CS294-204 Phase Transitions (Fall 2021)
Homework #3
Due Fr. 10/8

Superspreaders

In this exercise, we will see that for graphs with degree inhomogeneities, the average degree is not the right quantity to consider when calculating R_0.

To this end, we will study the so-called Chung-Lu random graph, defined in terms of a sequence $\mathbf{w}$ of n positive numbers, $\mathbf{w} = (w_1, \ldots, w_n)$ (called weights). We then connect vertex i and j with probability $p_{ij} = \frac{w_i w_j}{\ell_n}$, where $\ell_n = \sum w_i$. We assume that $\max_i w_i \leq \sqrt{\ell_n}$, to make sure this is a well defined probability. We furthermore assume that $c_1 n \leq \ell_n \leq c_2 n$ for some constants $0 < c_1 \leq c_2 < \infty$; as we will see below, this corresponds to a graph with average expected degree of order 1, which we take to be a natural assumption to model real world social networks.

We use

$$\rho(w) = \frac{n_w}{n} \quad \text{where} \quad n_w = |\{in \in [n] \colon w_i = w\}|$$

to denote the empirical distribution of the weights w_i, $i \in [n]$.

1 Overall Structure of this Exercise

This assignment will be organized as follows: Sections 2, 3 and 5 will contain exercises I will walk you through with many hints (so don't be scared of the length), and Section 4 will be one without exercises, in which I "derive" a certain SIR-process to be studied in Section 5.

Specifically, in a preliminary subsection, Section 2, I will ask you to establish a well known fact from statistics we will need later, namely, the fact that N independent $Be(p)$ trials on top of a Poisson number $N \sim Poi(c)$ of coins gives you a number of heads which is again Poisson, with scaled parameter pc instead of c.

Next, in Section 3, you will analyze the degree distribution of a fixed vertex i with given weight w_i in the Chung-Lu graph, and show that it is approximately $Poi(w_i)$; you will also show that the weights of the neighbor of any given vertex are not given by the empirical distribution ρ, but by a modified distribution ρ^*, see (1) below.

In Section 4, I will use what you established in Section 3 to "derive" a branching process describing an infection tree on the Chung-Lu graph (I won't ask you to derive anything, just to try to follow my derivation). As we will see, the random tree describing (the early stages of) an infection on the Chung-Lu graph turns out to be a two stage birth process, where the root has the degree of a random vertex in the graph, some of which will then be infected. But after the root, the process looks different: instead of starting with a number of children which is just the degree of a random vertex, we will have to take into account that the degree distribution of an infected vertex tends to be skewed towards higher degrees. This will lead to two different off-spring distributions describing the number of infected children: a random off-spring X for the root, and a modified random off-spring X^* for anyone below the root.

In Section 5, you will analyze this two stage branching process, and will show that the infection threshold is not given by the naive one involving the average degree of a random vertex, but rather by a modified one, involving the average degree of an infected vertex.

Remark: *I have written the problem set in such a way that you don't really need to do the problems from Section 3 to solve the rest - all you need are the results, which are summarized in Section 4. I suggest you still solve the problems from Section 2 and the first two or three problems from Section 4, before going to Section 5. But you at first could skip Problem 4, which shows that the degree distribution of vertex i is approximately $Poi(w_i)$, and Problem 5, which shows that the distribution of the weights w of the neighbors of any vertex is given by the re-weighted distribution defined in* (1)*, and just accept these results to solve the remaining problems.*

2 Bernoulli trials on top of Poisson [2 Points]

Consider a random variable Y obtained by first choosing $N \sim Poi(c)$ and then $Y \sim Bin(N, p)$. Prove that $Y \sim Poi(cp)$. To do this, all you have to do is to write out the probability that $Y = k$ and simplify the resulting formula to obtain that this is equal to $\frac{1}{k!}(pc)^k e^{-pc}$.

3 Degree distribution of a fixed vertex and its neighbors [10 Points]

In a first step, we will analyze the degree distribution of a fixed vertex i with given weight w_i, the degree distribution of its neighbors, and the degree distribution of a random vertex in $[n]$. Recall that we assumed that $w_i \leq \sqrt{\ell_n}$ and $c_1 n \leq \ell_n \leq c_2 n$, which in particular shows that w_j is much smaller than $\ell_n = \sum_j w_j$, allowing us to replace a sum over w_j which goes over all but a finite number of terms by the full sum $\sum_j w_j = \ell_n$. In addition, we will assume that $p_{ij} = \frac{w_i w_j}{\ell_n}$ is small enough to justify replacing terms like $1 - p_{ij}$ by $e^{-p_{ij}}$.

1. Calculate the expected degree of vertex i by writing the degree d_i as a sum of $n-1$ random variables X_{ij} and taking expectations. How much does the result differ from w_i? [1 Point]

2. Calculate the expected number of edges, and verify that the condition $c_1 n \leq \ell_n \leq c_2 n$ does indeed imply that the average degree is of order 1. [1 Point]

3. In Item 5 below, we will show that conditioned on vertex i having k neighbors, the probability distributions of the labels of these k neighbors are *iid*, with the weights drawn from the re-weighted distribution

$$\rho^*(w) = \frac{w\rho(w)}{\mathbb{E}[w]} \qquad \text{where} \qquad \mathbb{E}[w] = \frac{1}{n}\sum_i w_i = \sum_w w\rho(w), \tag{1}$$

   Relate this result to the popular statement **"On average, your friends have more friends than you have, on average"** [2 Points]. To do so,

   (a) Recall your results from Item 1, which says that the expected degree of a vertex with weight w is approximately w. According to Item 5 below, the distribution of w for any neighbors of a given vertex i is ρ^*, so taking expectations with respect to ρ^* gives you the expected degree for all of the neighbors of i, and hence also the expected average degree. Express this in terms of the first and second moment of the distribution ρ.

   (b) Next, express the average expected degree of a random vertex i in terms of the distribution ρ, and compare the two. Show that the latter is always smaller, unless the distribution $\rho(w)$ is degenerated (i.e., has variance zero).

4. Show that the probability distribution of the degree d_i of a fixed vertex i is approximately $Poi(w_i)$. [3 Points]

   One way to do this is to observe that d_i is the sum of $n-1$ independent (but not iid) random variables as above, which will imply that the generating function factors over expectations with respect to these random variables. Approximating terms of the form $1-(1-x)p_{ij}$ by $e^{-(1-x)p_{ij}}$ should then allow you to show that the generating function of d_i is approximately that of a $Poi(w_i)$ random variable.

   Another way is the following (you might find the first version easier, and can choose either one, but you probably will need to understand the second one to do the last part of this exercise, Item 5 below):

(a) First calculate the probability $P_i(v_1, \ldots, v_k)$ that the neighbors of i are precisely the vertices $v_1, \ldots v_k$ (where $v_1, \ldots, v_k$ are, of course, pairwise distinct, and different from i). Do this exactly (no approximations yet). If you want a motivation for why we do this, look at Item 4c below.

(b) You should get a product of p_{iv}'s for the v's that are connected to i, and a product of $1 - p_{iu}$'s for the u's that are not connected to i. Approximate the product over the $1 - p_{iu}$ by $\exp(-\sum_u p_{iu})$ and approximate the sum over the u's that are not connected to i by a sum over all u's. If you do this right, you should get

$$P_i(v_1, \ldots, v_k) \approx e^{-w_i} \prod_{t=1}^{k} p_{iv_t}.$$

(c) The previous calculation holds for any fixed set of neighbors $\{v_1, \ldots, v_k\}$ of pairwise distinct vertices. To get the probability that the degree, d_i of vertex i is k, we need to sum over different sets $\{v_1, \ldots, v_k\}$ not containing i. Dividing by a factor of $k!$, we can equivalently sum of sequences $v_1, \ldots v_k$,

$$\Pr(d_i = k) = \frac{1}{k!} \sum_{\substack{v_1, \ldots, v_k \\ \text{pairwise distinct, different from } i}} P_i(v_1, \ldots, v_k)$$

where the sum goes over sequences such that $v_t \neq i$ for all t, $v_t \neq v_s$ for $s \neq t$.

Neglect both constraints, to get a sum over unconstrained sequences, and insert the previous approximation. You should get a sum over sequences of a product, which can be written as a product over sums of the form $\sum_{v_t} p_{iv_t}$. Note that these sums are actually all the same, and equal to $\sum_v p_{iv}$ — we used this principle in class already several times. If you do things correctly, you should get the probability that a $Poi(w_i)$ random variable is equal to k, as claimed.

5. Next we condition on i having k neighbors, and show that the probability distributions of the labels are approximately iid with distribution $\rho^*(w)$ (see (1)), independently of the weight w_i for i. [3 Points]

(a) We again start from the probability of seeing a particular sequence of vertices, which by the previous part is

$$\frac{1}{k!} P_i(v_1, \ldots, v_k) \approx \frac{1}{k!} e^{-w_i} \prod_{t=1}^{k} p_{iv_t}.$$

If instead of summing over all sequences, we only sum over those for which the weights are a particular sequence of weights $w_1, \ldots, w_k$ (i.e if we restrict the sum to v_t such that $w_{v_t} = v_t$ for all t), we will get the joint probability that i has degree k, and its neighbors have weights $w_1, \ldots, w_k$.

Write this out (again neglecting the constrains of pairwise distinctness, and distinctness from i), and use that again, the sum over the product can be written as a product over sums. You should get products of sums of the form $\sum_{v_t : w_{v_t} = w_t} p_{iv_t}$ (as a control, if you sum over the w_t's, which is the same as omitting the constraint $w_{v_t} = w_t$, you should get your previous result back).

(b) Show that the sum $\sum_{v_t : w_{v_t} = w_t} p_{iv_t}$ can be written as $\frac{w_i w_t}{\ell_n} n_{w_t}$ and express this in terms of w_i and the $\rho^*(w_t)$.

(c) To get the final, conditional probability, you need to divide by the probability of i having degree k. Inserting what you got in the previous part, this should yield the product $\prod_t \rho^*(w_t)$, as required.

4 A two stage SIR branching process

Here I will describe the consequences of the above for an SIR model on the Chung-Lu random graph.

Consider an epidemic started by infecting a random vertex. As before, an infected vertex i will draw a recovery time $T_i \sim exp(1)$, and then independently infect all its neighbors with probability $p_{T_i} = 1 - e^{-\lambda T_i}$ (as in the previous exercise, we neglect that we can only infect vertices which are still susceptible, so we just allow a vertex to infect all neighbors). Given what we learned so far, in the branching process approximation, this will now lead to the following birth process describing the infection tree:

2. For the first vertex, we choose $i \in \{1, \ldots, n\}$ at random, which means its weight $w = w_i$ will be distributed according to the empirical distribution, $w \sim \rho$. Then its degree d will be chosen from $Poi(w)$, then the recovery time $T \sim exp(1)$, and finally the number of infected children X from $Bin(d, p_T)$. Using what you showed in Section 2, the off-spring X of the root in the infection tree is thus distributed as follows:

3. Draw $w \sim \rho$ and $T \sim exp(1)$. Then draw $X \sim Poi(wp_T)$.

4. Consider a vertex v_t that is somewhere further down the infection tree. It is then the neighbor of some vertex i, thus we know that its degree distribution is obtained by first drawing w^* from ρ^* and then choosing $d \sim Poi(w^*)$. Arguing as above, the infection tree below any vertex except for the root will have an off-spring X^* distributed as follows:

5. Draw $w^* \sim \rho^*$ and $T \sim \exp(1)$. Then draw $X^* \sim Poi(w^* p_T)$.

In other words, the infection tree is given by a two-stage SIR process, where the distribution of the off-spring in later generation is different, and encodes the fact that infected vertices tend to have higher degrees than random vertices.

5 Analyzing the two-step SIR process [6 Points]

In this exercise, we analyze a birth-process T_{SIR} with two off-spring distributions

1. A random number X of infected children for the root, obtained by first drawing $w \sim \rho$ and $T \sim exp(1)$, and then drawing $X \sim Poi(wp_T)$, where $p_T = 1 - e^{-\lambda T}$.

2. A random number X^* of infected children for every other node in the tree, obtained by first drawing $w^* \sim \rho^*$ and $T \sim exp(1)$, and then drawing $X^* \sim Poi(w^* p_T)$.

To this end, we will first consider a branching process T^*_{SIR} with off-spring distribution X^*. Since vertices further down in this branching process have the same off-spring distribution as the root, our previous results for branching processes apply, so its extinction probability is the smallest solution η^* of

$$\eta^* = G^*(\eta^*),$$

where G^* is the generating function for X^*. The threshold for survival of T^*_{SIR} is also determined as before, namely by whether

$$R_0 = \mathbb{E}[X^*]$$

is smaller or larger than 1.

1. Express R_0 in terms of $\mathbb{E}[w^*]$ and $p = \mathbb{E}[p_T] = \frac{\lambda}{1+\lambda}$. Compare this to the naive expression $\mathbb{E}[d]p$ where d is the degree of a randomly chosen vertex, and the averages are taken over a random initial vertex i, and the degree distribution $Poi(w_i)$ of i derived in the previous part. Using that for a Poisson random variable $Y \sim Poi(c)$, $\mathbb{E}[Y(Y-1)] = c^2$, show that R_0 can be expressed as

$$R_0 = \frac{\mathbb{E}[d(d-1)]}{\mathbb{E}[d]} p$$

   [3 Points]

2. Next consider the full tree T_{SIR} under the root. Prove that $R_0 = 1$ is the threshold for survival of the infection on the Chung-Lu model, and derive an expression for its survival probability P. [3 Points]

   (a) To this end, first condition on the root having k infected children. What is the probability of this happening? Then observe that as in the standard birth-process discussed in the course, all the trees under these children must die out for the infection to die out. Finally, observe that the trees under the children of the root have the same distribution as the tree T^*_{SIR}, and thus go extinct with probability η^*. Denoting the generating function of the off-spring distribution X of the root by $G(y)$, $G(y) = \sum_{k=0}^{\infty} y^k \Pr(X = k)$, this should allow you to prove that the extinction probability of T is given by $\eta = G(\eta^*)$, or equivalently, that its survival probability is

$$P = 1 - G(\eta^*).$$

   (b) From this, further conclude that $P > 0$ if and only if $\eta^* < 1$, which shows that $R_0 = 1$ is the threshold for survival of T_{SIR}, i.e., for survival of an infection starting at a random vertex i.

3. **[2 Bonus Points]**: Give an example of a distribution for w where $R_0 = \infty$ for all $\lambda > 0$, while the naive prediction $p\mathbb{E}[d]$ would predict $R_0 < 1$ for sufficiently small λ.