Complex Networks and Statistical Learning
Homework 4

Chenghua Liu
liuch18@mails.tsinghua.edu.cn
Department of Computer Science
Tsinghua University

1 Erdos-Renyi Model

1.1 "Networks", Exercise 11.1

Consider the random graph $G(n,p)$ with mean degree c.

a) Show that in the limit of large n the expected number of triangles in the network is $\frac{1}{6}c^3$. This means that the number of triangles is constant, neither growing nor vanishing in the limit of large n.

b) Show that the expected number of connected triples in the network (as defined on page 184) is $\frac{1}{2}nc^2$.

c) Hence calculate the clustering coefficient C, as defined in Eq. (7.28), and confirm that it agrees for large n with the value given in Eq. (11.11).

Solution:

(a)

The random graph $G(n,p)$ with mean degree c implys that $c = (n-1)\times p$.Namely, $p = \frac{c}{n-1}$.

There are $\binom{n}{3}$ triples of vertices. Each triple has probability $\left(\frac{c}{n-1}\right)^3$ of being a triangle. Let Δ_{ijk} be the indicator variable for the triangle with vertices i, j, and k being present. Then the number of triangles is $x = \sum_{ijk} \Delta_{ijk}$. By linearity of expectation, we have

$$E(x) = E\left(\sum_{ijk} \Delta_{ijk}\right) = \sum_{ijk} E\left(\Delta_{ijk}\right) = \binom{n}{3}\left(\frac{c}{n-1}\right)^3 \to \frac{c^3}{6}$$

(b)

The connected triple means three nodes uvw with edges (u,v) and (v,w). (The edge (u,w) can be present or not.) Doing the same as (a) ,each triple has probability $\binom{3}{2}\left(\frac{c}{n-1}\right)^2$ of being

a connected triple. Let $\tilde{\Delta}_{ijk}$ be the indicator variable for the connected triple with vertices i, j, and k being present. Then the number of connected triples is $\tilde{x} = \sum_{ijk} \tilde{\Delta}_{ijk}$. By linearity of expectation, we have

$$E(\tilde{x}) = E\left(\sum_{ijk} \tilde{\Delta}_{ijk}\right) = \sum_{ijk} E\left(\tilde{\Delta}_{ijk}\right) = \binom{n}{3} \times 3 \left(\frac{c}{n-1}\right)^2 \to \frac{nc^2}{2}$$

(c)

$$C = \frac{(\text{ number of triangles }) \times 3}{(\text{ number of connected triples })} = \frac{c}{n}$$

1.2 "Networks", Exercise 11.2

Consider the random graph $G(n, p)$ with mean degree c.

a) Argue that the probability that a node of degree k belongs to a small component is $(1-S)^k$, where S is the fraction of the network occupied by the giant component.

b) Thus, using Bayes' theorem (or otherwise) show that the fraction of nodes in small components that have degree k is $\mathrm{e}^{-c}c^k(1-S)^{k-1}/k!$

Solution:

Accounting for all the possible sets of these k vertex among the $n-1$ possible neighbors we get the total probability of being connected to exactly k vertices:

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

$G(n, p)$ has a binomial degree distribution.And at the limit of $n \to \infty$, the degree distribution follows a Poisson distribution

$$P(k) = e^{-c}\frac{c^k}{k!}$$

We define the degree generating function of $P(k)$, namely

$$G_0(x) = \sum_{k=0}^{\infty} x^k P(k)$$

S represents the probability of a random node in the network resides on the giant component. We define the probability $\tilde{S}$ that a random neighbor of a random node, i, belongs to the giant component of the reduced network, which does not include the node i. In the thermodynamic limit, $N \to \infty$, we have([2][3])

$$S = 1 - G_0(1 - \tilde{S})$$

The probability, g, that a randomly selected node belongs to the giant component can be expressed in the form

$$g = \sum_{k=0}^{\infty} g_k P(k)$$

where g_k is the conditional probability that a random node belongs to the giant component, given that its degree is k. Comparing this expression to equation, we find that

$$g_k = 1 - (1 - \tilde{S})^k$$

To make the proof explicit, we introduce an indicator variable $\Lambda \in \{0, 1\}$, with $\Lambda = 1$ indicating that an event happens on the giant component, whereas $\Lambda = 0$ indicates that it happens on one of the finite components of the network. The probability that a random node resides on the giant component is given by $P(\Lambda = 1) = S$, while the probability that it resides on one of the finite components is $P(\Lambda = 0) = 1 - S$. The probability that a random node of a given degree k resides on the giant component is given by

$$P(\Lambda = 1 \mid K = k) = g_k = 1 - (1 - \tilde{S})^k$$

while the probability that it resided on one of the finite components is

$$P(\Lambda = 0 \mid K = k) = 1 - g_k = (1 - \tilde{S})^k$$

Using Bayes' theorem

$$P(K = k \mid \Lambda = \lambda) = \frac{P(\Lambda = \lambda \mid K = k)}{P(\Lambda = \lambda)} P(K = k)$$

The conditional degree distribution of nodes which reside on the giant component is given by

$$P(k \mid 1) = \frac{1 - (1 - \tilde{S})^k}{S} P(k)$$

while the conditional degree distribution for nodes which reside on the finite tree components is

$$P(k \mid 0) = \frac{(1 - \tilde{S})^k}{1 - S} P(k).$$

In the ER Graph, we have $\tilde{S} = S$.Therefore, we have proved the proposition.

2 Configuration Model

2.1 "Networks", Exercise 12.3

The "friendship paradox" of Section 12.2 says that your friends tend to have more friends than you do on account of the fact that they are reached by following an edge and hence have higher-than-average expected degree. The generalized friendship paradox is a related phenomenon in which your friends are richer/smarter/happier than you are (or any other characteristic). This happens when characteristics are correlated with degree. If, for instance, rich people tend to have more friends, then wealth and degree will be positively correlated in the friendship network. Since your friends have higher degree than you on average, we can then expect them also to be richer.

a) Suppose we have some value, such as wealth, on each node in a configuration model network. Let us denote the value on node i by x_i. Show that if we follow any edge in the network, the average value of x_i on the node at its end will be

$$\langle x \rangle_{\text{edge}} = \frac{1}{2m} \sum_i k_i x_i$$

b) Hence show that the equivalent of Eq. (12.14) — the difference between the average value of x_i for a friend and the average value for the network as a whole-is

$$\langle x \rangle_{\text{edge}} - \langle x \rangle = \frac{\text{cov}(k, x)}{\langle k \rangle}$$

where $\text{cov}(k, x) = \langle kx \rangle - \langle k \rangle \langle x \rangle$ is the covariance of k and x over nodes. Equation (12.14) is the special case of this result where x_i is the degree of the node, so that $\text{cov}(k, x) = \sigma_{k'}^2$ the variance of the degree distribution.

Solution:

(a)

Note that in andirected graph$\sum_i k_i = 2m$ .Then,

$$\langle x \rangle_{\text{edge}} = \frac{\sum_i k_i x_i}{\sum_i k_i} = \frac{1}{2m} \sum_i k_i x_i$$

(b)

$$\begin{aligned} &\langle x \rangle_{\text{edge}} - \langle x \rangle = \frac{\text{cov}(k, x)}{\langle k \rangle} \\ \Leftrightarrow &\langle x \rangle_{\text{edge}} \langle k \rangle - \langle x \rangle \langle k \rangle = \langle kx \rangle - \langle k \rangle \langle x \rangle \\ \Leftrightarrow &\langle x \rangle_{\text{edge}} \langle k \rangle = \langle kx \rangle \end{aligned}$$

The last equation is obviously right by easily substitution and $\sum_i k_i = 2m$.

3 Preferential Attachment

3.1 "Networks", Exercise 13.2

Within a set of papers on a particular topic it is found that the average paper cites 30 others in the set. Moreover the network of citations among the papers appears to be scale-free, with power-law exponent $\alpha = 3$. Let us assume that the network is well described by the preferential attachment model of Price, discussed in Section 13.1.

a) What is the average number of citations received by a paper?

b) On average what fraction of papers receive no citations at all?

c) On average what fraction of papers receive 100 or more citations?

d) The set consists of 10000 papers. What is the probability that the 100 th paper published has no citations? What is the probability that the 100 th-to-last paper published has no citations?

Solution:

(a)

The total of number of citations received by a paper equals to the total of citation.So, the anwser is 30.

(b)

Noted that $\alpha = 2 + \frac{a}{c}$,we have $a = c \times (\alpha - 2) = 30$.The formula to calculate the fraction of nodes in the network that have in-degree q when the network contains n nodes at $n \to \infty$

$$p_q = \frac{\mathrm{B}(q + a, 2 + a/c)}{\mathrm{B}(a, 1 + a/c)} = \frac{\mathrm{B}(q + 30, 3)}{\mathrm{B}(30, 2)}$$

where the $\mathrm{B}(x, y)$ is the Euler's beta function $\mathrm{B}(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$

Then we get $p_0 = \frac{1}{16}$ by mathematica.

(c)

We calculate the sum of p_q from $p = 0$ to 99

$$\sum_{q=0}^{99} \frac{\mathrm{B}(q + 30, 3)}{\mathrm{B}(30, 2)} = \frac{1610}{1703}$$

Then, we get the result $1 - \frac{1610}{1703} = \frac{93}{1703}$.We can also get the same result by

$$\begin{aligned} P_q = \sum_{q'=q}^{\infty} p_{q'} &= \frac{1}{\mathrm{B}(a, 1 + a/c)} \sum_{q'=q}^{\infty} \int_0^1 u^{q'+a-1}(1-u)^{1+a/c}\, \mathrm{d}u \\ &= \frac{1}{\mathrm{B}(a, 1 + a/c)} \int_0^1 u^{a-1}(1-u)^{1+a/c} \sum_{q'=q}^{\infty} u^{q'} \mathrm{d}u \\ &= \frac{1}{\mathrm{B}(a, 1 + a/c)} \int_0^1 u^{q+a-1}(1-u)^{a/c}\, \mathrm{d}u \\ &= \frac{\mathrm{B}(q + a, 1 + a/c)}{\mathrm{B}(a, 1 + a/c)} \end{aligned}$$

(d)

$n = 10000$ is large .We can consider it with $n \to \infty$. Then,

$$\pi_0(\tau) = \tau^{ca/(c+a)} = \tau^{15}$$

Therefore,

$$\pi_0(\frac{100}{10000}) = 1 \times 10^{-30}$$
$$\pi_0(\frac{9900}{10000}) = 0.860058$$

4 Navigable Small-World

4.1 "Networks", Exercise 18.3

The network navigation model of Kleinberg described in Section 18.3.1 is a onedimensional version of what was, originally, a two-dimensional model. In Kleinberg's original version, the model was built on a two-dimensional square lattice with nodes connected by shortcuts with probability proportional to $r^{-\alpha}$ where r is the "Manhattan" distance between the nodes, i.e., the network distance in terms of number of edges traversed (rather than the Euclidean distance). Following the outline of Section 18.3.1, sketch an argument to show for this variant of the model that it is possible to find a target node in $\mathrm{O}\left(\log^2 n\right)$ steps, but only if $\alpha = 2$.

Solution:

we design a twodimensional grid and allow for edges to be directed. The set of nodes are identified with the set of lattice points in an $n \times n$ square, $\{(i,j) : i \in \{1,2,\ldots,n\}, j \in \{1,2,\ldots,n\}\}$, and we define the lattice distance between two nodes (i,j) and (k,ℓ) as "Manhattan" distance : $d((i,j),(k,\ell)) = |k-i| + |\ell - j|$. For a universal constant $p \geq 1$, the node u has a directed edge to every other node within distance p . For universal constants $q \geq 0$ and $\alpha \geq 0$, we also construct directed edges from u to q other nodes (the long-range contacts) using independent random trials; the i^{th} directed edge from u has endpoint v with probability proportional to $[d(u,v)]^{-\alpha}$.

We set $p = 1$, $q = 1$ and $\alpha = 2$. The probability that u chooses v as its long-range contact is $d(u,v)^{-2} / \sum_{v \neq u} d(u,v)^{-2}$, and we have

$$\begin{aligned}\sum_{v \neq u} d(u,v)^{-2} &\leq \sum_{j=1}^{2n-2} (4j)\left(j^{-2}\right) \\ &= 4 \sum_{j=1}^{2n-2} j^{-1} \\ &\leq 4 + 4\ln(2n-2) \leq 4\ln(6n)\end{aligned}$$

The first inequality is from considering all values of $d(u,v)$ from 1 to $2n-2$.Each value has most $4 \times d(u,v)$ combination. Thus, the probability that v is chosen is at least $[4\ln(6n)d(u,v)^2]^{-1}$.

The decentralized algorithm $\mathcal{A}$ is defined as follows: in each step, the current messageholder u chooses a contact that is as close to the target t as possible, in the sense of lattice distance. For $j > 0$, we say that the execution of $\mathcal{A}$ is in phase j when the lattice distance from the current node to t is greater than 2^j and at most 2^{j+1}. We say $\mathcal{A}$ is in phase 0 when the lattice distance to t is at most 2 . Thus, the initial value of j is at most $\log n$. Now, because the distance from the message to the target decreases strictly in each step, each node that becomes the message holder has not touched the message before; thus, we may assume that the long-range contact from the message holder is generated at this moment.

Suppose we are in phase $j, \log(\log n) \leq j < \log n$, and the current message holder is u. What is the probability that phase j will end in this step? This requires the message to enter the set B_j of nodes within lattice distance 2^j of t. There are at least

$$1 + \sum_{i=1}^{2^j} i = \frac{1}{2}2^{2j} + \frac{1}{2}2^j + 1 > 2^{2j-1}$$

nodes in B_j, each is within lattice distance $2^{j+1} + 2^j < 2^{j+2}$ of u, and hence each has a probability of at least $(4\ln(6n)2^{2j+4})^{-1}$ of being the long-range contact of u. If any of these nodes is the long-range contact of u, it will be u 's closest neighbor to t; thus the message enters B_j with probability at least

$$\frac{2^{2j-1}}{4\ln(6n)2^{2j+4}} = \frac{1}{128\ln(6n)}$$

Let X_j denote total number of steps spent in phase $j, \log(\log n) \leq j < \log n$. We have

$$\begin{aligned} EX_j &= \sum_{i=1}^{\infty} \Pr[X_j \geq i] \\ &\leq \sum_{i=1}^{\infty} \left(1 - \frac{1}{128\ln(6n)}\right)^{i-1} \\ &= 128\ln(6n). \end{aligned}$$

An analogous set of bounds shows that $EX_j \leq 128\ln(6n)$ for $j = \log n$ as well. Finally, if $0 \leq j \leq \log(\log n)$, then $EX_j \leq 128\ln(6n)$ holds for the simple reason that the algorithm can spend at most $\log n$ steps in phase j even if all nodes pass the message to a local contact. Now, if X denotes the total number of steps spent by the algorithm, we have

$$X = \sum_{j=0}^{\log n} X_j$$

and so by linearity of expectation we have $EX \leq (1 + \log n)(128\ln(6n)) \leq \beta(\log n)^2$ for a suitable choice of β.

5 Reference

[1] Shai Carmi: Generating Function Method

[2] M. Molloy and A. Reed, The size of the giant component of a random graph with a given degree sequence, Combin., Prob. and Comp. 7, 295-305 (1998).

[3] M. Molloy and A. Reed, A critical point for random graphs with a given degree sequence, Random Structures and Algorithms 6, 161-180 (1995).

[4] Tishby I, Biham O, Katzav E, Kühn R. Revealing the microstructure of the giant component in random graph ensembles. Phys Rev E. 2018 Apr;97(4-1):042318. doi: 10.1103/PhysRevE.97.042318. PMID: 29758739.

[5] J. Kleinberg, The small-world phenomenon: An algorithmic perspective, in: Proceedings of the 32nd ACM Symposium on Theory of Computing, 2000, pp. 163–170 (2000)