Social and Technological Networks Edinburgh, 2017

Exercises 2 solutions. Random graphs.

TA: Benedek Rozemberczki *Exercises*

Note that the questions are meant to help you in exploring relevant ideas, and to give you practise in formalising problems in mathematical terms. Accordingly, they are sometimes vague. They may also have intentional or unintentional errors or inconsistencies!

Exercise 0.1. *Show that* $\ln n = \Theta(\lg n)$, *and* $\lg n = \Theta(\log n)$.

Answer. We can use the identity related to the change of base: $\log_y x = \frac{\log_z x}{\log_z y}$. Using this we know that $\ln n = \lg n / \lg e$ and $\lg n = \log n / \log e$. We know that $1/\lg e$ and $1/\log e$ are both constants. From this it follows that $\ln n = \Theta(\lg n)$, and $\lg n = \Theta(\log n)$. [log usually means base 10.]

The point of this exercise is that logs are within constant factors of each-other as long as the bases are constants.

Exercise 0.2. *Set up the ipython notebook on a system of your choice with networkx. Try it out.*

Exercise 0.3. *Write code to create plots showing the threshold phenomenon for existence of isolated vertices.*

```
# We import networkx, matplotlib and numpy for creating probabilities on a log scale.
import networkx as nx
import matplotlib.pyplot as plt
import numpy as np
# We set the number of sampled ER graphs, number of nodes and edge existence probabilities.
replications = 100
vertex_set_size = 100
probabilities = np.logspace(-3, 0, 100, endpoint=True)
# The data generation.
expected_isolates = []
for probability in probabilities:
    num_of_isolates = 0
    for replication in range(0,replications):
        sample_graph = nx.gnp_random_graph(vertex_set_size, probability)
        num_of_isolates = num_of_isolates + len(nx.isolates(sample_graph))
    num_of_isolates = num_of_isolates / replications
    expected_isolates.append(num_of_isolates)

# We plot the data - the horixontal axis is on a log scale.
plt.semilogx(probabilities,expected_isolates)
plt.show()
```

Exercise 0.4. Coupon collector problem. *Suppose they are giving out one coupon in each cereal boxes. There are n different types of coupons. You have to collect all n types to win a prize. Show that in expectation you need to buy $n \ln n$ boxes to to win the prize.*

Answer. Let X be the number of boxes that you have to buy to collect all n unique boxes. Let us divide the buying process into phases $1, 2, 3 \ldots$ where at phase i we have collected $i-1$ unique coupons, and are buying boxes to get our i^{th} unique coupon. Let X_i be the number of boxes we buy at stage i. So, $X = X_1 + X_2 + \ldots X_n$.

Now, the first coupon is obtained on the first try, so $E(X_1) = 1$. When buying the second box, the probability of getting as unique coupon is $\frac{n-1}{n}$. So, the expected number of tries is $E(X_2) = \frac{n}{n-1}$. In general, the expected number of boxes at stage i, is $E(X_i) = \frac{n}{n-i+1}$.

By linearity of expectation, $E(X) = E(X_1) + E(X_2) + \ldots E(X_n)$. Therefore,

$$\begin{aligned} E(X) &= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{n-i+1} + \cdots + \frac{n}{1} \\ &= n\left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n}\right) \\ E(X) &= n \cdot \sum_{i=1}^{n} \frac{1}{i}. \end{aligned}$$

The sum $\sum_{i=1}^{n} \frac{1}{i}$ is called the n^{th} harmonic number, and $H_n \approx \ln n$. So the expected number of boxes that we have to buy is:

$$E(X) = n \cdot H_n \approx n \cdot \ln n.$$

Exercise 0.5. *Show that for a suitable constant c, buying $cn \ln n$ boxes suffices to guarantee that you get at least one coupon of each type with high probability.*

Answer. Let A_i be the event that i^{th} type of coupon is not collected after buying k boxes. Thus, $P(A_i) = (1 - \frac{1}{n})^k$. Probability of getting at least one of all types of coupons is $(1 - P(\bigcup_{i=1}^{n} A_i))$.

From union bound rule we know that the probability of not getting some types of coupons is:

$$\begin{aligned} P(\bigcup_{i=1}^{n} A_i) &\leq \sum_{i=1}^{n} P(A_i) \\ &= \sum_{i=1}^{n} (1 - \frac{1}{n})^k \\ &= \sum_{i=1}^{n} (1 - \frac{1}{n})^{cn \ln n} \\ &\leq \sum_{i=1}^{n} e^{-c \ln n} \\ &= \sum_{i=1}^{n} n^{-c} \\ &= n^{1-c} \end{aligned}$$

Therefore, with high probability, we get at least one coupon of each type.

Exercise 0.6. *Show that a connected graph has at least $\Omega(n)$ triads.*

Answer. As per the definition of $\Omega(n)$, we need to show that for some constant $c > 0$, number of triad in a connected graph $T > c.n$ for $n > n_0$.

We would prove this by induction. In a minimally connected graph, a tree with three vertices, the number of triad is one. This constitutes the base case with $c = 1/3$.

For induction, we are assuming there is a graph G with n nodes and T triads. In G, $T \geq \frac{n}{3}$. Now, we'll show that after adding nodes and edges to G, it still holds this property.

- Adding an edge to G: Adding an edge to G only increases the number of triads. Thus, the number of triads in the new graph $T' \geq T \geq \frac{n}{3}$.
- Adding an vertex to G: As the new graph is to be connected, there should be at least an edge connecting the newly added vertex (i) to one of the vertices (j) in G. As, G was connected there was at least an edge jk. Thus, ijk is a triad in the new graph. Thus, $T' \geq T + 1 \geq \frac{n}{3} + 1 > \frac{1}{3}(n + 1)$.

Therefore, for $n \geq 3$, $T \geq c.n$ where $c = \frac{1}{3}$.

Exercise 0.7. *Consider an infinite 2D grid graph in the plane where each edge has length 1. Now consider the graph growth (measured in number of vertices) around any vertex. We are interested in the asymptotic growth measured in as $\Theta(\cdot)$. Remember that the growth depends upon the metric used. So, what happens when:*

- *We measure distances in the extrinsic Euclidean metric.*

- *We measure distances in the intrinsic graph metric.*

Do any of the answers change when we take a finite grid graph?

Answer. Euclidean metric. Let us count the number of grid squares that can be contained in a disk. Each square has side 1 and area 1. Thus a disk B of radius r can contain at most πr^2 squares. Therefore the circle has $O(r^2)$ squares – which give us an upper bound.

To get a lower bound, consider a disk B' of radius $(r - \sqrt{2})$ with the same center. Any square that intersects B' must be completely inside B. The area of B' is the area of its intersections with squares. Since the area of B' is $\Omega(\pi(r - \sqrt{2})^2)$, the number of squares inside B and intersecting with B' is $\Omega(\pi(r - \sqrt{2})^2) = \Omega(\pi r^2)$. Counting at least one vertex per square, and together with the upper bound, the number of vertices is $\Theta(r^2)$.

This analysis assumes that we are using the natural or canonical embedding of e grid in the plane that we normally use. The extrinsic metric will depend very much on what embedding we use. For an unknown embedding, we cannot say anything.

Graph metric. The growth in the intrinsic metric is also $\Theta(d^2)$. This can be seen using the canonical embedding, where we can consider the Euclidean circle of radius r enclosing the intrinsic ball, and the circle of radius $r/\sqrt{2}$ contained in the area of the intrinsic ball. We omit the full proof here.

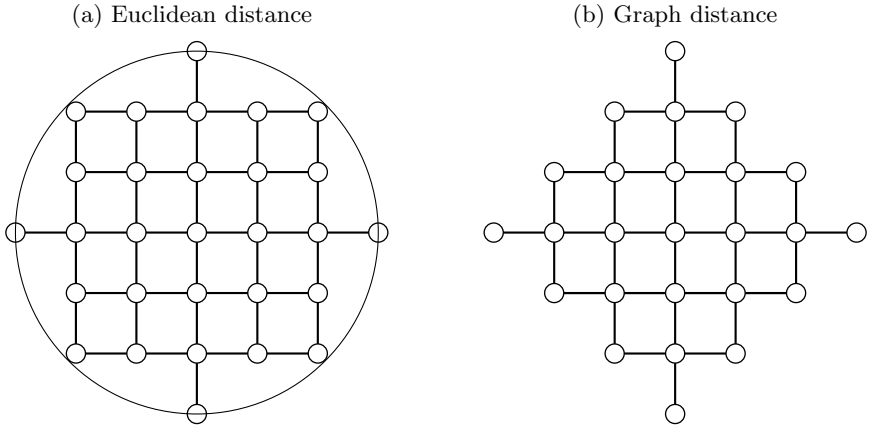

Figure 1: Vertices within distance 3 of a node.

The fact that we consider a finite grid only affects nodes that are at the borders of the grid. Even for those nodes we have an asymptotic growth that is described by $\Theta(d^2)$.

*** Exercise 0.8.** *Can you answer the same question for:*

- *A balanced binary tree.*
- *An infinite grid where each node may be absent with a probability $p = 0.01$ and $p = 0.5$.*

Answer.

1. (a) A binary tree has no obvious canonical embedding in the Euclidean space, and the asymptotic growth depends on this embedding. This there is no clear answer to this question.

   (b) When we use graph distances we can give an answer. Every node has 2 children – the asymptotic growth is $\Theta(2^d)$.

2. (a) The random removal of nodes does not affect the asymptotic growth of the Euclidean metric itself, but given that only $(1-p)$ fraction of the vertices remain, the growth measured in number of vertices will be $\Theta((1-p)d^2)$.

   (b) When the intrinsic graph metric is used, the absence of some nodes and edges will cause the distances to become longer, and as a result the growth will be slower, and for larger values of p the network will become disconnected. This is difficult to analyse. You can try it if you wish!