Chapter 4

Generalizations of the Erdős–Rényi random graphs

Erdős–Rényi random graphs have two incarnations: The first one that we studied in much details is $\mathcal{G}(n,p)$ when the probability of each edge is specified. And the second one is $\mathcal{G}(n,m)$ when a fixed number of edges m is distributed throughout the graph (this we almost did not discuss). The two models have very similar properties in the case $m = \binom{n}{2}p$, which can be rigorously proved. These two models are also important for understanding of two main generalizations of the Erdős–Rényi random graph: inhomogeneous random graphs and configuration model. In the former case the weights $\boldsymbol{w}$ for each vertex are specified, and the probability that an edge connects vertices i and j is defined using the weights w_i and w_j corresponding to the vertices i and j. This model, which we denote $\mathcal{G}(n,\boldsymbol{w})$, is a generalization of $\mathcal{G}(n,p)$. The second model, which is a generalization of $\mathcal{G}(n,m)$, is the configuration model, when the degrees of each vertex are defined. This will be denoted $\mathcal{G}(n,\boldsymbol{d})$, where $\boldsymbol{d} = (d_1,\dots,d_n)$ are a graphical sequence of the vertex degrees. It turns out that these two models have similar properties when $\boldsymbol{w}$ and $\boldsymbol{d}$ are related, and actually it is enough to carefully analyze one model to transfer the results onto the other one. Both of these models are capable of producing random graphs with predetermined degree distribution, that is why they generalize the Erdős–Rényi random graphs, they both are small worlds, however, there are two main drawbacks of these models. First, the degree distribution is given as the model parameter and does not appear as an evolved property of the system. A different approach is to evolve a network, and the preferential attachment model allows exactly this: to produce power law distribution starting with quite plausible evolutionary graph process. Second, being "very random", these models (including the preferential attachment model) do not allow for additional network structure, in particular their clustering coefficients approach zero as $n \to \infty$. To tackle this obstacle I will finish this chapter by analyzing a mathematical model of what is conventionally called a small world graph.

4.1 Inhomogeneous random graphs

The first model I consider is the so-called the inhomogeneous random graph model, which is specified by the number of vertices n and the sequence $\boldsymbol{w} = (w_1,\dots,w_n)$ of weights of every vertex. I will denote this model as $\mathcal{G}(n,\boldsymbol{w})$. To be able to prescribe an arbitrary degree distribution to $\mathcal{G}(n.\boldsymbol{w})$, I set the probability that vertices i and j are connected as

$$p_{ij} = \frac{w_i w_j}{\sum_k w_k + w_i w_j}\,.$$

I can always assume that $w_i > 0$ for all i since otherwise the vertex i is isolated and can be disregarded.

Problem 4.1. Show that if $w_i = n\lambda/(n-\lambda)$ then $\mathcal{G}(n,\boldsymbol{w})$ becomes $\mathcal{G}(n,\lambda/n)$.

Problem 4.2. Consider an inhomogeneous random graph with n_1 vertices with the weights m_1 and n_2 vertices with the weights m_2. Show that the expected degree of the former is approximately m_1 and of the latter is approximately m_2 if $m_1^2 + m_2^2 = o(\sum_k w_k)$.

Problem 4.3. In general, if $w_i w_j = o(\sum_k w_k)$ for any i,j then the expected degree of vertex i is

$$\mathsf{E}D_i \approx w_i.$$

Note that the previous problem tells only about the expected degree of a given vertex. It is possible to prove[1] that the degree distribution of a particular picked vertex i is close to a Poisson one with parameter w_i. This implies that it is reasonable to expect that the distribution of degree sequence of $\mathcal{G}(n,\boldsymbol{w})$ is close to a mixed Poisson distribution, where the mixing function is given by the a probability distribution function for the random variable W, which is defined in terms on the sequence $\boldsymbol{w}$.

Recall that a random variable X has a *mixed Poisson distribution* with mixing distribution F when, for every $k \in \mathbf{N}$

$$\mathsf{P}(X=k) = \mathsf{E}\left(e^{-W}\frac{W^k}{k!}\right),$$

[1] for all the details see Remco van der Hofstad, Random Graphs and Complex Networks, Vol. I, Chapter 6

where W is a random variable with distribution function F. In terms of $\boldsymbol{w}$ can be defined as

$$F(x) = \mathsf{P}(X \le x) = \frac{1}{n} \sum_{i \in [n]} \mathbf{1}_{\{w_i \le x\}}.$$

There exist a lot of various models that are defined in terms of weights of the vertices. It can be shown that they are equivalent (in some rigorous sense) when $n \to \infty$.

Chung–Lu model. In this model the probabilities are taken as

$$p_{ij} = \frac{w_i w_j}{\sum_k w_k} \wedge 1,$$

where $a \wedge b$ means $\min\{a, b\}$. This model is equivalent to $\mathcal{G}(n, \boldsymbol{w})$ is $\sum_k w_k^3 = o(n^{3/2})$.

Norris–Reittu model of the Poisson graph process. This model is defined as a random graph process, where at each time t a new vertex is born with the weight w_t, and it is connected to any existing vertex i by the number of edges that have Poisson distribution with the parameter $w_i w_t / \sum_k w_k$. Furthermore, at each time each of the older edges is erased with probability $w_t / \sum_k w_k$. It turns out that the number of edges between i and j is a Poisson random variable with parameter $w_i w_j / \sum_k w_k$. Note that we essentially deal with a multigraph in this case. However, if the weights are bounded then the probability that the resulting graph is simple is positive.

Problem 4.4. Calculate this probability.

If one erases all self-loops in the Norris–Reittu graph and merges multiple edges into one, then the resulting random graph is equivalent to $\mathcal{G}(n, \boldsymbol{w})$.

4.2 Configuration model

Our goal in this section is to formulate a model that produces a *uniform* random graph with prescribed degree distribution.

4.2.1 Definition. Basic properties

Assume that the vector $\boldsymbol{d} = (d_1, \dots, d_n)$ is *graphical*, i.e., there exits a graph on n vertices such that vertex 1 has degree d_1, vertex 2 has degree d_2, and so on. We would like to generate a random graph having exactly prescribed degree sequence $\boldsymbol{d}$. To accomplish this goal we consider $2m$ half-edges, $2m = \sum_{i \in [n]} d_i$, where $[n] = \{1, 2, \dots, n\}$, of our graph and perform a random matching of these half-edges. I will denote the resulting model $\mathcal{G}(n, \boldsymbol{d})$.

Problem 4.5. Show that among $2m$ half-edges there are

$$(2m-1)!! = (2m-1)(2m-3)\dots 3 \cdot 1$$

possible matchings.

If we perform uniform pairing of half-edges, then we obtain a random graph with exactly the degree sequence $\boldsymbol{d}$, which is a realization of the *configuration model* $\mathcal{G}(n, \boldsymbol{d})$ *with degree sequence* $\boldsymbol{d}$.

Here are some basic properties of the configuration model.

- The outcomes of the configuration model are multigraphs, since self-loops and multiple edges are possible while performing random matching of the half-edges.

- The name *configuration model* comes from the construction of another graph: Assume that we have $2m$ vertices and perform uniform pairing of them, the result is called configuration.

- We can identify a graph $G \in \mathcal{G}(n, \boldsymbol{d})$ with the matrix $(x_{ij})_{i,j\in[n]}$, where x_{ij} is the number of edges connecting vertices i and j, and x_{ii} is the number of self-loops of the vertex i. We have

$$d_i = x_{ii} + \sum_{j\in[n]} x_{ij},$$

  where x_{ii} is counted twice, since each self-loop adds two to the vertex degree.

- The probability that we observe the multigraph $G = (x_{ij})_{i,j\in[n]}$ is

$$\mathsf{P}(\{G \in \mathcal{G}(n, \boldsymbol{d})\}) = \frac{1}{(2m-1)!!} \frac{\prod_i d_i!}{\prod_i 2^{x_{ii}} \prod_{1\le i<j\le n} x_{ij}!},$$

  where we have to account the facts that random permutations of the half-edges adjacent to some vertices produce the same graphs, and permutations of half-edges that contribute to self-loops and multiple edges should not be counted, since these are the same matchings.

  This formula shows that the probability distribution on the configuration model is not uniform (it is outcome dependent). However, the same formula shows that conditioned on the event that G is simple, we obtain a uniform distribution.

- It is usually more convenient to deal with the sequence n_k, which is the number of vertices of degree k, then with $\boldsymbol{d}$. We have $p_k = \frac{n_k}{n}$ is the degree distribution of our the configuration model. Having $(p_k)_{k=0}^{\infty}$ we can speak of a "random variable" D (actually, there is no random variable, the sequence $\boldsymbol{d}$ is deterministic) which takes its values on $\mathcal{G}(n, \boldsymbol{d})$. Note that

$$2m = \sum_i d_i = \sum_k k n_k = n \sum_k k p_k = n\, \mathsf{E}D,$$

  where $\mathsf{E}D$ is the average degree. Similarly, we can define $\mathsf{E}D^2 = \sum_k k^2 p_k = \sum_i \frac{d_i^2}{n}$. Usually we assume that the degree distribution is such that both $\mathsf{E}D$ and $\mathsf{E}D^2$ are defined and finite.

The fact that outcomes of the configuration model are multigraphs is not very disappointing because it can be fixed in two ways. First, after a random graph is produced, self-loops and multiple edges can be erased. The important fact is that the degree distribution of the obtained simple graph converges to the fixed at the beginning degree distribution of D. We show this by starting with the following proposition.

Proposition 4.1. *Let M_n, S_n be the random variables that denote the number of multiple edges and self-loops in $\mathcal{G}(n, \boldsymbol{d})$ respectively. Then*

$$\mathsf{E}S_n \sim \frac{\nu}{2}, \quad \mathsf{E}M_n \le \frac{\nu^2}{4},$$

where

$$\nu = \frac{\mathsf{E}D^2 - \mathsf{E}D}{\mathsf{E}D} = \frac{\mathsf{E}(D)_2}{\mathsf{E}D}.$$

Proof. Consider the events

$$\tau_{st}^i = \{ \text{ half-edge } s \text{ is paired with half-edge } t, \text{ both belonging to vertex } i \ \}.$$

Then

$$S_n = \sum_{i\in[n]} \sum_{1\le s\le t\le d_i} \mathbf{1}_{\tau_{st}^i},$$

which implies

$$\begin{aligned}\mathsf{E}S_n &= \sum_{i\in[n]}\sum_{1\le s\le t\le d_i}\mathsf{P}(\tau^i_{st}) = \frac{1}{2}\sum_{i\in[n]} d_i(d_i-1)\,\mathsf{P}(\tau^i_{12})\\ &= \frac{1}{2}\sum_{i\in[n]}\frac{d_i(d_i-1)}{2m-1} \approx \frac{1}{2}\sum_{i\in[n]}\frac{d_i(d_i-1)}{2m}\\ &= \frac{\nu}{2}\,.\end{aligned}$$

Similarly, for the events

$$\tau^{ij}_{s_1t_1,s_2t_2} = \{\ s_1 \text{ is paired with } t_1,\ s_2 \text{ is paired with } t_2,\ s_k \in \text{vertex } i,\ t_k \in \text{vertex } j\ \},$$

one has (the factor $1/2$ in order not to count twice the edges from i to j and from j to i)

$$M_n \le \frac{1}{2}\sum_{1\le i\ne j\le j}\ \sum_{1\le s_1\le s_2\le d_i}\ \sum_{1\le t_1\ne t_2\le d_j} \mathbf{1}_{\tau^{ij}_{s_1t_1,s_2t_2}}.$$

Here we have $\le$ because the right hand side gives overestimate when, e.g., there are exactly three edges between vertices i and j. Hence,

$$\begin{aligned}\mathsf{E}M_n &\le \frac{1}{2}\sum_{i,j\in[n]}\ \sum_{1\le s_1\le s_2\le d_i}\ \sum_{1\le t_1\ne t_2\le d_j}\mathsf{P}(\tau^{ij}_{s_1t_1,s_2t_2})\\ &= \frac{1}{4}\sum_{i,j\in[n]}\frac{d_i(d_i-1)d_j(d_j-1)}{(2m-1)(2m-3)} \approx\\ &\approx \frac{\nu^2}{4}\,.\end{aligned}$$

■

As a simple corollary of the last proposition we have that, provided ν is fixed and $n\to\infty$, the proportion of the self-loops and multiple edges approaches zero. This can be used to prove the following important theorem:

Theorem 4.2. *Let $P_k^{(er)}$ be the proportion of vertices of degree k in the configuration model after self-loops and multiple edges are erased, and p_k be the degree distribution of $D = d_U$, where U is a uniform random variable. Then for any $\epsilon > 0$*

$$\mathsf{P}\Big(\sum_{k=1}^{\infty}|P_k^{(er)} - p_k| \ge \epsilon\Big) \to 0.$$

Moreover, it can be proved, using the method of moments, that the pair of the random variables (S_n, M_n) converges to two independent random variables with the Poisson distribution with the parameters $\nu/2$ and $\nu^2/4$ respectively. This fact has a remarkable corollary that

Corollary 4.3. *Probability that there are no self loops and multiple edges in a realization of the configuration model is given by*

$$\mathsf{P}(\{\mathcal{G}(n,\boldsymbol{d}) \text{ is simple}\}) = e^{-\nu/2-\nu^2/4}(1+o(1)).$$

Therefore, the number of simple graphs in $\mathcal{G}(n,\boldsymbol{d})$ can be approximated as

$$e^{-\nu/2-\nu^2/4}\frac{(2m-1)!!}{\prod_i d_i!}(1+o(1)).$$

Problem 4.6. Use the last formula to approximate the number of r-regular simple graphs.

4.2.2 Excess degree distribution. Generating functions for the configuration model. Diameter and the size of the giant component

Let $(p_k)_{k=0}^{\infty}$ be the degree distribution of the random variable D defined on $\mathcal{G}(n, \boldsymbol{d})$. For the following we assume that two first moments $\mathsf{E}D$ and $\mathsf{E}D^2$ are finite. The probabilities p_k answer the question "What is the probability that randomly picked vertex has degree k?" Now let us ask another question: Assume that we randomly pick a vertex and assume also that this vertex has degree bigger than zero. We choose any of the edges incident to this vertex and along this edge we approach another vertex, a neighbor of the initially chosen node. Here is the question: What is the probability that this randomly chosen neighbor has degree k? After a thought it should be clear that these probabilities has to be different from $(p_k)_{k\geq 0}$, because, for instance, we already know for sure that our vertex has degree at least one, and hence any non-zero p_0 will be at odds with this fact. To answer this question we note that there are k half-edges along which we can approach a neighbor of degree k, the total number of half-edges is $2m-1$, and the total number of such neighbors in the network is n_k, hence

$$\frac{kn_k}{2m-1} \approx \frac{nkp_k}{2m} = \frac{kp_k}{\mathsf{E}D}$$

is the probability that a randomly chosen neighbor of a randomly chosen vertex in the configuration model has degree k. Note that it sums, how it should, to one, if you go from $k=1$ to ∞. Let us see a not very obvious aspect of this distribution: To find the expectation of this distribution, we need to evaluate

$$\sum_{k=1}^{\infty} k\frac{kp_k}{\mathsf{E}D} = \frac{\mathsf{E}D^2}{\mathsf{E}D} \geq \mathsf{E}D,$$

where the last inequality follows, for instance, from $0 \leq \operatorname{Var} D = \mathsf{E}D^2 - (\mathsf{E}D)^2$. This proves that on average a neighbor has higher degree: "Your friend has more friends than you." This may look counterintuitive, but actually is the consequence of the fact that it is much higher chance to pick an edge that ends up at a vertex of high degree than to pick up a lonely edge of a vertex degree 1, and of the set up of the configuration model in which all pairing are uniformly random (this is not true for real world networks, where we should often expect some correlations between vertices).

Now consider degree of a randomly chosen neighbor of a randomly picked vertex minus one (i.e., we do not count the edge along which we approach our neighbor). This is called *the excess degree*. The probabilities that excess degree is exactly k will be denoted q_k. We have, from the previous,

$$q_k = \frac{(k+1)p_{k+1}}{\mathsf{E}D},$$

because we do not count one of the half-edges. $(q_k)_{k=0}^{\infty}$ is called the *excess degree distribution* or, sometimes, *size-biased distribution.*

Define the generating functions of the degree distribution and excess degree distribution:

$$\varphi_0(s) = \sum_{k=0}^{\infty} p_k s^k,$$
$$\varphi_1(s) = \sum_{k=0}^{\infty} q_k s^k.$$

First note that if we know $\varphi_0(s)$ then we also know $\varphi_1(s)$:

$$\varphi_1(s) = \sum_{k=0}^{\infty} \frac{(k+1)p_k}{\mathsf{E}D} s^k = \frac{1}{\mathsf{E}D} \sum_{k=1}^{\infty} kp_k s^{k-1} = \frac{\varphi_0'(s)}{\varphi_0'(1)},$$

since $\mathsf{E}D = \varphi_0'(1)$.

Problem 4.7. Show that $\varphi_0(s) \equiv \varphi_1(s)$ if and only if $D \sim \text{Poisson}\,(\lambda)$.

Now, using the generating functions, we can calculate probabilities that there are k neighbors at the distance l. Let us start with $l = 2$, i.e., how many neighbors at the distance 2 a randomly chosen vertex in the configuration model has. Denote

$$p_k^{(2)} = \mathsf{P}(\{\text{a vertex has } k \text{ second neighbors}\}).$$

We have the generating function for $p_k^{(2)}$:

$$\varphi^{(2)}(s) = \sum_{k=0}^{\infty} p_k^{(2)} s^k = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} p_j \,\mathsf{P}(k \mid j) s^k,$$

where $\mathsf{P}(k \mid j)$ is the conditional probability that there are k second neighbors provided that there are j first neighbors (with probability p_j). Further, using the property that the generating function of a sum of independent random variables is equal to the product of the generating functions, we obtain

$$\varphi^{(2)}(s) = \sum_{j=0}^{\infty} p_j \sum_{k=0}^{\infty} \mathsf{P}(k \mid j) s^k = \sum_{j=0}^{\infty} p_j (\varphi_1(s))^j = \varphi_0(\varphi_1(s)).$$

Here I have used the fact that for each first neighbor the number of second neighbors is defined by the excess degree distribution with the generating function $\varphi_1(s)$, and there are exactly j of these neighbors, hence the total number of the second neighbors is defined by the sum of random variables with excess degree distribution. Here you can also see what kind of mistake we make in these calculations: The second neighbors can be counted several times depending on to how many first neighbors they are adjacent. For large n and short distances however, this error is negligible.

Similarly, we can find (fill in the details) that

$$\varphi^{(l)}(s) = \varphi^{(l-1)}(\varphi_1(s)).$$

Let us find the average number of neighbors at the distance l, which I denote ρ_l. We have

$$\frac{\mathrm{d}}{\mathrm{d}s}\varphi^{(l)}(s)\bigg|_{s=1} = \frac{\mathrm{d}}{\mathrm{d}s}\varphi^{(l-1)}(\varphi_1(s))|_{s=1}\varphi_1'(s)|_{s=1},$$

or

$$\rho_l = \rho_{l-1}\varphi_1'(1) = \rho_{l-1}\frac{\mathsf{E}D^2 - \mathsf{E}D}{\mathsf{E}D}\,.$$

Plus we would need the initial condition $\rho_1 = \mathsf{E}D$. Finally, denoting $\nu = (\mathsf{E}D^2 - \mathsf{E}D)/\,\mathsf{E}D$,

$$\rho_l = \nu^{l-1}\,\mathsf{E}D.$$

We have a very important conclusion: The number of neighbors at the distance l is growing if and only if $\nu > 1$ and decreasing if $\nu < 1$. This is actually the condition for the appearance of the giant component:

$$\nu > 1 \implies \mathsf{E}D^2 - 2\,\mathsf{E}D > 0.$$

Moreover, we cannot have more neighbors than the total number of vertices n. This suggests that the diameter of the configuration model is given by the expression

$$\nu^l = n \implies l = \frac{\log n}{\log \nu}\,.$$

To find the size of the giant component, let us use again heuristic derivation similar to the one used when we discussed the Erdős–Rényi random graph. A randomly picked vertex does not belong to

the giant component if and only if none of its neighbors belong to the giant component. Define u as the probability that a vertex is not connected to the giant component through its neighbors. If this vertex has k neighbors, then u^k is the probability that it is not connected to the giant component, and $\sum_k p_k u^k = \varphi_0(u)$ is the probability that this vertex does not belong to the giant component, hence

$$v = 1 - \varphi_0(u)$$

is the equation for the fraction of the nodes that are in the giant component. To find u let us reason in a similar way. The probability that a vertex is not connected to the giant component through any of its neighbors equal the average of the probabilities that none of the neighbors connected to the giant component through their neighbors, or $u = \sum_k q_k u^k = \varphi_1(u)$. Therefore, finally we have

$$v = 1 - \varphi_0(u), \quad u = \varphi_1(u). \tag{4.1}$$

Problem 4.8. Show that the last equation has a positive root $0 < u < 1$ if and only if

$$\nu = \frac{\mathsf{E}D^2 - \mathsf{E}D}{\mathsf{E}D} > 1.$$

Problem 4.9. Assume that we have a network whose vertices have degrees only $0, 1, 2, 3$ with probabilities p_0, p_1, p_2, p_3. Show that $u = p_1/(3p_3)$ is the solution to $u = \varphi_1(u)$ when the giant component exists. Show that the size v of the giant component actually depends on all the probabilities p_i and find explicit solution.

4.2.3 Configuration model with the power law distribution

As an illustration of the obtained results consider the degree sequence $\boldsymbol{d}$, which follows the discrete power law:

$$p_k = Ck^{-\alpha}, \quad k \geq 1,$$

where C is a normalization constant, which can be calculated as

$$C = \frac{1}{\zeta(\alpha)}, \quad \zeta(\alpha) = \sum_{k=1}^{\infty} k^{-\alpha}.$$

We calculate

$$\mathsf{E}D = \sum_{k=1}^{\infty} kp_k = \frac{\zeta(\alpha - 1)}{\zeta(\alpha)},$$

and

$$\mathsf{E}D^2 = \sum_{k=1}^{\infty} k^2 p_k = \frac{\zeta(\alpha - 2)}{\zeta(\alpha)}.$$

Therefore, the condition for the existence of the giant component becomes

$$\zeta(\alpha - 2) > 2\zeta(\alpha - 1),$$

which becomes true only if $\alpha < 3.4788$. This result will change if one considered any power law distribution with behavior that is different from the power law for small values of k, however, we still can get a general result.

Recall that if $2 < \alpha \leq 3$ then $\mathsf{E}D^2$ does not exits (the corresponding series diverges), hence the condition

$$\mathsf{E}(D)_2 > \mathsf{E}D.$$

is true for any power law distribution, therefore there exists the giant component.

If the specific form of the power law is given then we can calculate the size of the giant component. For the pure power law one has

$$\varphi_0(s) = \frac{\mathrm{Li}_\alpha(s)}{\zeta(s)}\,, \quad \varphi_1(s) = \frac{\mathrm{Li}_{\alpha-1}(s)}{s\zeta(s)}\,,$$

where

$$\mathrm{Li}_\alpha(s) = \sum_{k=1}^{\infty} k^{-\alpha} s^k$$

is the polylogarithm function. Now we have

$$u = \varphi_1(u) \implies u = 0$$

if $\alpha \le 2$ since $\zeta(\alpha - 1)$ diverges in this case. Therefore, $v = 1 - \varphi_0(0) = 1 - p_0 = 1$ because we assume that $p_0 = 0$ (there are no isolated vertices), and hence the giant component coincides with the whole network whp.

4.2.4 Relation of $\mathcal{G}(n, \boldsymbol{w})$ and $\mathcal{G}(n, \boldsymbol{d})$

`to be added`

Problem 4.10. Prove that in the configuration model $\mathcal{G}(n, \boldsymbol{d})$ the global clustering coefficient tends to zero as $n \to \infty$.

Problem 4.11. Assume that we have a network whose vertices have degrees only 0,1,2,3 with probabilities p_0, p_1, p_2, p_3. Show that the probability u that a randomly picked vertex does not belong to the giant component of the configuration model is given by $u = p_1/(3p_3)$, when the giant component exists. Find the size of the giant component in this case.

Problem 4.12. Consider the configuration model with the exponential degree distribution

$$p_k = (1 - e^{-\lambda})e^{-\lambda k}$$

with $\lambda > 0$.

Find the condition for the giant component to exist and the size of the giant component.

Problem 4.13. Can you give any heuristic arguments that the configuration model is a small world?