{(\sum_{i \in I} x_i)^2}{\|x\|_1} \;=\; \frac{O(|I^{(n)}|^2)}{\|x\|_1^{(n)}} \;\sim\; \frac{O(|I^{(n)}|^2)}{n\mathbb{E}X_n}$$

Chapter 7

Subgraph statistics

7.1 Functions

A *function* ϕ is a set of ordered pairs such that $(x, y) \in \phi$ and $(x, y') \in \phi$ imply $y = y'$. The *domain* of ϕ is the set of x such that $(x, y) \in \phi$ for some y, and the *range* of ϕ is the set of y such that $(x, y) \in \phi$ for some x. We say that a function ϕ is an *injective* if $(x, y) \in \phi$ and $(x', y) \in \phi$ imply $x = x'$; an injective function is also called an *injection*. For any x in the domain of ϕ we denote by $\phi(x)$ the unique element y such that $(x, y) \in \phi$. We use the terms *map* and *mapping* as synonyms of the term function.

We say that ϕ is a *function from A to B* and denote $\phi : A \to B$ if the domain of ϕ equals A and the range of ϕ is a subset of B. Hence functions $\phi_1 : A_1 \to B_1$ and $\phi_2 : A_2 \to B_2$ are equal if and only if $A_1 = A_2$ and $\phi_1(x) = \phi_2(x)$ for all $x \in A_1$. A *surjection from A to B* is a function with domain A and range B, and a *bijection from A to B* is an injective function with domain A and range B.

For functions ϕ and ψ such that the range of ϕ is contained in the domain of ψ, the *composition* $\psi \circ \phi$ is the set of ordered pairs (x, z) such that $(x, y) \in \phi$ and $(y, z) \in \psi$ for some y. The composition is a function with the same domain as ϕ and range contained in the range of ψ, and $\psi \circ \phi(x) = \psi(\phi(x))$ for all x in the domain. If $\phi : A \to B$ and $\psi : B \to C$, then $\psi \circ \phi : A \to C$.

For any injection ϕ we define the *inverse* ϕ^{-1} as the set of ordered pairs (x, y) such that $(y, x) \in \phi$. Then ϕ^{-1} is a function with domain being the range of ϕ, and it follows that $\phi(x) = y$ if and only if $x = \phi^{-1}(y)$.

For any subset X of the domain of ϕ we denote by $\phi(X)$ the set elements y such that $(x, y) \in \phi$ for some $x \in X$. For any function ϕ we define the *set-to-set extension* as the set of ordered pairs $(X, \phi(X))$ such that X is a subset of the domain of ϕ. The set-to-set extension is a function with domain being the power set of the domain of ϕ. We use the same symbol ϕ to denote both the original function and the set-to-set extension. In the context of graphs, we usually restrict the domain of the set-to-set extension to the subsets of cardinality two, corresponding to node pairs of graph.

Remark 7.1. Note that "ϕ is an injection" is a property of a function ϕ, whereas "ϕ is a surjection" and "ϕ is a bijection" are not; the latter two are properties of the triple (ϕ, A, B). Hence *injective* is a proper adjective to describe a function, but *surjective* and *bijective* are not.

Remark 7.2. Many authors define a function to be a triple (ϕ, A, B) with ϕ being a subset of $A \times B$ such that $(x, y) \in \phi$ and $(x, y') \in \phi$ imply $y = y'$, calling ϕ the graph, A the domain, and B the codomain of the function (A, B, ϕ). In most applications and most mathematical theories it does not matter which of the definitions is used. However, when we are interested in counting cardinalities of function sets, then it does make a difference how functions are defined. For example, consider the set S of functions $\phi : \{1, 2, 3\} \to \{1, 2, \dots\}$ such that $\phi(x) \leq 10$ for all x. According to common sense, S should be finite with size 10^3. Using the first definition this is the case. Using the alternative definition as a triple, the size of S infinite. This is why here we will use the first definition, and we do not associate the notion of a codomain to a function.

7.2 Graph embeddings

Let F and G be (simple, undirected, loopless) graphs. A function $\phi : V(F) \to V(G)$ is called a *homomorphism* from F to G if

$$e \in E(F) \quad \Longrightarrow \quad \phi(e) \in E(G),$$

and a *strong homomorphism* if

$$e \in E(F) \quad \Longleftrightarrow \quad \phi(e) \in E(G).$$

Homomorphisms from F to G are denoted $\operatorname{Hom}(F, G)$. An injective homomorphism is called an *embedding* and the set such functions is denoted $\operatorname{Emb}(F, G)$; an injective strong homomorphism is called a *strong embedding* and the set of such functions is denoted $\operatorname{Emb}_i(F, G)$. A strong embedding of F to G with range $V(G)$ is called an *isomorphism* from F to G and the set of such functions is denoted $\operatorname{Iso}(F, G)$. Isomorphisms from a graph to itself are called *automorphisms* and denoted $\operatorname{Aut}(G)$. We use lowercase symbols

$\hom(F,G)$, $\mathrm{emb}(F,G)$, etc. to denote the cardinalities of the above function sets.

For any graph G and any injection ϕ with domain $V(G)$, the *image* of G by ϕ is defined as

$$\phi(G) \ := \ \Big(\{\phi(v) : v \in V(G)\},\ \{\phi(e) : e \in E(G)\}\Big).$$

This is a graph with node set being the range of ϕ. By construction, ϕ is an isomorphism from G to $\phi(G)$. For any graphs F and G and for any injection ϕ from $V(F)$ to $V(G)$, one can verify that

exercise

$$\begin{aligned} \phi \in \mathrm{Emb}(F,G) \quad &\iff \quad \phi(F) \subset G, \\ \phi \in \mathrm{Emb}_i(F,G) \quad &\iff \quad \phi(F) \subset_i G, \\ \phi \in \mathrm{Iso}(F,G) \quad &\iff \quad \phi(F) = G, \end{aligned}$$

Example 7.3. Let $K_{1,2}$ be the 2-star on $[3]$ with links $\{1,2\}$ and $\{1,3\}$, and let K_3 be the complete graph on $[3]$. Then $\mathrm{Hom}(K_{1,2},K_3)$ is the set of functions $\phi : [3] \to [3]$ such that $\phi(x) \neq \phi(1)$ for $x \neq 1$, and none of the homomorphisms are strong. Hence $\mathrm{Emb}_i(K_{1,2},K_3) = \emptyset$, $\mathrm{Emb}(K_{1,2},K_3) = \mathrm{Bij}([3],[3])$.

Exercise: Fill in details, count cardinalities

Two graphs are called *isomorphic* if there exists an isomorphism from one to the other. The set of R-isomorphic subgraphs (resp. induced subgraphs) of G is denoted by $\mathrm{Sub}(R,G)$ (resp. $\mathrm{Sub}_i(R,G)$).

Remark 7.4. Because every homomorphism from a complete graph K_n to any other graph G is injective and also a strong homomorphism, it follows that $\mathrm{Hom}(K_n,G) = \mathrm{Emb}(K_n,G) = \mathrm{Emb}_i(K_n,G)$.

In terms of adjacency matrices, a map ϕ is a homomorphism if and only if

$$F_{x,y} \ \le \ G_{\phi(x),\phi(y)} \quad \text{for all } x,y,$$

and a strong isomorphism if and only if

$$F_{x,y} \ = \ G_{\phi(x),\phi(y)} \quad \text{for all } x,y.$$

Lemma 7.5. *If the set* $\mathrm{Emb}(R,G,R') = \{\phi \in \mathrm{Emb}(R,G) : \phi(R) = R'\}$ *is nonempty, then it can be represented as*

$$\mathrm{Emb}(R,G,R') \ = \ \{\phi_0 \circ \psi : \psi \in \mathrm{Aut}(R)\}$$

where $\phi_0 \in \mathrm{Emb}(R,G,R')$ *is arbitrary.*

Proof. Fix some $\phi_0 \in \mathrm{Emb}(R, G, R')$. If $\phi = \phi_0 \circ \psi$ for some $\psi \in \mathrm{Aut}(R)$, then obviously ϕ is injective. Moreover, the fact that ψ is a bijection from $V(R)$ to itself implies that

$$\{\phi(v) : v \in V(R)\} \;=\; V(R').$$

Analogously, the fact that ψ considered as a set-to-set function with domain $E(R)$ is a bijection from $E(R)$ to itself shows that

$$\begin{aligned}
\{\phi(e) : e \in E(R)\} &= \{\phi_0((\psi(e)) : e \in E(R)\} \\
&= \{\phi_0((\psi(e)) : \psi(e) \in E(R)\} \\
&= \{\phi_0(e) : e \in E(R)\} \\
&= E(R').
\end{aligned}$$

Hence we conclude that $\phi(R) = R'$. Furthermore, by noting that

$$e \in E(R) \implies \psi(e) \in E(R) \implies \phi_0(\psi(e)) \in E(G)$$

we see that $\phi \in \mathrm{Emb}(R, G, R')$.

Assume next that $\phi \in \mathrm{Emb}(R, G, R')$. Then ϕ is an isomorphism from R to R', and so is ϕ_0. It follows that $\psi = \phi_0^{-1} \circ \phi$ is an automorphism of R. Moreover, $\phi_0 \circ \psi = \phi$. □

Remark 7.6. The number of r-cliques in a graph G equals $|\mathrm{Sub}(K_r, G)| = |\mathrm{Sub}_i(K_r, G)| = \frac{|\mathrm{Hom}(K_r,G)|}{r!}$.

Proposition 7.7. *For any graphs R and G,*

$$\begin{aligned} |\mathrm{Emb}(R,G)| &= |\mathrm{Sub}(R,G)| \cdot |\mathrm{Aut}(R)|, && (7.1) \\ |\mathrm{Emb}_i(R,G)| &= |\mathrm{Sub}_i(R,G)| \cdot |\mathrm{Aut}(R)|. && (7.2) \end{aligned}$$

If R and G are isomorphic, then

$$|\mathrm{Iso}(R,G)| = |\mathrm{Aut}(R)| = |\mathrm{Aut}(G)|. \tag{7.3}$$

Proof. (i) Because $\phi(R)$ is an R-isomorphic subgraph of G for every $\phi \in \mathrm{Emb}(R,G)$, we may represent the set of embeddings from R to G as a disjoint union

$$\mathrm{Emb}(R,G) = \bigcup_{R' \in \mathrm{Sub}(R,G)} \mathrm{Emb}(R,G,R'),$$

where $\{\phi \in \mathrm{Emb}(R,G) : \phi(R) = R'\}$. Observe next that if $R' \in \mathrm{Sub}(R,G)$, then there exists an isomorphism ϕ from R to R'. Because such ϕ is also an embedding of R into G, it follows that the set $\mathrm{Emb}(R,G,R')$ is nonempty, and by Lemma 7.5, the cardinality of $\mathrm{Emb}(R,G,R')$ equals $\mathrm{aut}(R)$. As a consequence, (7.1) follows.

(ii) Note that $\mathrm{Emb}_i(R,G)$ equals the set of $\phi \in \mathrm{Emb}(R,G)$ such that $\phi(R)$ is an induced subgraph of G. Therefore, we may represent the set of inductive embeddings from R to G as a disjoint union

$$\mathrm{Emb}_i(R,G) = \bigcup_{R' \in \mathrm{Sub}_i(R,G)} \mathrm{Emb}(R,G,R'),$$

In the proof of (i) we saw that $\mathrm{Emb}(R,G,R')$ is nonempty for every $R' \in \mathrm{Sub}(R,G)$. Hence by Lemma 7.5, we obtain (7.2).

(iii) Note that $\mathrm{Iso}(R,G) = \mathrm{Emb}(R,G,G)$. When R and G are isomorphic, the set $\mathrm{Emb}(R,G,G)$ is nonempty, and by Lemma 7.5, its cardinality equals $\mathrm{aut}(F)$. Therefore $|\mathrm{Iso}(R,G)| = |\mathrm{Aut}(F)|$. The second equality of (7.3) follows by symmetry after noting that $|\mathrm{Iso}(R,G)| = |\mathrm{Iso}(G,R)|$. □

7.3 Upper bound on the number of homomorphisms

The following result shows that the number of homomorphisms from a tree T to any graph G is maximized by taking T to be a star. The proof is given in Section 7.3.2

Theorem 7.8 (Sidorenko's inequality [Sid94]). *For any tree T with k links and for any graph G,*

$$\hom(T, G) \;\le\; \hom(K_{1,k}, G) \;=\; \sum_{v \in V(G)} d_G(v)^k.$$

In the proof below we assumed G to be connected.

As a corollary of Theorem 7.8 we get the following result.

Theorem 7.9. *For any connected graph F with r nodes and for any graph G,*

$$\hom(F, G) \;\le\; \hom(K_{1,r-1}, G) \;=\; \sum_{v \in V(G)} d_G(v)^{r-1}.$$

Proof. Let F be a connected graph with r nodes. Then F contains a spanning tree T, and by Lemma 7.10 below we see that $\hom(H, G) \le \hom(T, G)$. Because T is a tree with $r - 1$ links, Sidorenko's inequality implies

$$\hom(F, G) \;\le\; \hom(T, G) \;\le\; \hom(K_{1,r-1}, G).$$

□

Lemma 7.10. *If F_1 is a subgraph of F_2 such that $V(F_1) = V(F_2)$, then*

$$\hom(F_2, G) \;\le\; \hom(F_1, G)$$

Proof. Exercise. □

7.3.1 Randomly marked trees

We follow the proof in [LP17], based on notions in [BP94]. Let T be a finite tree and let S be a countable space, and let P be a transition probability matrix on S with invariant distribution π. We define a random function $X : V(T) \to S$ which can be viewed as a *randomly marked rooted tree* induced by (π, P) as follows. First we select some node $\rho \in V(T)$ as a root, and we let $X_\rho \in S$ be π-distributed. Then we require for all nonroot $u \in V(T)$ the Markov-type property

$$\begin{aligned} \mathbb{P}(X_u = v \,|\, X_{u'} = x_{u'} : |u'| \le |u|, u' \ne u) \;&=\; \mathbb{P}(X_u = v \,|\, X_{u^\uparrow} = x_{u^\uparrow}) \\ &=\; P(x_{u^\uparrow}, v), \end{aligned}$$

where $|u|$ denotes the distance of u from the root in T and $u^\uparrow$ denotes the parent of u in T. The above property uniquely determines the distribution p of random function $X \in S^{V(T)}$, which can be written as

$$p(x) \;=\; \pi(x_\rho) \prod_{u \ne \rho} P(x_{u^\uparrow}, x_u), \quad x \in S^{V(T)}.$$

The definition implies that $(X_{u_0}, \dots, X_{u_\ell})$ is a Markov chain with initial distribution π and transition matrix P for any directed path $u_0, u_1, \dots, u_\ell$ in the tree T starting with $u_0 = \rho$. Because π is invariant for P, it hence follows that X_u is π-distributed for every $u \in V(T)$.

Now let us apply the above construction to the case where S is the node set of a finite graph G, and P is the symmetric random walk on G with invariant distribution $\pi(v) = \frac{d_G(v)}{2m}$ where $m = |E(G)|$. Assume first that G is connected. Then

$$p(x) \;=\; \frac{d_G(x_\rho)}{2m} \prod_{u \neq \rho} \frac{1}{d_G(x_{u^\uparrow})} \;=\; \frac{1}{2m} \prod_{u \neq \rho, \rho'} \frac{1}{d_G(x_{u^\uparrow})},$$

where ρ' is an arbitrary child of the root ρ. In the product on the right, the term $1/d_G(x_\rho)$ appears $d_T(\rho) - 1$ times. A moment's reflection reveals that the $1/d_G(x_u)$ occurs in the product $d_T(u) - 1$ also when u is a nonroot and nonleaf node, and when u is a leaf node. Hence the above formula can be written as

$$p(x) \;=\; \frac{1}{2m} \prod_{u \in V(T)} d_G(x_u)^{1 - d_T(u)},$$

and this shows that the distribution p is invariant to the initial choice of the root $\rho \in V(T)$. In a sense, this construction provides an exchangeable coupling of several stationary (and reversible) random walks on G simultaneously.

7.3.2 Proof of Sidorenko's inequality

We will now present a proof Theorem 7.8, as outlined in [LP17]. The construction in the previous section shows that the support of p equals $\mathrm{Hom}(T, G)$. Therefore, we may count the cardinality of $\mathrm{Hom}(T, G)$ using the importance sampling formula

$$|\mathrm{Hom}(T,G)| \;=\; \sum_{x \in \mathrm{Hom}(T,G)} p(x) \frac{1}{p(x)} \;=\; \mathbb{E} \frac{1}{p(X)},$$

where X is distributed according to p. Hence

$$|\mathrm{Hom}(T,G)| \;=\; 2m \mathbb{E} \prod_{u \neq \rho, \rho'} d_G(X_{u^\uparrow}).$$

Because $|V(T)| = k+1$, the product on the right side above has $k-1$ terms, and the arithmetic-geometric mean inequality implies that

$$\prod_{u \neq \rho, \rho'} d_G(X_{u^\uparrow}) \;=\; \left(\prod_{u \neq \rho, \rho'} d_G(x_{u^\uparrow})^{k-1} \right)^{1/(k-1)} \;\leq\; \frac{1}{k-1} \sum_{u \neq \rho, \rho'} d_G(X_{u^\uparrow})^{k-1}$$

Because X_u is π-distributed for all u, it follows by taking expectations that

$$|\mathrm{Hom}(T,G)| \;\le\; 2m\mathbb{E}d_G(X_\rho)^{k-1} \;=\; 2m \sum_{v\in V(G)} d_G(v)^{k-1}\frac{d_G(v)}{2m} \;=\; \sum_{v\in V(G)} d_G(v)^k.$$

This concludes the proof of Theorem 7.8 because the right side above equals $\mathrm{Hom}(K_{1,k},G)$; to see why, note that the number of homomorphisms from $K_{1,k}$ into G such that the hub of $K_{1,k}$ is mapped to $v \in V(G)$ equals $d_G(v)^k$.

7.4 Counting subgraphs

We will discuss local descriptive statistics of a finite undirected graph G on node set $V = [n]$. Basic quantities are the link count

$$|E(G)| \;=\; \sum_{(i,j):i<j} G_{ij},$$

the average degree

$$\frac{1}{n}\sum_{i=1}^n \deg_G(i) \;=\; \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n G_{ij},$$

and the empirical degree distribution $k \mapsto \frac{1}{n}\sum_i 1(\deg_G(i) = k)$. More detailed information about the local graph structure can be obtained for example by computing the number of triangles $\sum_{(i,j,k):i<j<k} G_{ij}G_{ik}G_{jk}$ contained in G.

When extending this from triangles to more general subgraphs we need to be a bit careful what we mean by saying that something is "contained in G". Recall that for undirected graphs F and G:

- F and G are *equal*, denoted $F = G$, if $V(F) = V(G)$ and $E(F) = E(G)$.
- F is a *subgraph* of G, denoted $F \subset G$, if $V(F) \subset V(G)$ and $E(F) \subset E(G)$. This is denoted $F \subset G$.
- F is an *induced subgraph* of G, denoted $F \subset_i G$ if $V(F) \subset V(G)$ and $E(F) = E(G) \cap \binom{V(F)}{2}$.

For any $A \subset V(G)$ the subgraph of G *induced by A* is the graph $G[A] = (A, E(G) \cap \binom{A}{2})$. The adjacency matrix of $G[A]$ is the adjacency matrix of G restricted to $A \times A$. These definitions are more convenient to express using adjacency matrices, see Table 7.1.

Notion	Node set condition	Adjacency matrix condition
$F = G$	$V(F) = V(G)$	$F_{ij} = G_{ij}$ for all $i, j \in V(F)$
$F \subset_i G$	$V(F) \subset V(G)$	$F_{ij} = G_{ij}$ for all $i, j \in V(F)$
$F \subset G$	$V(F) \subset V(G)$	$F_{ij} \le G_{ij}$ for all $i, j \in V(F)$

Table 7.1: Subgraph definitions.

Exercise 7.11. Show that F is an induced subgraph of G iff $F = G[S]$ for some set S.

Exercise 7.12. Is $F \subset G$ a partial order on the set of all finite graphs? Prove the claim true of false. What about $F \subset_i G$?

Graphs F and G are called *isomorphic*, if there exists a bijection $\phi : V(F) \to V(G)$ such that $\{\phi(i), \phi(j)\} \in E(G)$ if and only if $\{i, j\} \in E(F)$. In terms of adjacency matrices, graphs F and G are isomorphic if and only if $F_{i,j} = G_{\phi(i),\phi(j)}$ for some bijection $\phi : V(F) \to V(G)$.

We may study local properties of a large graph G by computing how frequently a copy of a small graph[1] R can be identified inside G. More precisely, for an arbitrary graph R we define the *R-matching density* of G this is denoted $t_{\mathrm{ind}}(R, G)$ in [Lov12] by

$$\hat{P}_G(R) = \frac{|\mathrm{Sub}_i(R, G)|}{|\mathrm{Sub}(R, K_n)|} = \frac{1}{|\mathcal{G}_n(R)|} \sum_{R' \in \mathcal{G}_n(R)} 1(G \supset_i R') \tag{7.4}$$

and the *R-covering density* of G this is denoted $t_{\mathrm{inj}}(R, G)$ in [Lov12] by

$$\hat{Q}_G(R) = \frac{|\mathrm{Sub}(R, G)|}{|\mathrm{Sub}(R, K_n)|} = \frac{1}{|\mathcal{G}_n(R)|} \sum_{R' \in \mathcal{G}_n(R)} 1(G \supset R'), \tag{7.5}$$

where $\mathcal{G}_n(R)$ denotes the set of all R-isomorphic subgraphs with node set contained in $[n]$. By definition, $0 \le \hat{P}_G(R) \le \hat{Q}_G(R) \le 1$, and $\hat{P}_G(R) = \hat{Q}_G(R)$ when R is a complete graph. Lemma 7.16 shows a formula for counting the denominator above.

When $R = K_2$, we note that the matching and covering densities are both equal to the usual link density of the graph

$$\hat{P}_G(R) = \hat{Q}_G(R) = \frac{|E(G)|}{\binom{n}{2}}$$

[1] sometimes called a *graphlet*

Example 7.13 (2-star densities)**.** Denote by $N_G(K_{12})$ (resp. $N^i_G(K_{12})$) the number of K_{12}-isomorphic subgraphs (resp. induced subgraphs) in G. For a three-node set $U \subset [n]$, let $N_G(K_{12}; U)$ be the number of K_{12}-isomorphic subgraphs of G with node set U. Then[2]

$$N_G(K_{12}) \;=\; \sum_{U \in \binom{V(G)}{3}} N_G(K_{12}; U).$$

But now

$$N_G(K_{12}; U) \;=\; N^i_G(K_{12}; U) + 3N^i_G(K_3; U).$$

By summing both sides across U, it follows that

$$N_G(K_{12}) \;=\; N^i_G(K_{12}) + 3N^i_G(K_3).$$

Because $|\mathcal{G}_n(K_{12})| = 3\binom{n}{3}$ and $|\mathcal{G}_n(K_3)| = \binom{n}{3}$, it hence follows that

$$\begin{aligned}\hat{Q}_G(K_{12}) \;&=\; \frac{N_G(K_{12})}{|\mathcal{G}_n(K_{12})|} \;=\; \frac{N^i_G(K_{12})}{|\mathcal{G}_n(K_{12})|} + \frac{3N^i_G(K_3)}{|\mathcal{G}_n(K_{12})|} \\ &=\; \frac{N^i_G(K_{12})}{|\mathcal{G}_n(K_{12})|} + \frac{N^i_G(K_3)}{|\mathcal{G}_n(K_3)|} \\ &=\; \hat{P}_G(K_{12}) + \hat{P}_G(K_3).\end{aligned}$$

Example 7.14 (Diamond)**.** Let G be a diamond on $[4]$ with links $E(G) = \{12, 23, 34, 41, 24\}$. When $R = K_2$, we see that

$$\hat{P}_G(\text{link}) = \hat{Q}_G(\text{link}) = \frac{|E(G)|}{\binom{4}{2}} = \frac{5}{6}.$$

Next, let R be a 2-star (i.e. 2-path). We can count the number of copies of the 2-star contained in G by going through all unordered triplets in $[4]$, so that

$$\sum_{R' : R' \cong R} 1(R' \subset G) \;=\; \sum_{V_0 \subset V : |V_0| = 3} \; \sum_{R' : V(R') = V_0, R' \cong R} 1(R' \subset G).$$

Note first that there exist three R-isomorphic graphs on $V_0 = \{1, 2, 3\}$, and one of them is a subgraph in G. Next, the number R-isomorphic graphs on $V_0 = \{1, 2, 4\}$ equals three, and each of them is a subgraph of G. Proceeding this way, we find that

$$\hat{Q}_G(\text{2-star}) \;=\; \frac{1 + 3 + 1 + 3}{3 + 3 + 3 + 3} \;=\; \frac{8}{12}.$$

[2] Similar formulas for 4-node graphlets (connected subgraphs) have been derived in [MS12].

A similar computation shows that

$$\hat{P}_G(\text{2-star}) \;=\; \frac{1+0+1+0}{3+3+3+3} \;=\; \frac{2}{12}.$$

For a triangle, we find that

$$\hat{P}_G(\text{triangle}) = \hat{Q}_G(\text{triangle}) = \frac{0+1+0+1}{1+1+1+1} = \frac{2}{4}.$$

Next, let R be a graph on a set of three nodes only containing one link. On the node set $\{1,2,3\}$ there exist three graphs isomorphic to R, two of which are subgraphs of G and none of which equals G. On the node set $\{1,2,4\}$ there exist three graphs isomorphic to R, all of which are subgraphs of G, and none of which are equal to G. This way, we find that

$$\hat{Q}_G(\text{lone link among three}) = \frac{2+3+2+3}{3+3+3+3} = \frac{10}{12}$$

and

$$\hat{P}_G(\text{lone link among three}) = \frac{0+0+0+0}{3+3+3+3} = 0.$$

R	$\hat{P}_G(R)$	$\hat{Q}_G(R)$
Empty graph of two nodes	$\frac{1}{6}$	1
Link	$\frac{5}{6}$	$\frac{5}{6}$
Empty graph of three nodes	0	1
Lone link among three nodes	0	$\frac{5}{6}$
2-star	$\frac{1}{6}$	$\frac{2}{3}$
Triangle	$\frac{1}{2}$	$\frac{1}{2}$

Table 7.2: Matching and covering densities of a diamond.

Proposition 7.15. *The matching and covering densities are related by*

$$\hat{Q}(R) \;=\; \sum_{S \supset R : V(S)=V(R)} \hat{P}(S).$$

Proof. Denote by $N_G(R)$ (resp. $N_G^i(R)$) the number of R-isomorphic subgraphs (resp. induced subgraphs) of G. Note that

$$N_G(R) \;=\; \sum_{U \in \binom{[n]}{r}} N_G(R,U), \qquad N_G^i(R) \;=\; \sum_{U \in \binom{[n]}{r}} N_G^i(R,U),$$

where $N_G(R,U)$ (resp. $N^i_G(R,U)$) denotes the number of R-isomorphic subgraphs (resp. induced subgraphs) of G on node set U. Now observe that a graph R' on node set U is R-isomorphic if and only if there exists a bijection $\phi : [r] \to U$ such that[3] $R' = M_\phi(R)$, and that for each such R-isomorphic graph R' the number of such bijections equals $|\mathrm{Aut}(R)|$. Now fix some $U \subset [n]$ of size r. Then

$$\begin{aligned} N_G(R,U) &= \frac{1}{|\mathrm{Aut}(R)|} \sum_{\phi \in \mathrm{Bij}([r],U)} 1(G \supset M_\phi(R)) \\ &= \frac{1}{|\mathrm{Aut}(R)|} \sum_{\phi \in \mathrm{Bij}([r],U)} 1(G[U] \supset M_\phi(R)) \\ &= \frac{1}{|\mathrm{Aut}(R)|} \sum_{\phi \in \mathrm{Bij}([r],U)} 1(M_\phi^{-1}(G[U]) \supset R). \end{aligned}$$

Observe next that for any graph $\tilde{G}$ on node set $V(R) = [r]$,

$$1(\tilde{G} \supset R) = \sum_{S \supset R : V(S)=[r]} 1(\tilde{G} = S).$$

By applying this formula to $\tilde{G} = M_\phi^{-1}(G[U])$, it follows that

$$\begin{aligned} 1(\phi^{-1}(G[U]) \supset R) &= \sum_{S \supset R : V(S)=[r]} 1(M_\phi^{-1}(G[U]) = S) \\ &= \sum_{S \supset R : V(S)=[r]} 1(G[U] = M_\phi(S)) \\ &= \sum_{S \supset R : V(S)=[r]} 1(G \supset_i M_\phi(S)), \end{aligned}$$

[3] We extend the definition of ϕ to a map M_ϕ from the set of graphs on $[r]$ into the set of graphs on U in a natural way. Mapping G into $M_\phi(G)$ corresponds to relabeling the nodes of G using $\{\phi(a_1), \phi(a_2), \ldots, \phi(a_n)\}$ instead of $A = \{a_1, \ldots, a_n\}$. This map is a bijection with $M_\phi^{-1} = M_{\phi^{-1}}$.

so that

$$\begin{aligned} N_G(R,U) &= \frac{1}{|\mathrm{Aut}(R)|} \sum_{\phi \in \mathrm{Bij}([r],U)} 1(M_\phi^{-1}(G[U]) \supset R) \\ &= \frac{1}{|\mathrm{Aut}(R)|} \sum_{\phi \in \mathrm{Bij}([r],U)} \sum_{S \supset R : V(S)=[r]} 1(G \supset_i M_\phi(S)) \\ &= \frac{1}{|\mathrm{Aut}(R)|} \sum_{S \supset R : V(S)=[r]} \sum_{\phi \in \mathrm{Bij}([r],U)} 1(G \supset_i M_\phi(S)) \\ &= \frac{1}{|\mathrm{Aut}(R)|} \sum_{S \supset R : V(S)=[r]} |\mathrm{Aut}(S)|\, N_G^i(S,U). \end{aligned}$$

We have hence shown that

$$N_G(R,U) = \sum_{S \supset R : V(S)=[r]} \frac{|\mathrm{Aut}(S)|}{|\mathrm{Aut}(R)|} N_G^i(S,U)$$

for all $U \subset [n]$ of size r. By summing both sides over U, we conclude that

$$N_G(R) = \sum_{S \supset R : V(S)=[r]} \frac{|\mathrm{Aut}(S)|}{|\mathrm{Aut}(R)|} N_G^i(S)$$

The claim now follows by noting that

$$\hat{Q}_G(R) = \frac{N_G(R)}{|\mathcal{G}_n(R)|} = \frac{N_G(R)}{\binom{n}{r}\frac{r!}{|\mathrm{Aut}(R)|}}$$

and

$$\hat{P}_G(R) = \frac{N_G^i(R)}{|\mathcal{G}_n(R)|} = \frac{N_G^i(R)}{\binom{n}{r}\frac{r!}{|\mathrm{Aut}(R)|}}.$$

□

Lemma 7.16. *For any graph R of $r \le n$ nodes, the number of R-isomorphic graphs with node set contained in $[n]$ equals*

$$|\mathrm{Sub}(R, K_n)| = \frac{(n)_r}{|\mathrm{Aut}(R)|}.$$

Proof. Observe that $\mathrm{Emb}(R, K_n)$ equals the set of all injective maps from $V(R)$ into $V(K_n)$. Hence $|\mathrm{Emb}(R, K_n)| = (n)_r$, and the claim follows by Proposition 7.7. □

7.5 Subgraph thresholds in random graphs

Suppose we have an observed a clique of 10 nodes all connected to each other in a large but relatively sparse graph. Is this a manifestation that the observed graph has some hidden structure or mechanism which produces 10-cliques, or is it just a moderately rare event which every now and then occurs in a large random graph? To study this question we need some theory to explain what we would expect to see in a random graph where node pairs are independently linked. We will first look at Erdős–Rényi random graphs.

7.5.1 Erdős–Rényi random graphs

Fix a graph R on node set $[r]$, and let G be the random graph on $[n]$ where each node pair is linked with probability p, independently of other node pairs. Let $N_G(R)$ be the number of R-isomorphic subgraphs of G. We may represent this number as

$$N_G(R) \;=\; \sum_{R' \in \mathcal{G}_n(R)} 1(G \supset R') \tag{7.6}$$

where $\mathcal{G}_n(R)$ denotes the set of all R-isomorphic subgraphs of the complete graph on $[n]$. The expectation is

$$\mathbb{E} N_G(R) \;=\; |\mathcal{G}_n(R)| \, \mathbb{P}(G \supset R).$$

With the help of Lemma 7.16, we find that

$$\mathbb{E} N_G(R) \;=\; \binom{n}{r} \frac{r!}{|\mathrm{Aut}(R)|} p^{|E(R)|}. \tag{7.7}$$

For a sequence of graphs G_n parametrized by n and p_n, it follows that

$$\mathbb{E} N_{G_n}(R) \;\sim\; \frac{1}{|\mathrm{Aut}(R)|} n^{|V(R)|} p_n^{|E(R)|} \tag{7.8}$$

as $n \to \infty$. From these observations we obtain the following result. Let us denote *average degree* of a graph R by

$$d_{\mathrm{avg}}(R) \;=\; \frac{2|E(R)|}{|V(R)|}.$$

Theorem 7.17. *If $p_n \ll n^{-\frac{2}{d_{\mathrm{avg}}(R)}}$, then the homogeneous random graph G_n contains no R-isomorphic subgraphs with high probability.*

Proof. If $p_n \ll n^{-\frac{2}{d_{\mathrm{avg}}(R)}}$, then as $n \to \infty$,

$$\left(n^{|V(R)|} p_n^{|E(R)|}\right)^{1/|E(R)|} \;=\; n^{\frac{2}{d_{\mathrm{avg}}(R)}} \, p_n \;\to\; 0$$

Therefore also $n^{|V(R)|} p_n^{|E(R)|} \to 0$, and Markov's inequality together with (7.8) shows that

$$\mathbb{P}(N_{G_n}(R) \ge 1) \;\le\; \mathbb{E} N_{G_n}(R) \;\to\; 0.$$

□

When $p_n \gg n^{-\frac{2}{d_{\mathrm{avg}}(R)}}$, then (7.8) implies that $\mathbb{E} N_{G_n}(R) \to \infty$. Can we then conclude that G_n contains R-isomorphic subgraphs with high probability? The answer is not as simple as one might expect. Let us consider the following example.

Example 7.18 (Lollipop counts)**.** Consider a random graph with n nodes and link probability $p_n = n^{-0.7}$. Consider a $(4,2)$-lollipop graph defined as the union of a complete graph on $\{1,2,3,4\}$ and a 2-path on $\{4,5,6\}$. This graph has average degree $8/3$, and hence

$$\mathbb{E} N_{G_n}((4,2)\text{-lollipop}) \;\to\; \infty \tag{7.9}$$

due to $p_n = n^{-0.7} \gg n^{-\frac{2}{8/3}}$. On the other hand, the 4-clique (complete graph with 4 nodes) has average degree 3, so that

$$\mathbb{E} N_{G_n}(4\text{-clique}) \;\to\; 0$$

due to $p_n = n^{-0.7} \ll n^{-\frac{2}{3}}$. Moreover, Theorem 7.17 tells that G_n does not contain 4-cliques, with high probability. But if the graph does not contain any 4-cliques, it cannot contains any (4,2)-lollipops either. This seems to conflict with (7.9), so what's going on?

The lollipop count in the above example provides a situation where the expectation does not give the correct picture about the stochastic phenomenon under study. The following theorem due to Bollobás [Bol81] gives a precise answer. The proof is essentially due to [RV86].

Theorem 7.19. *For a homogeneous Bernoulli graph G_n with $n \to \infty$ nodes and link probability p_n,*

$$\mathbb{P}(G_n \text{ contains an } R\text{-isomorphic subgraph}) \;\to\; \begin{cases} 0, & \text{if } p_n \ll n^{-2/d^*(R)}, \\ 1, & \text{if } p_n \gg n^{-2/d^*(R)}, \end{cases}$$

where $d^*(R)$ is the maximum average degree over nonempty subgraphs of R.

Proof. (i) Assume that $p_n \ll n^{-2/d^*(R)}$. Then there exists a subgraph $H \subset R$ such that $p_n \ll n^{-\frac{2}{d_{\mathrm{avg}}(H)}}$, and by Theorem 7.17 we conclude that with high probability, G_n does not contain H-isomorphic subgraphs. Hence with high probability, G_n does not contain R-isomorphic subgraphs either.

(ii) Assume next that $p_n \gg n^{-2/d^*(R)}$. We will now compute the variance of $N_{G_n}(R)$. By applying the representation (7.6), we find that

$$\begin{aligned}
\mathrm{var}(N_G(R)) &= \mathrm{cov}\left(\sum_{R'} 1(G \supset R'),\ \sum_{R''} 1(G \supset R'')\right) \\
&= \sum_{R'}\sum_{R''} \mathrm{cov}\left(1(G \supset R'), 1(G \supset R'')\right) \\
&= \sum_{R'}\sum_{R''} \Big(\mathbb{P}(G \supset R', G \supset R'') - \mathbb{P}(G \supset R)^2\Big),
\end{aligned}$$

where R' and R'' are summed over the set $\mathcal{G}_n(R)$ of all R-isomorphic subgraphs of the complete graph on $[n]$.

Denote $r = |V(R)|$ and $s = |E(R)|$. Observe next that

$$\begin{aligned}
\mathbb{P}(G_n \supset R',\, G_n \supset R'') &= \mathbb{P}(G_n \supset R' \cup R'') \\
&= p_n^{|E(R') \cup E(R'')|} \\
&= p_n^{2s - |E(R') \cap E(R'')|}.
\end{aligned}$$

Because $R' \cap R''$ is a subgraph of R', and R' is isomorphic to R, it follows that

$$\begin{aligned}
|E(R' \cap R'')| &= \frac{2|E(R' \cap R'')|}{|V(R' \cap R'')|}\,\frac{1}{2}|V(R') \cap V(R'')| \\
&= d_{\mathrm{avg}}(R' \cap R'')\,\frac{1}{2}|V(R') \cap V(R'')| \\
&\le d^*(R)\,\frac{1}{2}|V(R') \cap V(R'')|,
\end{aligned}$$

and we conclude that

$$\mathbb{P}(G_n \supset R',\, G_n \supset R'') \le p_n^{2s - \frac{d^*(R)}{2}|V(R') \cap V(R'')|}.$$

We will split the variance sum by the conditioning on how much the node sets of R' and R'' overlap.

$$\begin{aligned}
\mathrm{var}(N_{G_n}(R)) &= \sum_{j=2}^{r} \sum_{(R',R''):|V(R') \cap V(R'')| = j} \Big(\mathbb{P}(G \supset R', G \supset R'') - \mathbb{P}(G \supset R)^2\Big) \\
&\le \sum_{j=2}^{r} \sum_{(R',R''):|V(R') \cap V(R'')| = j} p_n^{2s - \frac{d^*(R)}{2} j}.
\end{aligned}$$

Because there are $\binom{n}{r}\binom{r}{j}\binom{n-r}{r-j}$ ways to choose the node sets $V(R')$ and $V(R'')$ so that they have j common nodes, by Lemma 7.16, it follows that the number of $(R', R'') \in \mathcal{G}_n(R)$ such that $|V(R') \cap V(R'')| = j$ equals

$$\binom{n}{r}\binom{r}{j}\binom{n-r}{r-j}\left(\frac{r!}{|\mathrm{Aut}(R)|}\right)^2.$$

Because $\binom{n}{r} \le n^r/r!$, $\binom{r}{j} \le r!$, and $\binom{n-r}{r-j} \le \binom{n}{r-j} \le n^{r-j}/(r-j)!$, it follows that

$$\binom{n}{r}\binom{r}{j}\binom{n-r}{r-j} \;\le\; n^{2r-j},$$

and

$$\mathrm{var}(N_{G_n}(R)) \;\le\; \left(\frac{r!}{|\mathrm{Aut}(R)|}\right)^2 n^{2r} p_n^{2s} \sum_{j=2}^{r} n^{-j} p_n^{-\frac{d^*(R)}{2}j}.$$

Especially, by (7.7) and the fact that $\binom{n}{r} \;\sim\; n^r/r!$,

$$\frac{\mathrm{var}(N_{G_n}(R))}{(\mathbb{E}N_{G_n}(R))^2} \;\le\; \binom{n}{r}^{-2} n^{2r} \sum_{j=2}^{r} n^{-j} p_n^{-\frac{d^*(R)}{2}j} \;\sim\; (r!)^2 \sum_{j=2}^{r} n^{-j} p_n^{-\frac{d^*(R)}{2}j}.$$

The right side above tends to zero because $p_n \gg n^{-2/d^*(R)}$. Hence by Chebyshev's inequality,

$$\begin{aligned}
\mathbb{P}(N_{G_n}(R) \le 0) \;&=\; \mathbb{P}(N_{G_n}(R) - \mathbb{E}N_{G_n}(R) \le -\mathbb{E}N_{G_n}(R)) \\
&\le\; \mathbb{P}((N_{G_n}(R) - \mathbb{E}N_{G_n}(R))^2 \ge (\mathbb{E}N_{G_n}(R))^2) \\
&\le\; \frac{\mathrm{var}(N_{G_n}(R))}{(\mathbb{E}N_{G_n}(R))^2} \\
&\to\; 0.
\end{aligned}$$

□

7.5.2 Stochastic block models

Consider a stochastic block model $\mathrm{SBM}(z^{(n)}, K^{(n)})$ on node set $[n]$ with node labeling $z^{(n)} = (z_1, \dots, z_n)$ and connectivity matrix

$$K_{u,v}^{(n)} \;=\; \rho_n K_{u,v} \wedge 1,$$

where $\rho_n > 0$ is a sparsity parameter and K is the normalized connectivity matrix. The following result states the subgraph threshold are quite similar to subgraph thresholds in Erdős–Rényi graphs.

Theorem 7.20. *For a random graph G_n corresponding to a* $\mathrm{SBM}(z^{(n)}, K^{(n)})$ *with $n \to \infty$, $\rho_n \to 0$, and where the limiting kernel is bounded by $c_1 \le K_{u,v} \le c_2$ for all u, v, for some constants $c_1, c_2 \in (0, \infty)$,*

$$\mathbb{P}(G_n \textit{ contains an } R\textit{-isomorphic subgraph}) \;\to\; \begin{cases} 0, & \textit{if } \rho_n \ll n^{-2/d^*(R)}, \\ 1, & \textit{if } \rho_n \gg n^{-2/d^*(R)}, \end{cases}$$

where $d^*(R)$ is the maximum average degree over nonempty subgraphs of R.

Proof. The proof is based on stochastic coupling method. We only analyze the case where $\rho_n \to 0$, and we restrict our attention to large enough values of n so that $c_2\rho_n \le 1$. Now G_n is a random graph where each node pair is linked with probability $p_{ij}^{(n)} = \rho_n K_{z_i z_j}$, independently of other node pairs. Let $G_n^{(1)}$ (resp. $G_n^{(2)}$) be a random graph where each node pair is independently linked with probability $c_1\rho_n$ (resp. $c_2\rho_n$). Then by Theorem 3.6,

$$G_n^{(1)} \;\le_{\mathrm{st}}\; G_n \;\le_{\mathrm{st}}\; G_n^{(2)}.$$

Denote by $\mathcal{F}_R$ the set of graphs on $[n]$ containing an R-isomorphic graph as a subgraph. Because $\mathcal{F}_R$ is an upper set (monotone graph property), it follows that

$$\mathbb{P}(G_n^{(1)} \in \mathcal{F}_R) \;\le\; \mathbb{P}(G_n \in \mathcal{F}_R) \;\le\; \mathbb{P}(G_n^{(2)} \in \mathcal{F}_R).$$

Now observe that $G_n^{(1)}$ is a standard Erdős–Rényi graph. Hence if $\rho_n \gg n^{-2/d^*(R)}$, then Theorem 7.19 tells that the leftmost term above tends to one as $n \to \infty$. Hence so does the lowermost term above. On the other hand, if $\rho_n \ll n^{-2/d^*(R)}$, then a similar argument shows that

$$\begin{aligned} &\mathbb{P}(G_n \text{ contains an } R\text{-isomorphic subgraph}) \\ &\le \mathbb{P}(G_n^{(2)} \text{ contains an } R\text{-isomorphic subgraph}), \end{aligned}$$

and by Theorem 7.19 we know that the rightmost term above tends to zero as $n \to \infty$. □

Chapter 8

Learning SBM parameters

8.1 Statistical network inference

Network data usually consists of relational information between a set of nodes that is represented by an n-by-n matrix (X_{ij}) with binary or numerical entries, and node attribute data represented by an n-vector (Z_i) with numerical or categorical entries, see Figure 8.1. Network inference problems concern *computing estimates, making predictions, and testing hypotheses* of network structure and node attributes based on partial or noisy observations of the network data matrix (X_{ij}), node attribute vectors (Z_i), and possibly some auxiliary data related to temporal dynamics (diffusions, random walks) on the network. → intro?

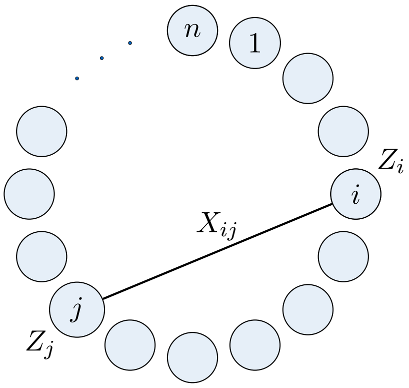

Figure 8.1: Node attributes and relationships.

This framework contains a rich class of applications, for example:

Example 8.1 (Community learning)**.** Estimate node attributes (Z_i) based on fully observed network structure (X_{ij}), up to a permutation of node labels.

This amounts to estimating a partition of the node set generated by the sets $V_s = \{i : Z_i = s\}$ called communities.

Example 8.2 (Phylogenetics)**.** Denote by Z_i a genetic trait of an individual or a group of organisms i. If the values of Z_i have been observed for a set of leaf nodes in an evolutionary tree with fully or partially observed structure (X_{ij}), the task is to infer the value Z_{i_0} of the initial ancestor corresponding to the root node i_0 of the evolutionary tree.

Example 8.3 (Epidemics)**.** Let Z_i be a binary variable indicating whether an individual i falls victim to an infectious disease, and let X_{ij} be a binary variable indicating whether the disease is transmitted through a direct contact between individuals i and j. An important statistical task is to estimate the size of the set $\{i : Z_i = 1\}$ of eventually infected individuals, based on observing values of Z_i for a typically small subset of nodes, and partial observations of the network structure (X_{ij}).

Network data is often given in bipartite form so that we observed relational information (X_{ij}) in the form of an m-by-n matrix between m nodes of a particular type having attributes (Z_i^L), and n nodes of a different type having attributes (Z_j^R), see Figure 8.2. Practical learning tasks involving bipartite data are common in crowdsourcing and collaborative filtering contexts, see the examples below.

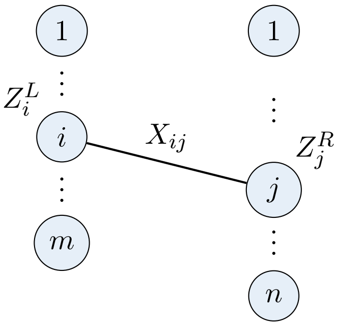

Figure 8.2: Bipartite network data.

Example 8.4 (Crowdsourcing)**.** In microtasking platforms such as Amazon Mechanical Turk, a set of m simple tasks are allocated to n workers who might provide unreliable answers. The unreliability is mitigated by allocating the same task to several workers. Denote by X_{ij} the outcome of task i performed by worker j, by Z_i^L the true outcome of task i, and by Z_j^R the inherent reliability of worker j. The inference problem is to estimate the true outcomes (Z_i^L) based on observed data (X_{ij}).

Example 8.5 (Collaborative filtering)**.** In online recommendation systems a common objective is to infer customer's preferences based on their own and other customers' rankings on a set of items. Let X_{ij} be a number indicating the level of preference of item i by user j. Having observed a partial set of entries of (X_{ij}), the challenge is to complete the matrix by estimating the unobserved remaining values. A famous example of this problem is the Netflix challenge[1]. This problem setup does not involve item attributes (Z_i^L) or customer attributes (Z_j^R), but they could be incorporated as auxiliary model variables.

For notational simplicity, these lecture notes restrict to the unipartite network setting corresponding to Figure 8.1. We will model the joint distribution of the network structure (X_{ij}) and the node attributes (Z_i) using a statistical model where the entries X_{ij} are mutually independent conditionally on the node attributes. This model is described in detail in next section.

8.2 Learning stochastic block models

Denote by $\mathcal{G}_n$ the set of undirected graphs on node set $[n]$, or equivalently, the set of all binary arrays (x_{ij}) indexed by $1 \le i < j \le n$. A stochastic block model with n nodes and m communities is a probability distribution on $\mathcal{G}_n \times [m]^n$ defined by

$$f(x,z) \;=\; \sum_{z \in [m]^n} f(x \,|\, z) \prod_{i=1}^n \alpha(z_i) \tag{8.1}$$

and

$$f(x \,|\, z) \;=\; \prod_{1 \le i < j \le n} (1 - K_{z_i z_j})^{1-x_{ij}} K_{z_i z_j}^{x_{ij}}, \tag{8.2}$$

where K is a symmetric m-by-m matrix with nonnegative entries, and α is a probability distribution on $[m]$. Formula (8.1) represents the joint distribution of a random graph (X_{ij}) and a random vector (Z_i) such that the entries of (Z_i) are independent and α-distributed, and conditionally on (Z_i), the entries X_{ij} are independent and Bernoulli distributed with mean $K(Z_i, Z_j)$.

The SBM is parametrized by $\theta = (m, n, \alpha, K)$, and we often write $f(x,z) = f_\theta(x,z)$ and $f(x \,|\, z) = f_\theta(x \,|\, z)$ to emphasize the dependence of the distribution on its parameters. When n and m are known, the conditional density (8.2) only depends on the parameter K, so we might write $f(x \,|\, z) = f_K(x \,|\, z)$.

[1] https://en.wikipedia.org/wiki/Netflix_Prize

8.3 Learning the kernel when node attributes are known

The easiest learning problem is to estimate the kernel K from observed graph sample $x = (x_{ij})$ when the number of communities m and the node attributes $z = (z_i)$ are known, or they have been first estimated using some other method. A *maximum likelihood estimate* of K is a symmetric nonnegative m-by-m matrix $\hat{K}$ which maximizes the likelihood $K \mapsto f_K(x \mid z)$ corresponding to formula (8.2).

Let us first introduce some helpful notation related to the block structure of the observed graph sample. First, let us represent the attribute vector (z_i) as an n-by-m binary matrix (z_{ij}) with entries $z_{is} = 1(z_i = s)$ indicating whether node i belongs to community s. Then the size of community s can be written as

$$n_s = \sum_{i=1}^n z_{is},$$

and the number of links between communities s and t as

$$e_{st} = \begin{cases} \sum_{i=1}^n \sum_{j=1}^n x_{ij} z_{is} z_{jt}, & s \neq t, \\ \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n x_{ij} z_{is} z_{js}, & s = t. \end{cases}$$

As a consequence, the link density between communities s and t in the observed graph can be written as

$$d_{st} = \frac{e_{st}}{n_{st}}, \tag{8.3}$$

where

$$n_{st} = \begin{cases} n_s n_t, & s \neq t, \\ \frac{1}{2} n_s (n_s - 1), & s = t. \end{cases}$$

Theorem 8.6. *The unique maximum likelihood estimate of the kernel K is the m-by-m matrix with entries being the observed block densities $\hat{K}_{st} = d_{st}$ defined by* (8.3).

Proof. Recall that maximizing a function is equivalent to maximizing its logarithm. We take logarithm of the likelihood to transform the product in (8.2) into a sum. The log-likelihood can be written as

$$\log f_K(x \mid z) = \sum_{1 \le i < j \le n} \Big\{ (1 - x_{ij}) \log(1 - K_{z_i, z_j}) + x_{ij} \log K_{z_i, z_j} \Big\}.$$

In the above sum there is a lot of redundancy in the sense that the only possible values of the terms on are $\log(1-K_{s,t})$ and $\log K_{s,t}$ for some $1 \le s \le t \le m$. By counting how many times these values occur in the sum, we see that the log-likelihood can be written as

$$\begin{aligned}\log f_K(x \,|\, z) &= \sum_{1\le s\le t\le m} \Big\{ (N_{st}-e_{st}) \log(1-K_{st}) + e_{st}\log(K_{st}) \Big\} \\ &= \sum_{1\le s\le t\le m} N_{st} \Big\{ (1-d_{st}) \log(1-K_{st}) + d_{st}\log(K_{st}) \Big\}.\end{aligned}$$

After brief algebraic manipulations, one can also verify that

$$\log f_K(x \,|\, z) = \sum_{1\le s\le t\le m} N_{st} \Big\{ -H(\mathrm{Ber}(d_{st})) - d_{\mathrm{KL}}(\mathrm{Ber}(K_{st}) || \mathrm{Ber}(d_{st})) \Big\},$$

where $H(f)$ denotes the entropy of probability distribution f, and $d_{\mathrm{KL}}(f||g)$ denotes the *Kullback–Leibler divergence* of f with respect to g. Because $d_{\mathrm{KL}}(f||g) \ge 0$ always, with equality holding if and only if $f = g$, it follows that the above quantity is maximized when $\mathrm{Ber}(K_{st}) = \mathrm{Ber}(d_{st}))$ for all s and t, that is, when $K_{st} = d_{st}$. □

8.4 Learning community frequencies and community link rates

We will discuss the article [BCL11]. A large random graph is modeled as a sequence of doubly stochastic block models $\mathrm{SBM}(\alpha, K^{(n)})$ on node set $[n]$ indexed by $n = 1, 2, \dots,$ where the label distribution α is a probability distribution on a set S, and the connectivity matrix is given by

$$K^{(n)}(s,t) = \rho_n K(s,t) \wedge 1, \tag{8.4}$$

where the *link density* ρ_n is a scalar such that $\rho_n \to 0$ as $n \to \infty$, and the BJR07
normalized kernel $K : S \times S \to [0, \infty)$ is a symmetric function[2] normalized assume
according to[3] $\rho_n = n^{-1}$

$$\sum_s \sum_t K(s,t)\, \alpha(s)\alpha(t) = 1.$$

[2] In the paper [BCL11] a different truncation $w1(w \le 1)$ was used in place of $w \wedge 1$, and $S = (0,1)$, but this should not make a difference.

[3] If S is an uncountable measurable space, then all sums over S involving $\alpha(u)$ should be replaced by integrals involving $\alpha(du)$.

As $n \to \infty$, one can verify (exercise) that any particular node pair is linked with probability $(1+o(1))\rho_n$, and the expected degree of any node equals $(1+o(1))n\rho_n$. The statistical learning problem is now to determine the label distribution α and the normalized kernel K from a graph sample $X^{(n)}$ obtained from the $\mathrm{SBM}(\alpha, K^{(n)})$ distribution.

An moment-based estimation approach for learning the model parameters is to compute the R-matching (or R-covering) densities defined in (7.4) of the observed graph sample $X^{(n)}$ for a suitable collection of small graphs R, and try to match the so-obtained empirical densities to the corresponding theoretical densities of the model. Because in the sparse setting with $\rho_n \to 0$, the empirical and model densities converge to zero, we need to work with normalized densities. For a graph R on node set $[r]$, the *normalized R-covering density* of the model is defined by

$$Q^*(R) \;=\; \sum_{z_1} \cdots \sum_{z_r} \alpha(z_1)\cdots\alpha(z_r) \prod_{ij \in E(R)} K(z_i, z_j),$$

and the normalized empirical R-covering density of the graph sample $X^{(n)}$ is defined by $\rho_n^{-|E(R)|}\hat{Q}_{X^{(n)}}(R)$ where $\hat{Q}_{X^{(n)}}(R)$ is defined in (7.5). The following result provides a sufficient condition for the normalized empirical R-covering density to be a consistent estimator of $Q^*_{\alpha,K}(R)$.

Theorem 8.7. *Assume that $cn^{-1} \le \rho_n \ll 1$, and that* Clarify me

$$\sum_s \sum_t K(s,t)^{2r} \alpha(s)\alpha(t) < \infty.$$

Then for any acyclic graph R with r nodes,

$$\rho_n^{-|E(R)|}\hat{Q}_{X^{(n)}}(R) \;\xrightarrow{\mathbb{P}}\; Q^*(R).$$

Sketch of proof. Because the distribution of the random graph $X = X^{(n)}$ is invariant with respect to node relabeling, we may relabel the nodes of R so that $V(R) = [r]$, and $\mathbb{P}(X \supset R') = \mathbb{P}(X \supset R)$ whenever R' is isomorphic to R. Hence the expected R-covering density of X equals

$$\mathbb{E}\hat{Q}_{X^{(n)}}(R) \;=\; \mathbb{E}\frac{\sum_{R' \in \mathcal{G}_n(R)} 1(X^{(n)} \supset R')}{|\mathcal{G}_n(R)|} \;=\; \mathbb{P}(X^{(n)} \supset R).$$

Because the the entries X_{ij} are conditionally independent given the node labeling Z, it follows that

$$\mathbb{P}(X^{(n)} \supset R \,|\, Z = z) \;=\; \prod_{ij \in E(R)} (\rho_n K(z_i, z_j) \wedge 1),$$

and

$$\rho_n^{-|E(R)|}\mathbb{P}(X^{(n)} \supset R \,|\, Z = z) \;=\; \prod_{ij\in E(R)} (K(z_i,z_j) \wedge \rho_n^{-1}) \;\to\; \prod_{ij\in E(R)} K(z_i,z_j).$$

After multiplying the left side above by $\alpha(z_1)\cdots\alpha(z_r)$ and summing over $z_1,\dots,z_r$, it follows (by Lebesgue's monotone convergence) that

$$\rho_n^{-|E(R)|}\mathbb{P}(X^{(n)} \supset R) \;\to\; Q^*(R),$$

and we conclude that

$$\mathbb{E}\,\rho_n^{-|E(R)|}\hat{Q}_{X^{(n)}}(R) \;\to\; Q^*(R).$$

To finish the proof by Chebyshev's inequality (i.e. the second moment method), it suffices to show that

$$\operatorname{Var}\left(\rho_n^{-|E(R)|}\hat{Q}_{X^{(n)}}(R)\right) \;\to\; 0.$$

This is done in [BCL11, Proof of Theorem 1] (see also [Bol01, Sec 4.1]). □

Using matching densities instead of covering densities

For sparse doubly stochastic block models, the empirical matching and covering densities behave roughly similarly. By similar arguments as for the R-covering density, it follows that the expected R-matching density of $X = X^{(n)}$ equals

$$\mathbb{E}\hat{P}_X(R) \;=\; \mathbb{P}(X[[r]] = R) \;=\; \mathbb{P}\Big(X_{ij} = R_{ij} \text{ for all } 1 \le i < j \le r\Big).$$

Observe that the difference between the covering and the matching densities is bounded by $\hat{Q}_X(R) - \hat{P}_X(R) \ge 0$ and

$$\begin{aligned}
\hat{Q}_X(R) - \hat{P}_X(R) \;&=\; \frac{1}{|\mathcal{G}_n(R)|} \sum_{R'\in\mathcal{G}_n(R)} 1(X \supset R')1(X_{k\ell} = 1 \text{ for some } k\ell \notin E(R')) \\
&=\; \frac{1}{|\mathcal{G}_n(R)|} \sum_{R'\in\mathcal{G}_n(R)} \prod_{ij\in E(R')} X_{ij} 1(X_{k\ell} = 1 \text{ for some } k\ell \notin E(R')) \\
&\le\; \frac{1}{|\mathcal{G}_n(R)|} \sum_{R'\in\mathcal{G}_n(R)} \left(\prod_{ij\in E(R')} X_{ij}\right)\left(\sum_{k\ell\notin E(R')} X_{k\ell}\right),
\end{aligned}$$

where $k\ell \notin E(R')$ refers to the $k\ell \in \binom{[r]}{2} \setminus E(R')$, so that

$$\begin{aligned}
\mathbb{E}|\hat{Q}_X(R) - \hat{P}_X(R)| &\leq \mathbb{E}\left(\prod_{ij \in E(R)} X_{ij}\right)\left(\sum_{k\ell \in \binom{[r]}{2} \setminus E(R)} X_{k\ell}\right) \\
&\leq \rho_n^{|E(R)|+1} \mathbb{E}\left(\prod_{ij \in E(R)} K(Z_i, Z_j)\right)\left(\sum_{k\ell \in \binom{[r]}{2} \setminus E(R)} K(Z_k, Z_\ell)\right) \\
&\leq c\rho_n^{|E(R)|+1}
\end{aligned}$$

under sufficient moment conditions on w. Hence by Markov's inequality,

$$\rho_n^{-|E(R)|}\hat{Q}_X(R) - \rho_n^{-|E(R)|}\hat{P}_X(R) \xrightarrow{\mathbb{P}} 0.$$

Hence by Theorem 8.7 it follows that also the normalized empirical matching density converges to $Q^*(R)$, according to

$$\rho_n^{-|E(R)|}\hat{P}_X(R) \xrightarrow{\mathbb{P}} Q^*(R).$$

8.5 Identifiability of the doubly stochastic block model from covering densities

The label distribution α can be viewed as a column vector of m numbers $\alpha_s \in [0,1]$ normalized according to

$$\sum_{s=1}^m \alpha_s = 1. \tag{8.5}$$

In a finite label we can ignore the truncation term in the kernel definition (8.4), and we can write $K^{(n)}(s,t) = \rho_n K_{s,t}$, where $\rho_n \in (0,1)$ is the overall link density and the limiting kernel K is now a symmetric m-by-m matrix with entries in $K_{st} \in [0,1]$ normalized by

$$\sum_{s=1}^m \sum_{t=1}^m K_{st}\alpha_s\alpha_t = 1. \tag{8.6}$$

To learn the model it is then sufficient to determine the m real numbers α_s and the $m(m+1)/2$ real numbers $K_{st}, 1 \leq s \leq t \leq m$. Actually, a bit less is sufficient. Namely, (8.5) and (8.6) imply that we can omit learning one entry of α and one entry of K. Therefore, the number of free parameters in the model equals $m(m+3)/2-2$.

The limiting normalized R-covering density of a doubly stochastic block model with label distribution α and kernel K was found to be

$$Q^*(R) \;=\; \sum_{z_1} \cdots \sum_{z_r} \prod_{ij \in E(R)} K_{z_i,z_j}\, \alpha_{z_1} \cdots \alpha_{z_r}.$$

The problem is now to determine α and K from $Q^*(R)$ for a collection of R.

Let us compute the normalized R-covering density for some simple graphs first. When R is a single link, we get

$$Q^*(\text{link}) \;=\; \sum_s \sum_t K_{st}\, \alpha_s \alpha_t \;=\; 1$$

due to the normalization constraint (8.6). For the triangle we obtain

$$Q^*(\text{triangle}) \;=\; \sum_s \sum_t \sum_u K_{st} K_{tu} K_{su}\, \alpha_s \alpha_t \alpha_u,$$

but this appears a complicated formula to analyze. To obtain simpler algebraic expressions, we will try computing covering densities for some acyclic graphs. For the 3-path with $V(R) = \{1,2,3,4\}$ and $E(R) = \{\{1,2\},\{2,3\},\{3,4\}\}$, we find that

$$\begin{aligned} Q^*(\text{3-path}) \;&=\; \sum_{u_1,u_2,u_3,u_4} K_{u_1u_2} K_{u_2u_3} K_{u_3u_4}\, \alpha_{u_1} \alpha_{u_2} \alpha_{u_3} \alpha_{u_4} \\ &=\; \sum_{u_1,u_2,u_3,u_4} \alpha_{u_1} L_{u_1u_2} L_{u_2u_3} L_{u_3u_4} \\ &=\; \sum_{u_1} \sum_{u_4} \alpha_{u_1} L^3_{u_1u_4} \\ &=\; \sum_u \alpha_u (L^3 e)_u, \end{aligned}$$

where $L_{uv} = K_{uv}\alpha_v$ is the matrix product of K and the diagonal matrix with entries $\alpha_1, \ldots, \alpha_m$, and e is the column vectors of m ones. For the 3-star with $V(R) = \{1, 2, 3, 4\}$ and $E(R) = \{\{1, 2\}, \{1, 3\}, \{1, 4\}\}$, we get

$$\begin{aligned}
Q^*(\text{3-star}) &= \sum_{u_1,u_2,u_3,u_4} K_{u_1u_2}K_{u_1u_3}K_{u_1u_4}\,\alpha_{u_1}\alpha_{u_2}\alpha_{u_3}\alpha_{u_4} \\
&= \sum_{u_1} \alpha_{u_1} \left(\sum_{u_2}\sum_{u_3}\sum_{u_4} K_{u_1u_2}K_{u_1u_3}K_{u_1u_4}\alpha_{u_2}\alpha_{u_3}\alpha_{u_4} \right) \\
&= \sum_{u} \alpha_u \left(\sum_{v} K_{uv}\alpha_v \right)^3 \\
&= \sum_{u} \alpha_u \left((Le)_u\right)^3.
\end{aligned}$$

The above computations can be generalized to (exercise)

$$\begin{aligned}
Q^*(k\text{-path}) &= \sum_u \alpha_u (L^k e)_u, \\
Q^*(\ell\text{-star}) &= \sum_u \alpha_u \left((Le)_u\right)^\ell.
\end{aligned}$$

Even more generally, one can verify (exercise) that

$$Q^*((k,\ell)\text{-star}) = \sum_u \alpha_u (L^k e)_u^\ell, \tag{8.7}$$

where a *(k,ℓ)-star* refers to a graph of radius k obtained by joining the endpoints of ℓ paths of length k at a common hub node in the center. Hence an $(1,\ell)$-star is the usual ℓ-star.

Can we identify α and K from the covering densities of paths and stars? The first claim is that we can identify $(\alpha_1, \dots, \alpha_m)$ from the ℓ-star covering densities with $\ell = 1, \dots, 2m-1$. Why? Let X_1 be a random variable which takes on value $(Le)_u$ with probability α_u for all $u = 1, \dots, m$. Then

$$\mathbb{E}X_1^\ell = \sum_u \alpha_u \left((Le)_u\right)^\ell = Q^*(\ell\text{-star}).$$

Then a classical theorem about the method of moments [Fel71] tells that the distribution (support and probabilities) of X_1 can be recovered from sufficiently many moments $\mathbb{E}X_1, \mathbb{E}X_1^2, \dots$ Hence, we may obtain the label distribution α and the rows sums $(Le)_1, \dots (Le)_m$ from the star covering densities. Next, let X_k be a random variable which takes on value $(L^k e)_u$ with probability α_u. Then by (8.7),

$$\mathbb{E}X_k^\ell = \sum_u \alpha_u (L^k e)_u^\ell = Q^*((k,\ell)\text{-star}),$$

and hence we may recover the rows sums $(L^k e)_1, \ldots, (L^k e)_m$ from the (k, ℓ)-star covering densities. Let us now defined the m-by-m square matrices

$$V^{(1)} = \begin{bmatrix} e & Le & \cdots & L^{m-1}e \end{bmatrix}$$

and

$$V^{(2)} = \begin{bmatrix} Le & L^2 e & \cdots & L^m e \end{bmatrix}.$$

Then

$$LV^{(1)} = V^{(2)},$$

and if the columns of $V^{(1)}$ are linearly independent, we obtain the matrix L from

$$L = V^{(2)}(V^{(1)})^{-1},$$

and thereafter the matrix K by $K_{uv} = L_{uv}\alpha_v^{-1}$. Hence we have proved the following result.

Theorem 8.8. *Assume that the vectors $e, Le, \ldots, L^{m-1}e$ are linearly independent and $\alpha_u > 0$ for all u. Then the label distribution α and the normalized kernel K can be identified from the normalized covering densities $Q^*((k, \ell)$-star) with $k, \ell \geq 1$.*

Combining Theorems 8.7 and 8.8 yields a consistent way to estimate the parameters α and K of a large and sparse doubly stochastic block model from the normalized empirical covering densities of (k, ℓ)-stars computed from a single large sample $X^{(n)}$.

Appendix A

Probability

Here are some miscellaneous facts from probability theory that are used in the text.

A.1 Inequalities

Proposition A.1 (Markov's inequality)**.** *For any random number* $X \geq 0$ *and any* $a > 0$*,*

$$\mathbb{P}(X \geq a) \ \leq \ a^{-1}\mathbb{E}X.$$

Proof. First, note that $\mathbb{P}(X \geq a) = \mathbb{E}1(X \geq a)$ where $1(A)$ in general denotes the *indicator* of the event A. Hence by the linearity of expectation,

$$a\mathbb{P}(X \geq a) \ = \ a\mathbb{E}1(X \geq a) \ = \ \mathbb{E}a1(X \geq a).$$

Next, the inequalities

$$a1(X \geq a) \ \leq \ X1(X \geq a) \ \leq \ X$$

which are valid for any realization of X, and the monotonicity of the expectation imply that

$$\mathbb{E}a1(X \geq a) \ \leq \ \mathbb{E}X.$$

Hence $a\mathbb{P}(X \geq a) \leq \mathbb{E}X$, and the claim follows. □

Proposition A.2 (Chebyshev's inequality)**.** *For any random number* X *with a finite mean* $\mu = \mathbb{E}X$ *and any* $a > 0$*,*

$$\mathbb{P}(|X - \mu| \geq a) \ \leq \ a^{-2}\,\mathrm{Var}(X).$$

Proof. By applying Markov's inequality for $Y = (X - \mu)^2$, we find that

$$\begin{aligned} \mathbb{P}(|X - \mu| \ge a) &= \mathbb{P}((X - \mu)^2 \ge a^2) \\ &\le (a^2)^{-1}\mathbb{E}(X - \mu)^2 \;=\; a^{-2}\,\mathrm{Var}(X). \end{aligned}$$

□

The following inequality is due to the Finnish-born Wassily Hoeffding.

Proposition A.3 (Hoeffding's inequality)**.** *Let $S_n = \sum_{i=1}^n X_i$ where the summands are independent and bounded by $a_i \le X_i \le b_i$. Then for any $t > 0$,*

$$\begin{aligned} \mathbb{P}(S_n \ge \mathbb{E}S_n + t) &\le e^{-\frac{2t^2}{\sum_i (b_i - a_i)^2}}, \\ \mathbb{P}(S_n \le \mathbb{E}S_n - t) &\le e^{-\frac{2t^2}{\sum_i (b_i - a_i)^2}}, \end{aligned}$$

and

$$\mathbb{P}(|S_n - \mathbb{E}S_n| \;\ge\; t) \;\le\; 2e^{-\frac{2t^2}{\sum_i (b_i - a_i)^2}}.$$

Proof. A well-written proof of the first inequality, based on an extremality property related to convex stochastic orders, is available in the original research article Hoeffding [Hoe63]. The second inequality follows by applying the first inequality to $\tilde{S}_n = -S_n$ and the third inequality follows from the first two by the union bound. □

A.2 Convergence of discrete probability distributions

For probability distributions on a countable set S we say that $\mu_n \to \mu$ *weakly* if $\sum_x \phi(x)\mu_n(x) \to \sum_x \phi(x)\mu(x)$ for every bounded function $\phi : \mathbb{R} \to \mathbb{R}$. For random variables distributed according to μ_n and μ, this is denoted as $X_n \xrightarrow{d} X$.

Theorem A.4. *The following are equivalent for random sequences in a countable set S:*

(i) $X_n \xrightarrow{d} X$.

(ii) $\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$ *for all* $A \subset S$.

(iii) $\mathbb{P}(X_n = s) \to \mathbb{P}(X = s)$ *for all* $s \in S$.

(iv) $X_n \xrightarrow{tv} X$.

Proof. Sorry, the current proof is in Finnish, and part (iv) is missing. Todistetaan lause kolmessa vaiheessa näyttämällä toteen seuraamukset (i) $\implies$ (ii) $\implies$ (iii) $\implies$ (i). Kaksi ensimmäistä vaihetta on helppoja, kun taas kolmas vaatii vähän enemmän työtä.

(i) $\implies$ (ii). Oletetaan, että (i) pätee ja valitaan mielivaltainen $A \subset S$. Funktio $f(s) = 1_A(s)$ on nyt rajoitettu, joten (i):n nojalla

$$\mathbb{P}(X_n \in A) = \mathbb{E}1_A(X_n) \to \mathbb{E}1_A(X) = \mathbb{P}(X \in A).$$

(ii) $\implies$ (iii). Oletetaan, että (ii) pätee ja valitaan mielivaltainen $s \in S$. Määritellään $A = \{s\}$. Tällöin (ii):n nojalla

$$\mathbb{P}(X_n = s) = \mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A) = \mathbb{P}(X = s).$$

(iii) $\implies$ (i). Oletetaan, että (iii) pätee ja valitaan jokin rajoitettu funktio $f : S \to \mathbb{R}$. Valitaan myös jokin mielivaltaisen pieni $\epsilon > 0$. Esitetään S numeroimalla sen alkiot muodossa $S = \{s_1, s_2, \dots\}$ ja merkitään $C_k = \{s_1, \dots, s_k\}$. Koska $\sum_{j=1}^{\infty} P_X(s_j) = 1$, voidaan valita luku k siten, että

$$\mathbb{P}(X \in C_k^c) = \sum_{j=k+1}^{\infty} P_X(s_j) \le \epsilon.$$

Kirjoitetaan tarkasteltavana olevien odotusarvojen erotus muodossa

$$\mathbb{E}f(X_n) - \mathbb{E}f(X) \;=\; \Delta_n + \mathbb{E}f(X_n)1_{\{X_n \in C_k^c\}} - \mathbb{E}f(X)1_{\{X_n \in C_k^c\}}, \tag{A.1}$$

missä

$$\Delta_n = \mathbb{E}f(X_n)1_{\{X_n \in C_k\}} - \mathbb{E}f(X)1_{\{X_n \in C_k\}},$$

Oletuksen (iii) nojalla

$$\Delta_n = \sum_{s \in C_k} f(s)(P_{X_n}(s) - P_X(s)) \to 0,$$

kun $n \to \infty$. Seuraavaksi nähdään, että yhtälön (A.1) oikean puolen toinen termi toteuttaa

$$\left|\mathbb{E}f(X_n)1_{\{X_n \in C_k^c\}}\right| \le ||f||\,\mathbb{P}(X_n \in C_k^c) = ||f||\left(\mathbb{P}(X \in C_k^c) + \Delta_n'\right),$$

missä $||f|| = \sup_{s \in S} |f(s)|$ ja

$$\begin{aligned}\Delta_n' &= \mathbb{P}(X_n \in C_k^c) - \mathbb{P}(X \in C_k^c)\\ &= \mathbb{P}(X \in C_k) - \mathbb{P}(X_n \in C_k)\\ &= \sum_{s \in C_k} (P_X(s) - P_{X_n}(s)) \to 0,\end{aligned}$$

kun $n \to \infty$ oletuksen (iii) nojalla. Näin ollen siis

$$\begin{aligned}|\mathbb{E}f(X_n) - \mathbb{E}f(X)| &\le |\Delta_n| + ||f||\,(2\mathbb{P}(X_n \in C_k^c) + |\Delta_n'|) \\ &\le |\Delta_n| + ||f||\,(2\epsilon + |\Delta_n'|)\,.\end{aligned}$$

Koska $|\Delta_n| \le \epsilon$ ja $|\Delta_n'| \le \epsilon$ kaikilla riittävän suurilla n, nähdään että

$$|\mathbb{E}f(X_n) - \mathbb{E}f(X)| \le (1 + 3||f||)\epsilon$$

kaikilla riittävän suurilla n. Koska ϵ oli mielivaltaisen pieni, (i) seuraa. □

A.3 Weak convergence of probability measures

Let $\mu, \mu_1, \mu_2, \dots$ be probability distributions on $\mathbb{R}$. We say that $\mu_n \to \mu$ *weakly* if $\int \phi(x)\mu_n(dx) \to \int \phi(x)\mu(dx)$ for every bounded continuous function $\phi : \mathbb{R} \to \mathbb{R}$. We say that $\mu_n \to \mu$ *weakly and with k-th moments*, if in addition μ_n and μ have finite k-th moments and $\int |x|^k \mu_n(dx) \to \int |x|^k \mu(dx)$. The sequence (μ_n) is called *uniformly integrable* if $\sup_n \int |x|\mu_n(dx) 1(|x| > K) \to 0$ as $K \to \infty$. Let $X, X_1, X_2, \dots$ be real-valued random variables. We say that $X_n \to X$ weakly (resp. with weakly with k-th moments) if the corresponding probability distributions converge weakly (resp. weakly with k-th moments). We say that (X_n) is uniformly integrable if the collection of corresponding probability distributions is uniformly integrable.

Lemma A.5. *Let X_n and X be random numbers such that $X_n \to X$ weakly with 1st moments. Then the sequence (X_n) is uniformly integrable.*

Proof. Given $\epsilon > 0$, by Lebesgue's dominated convergence we may choose $K > 0$ such that $EX1(X > K) \le \epsilon/3$. Then let ϕ_K be a continuous bounded function such that $\phi_K(x) = x$ for $x \le K$ and $\phi_K = 0$ for $x \ge K + 1$. Then

$$x1(x \le K) \;\le\; \phi_K(x) \;\le\; x1(x \le K + 1),$$

so that

$$\begin{aligned}\mathbb{E}X_n 1(X_n > K + 1) &= \mathbb{E}X_n - \mathbb{E}X_n 1(X_n \le K + 1) \\ &\le \mathbb{E}X_n - \mathbb{E}\phi_K(X_n) \\ &= \mathbb{E}X_n - \mathbb{E}\phi_K(X) + \mathbb{E}\phi_K(X) - \mathbb{E}\phi_K(X_n) \\ &\le \mathbb{E}X_n - \mathbb{E}X1(X \le K) + \mathbb{E}\phi_K(X) - \mathbb{E}\phi_K(X_n) \\ &= \mathbb{E}X1(X > K) + \mathbb{E}X_n - \mathbb{E}X + \mathbb{E}\phi_K(X) - \mathbb{E}\phi_K(X_n) \\ &\le \epsilon/3 + |\mathbb{E}X_n - \mathbb{E}X| + |\mathbb{E}\phi_K(X_n) - \mathbb{E}\phi_K(X)|.\end{aligned}$$

Then we may choose n_0 so large that $|\mathbb{E}X_n - \mathbb{E}X| \leq \epsilon/3$ and $|\mathbb{E}\phi_K(X_n) - \mathbb{E}\phi_K(X)| \leq \epsilon/3$ for all $n > n_0$. Hence $\mathbb{E}X_n 1(X_n > K + 1) \leq \epsilon$ for all $n > n_0$. Furthermore, for every $1 \leq m \leq n_0$ we may choose, again by Lebesgue's dominated convergence, K_m such that $\mathbb{E}X_m 1(X_m > K_m) \leq \epsilon$. Now if we choose $L = \max\{K + 1, K_1, \ldots, K_{n_0}\}$, it follows that $\sup_n \mathbb{E}X_n 1(X_n > L) \leq \epsilon$. $\square$

Appendix B

Locally finite rooted graphs

We discuss the concepts introduced by Itai Benjamini and Oded Schramm in [BS01]. A *rooted graph* is a pair $G^\bullet = (G, o)$ where G is a graph (undirected, no parallel links, loopless) and $o \in V(G)$ is a node of the graph called the *root*. The root of $G^\bullet$ is usually denoted by $o(G^\bullet)$. The *r-neighborhood* of a rooted graph $G^\bullet = (G, o)$ is the rooted graph $T_r G^\bullet = (G_{o,r}, o)$ where $G_{o,r}$ is the subgraph of G induced by the set of nodes reachable from o by a path of length at most r. A rooted graph is *connected* if every node is reachable from the root via some finite path. A rooted graph is *locally finite* if every node has a finite degree.

Remark B.1 (Graphs with loops)**.** Homomorphisms are conceptually easier to define for graphs with loops. Fix a set V and let $\mathcal{G}_s$ be the set of finite, simple (no parallel links, loops are allowed), undirected graphs. Let $\mathcal{G}_s(V)$ be the set of graphs in $\mathcal{G}_s$ with node set V. Here $E(G)$ is a collection of subsets of $V(G)$ of cardinality one or two, and singleton sets in $E(G)$ are called loops. Note that any function $\phi : V \to V'$ induces a map $\mathrm{Pow}(V) \to \mathrm{Pow}(V')$ such that $e \subset V$ mapsto $\phi(e) := \{\phi(v) : v \in e\}$. Hence any function $\phi : V \to V'$ also induces a graph-to-graph map such that $V(G) \mapsto \phi(V(G)) = \{\phi(v) : v \in V(G)\}$ and $E(G) \mapsto \phi(E(G)) := \{\phi(e) : e \in E(G)\}$. Now loops map to loops, but also nonloops can map into loops. Now such a map ϕ is a homomorphism from G into H just means that $\phi : V(G) \to V(H)$ as a graph-to-graph map maps G into a subgraph of H in the sense that $\phi(V(G)) \subset V(H)$ and $\phi(E(G)) \subset E(H)$. It is easy to verify that a map ϕ is injective as map $V \to V'$ if and only if ϕ is injective as a map $\mathcal{G}_s(V) \to \mathcal{G}_s(V')$. Strong homomorphism means that $\phi(G)$ is an induced subgraph of H, a subgraph of H induced by $\phi(V(G))$. Embedding means an injective (as a function on V or equivalently on $\mathcal{G}_s$) homomorphism, and strong (or injective) embedding means an injective strong homomorphism.

If we insisted on working on the space of loopless simple graphs, then not every ϕ on V extends to a function on the space of loopless simple graphs — only injections do — but we often want to deal with noninjective homomorphisms. Note that in the above sense, if ϕ is a homomorphism from G into a loopless graph H, then also G must be loopless. Note that when we want to count the number of embeddings from G to H, we look at functions $\phi : V(G) \to V(H)$.

A *homomorphism* from $G^\bullet$ into $H^\bullet$ is map $\phi : V(G^\bullet) \to V(H^\bullet)$ such that $\phi(E(G^\bullet)) \subset E(H^\bullet)$ and $\phi(o(G^\bullet)) = \phi(o(H^\bullet))$. Equivalently, ϕ is a homomorphism of the unrooted graphs which preserves the root.

Rooted graphs $G^\bullet$ and $H^\bullet$ are *isomorphic*, denoted $G^\bullet \cong H^\bullet$, if there exists a bijection $\phi : V(G) \to V(H)$ which satisfies $\phi(o(G^\bullet)) = o(H^\bullet)$ and the property that $\{\phi(u), \phi(v)\} \in E(H)$ if and only if $\{u, v\} \in E(G)$.

B.1 Identification from local neighborhoods

The following technical result is later needed in verifying that the local distance defined on the isomorphism classes of connected locally finite rooted graphs is a metric.

Lemma B.2. *Let $F^\bullet$ and $G^\bullet$ be connected locally finite rooted graphs such that $T_r F^\bullet \cong T_r G^\bullet$ for all $r \geq 0$. Then $F^\bullet \cong G^\bullet$.*

Proof. We need to prove the existence of an isomorphism between $F^\bullet$ and $G^\bullet$. The proof is based on a nonconstructive subsequence argument. The assumption implies that for any r there exists an isomorphism ϕ_r between $T_r F^\bullet$ and $T_r G^\bullet$. We will extend this to a function $\psi_r : V(F^\bullet) \to V(G^\bullet)$ by defining

$$\psi_r(v) \;=\; \begin{cases} \phi_r(v), & v \in V(T_r F^\bullet), \\ o_G, & \text{otherwise.} \end{cases}$$

Now we have a sequence $(\psi_0, \psi_1, \psi_2, \dots)$ of functions from $V(F^\bullet)$ into $V(G^\bullet)$. We will next describe how we can extract a rooted graph isomorphism between $\mathcal{F}^\bullet$ and $\mathcal{G}^\bullet$ from this sequence. For convenience, denote by $V_r = V(T_r F^\bullet)$ the set of nodes within distance at most r from the root in $F^\bullet$.

Observe first that for any r, the restriction $\psi_r|_{V_0}$ of ψ_r to domain V_0 is an isomorphism between $T_0 F^\bullet$ and $T_0 G^\bullet$. Because the set of such isomorphisms is finite, there exists at least one such isomorphism, call it ϕ_0', which is repeated infinitely many times in the sequence $(\psi_r|_{V_0} : r \in \mathbb{N})$. Hence there exists an infinite subset $\mathbb{N}_0 \subset \mathbb{N}$ such that

$$\psi_r|_{V_0} \;=\; \phi_0' \quad \text{for all } r \in \mathbb{N}_0.$$

Let us now repeat this argument. Note that for any $r \in \mathbb{N}_0$, the restriction $\psi_r|_{V_1}$ is an isomorphism between $T_1 F^\bullet$ and $T_1 G^\bullet$. Again, because the set of such isomorphisms is finite, there is at least one, call it ϕ'_1, which is repeated infinitely many times in the sequence $(\psi_r|_{V_1} : r \in \mathbb{N}_0)$. Hence there exists an infinite subset $\mathbb{N}_1 \subset \mathbb{N}_0$ such that $\psi_r|_{V_1} = \phi'_1$ for all $r \in \mathbb{N}_1$. By continuing this way we see that there exist a nested sequence of infinite sets $\mathbb{N} \supset \mathbb{N}_0 \supset \mathbb{N}_1 \supset \cdots$ such that for all $k \ge 0$,

$$\psi_r|_{V_k} \;=\; \phi'_k \quad \text{for all } r \in \mathbb{N}_k,$$

where ϕ'_k is an isomorphism between $T_k F^\bullet$ and $T_k G^\bullet$.

Let us define ψ'_ℓ as the first member of the sequence $(\psi_r : r \in \mathbb{N}_\ell)$. Because ψ'_ℓ is also a member of $(\psi_r : r \in \mathbb{N}_k)$, it follows that for $v \in V_k$,

$$\psi'_\ell(v) \;=\; \phi'_k(v) \quad \text{for all } \ell \ge k.$$

Now because $F^\bullet$ is connected, we have $V(F^\bullet) = \cup_{k=0}^\infty V_k$, and the above equation shows that the function sequence $(\psi'_\ell : \ell \in \mathbb{N}_0)$ converges pointwise to a limiting function

$$\psi'_\infty \;=\; \sum_{k=0}^\infty \phi'_k 1_{U_k}$$

where $U_0 = V_0$ and $U_k = V_k \setminus V_{k-1}$ for $k \ge 1$.

Let us finally verify that ψ' is a rooted graph isomorphism between $F^\bullet$ and $G^\bullet$. The above representation makes it clear that ψ' is injective. It is also surjective because both $F^\bullet$ and $G^\bullet$ are connected. If $u, v \in V(F^\bullet)$, then for $k = \max\{d_F(o_F, u), d_F(o_F, v)\}$ we have $u, v \in V_k$. Because ϕ'_k is an isomorphism between $T_k F^\bullet$ and $T_k G^\bullet$, it follows that $\phi'_k(u), \phi'_k(v) \in V(T_k G^\bullet)$, and $\{u, v\} \in E(T_k F^\bullet)$ iff $\{\phi'_k(u), \phi'_k(v)\} \in E(T_k G^\bullet)$. Because $\psi'_\infty|_{V_k} = \phi'_k$, it follows that

$$\{u, v\} \in E(T_k F^\bullet) \iff \{\psi'(u), \psi'(v)\} \in E(T_k G^\bullet),$$

which now is equivalent to

$$\{u, v\} \in E(F^\bullet) \iff \{\psi'(u), \psi'(v)\} \in E(G^\bullet).$$

Finally, $\psi'(o_F) = \psi'|_{V_0}(o_F) = \phi'_0(o_F) = o_G$ concludes the claim. □

B.2 Completeness

The following important result tells that any compatible sequence of rooted graphs of increasing diameter admits a limiting graph which is unique up to isomorphism. Here compatibility means that truncated versions of the members of the sequence match with earlier members of the sequence.

Lemma B.3. *Let $G_0^\bullet, G_1^\bullet, \dots$ be connected locally finite rooted graphs which are compatible in the sense that $T_r G_s^\bullet \cong G_r^\bullet$ for all $r \le s$. Then there exists a connected locally finite rooted graph $H^\bullet$ such that*

$$G_r^\bullet \cong T_r H^\bullet \quad \textit{for all } r \ge 0. \tag{B.1}$$

Moreover, such $H^\bullet$ is unique up to isomorphism.

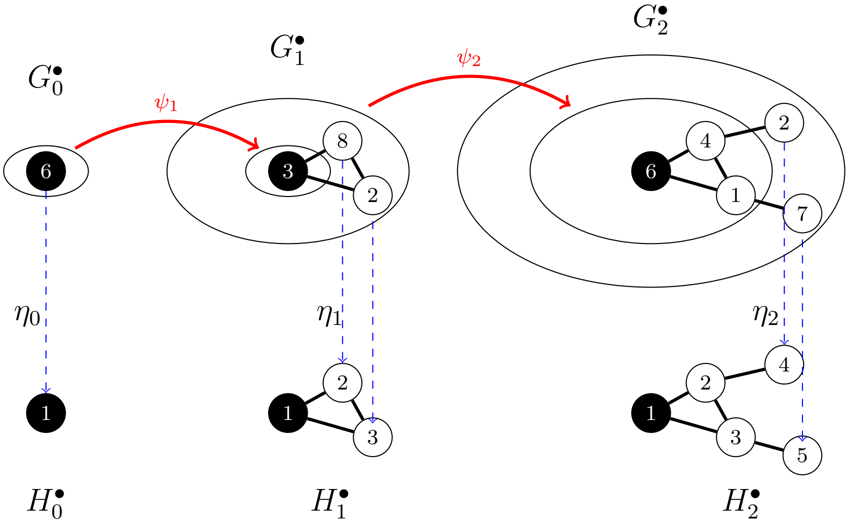

Figure B.1: Compatible construction of isomorphic copies of $G_0^\bullet, G_1^\bullet, \dots$

Proof. To prove existence, we will first construct isomorphic copies of $G_0^\bullet, G_1^\bullet, \dots$ on a common node set, in a compatible way. To do this, denote $V_r = \{v \in \mathbb{N} : v \le N_r\}$ where $N_r = |V(G_r^\bullet)|$. We will define a sequence of bijections $\phi_r : V(G_r^\bullet) \to V_r$ recursively as follows. First, we let $\phi_0 = \eta_0$ where η_0 is the unique bijection from $V(G_0^\bullet) = \{o(G_0^\bullet)\}$ onto the singleton set $V_0 = \{1\}$. For any $r \ge 1$, we define

$$\phi_r(v) \;=\; \begin{cases} \phi_{r-1}(\psi_r^{-1}(v)), & v \in V(T_{r-1}G_r^\bullet), \\ \eta_r(v), & v \in V(G_r^\bullet) \setminus V(T_{r-1}G_r^\bullet), \end{cases}$$

where ψ_r is an isomorphism between $G_{r-1}^\bullet$ and $T_{r-1}G_r^\bullet$, and η_r is an arbitrary bijection from $V(G_r^\bullet) \setminus V(T_{r-1}G_r^\bullet)$ onto $V_r \setminus V_{r-1}$, see Figure B.1. Next we define $H_r^\bullet = \phi_r(G_r^\bullet)$ as the image[1] of $G_r^\bullet$ with respect to ϕ_r. Because ϕ_r is a

[1] The *image* $\phi(G^\bullet)$ of a rooted graph $G^\bullet$ by a function ϕ with domain $V(G^\bullet)$ is defined as the rooted graph $H^\bullet = \phi(G^\bullet)$ with node set $V(H^\bullet) = \{\phi(v) : v \in V(G^\bullet)\}$, link set $E(H^\bullet) = \{\{\phi(u), \phi(v)\} : \{u, v\} \in E(G^\bullet)\}$, and root $o(H^\bullet) = \phi(o(G^\bullet))$.

bijection, it follows that

$$G_r^\bullet \;\cong\; H_r^\bullet \quad \text{for all } r \geq 0.$$

Moreover, the construction guarantees that $o(H_r^\bullet) = 1$ for all $r \geq 0$. We define a rooted graph $H^\bullet$ with node set $V(H^\bullet) = \cup_{r=0}^\infty V(H_r^\bullet)$, link set $E(H^\bullet) = \cup_{r=0}^\infty E(H_r^\bullet)$ and root $o(H^\bullet) = 1$. Then it follows that

$$T_r H^\bullet \;=\; H_r^\bullet \quad \text{for all } r \geq 0.$$

Now (B.1) follows from the fact that $G_r^\bullet \cong H_r^\bullet$, and from this we also conclude that $H^\bullet$ is locally finite and connected.

To verify uniqueness, assume that $H^\bullet$ and $I^\bullet$ are locally finite rooted graphs satisfying (B.1). Then $T_r H^\bullet \cong T_r I^\bullet$ for all r, and it follows that $d(H^\bullet, I^\bullet) = 0$. Hence by Lemma B.2, it follows that $H^\bullet \cong I^\bullet$.

□

B.3 Local distance

Let $\mathcal{G}^\bullet$ be the set of connected, locally finite, rooted graphs with node set contained in $\mathbb{N} = \{1, 2, \dots\}$. Note that this set is uncountably infinite. The *local distance* on $\mathcal{G}^\bullet$ is defined by

$$d_{\mathrm{loc}}(F^\bullet, G^\bullet) \;=\; 2^{-\sup\{r \geq 0 : T_r F^\bullet \cong T_r G^\bullet\}}. \tag{B.2}$$

The following result implies that d_{loc} is a pseudometric on $\mathcal{G}^\bullet$ (see for example [Kel75, Chapter 4]). We equip $\mathcal{G}^\bullet$ with the pseudometric topology, that is, the topology generated by the open balls $\{G^\bullet \in \mathcal{G}^\bullet : d_{\mathrm{loc}}(G^\bullet, F^\bullet) < \epsilon\}$.

Proposition B.4. *The map $d_{\mathrm{loc}} : \mathcal{G}^\bullet \times \mathcal{G}^\bullet \to [0,1]$ satisfies*

(i) $d_{\mathrm{loc}}(F^\bullet, G^\bullet) = 0$ *if and only if* $F^\bullet \cong G^\bullet$,

(ii) $d_{\mathrm{loc}}(F^\bullet, G^\bullet) = d_{\mathrm{loc}}(G^\bullet, F^\bullet)$,

(iii) $d_{\mathrm{loc}}(F^\bullet, H^\bullet) \leq \max\{d_{\mathrm{loc}}(F^\bullet, G^\bullet), d_{\mathrm{loc}}(G^\bullet, H^\bullet)\}$.

Moreover,

(iv) $d_{\mathrm{loc}}(\hat{F}^\bullet, \hat{G}^\bullet) = d_{\mathrm{loc}}(F^\bullet, G^\bullet)$ *whenever* $\hat{F}^\bullet \cong F^\bullet$ *and* $\hat{G}^\bullet \cong G^\bullet$.

Proof. (i) Assume that $d_{\mathrm{loc}}(F^\bullet, G^\bullet) = 0$. Then $T_r F^\bullet \cong T_r G^\bullet$ for all $r \geq 0$, and by Lemma B.2 it follows that $F^\bullet \cong G^\bullet$. The converse is immediate.

(ii) Immediate.

(iii) Let $r_{FH} = \sup\{r \ge 0 : T_r F^\bullet \cong T_r H^\bullet\}$, and define r_{FG}, r_{GH} similarly. Then $T_r F^\bullet \cong T_r G^\bullet$ and $T_r G^\bullet \cong T_r H^\bullet$ for all $r \le \min\{r_{FG}, r_{GH}\}$. This implies that $T_r F^\bullet \cong T_r H^\bullet$ for all $r \le \min\{r_{FG}, r_{GH}\}$. We conclude that $r_{FH} \ge \min\{r_{FG}, r_{GH}\}$, or equivalently, $2^{-r_{FH}} \le \max\{2^{-r_{FG}}, 2^{-r_{GH}}\}$, and hence the inequality in (iii) is valid.

(iv) This follows by noting that if $\hat{F}^\bullet \cong F^\bullet$ and $\hat{G}^\bullet \cong G^\bullet$, then $T_r F^\bullet \cong T_r G^\bullet$ if and only if $T_r \hat{F}^\bullet \cong T_r \hat{G}^\bullet$. □

For a rooted graph $G^\bullet$, the equivalence class

$$[G^\bullet] = \{H^\bullet \in \mathcal{G}^\bullet : H^\bullet \cong G^\bullet\}$$

is called an *unlabelled* rooted graph with *representative* $G^\bullet$. For a collection of rooted graphs $\mathcal{A}$ we denote $[\mathcal{A}] = \{[A^\bullet] : A^\bullet \in \mathcal{A}\}$. Then $[\mathcal{G}^\bullet]$ denotes the set of unlabelled rooted graphs which are connected and locally finite. Proposition B.4:(iv) implies that $d_{\mathrm{loc}}(F^\bullet, G^\bullet)$ depends on its inputs only via their equivalence classes. Hence we may define the *local distance* on $[\mathcal{G}^\bullet]$ by $d_{\mathrm{loc}}([F^\bullet], [G^\bullet]) = d_{\mathrm{loc}}(F^\bullet, G^\bullet)$.

Then the projection map $i : G^\bullet \mapsto [G^\bullet]$ from $\mathcal{G}^\bullet$ onto $[\mathcal{G}^\bullet]$ is an isometry of pseudometric spaces: $d_{\mathrm{loc}}([F^\bullet], [G^\bullet]) = d_{\mathrm{loc}}(F^\bullet, G^\bullet)$ for all $F^\bullet, G^\bullet \in \mathcal{G}^\bullet$. Hence i is a surjection that maps open balls of radius r in $\mathcal{G}^\bullet$ into open balls of radius r in $[\mathcal{G}^\bullet]$, conversely preimages of open balls of radius r in $[\mathcal{G}^\bullet]$ are open balls of radius r in $\mathcal{G}^\bullet$. Hence $\mathcal{A}$ is open in $\mathcal{G}^\bullet$ if and only if $[\mathcal{A}]$ is open in $[\mathcal{G}^\bullet]$.

Proposition B.5. *$([\mathcal{G}^\bullet], d_{\mathrm{loc}})$ is a complete separable metric space.*

Proof. Let us first verify that d_{loc} is a metric. Positivity and symmetry follow from Proposition B.4:(i)–(ii). The triangle inequality follows from Proposition B.4:(iii) which is indeed a stronger property, called an ultrametric inequality.

To verify separability, note that the set of all *finite* rooted graphs $\mathcal{G}^\bullet_{<\infty} \subset \mathcal{G}^\bullet$ is a countable set. We will verify that $[\mathcal{G}^\bullet_{<\infty}]$ is dense in $[\mathcal{G}^\bullet]$. This is easy because for any $G \in \mathcal{G}^\bullet$, and for any $\epsilon > 0$, $d_{\mathrm{loc}}(G_r, G) \le 2^{-r} < \epsilon$ when r is large enough. Hence every open d_{loc}-ball in $[\mathcal{G}^\bullet]$ also contains elements of $[\mathcal{G}^\bullet_{<\infty}]$.

To verify completeness, fix a Cauchy sequence $([G^\bullet_n])_{n\ge 1}$ with representative rooted graphs $G^\bullet_n$. Then for any $\epsilon > 0$ there exists n_ϵ such that $d_{\mathrm{loc}}(G^\bullet_m, G^\bullet_n) < \epsilon$ for all $m, n \ge n_\epsilon$. Equivalently, for any integer $r \ge 0$, there exists n_r such that $T_r G^\bullet_n \cong T_r G^\bullet_m$ for all $m, n \ge n_r$, which implies that

$$T_r G^\bullet_n \;\cong\; T_r G^\bullet_{n_r} \quad \text{for all } n \ge n_r.$$

Above we may select the threshold values so that $1 = n_0 < n_1 < \cdots$ Now define $H_r^\bullet = T_r G_{n_r}^\bullet$. Then $H_0^\bullet, H_1^\bullet, \dots$ form a compatible sequence in the sense that for $r \le s$,

$$T_r H_s^\bullet \;=\; T_r T_s G_{n_s}^\bullet \;=\; T_r G_{n_s}^\bullet \;\cong\; T_r G_{n_r}^\bullet \;=\; H_r^\bullet,$$

and by Lemma B.3 there exists a locally finite connected rooted graph $G^\bullet$ such that $T_r G^\bullet \cong H_r^\bullet$ for all r. Now

$$T_r G_n^\bullet \;\cong\; T_r G_{n_r}^\bullet \;=\; H_r^\bullet \;\cong\; T_r G^\bullet \quad \text{for all } n \ge n_r,$$

which implies that

$$d_{\text{loc}}([G_n^\bullet], [G^\bullet]) \;=\; d_{\text{loc}}(G_n^\bullet, G^\bullet) \;\le\; 2^{-r} \quad \text{for all } n \ge n_r.$$

Hence $[G_n^\bullet] \to [G^\bullet]$ in the topology induced by the local metric on $[\mathcal{G}^\bullet]$. □

B.4 Topological properties

We discuss topological properties of the metric space $([\mathcal{G}^\bullet], d_{\text{loc}})$, some of which differ quite a lot from what is common in usual topological spaces. We first note that this is a bounded metric space with diameter one, and $d_{\text{loc}}([F^\bullet], [G^\bullet]) = 1$ if and only if the 1-neighborhoods of the roots of $F^\bullet$ and $G^\bullet$ are nonisomorphic. Observe next that

$$d_{\text{loc}}([F^\bullet], [G^\bullet]) < \epsilon$$

if and only if $T_r F^\bullet \cong T_r G^\bullet$ for $r = \lfloor \log_2(1/\epsilon) \rfloor + 1$. Hence

$$[G_n^\bullet] \to [G^\bullet]$$

if and only if for all $r \ge 0$ there exists n_0 such that $T_r G_n^\bullet \cong T_r G^\bullet$ for all $n \ge n_0$.

The open ball of radius $\epsilon = 2^{-r}$ centered at $[F^\bullet]$ equals

$$B([F^\bullet], 2^{-r}) \;=\; \{[G^\bullet] \in [\mathcal{G}^\bullet] : T_{r+1} G^\bullet \cong T_{r+1} F^\bullet\}. \tag{B.3}$$

For example, all singleton sets corresponding to *finite* graphs are open, because the open ball of radius 2^{-r-1} around $G^\bullet \in \mathcal{G}^\bullet$ of diameter r equals $\{G^\bullet\}$. As a consequence the set $[\mathcal{G}_{<\infty}^\bullet]$ of all finite graphs is open, and so are all of its subsets. The local topology restricted to finite graphs is hence the discrete topology. Examples of open sets consisting of infinite rooted graphs include for example

- The set of all rooted graphs where the neighborhood of the root is a 3-clique.

About closed sets. Being a metric space, the space is Hausdorff. This implies that all singleton sets are closed.

We will next discuss compactness. We denote by

$$\Delta(G) \;=\; \sup_{v \in V(G)} \deg_G(v)$$

the *maximum degree* (possibly infinite) of a graph G, and this definition naturally extends to rooted graphs and unlabeled rooted graphs by setting $\Delta([G^\bullet]) = \Delta(G^\bullet) = \Delta(G)$. A set of called *relatively compact* if it has a compact closure.

Theorem B.6. *A set of unlabelled rooted graphs $[\mathcal{A}] \subset [\mathcal{G}^\bullet]$ is relatively compact if and only if*

$$\sup_{A^\bullet \in \mathcal{A}} \Delta(T_r A^\bullet) < \infty \quad \textit{for all } r \geq 0. \tag{B.4}$$

Proof. Because the metric space $([\mathcal{G}^\bullet], d_{\mathrm{loc}})$ is complete, we know that relative compactness is equivalent to total boundedness [Rud73, Theorem A4]). Recalling the shape of open balls (B.3), this means that $[\mathcal{A}]$ is relatively compact if and only if for all $r \geq 0$, the set $[\mathcal{A}]$ can be covered by finitely many open balls

$$\{[G^\bullet] : T_r G^\bullet \cong T_r F_i^\bullet\}, \quad i = 1, \dots, n,$$

with radii 2^{-r+1} and centers $[F_i^\bullet]$. Equivalently, every rooted graph $A^\bullet$ in $\mathcal{A}$ satisfies $T_r A^\bullet \cong T_r F_i^\bullet$ for some $i = 1, \dots, n$. This is further equivalent to saying that the set

$$[T_r \mathcal{A}] \;:=\; \{[T_r A^\bullet] : A^\bullet \in \mathcal{A}\}$$

is finite for every r.

We will now verify that $[T_r \mathcal{A}]$ is finite if and only if (B.4) holds. If $C_r = \sup_{A^\bullet \in \mathcal{A}} \Delta(T_r A^\bullet)$ is finite, then because any $A^\bullet \in \mathcal{A}$ is connected, the graph $T_r A^\bullet$ can have at most $1 + C_r + C_r^2 + \cdots C_r^r$ nodes, and hence $[T_r \mathcal{A}]$ must be finite. On the other hand, if C_r is infinite, then $\mathcal{A}$ contains a sequence of rooted graphs $A_1^\bullet, A_2^\bullet, \dots$ such that the cardinality of $V(T_r A_n^\bullet)$ converges to infinity as $n \to \infty$. Because graphs with differing node counts are nonisomorphic, it follows that $[T_r \mathcal{A}]$ then contains infinitely many elements. □

Example B.7 (Open relatively compact set)**.** Let $\mathcal{A} = \{T_r G^\bullet : r \geq 0\}$ where $G^\bullet$ is the infinite line graph with node set $V(G^\bullet) = \mathbb{N}$, link set $E(G^\bullet) = \{\{k, k+1\}, k \in \mathbb{N}\}$, and root $o(G^\bullet) = 1$. Then every graph in $\mathcal{A}$ has maximum degree 2, and therefore $[\mathcal{A}]$ is relatively compact. The set $[\mathcal{A}]$ is open because it is a union of open singletons $\{[T_r G^\bullet\}$, but not closed because $[T_r G^\bullet] \to [G^\bullet] \notin [\mathcal{A}]$.

Example B.8 (Compact set)**.** Let $[\mathcal{A}] \subset [\mathcal{G}^\bullet]$ be a set of unlabelled rooted graphs with maximum degree at most 100. Then $[\mathcal{A}]$ is relatively compact. This set is also closed because $[G^\bullet] \mapsto \Delta([G^\bullet])$ is continuous. Hence $[\mathcal{A}]$ is compact.

Example B.9 (Set which is not relatively compact)**.** Let $\mathcal{A} = \{K_{1,n}^\bullet : n \geq 1\}$ where $K_{1,n}^\bullet$ is a star graph with n leaves, rooted at the hub node. Then $\sup_{A^\bullet \in \mathcal{A}} \Delta(T_1 A^\bullet)$ is infinite because $T_1 K_{1,n}^\bullet = K_{1,n}^\bullet$ has maximum degree n. Hence the set $[\mathcal{A}]$ of finite unlabelled stars is not relatively compact.

B.5 Continuity

What about continuous functions from $[\mathcal{G}^\bullet]$ to $\mathbb{R}$? If $\phi([G^\bullet]) = \phi([T_r G^\bullet])$ depends on its input only via some r-neighborhood of the origin, then ϕ is continuous. The same is true for any function of the form

$$\phi([G^\bullet]) \;=\; \sum_{r=0}^{\infty} a_r \phi_r([T_r G^\bullet])$$

when $\sum_{r=0}^{\infty} |a_r| < \infty$. Intuitively, a function on $[\mathcal{G}^\bullet]$ is continuous if it depends on its input mainly by the structure near the origin, and the dependence on far-away structure vanishes the farther we look.

B.6 Random rooted graphs

A *random, unlabelled, locally finite, connected, rooted graph* is a random variable X in $[\mathcal{G}^\bullet]$ equipped with the Borel sigma-algebra induced by the local metric, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A *random, locally finite, connected, rooted graph* is a random variable X in $\mathcal{G}^\bullet$ equipped with the Borel sigma-algebra induced by the local pseudometric.

The distribution of X is the probability measure $\mathrm{Law}(X) = \mathbb{P} \circ X^{-1}$ on the Borel sets of $[\mathcal{G}^\bullet]$.

Proposition B.10. *Every random variable X in $[\mathcal{G}^\bullet]$ can be represented as $X = [Y]$ where Y is a random variable in $\mathcal{G}^\bullet$.*

Proof. Let $X : \Omega \to [\mathcal{G}^\bullet]$ be measurable. Then for any $\omega \in \Omega$, the realization $X(\omega)$ can be written as $X(\omega) = [Y(\omega)]$ for some $Y(\omega) \in \mathcal{G}^\bullet$. When we choose[2] one such $Y(\omega)$ for each ω, we obtain a function $Y : \Omega \to \mathcal{G}^\bullet$. We will next show that such Y is measurable. To do this, choose an arbitrary open ball $B(F^\bullet, \epsilon)$ in $\mathcal{G}^\bullet$. Because $d_{\mathrm{loc}}([F^\bullet], [G^\bullet]) = d_{\mathrm{loc}}(F^\bullet, G^\bullet)$ for all $F^\bullet, G^\bullet$, it follows that $G^\bullet \in B(F^\bullet, \epsilon)$ if and only if $[G^\bullet] \ \in B([F^\bullet], \epsilon)$. Especially,

$$\begin{aligned} \{\omega \in \Omega : Y(\omega) \in B(F^\bullet, \epsilon)\} &= \{\omega \in \Omega : [Y(\omega)] \in B([F^\bullet], \epsilon)\} \\ &= \{\omega \in \Omega : X(\omega) \in B([F^\bullet], \epsilon)\}. \end{aligned}$$

Because X is measurable, the above equality shows that the preimage of Y for any open ball in $\mathcal{G}^\bullet$ is a measurable set in Ω. Because the open balls generate the Borel sigma-algebra of $\mathcal{G}^\bullet$, it follows that Y is a measurable function. □

What about random variable $X = [G^\bullet]$? We would like to say that every random variable X can be represented as $[G^\bullet]$ where $G^\bullet$ is a random variable in the space (with sigma-algebra induced by the pseudometric d_{loc}) of labeled graphs $\mathcal{G}^\bullet$ with node set being a subset of $\mathbb{N}$.

Let us think of $(\mathcal{G}^\bullet, d_{\mathrm{loc}})$ as a pseudometric space with topology induced by the pseudometric d_{loc}. This is a separable topological space: the set $\mathcal{G}^\bullet_{<\infty}$ of finite rooted graphs with node set in $\mathbb{N}$ is countable, and it is also dense by the same argument that was done earlier. We equip $\mathcal{G}^\bullet$ with the Borel sigma-algebra induced by the local topology on $\mathcal{G}^\bullet$. A random variable in $\mathcal{G}^\bullet$ is a measurable function from a probability space into the $\mathcal{G}^\bullet$. If $G^\bullet$ is a random variable in $\mathcal{G}^\bullet$, then $[G^\bullet] = i \circ G^\bullet$ is a random variable in $[\mathcal{G}^\bullet]$ because the quotient map i is continuous and hence Borel measurable.

Proposition B.11. $\mathrm{Law}(X) = \mathrm{Law}(Y)$ *if and only if ...*

B.7 Convergence in distribution

Theorem B.12. *Let X_n, X be unlabelled random rooted graphs in $[\mathcal{G}^\bullet]$. Then $X_n \to X$ in distribution iff for any corresponding random variables $\mathbb{P}([T_r X_n^\bullet] = [G^\bullet]) \to \mathbb{P}([T_r X^\bullet] = [G^\bullet])$ for all $F^\bullet \in \mathcal{G}^\bullet$ and all integers $r \geq 0$. This is also equivalent to $[T_r X_n^\bullet] \to [T_r X^\bullet]$ in distribution, for every r.*

[2] We do not need the axiom of choice if take $Y(\omega)$ as the element of $X(\omega)$ with smallest label, and keeping in mind that we restrict $[\mathcal{G}^\bullet]$ to be the equivalence classes of a rooted graph with a node set contained in $\mathbb{N}$.

Proof. This argument follows Curien 2018 lecture notes, Proposition 4. Let $\mathcal{S}$ be the collection of sets of the form $S_{\mathcal{A},r} = \{[G^\bullet] : [T_r G^\bullet] \in \mathcal{A}\}$, with $r \geq 0$ and $\mathcal{A} \subset [\mathcal{G}^\bullet]$ being a Borel set. Note that

$$\begin{aligned} \mu_n(S_{\mathcal{A},r}) &= \mu_n([G^\bullet] : [T_r G^\bullet] \in \mathcal{A}) = \mathbb{P}([T_r X_n^\bullet] \in \mathcal{A}_r) \\ &= \sum_{[G^\bullet] \in \mathcal{A}_r} \mathbb{P}([X_n^\bullet] = [G^\bullet]), \end{aligned}$$

where $\mathcal{A}_r$ denotes the unlabelled rooted graphs in $\mathcal{A}$ of radius at most r. Because $\mathcal{A}_r$ is countable, a standard argument implies that $\mu_n(S_{\mathcal{A},r}) \to \mu(S_{\mathcal{A},r})$ for all $S_{\mathcal{A},r}$ in $\mathcal{S}$. The claim then follows from [Bil99, Theorem 2.2] after verifying that $\mathcal{S}$ is a π-system and that all open sets in $[\mathcal{G}^\bullet]$ are countable unions of sets in $\mathcal{S}$.

Note than any open ball in $[\mathcal{G}^\bullet]$ can be written in the form

$$B([F^\bullet], 2^{-r}) = \{[G^\bullet] : T_r G^\bullet \cong T_r F^\bullet\} = S_{\mathcal{A},r}$$

with $\mathcal{A} = \{[T_r F^\bullet]\}$. Furthermore, because $[\mathcal{G}^\bullet]$ is separable, it follows that every open set in $[\mathcal{G}^\bullet]$ can be written as a countable union of open balls. Hence every open set in $[\mathcal{G}^\bullet]$ can be written as a countable union of sets in $\mathcal{S}$. Next, observe that for $r \leq s$,

$$S_{\mathcal{A},r} \cap S_{\mathcal{B},s} = \{[G^\bullet] : [T_r G^\bullet] \in \mathcal{A}, [T_s G^\bullet] \in \mathcal{B}\} = S_{\mathcal{C},s},$$

where $\mathcal{C} = \{[G^\bullet] : [T_r G^\bullet] \in \mathcal{A}, [T_s G^\bullet] \in \mathcal{B}\}$. Hence $\mathcal{S}$ is a π-system. □

Bibliography

[AFT$^+$18] Avanti Athreya, Donniell E. Fishkind, Minh Tang, Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, Keith Levin, Vince Lyzinski, Yichen Qin, and Daniel L Sussman. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018.

[BCL11] Peter J. Bickel, Aiyou Chen, and Elizaveta Levina. The method of moments and degree distributions for network models. *Ann. Statist.*, 39(5):2280–2301, 2011.

[Bil99] Patrick Billingsley. *Convergence of Probability Measures.* Wiley, second edition, 1999.

[BJR07] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Struct. Algor.*, 31(1):3–122, 2007.

[Bol81] Béla Bollobás. Threshold functions for small subgraphs. *Mathematical Proceedings of the Cambridge Philosophical Society*, 90(2):197–206, 1981.

[Bol01] Béla Bollobás. *Random Graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics.* Cambridge University Press, second edition, 2001.

[BP94] Itai Benjamini and Yuval Peres. Markov chains indexed by trees. *Ann. Probab.*, 1994.

[BS01] Itai Benjamini and Oded Schramm. Recurrence of distributional limits of finite planar graphs. *Electron. J. Probab.*, 6:13 pp., 2001.

[Die17] Reinhard Diestel. *Graph theory.* Springer, fifth edition, 2017.

[ER59] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[Fel71] William Feller. *An Introduction to Probability Theory and Its Applications Vol 2.* Wiley, 1971.

[Gil59] Edgar Nelson Gilbert. Random graphs. *Ann. Math. Statist.*, 30:1141–1144, 1959.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[Jan10] Svante Janson. Asymptotic equivalence and contiquity of some random graphs. *Random Structures & Algorithms*, 2010.

[Kal02] Olav Kallenberg. *Foundations of Modern Probability.* Springer, second edition, 2002.

[Kel75] John L. Kelley. *General Topology.* Springer, 1975.

[Lov12] László Lovász. *Large Networks and Graph Limits.* American Mathematical Society, 2012.

[LP17] David A. Levin and Yuval Peres. Counting walks and graph homomorphisms via markov chains and importance sampling. *Amer. Math. Monthly*, 2017.

[MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 2015.

[MS12] D. Marcus and Y. Shavitt. RAGE – A rapid graphlet enumerator for large networks. *Computer Networks*, 56(2):810–819, 2012.

[Pre74] Christopher J. Preston. A generalization of the FKG inequalities. *Comm. Math. Phys.*, 36:233–241, 1974.

[Rud73] Walter Rudin. *Functional Analysis.* McGraw–Hill, 1973.

[RV86] Andrzej Ruciński and Andrew Vince. Strongly balanced graphs and random graphs. *Journal of Graph Theory*, 10(2):251–264, 1986.

[Sid94] Alexander Sidorenko. A partially ordered set of functionals corresponding to graphs. *Discrete Math.*, 131(263–277), 1994.

[Str65] Volker Strassen. The existence of probability measures with given marginals. *Ann. Math. Statist.*, 36(2):423–439, 1965.

[vdH17] Remco van der Hofstad. *Random Graphs and Complex Networks - Vol. I.* Cambridge University Press, 2017.

[vdH18] Remco van der Hofstad. Random graphs and complex networks - Vol. II, 2018.

[WG75] Henry William Watson and Francis Galton. On the probability of the extinction of families. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 4:138–144, 1875.