

## Assignment #1 – Warming up with Graphs and Randomness

Chris J. Riederer (TA)

A. Chaintreau (instructor)

**Why there is three parts in this assignment:** Each part fulfills one of the objectives of the class:

- **Manipulate concepts:** Getting Familiar with the technical concepts used in class, by reproducing similar arguments. Being proficient by manipulating the object to answer some small-size problem.  
You are expected to answer this question rigorously, the answer can be quite short as long as it contains all the required argument to justify your answer.
- **Experience the concepts in real case:** Being able to reproduce these concepts on real or synthetic data. Study their properties in real examples.
- **Connect the concepts to real-life:** Interpret a problem you find in light of the notions you have learned. Develop some critical eye to determine how the concepts introduced are useful in practice.

**How to read this assignment :** Exercise levels are indicated as follows

( $\rightarrow$ ) “elementary”: the answer is not strictly speaking obvious, but it fits in a single sentence, and it is an immediate application of results covered in the lectures.

*Use them as a checkpoint: it is strongly advised to go back to your notes if the answer to one of these questions does not come to you in a few minutes.*

( $\curvearrowright$ ) “intermediary”: The answer to this question is not an immediate translation of results covered in class, it can be deduced from them with a reasonable effort.

*Use them as a practice: how far are you from the answer? Do you still feel uncomfortable with some of the notions? which part could you complete quickly?*

( $\nrightarrow$ ) “tortuous”: this question either requires an advanced notion, a proof that is long or inventive, or it is still open.

*Use them as an inspiration: can you answer any of them? does it bring you to another problem that you can answer or study further? It is recommended to work on this question only AFTER you are done with the rest!*

## PART A — MANIPULATING THE CONCEPTS

**Exercise 1: Exercise 1: Cliques in random graphs (15pt)** A  $k$ -clique in a graph is a subset of vertexes such that any two of them are directly connected by an edge. Cliques are important objects in many computing problems and they also have simple combinatorial properties that allows to analyze them well. This exercise guides you towards understanding how they appear in a random environment.

In this exercise, we always work with a sequence of undirected uniform random graphs,  $G_n = (\{1, \dots, n\}, E_n)_{n \geq 0}$ . For each  $n \geq 0$  the graph  $G_n$  has  $n$  vertexes and a random collection of edges such that every edge  $e = \{i, j\}$ , where  $i \neq j$ , appears independently from others with a probability  $p(n)$ .

We wish to find, for any clique size  $k$ , a threshold for  $p(n)$ , defined as a function  $t(n)$  such that:

(i) When  $\lim_{n \rightarrow \infty} \frac{p(n)}{t(n)} = 0$ , then  $\mathbb{P}[G_n \text{ contains a clique of size } k] \rightarrow_{n \rightarrow \infty} 0$ .

(ii) When  $\lim_{n \rightarrow \infty} \frac{p(n)}{t(n)} = \infty$ , then  $\mathbb{P}[G_n \text{ contains a clique of size } k] \rightarrow_{n \rightarrow \infty} 1$ .

1. ( $\rightarrow$ ) What can you say about cliques of size 1 and 2? Can you define a threshold function for them.

We now consider the case  $k = 4$ , which means 4 vertexes connected by 6 edges together. It is unfortunately not as easy to characterize directly the probability of having at least one 4 cliques. Hence we need to use the probabilistic second moment method.

We denote by  $N_n$  the number of 4-cliques in the graph  $G_n$ . Since the edges of  $G_n$  appears randomly,  $N_n$  is a random variable (with values in  $\{0,1,2,\dots\}$ ). To be more precise, there are  $L_n = \binom{n}{4}$  number of choices of 4 elements in the  $n$  vertexes of  $G_n$ , and any of these can potentially form a clique in  $G_n$ , provided that the associated edges happen to be present, which is a probability event.

Let us denote  $C_1, C_2, \dots, C_{L_n}$  all these possible choices of 4 elements and for a subset  $C_l$ , let  $X_l$  denote the following random variable:  $X_l = \begin{cases} 1 & \text{if } C_l \text{ is a clique in } G_n \\ 0 & \text{otherwise} \end{cases}$ .

$$\text{We can now rewrite the variables } N_n \text{ as a sum: } N_n = \sum_{l=1, \dots, L_n} X_l. \quad (1)$$

2. ( $\rightarrow$ ) Are variables  $(X_l)_{l=1, \dots, L_n}$  mutually independent?
3. ( $\rightarrow$ ) What is the expectation of  $X_l$  (for a given  $l$ )? What is the expectation of  $N_n$ ?
4. ( $\curvearrowright$ ) Using Markov's inequality, show that when choosing  $t(n) = \frac{1}{n^{2/3}}$  the property (i) of threshold is satisfied.
5. ( $\rightarrow$ ) For this value of the threshold and if we assume that  $\lim_{n \rightarrow \infty} \frac{p(n)}{t(n)} = \infty$ , what can you say about the expectation of  $N_n$  as  $n$  goes large? Why is that not sufficient to conclude that the property (ii) is satisfied?

We recall the property of the variance of the sum of 0-1 variable, which implies from Eq.(1) that

$$\text{Var}[N_n] \leq \mathbb{E}[N_n] + \sum_{l, m | l \neq m} \text{Cov}[X_l, X_m]; \quad (\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]).$$

And the inequality  $\text{Cov}[X_l, X_m] \leq \mathbb{E}[X_l X_m]$  true for any non-negative variables.

6. ( $\rightarrow$ ) Show that when  $C_l$  and  $C_m$  are either disjoint or share a single vertex, then  $X_l, X_m$  are independent? What can you conclude on  $\text{Cov}[X_l, X_m]$ ?
7. ( $\curvearrowright$ ) Assuming that  $C_l$  and  $C_m$  share exactly two vertexes, give a bound on  $\text{Cov}[X_l, X_m]$ ? what if these subsets share exactly three vertexes?
8. ( $\curvearrowright$ ) Using the second moment method, conclude that the property (ii) of threshold is satisfied.
9. ( $\curvearrowright$ ) Without providing too much details, can you explain how you would characterize threshold for  $k = 5, 6, \dots$ . Assuming that a concentration argument holds (a variable is most likely around its mean) propose a reasonable candidate for threshold function  $t(n)$  for large value of  $k$ .
10. ( $\leftrightarrow$ ) Imagine we wish to apply the same methods to determine when *chordless cycles* appear in random environment. A cycle is a sequence of edges in the graph forming a path starting and ending in the same node, it is chordless if there does not exist an edge that connect nodes in the cycle except immediate neighbors (i.e. this cycle cannot be made shorter). Would it work?

## PART B — EXPERIENCING THE CONCEPTS

**Summary:** First, we are going to install Python and NetworkX. After this, we'll do some simple Python programming. We'll then get used to NetworkX, trying it out on a collaboration network. Finally, we'll examine strong and weak ties in the collaboration network.

**Exercise 2: Installing Python and NetworkX (0pt)** This is for no points, but is required for the rest of the assignment!

**Installing Python** First, install Python on your computer (or make sure it's already installed!). If you're using a Mac or running Linux, try opening up the command prompt and typing in "python". If it doesn't give you an error message, and you see a command prompt where you're able to run Python, everything is working! It may be useful to type in "python --version" to see what version you're running. Make a mental note of this version. Since you have Python running, you may skip to the section on NetworkX.

If you don't have Python preinstalled, you may install Python at this link: <http://www.python.org/getit/> Please use Python version 2.7.3 unless you have special permission from Chris. If you're running Windows, you should now be able to see Python and IDLE when you search for Python in the start menu.

Try opening up your command prompt and typing "python". If you get an error message about being unable to find Python, you will need to edit your Path variable. Be careful when doing so, as deleting something in your path could be bad for your system's stability. Instructions to do so are located here: <http://www.itechtalk.com/thread542.html> If you're using Windows 7, just use the instructions for Windows Vista.

**Installing NetworkX** After Python is set up, we'll install NetworkX. First, you should check if `easy_install` is on your computer. Open up a command prompt and type `easy_install networkx`. If this seems to download and install something, great! You're done. Otherwise, you need to download `setuptools` so we can use `easy_install`. This will make it extremely easy to download and install Python modules that aren't included in the original install. Follow the instructions here: <http://pypi.python.org/pypi/setuptools/> Now, open up your command prompt, and type in "easy\_install networkx". If you don't see any error messages, you should be good to go!

**Exercise 3: First steps using Python (3pt)** This question is making sure you have Python up and running and you understand how to write a function.

Please write all of your Python problems in one file.

NOTE: Please test your code on your own to make sure it works! The assignment says to write functions, but you should test these functions yourself!

1. (→) Write a Python function named `problem_one_a` that takes no inputs and returns the number 1337.
2. (→) Write a Python function named `problem_one_b` that prints out the string "Hello, world!"
3. (→) Write a Python function named `problem_one_c` that prints out all odd numbers between 50 and 150.

**Exercise 4: Introduction to NetworkX (6 pt)** This question will introduce us to NetworkX and how to use it. The NetworkX website is quite good and has a lot of information: <http://networkx.lanl.gov/> If you have a question, please look at the website before contacting the instructor.

To play with NetworkX, we will be using a collaboration network. This is a real dataset that shows coauthorship in papers about General Relativity and Quantum Cosmology (sounds complicated!). The nodes in this social network are authors. The nodes are connected by an edge if the authors are both listed as authors on the same paper. You should download this network here: <http://snap.stanford.edu/data/ca-GrQc.html>. Download the only file under the “files” section, `ca-GrQc.txt.gz`. The format `.txt.gz` is a compression format that can be unzipped with 7-Zip. You can get 7-Zip here: <http://www.7-zip.org/>

Feel free to open the file and glance at it. Each line of the file (except the comments!) corresponds to an edge in the social graph.

READ BEFORE CODING:

- To use NetworkX in a python script, you will need to import it. For instance, you can put `import networkx as nx` at the top of your script.
- The function `read_adjlist` can be used to read this file, e.g. `G = read_adjlist('ca-GrQc.txt.gz')`
- NetworkX includes a function `degree(G)` which returns a Python dictionary mapping node names to node degree.
- The `degree_histogram(G)` function may also be useful.
- Calling the `values()` method on a dictionary returns a list of just the values found, without any of the keys. The `keys()` method is similar. For iterating over both the keys and values in a dictionary, you may use `for (myKey, myValue) in dict.iteritems():`.

1. ( $\curvearrowright$ ) **Maximum and Minimum Degree** Write a function named `problem_two_a` that takes no inputs. This function should read the collaboration network into a NetworkX graph object. It should print the maximum degree and minimum degree found in the network.
2. ( $\curvearrowright$ ) **Who is Popular?** Write a function named `problem_two_b` that takes no inputs. This function should read the collaboration network into a NetworkX graph object. It should find the “name” of the author with the most coauthors, and print it. Note that the name will actually be a number, since this graph uses anonymized names.
3. ( $\curvearrowright$ ) **Coauthorship Percentages** Write a function named `problem_two_c` that takes no inputs. This function should read the collaboration network into a NetworkX graph object. This function will give us some insight into the distribution of coauthorship. The function should read the collaboration network into a NetworkX graph object. It should use the object to find and print the following numbers:
  - Percentage of authors with only 1 coauthor.
  - Percentage of authors with 10 or fewer coauthors.
  - Percentage of authors with 20 or fewer coauthors.
  - Percentage of authors with 40 or fewer coauthors.
  - Percentage of authors with 80 or fewer coauthors.

**Exercise 5: Weak and Strong Ties (6pt)** We will define a *strong tie* in the following way: for authors A and B, an edge is a strong tie if A and B share a coauthor. That is, there exists an author C such that A and B both have edges to C (A and B have written a paper with C). An edge that is not a strong tie is a *weak tie*. Note the following:

- Even though the nodes have labels that are numbers,
- For a graph G, you can find the list of neighbors of a node n by typing G[n]
- You can also use the G.neighbors(n) function if you prefer.
- You can do G.edges() or G.edges\_iter() to get a list of all edges in a graph. This will return a tuple. For now, you can pretend a tuple is a list with two elements. You can access these elements in the usual way, t[0] and t[1].
- You can use G.remove\_edge(a,b) to remove an edge between a and b.
- Check out <http://networkx.lanl.gov/reference/algorithms.component.html> to find NetworkX functions related to connected components.
- Note that “weak” and “strong” connectivity on the above page has nothing to do with our notion of strong and weak ties.

1. ( $\curvearrowright$ )**Strong Tie Function:** Write a function named `problem_three_a` that takes two inputs. This function should read the collaboration network into a NetworkX graph object. The inputs will be the labels of two nodes, and will output a Boolean value. The output should be True if the two nodes have a strong tie. The output should be False if the two nodes don't share an edge or if the two nodes have a weak tie. If you want, you can make sure the nodes are in the graph, and return False if either one is not.
2. ( $\curvearrowright$ )**How Strong?** Write a function named `problem_three_b` which takes no inputs. This function should read the collaboration network into a NetworkX graph object. The function should print, in order, the number of edges, the number of strong ties, and the number of weak ties. It may be useful to use the function you just wrote.
3. ( $\curvearrowright$ )**The Strength of our Weaknesses:** Write a function that takes no inputs. This function should read the collaboration network into a NetworkX graph object. The function should remove all weak ties from the graph object (or make a new graph object with only the strong ties of the original graph).

The function should print in order:

- The number of connected components originally
- The number of connected components without weak ties
- The number of nodes in the largest connected component originally.
- The number of nodes in the largest connected component without weak ties.

## PART C — CONCEPTS AT LARGE

### Exercise 6: Efficiency of social marketing (5 pt)

Imagine you are working for a company that sells online data storage like dropbox, which is cheaper and has better interface than previous products. Your product is already well recognized in a small community of early adopters, and now you would like it to get known to the most people. You decide to follow a social recommendation approach:

You start the following promotion: every current active user will be offered 3 invitations that can be forwarded to Facebook friend. When one invitation is accepted by this friend, and she sign up for the product, both users receives 2GB of storage for free. After one week, you see a significant increase of the product adoption through this technique, but it remains moderate. You would like to boost this sale, especially as the competition is quickly getting up to speed.

Your marketing team proposes to you two strategies:

- One called "grass root" in which, in addition, when two invitations get accepted and these users are friends, the three of you receives an additional 500MB free storage, to recognize you are a promoting cluster.
  - One called "forward" in which, in the same situation, the 500MB free storage is only given to the originator when two invitations are accepted and these people are not friends in Facebook.
1. (↷) Using concept seen in class, which policy would you suggest to use? How would you predict the ranking between "grass root", "forward" and the baseline strategy?
  2. (↷) Your company is also developing an app for online social reviews and recommendations on health, housework and babysitting. You would like your product to be known and plan on offering discount on these products via a similar invitation scheme. Would you provide the same recommendation for which rewarding strategy to use?

NB: You do not need to provide a very long answer, a short paragraph is sufficient, but it is important that your answer clearly describes your argument(s) in English without to someone who would not know the concepts in advance. You are welcome to reuse class readings by citing relevant facts.

### Exercise 7: Getting to know your online social environment (5 pt)

Choose a number  $N$  between 40 and 80. Look at the last  $N$  posts on your Facebook wall. For each post, give a point to the friend who who posted it. For each comment or like, give 0.5 points to the friend who posted it. Now, for each of these friends, look at your number of mutual friends.

NB: We recommend you take  $N$  sufficiently high so that you see a diversity of people (including some with more points than others). It should overall take no more than 15mn.

NB: We assumed that you can use a Facebook account. If that's more appropriate you can do the same exercise with RenRen, Google+, Twitter. Please contact us if that creates any difficulty.

1. (↷) Do you see any correlation among these two metrics? Can you interpret it?
2. (↷) Repeat the same experiment but counting points for the last  $N$  posts on your newsfeed, and for the last  $N$  messages in your Facebook inbox. Can you comment on any common patterns or differences you observe?

Please report these two metrics (in an anonymous manner) in a file. The file has one line per friends, with four columns separated by comma: The first columns contains the number of friends in common, The second columns contains how many points for this friend on the wall The third columns contains how many points for this friend on your newsfeed The second columns contains how many points for this friend on your Facebook inbox

For instance, if one of your friend A has three friends in common with you, 5 post in your wall, 1 comment in your newsfeed and 1 message in your inbox, one line in the file will be:

3, 5, 0.5, 1