COMS W4995-1: Intro to Social Networks — Due on September 22nd, 2011

Assignment #1 – Connect

A. Chaintreau (instructor), A. May (teaching assistant)

Why there is two parts in this assignment: Each part fulfills one of the two 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.
- **Connect the concepts to real-life:** Interpret a problem you find in light of the notions you have learned. Develop some critical eye w.r.t. how the concepts introduced are useful in practice.

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

($\rightarrowtail$) "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 a simple application 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?

($\rightsquigarrow$) "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 when you are done with the rest.

Part A

Practicing the concepts

Exercise 1: Cliques in random graphs 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 $\lim_{n\to\infty} \frac{p(n)}{t(n)} = 0$, then $\mathbb{P}\left[G_n \text{ contains a clique of size } k\right] \to_{n\to\infty} 0$.

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

1. ($\rightarrowtail$)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 $C_1, C_2, \cdots, C_{L_n}$ all these possible choices of 4 elements and for a subset C_l, let X_l denote the following random variable:

$$X_l = \begin{cases} 1 & \text{if } C_l \text{ is a clique in } G_n \\ 0 & \text{otherwise} \end{cases} .$$

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

2. ($\rightarrowtail$)Are variables $(X_l)_{l=1,\dots,L_n}$ mutually independent?

3. ($\rightarrowtail$)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. ($\rightarrowtail$)For this value of the threshold and if we assume that $\lim_{n\infty} \frac{p(n)}{(t(n)} = \infty$, 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

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

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

6. ($\rightarrowtail$)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 $\mathrm{Cov}\left[X_l, X_m\right]$?

7. ($\curvearrowright$)Assuming that C_l and C_m share exactly two vertexes, give a bound on $\mathrm{Cov}\left[X_l, X_m\right]$? 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 $k = 5, 6, \cdots$. 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. ($\nrightarrow$)Does the method generalize to $k = 3$? Do you think a threshold holds for triangles in graphs?

Exercise 2: Norm and number of neighbors on lattices in dimension $k \geq 1$: This exercise establishes an important step to answer the question in the next exercise. It is not directly related to proof seen in class, but it deals with a fundamental property of lattice that you ought to master.

A norm on the vector space $\mathbb{R}^k$ is any function $||.||$ that assigns to all elements x of $\mathbb{R}^k$ a real number $||x||$ in $\mathbb{R}$ such that:

$$\forall a \in \mathbb{R}\,,\, x \in \mathbb{R}^k\,,\, ||a \cdot x|| = |a| \cdot ||x||\,,$$
$$\forall x \in \mathbb{R}^k\,,\, y \in \mathbb{R}^k\,,\, ||x + y|| \leq ||x|| + ||y||\,,$$
$$\forall x \in \mathbb{R}^k\,,\, ||x|| = 0 \implies x = 0\,,$$

We recall the classical result all norms are equivalent (for finite dimensional vector space like $\mathbb{R}^k$). Hence, for any two norms $\|.\|_1$ and $\|.\|_2$, there always exist two constants $\alpha > 0$ and $\beta > 0$ such that

$$\forall x \in \mathbb{R}^k, \alpha\|x\|_1 \leq \|x\|_2 \leq \beta\|x\|_1\ .$$

The following functions defined on $\mathbb{R}^k$ are all norms:

- The $\mathbb{L}^1$ norm defined as: $\|x\|_{\mathbb{L}^1} = \sum_{i=1}^k |x_i|$; The $\mathbb{L}^2$ norm (or euclidean norm): $\|x\|_{\mathbb{L}^2} = \sqrt{\sum_{i=1}^k x_i^2}$.
- More generally, the $\mathbb{L}^j$ norm for $j \geq 1$ defined as: $\|x\|_{\mathbb{L}^j} = \left(\sum_{i=1}^k |x_i|^j\right)^{\frac{1}{j}}$.
- The maximum norm, also called the $\mathbb{L}^\infty$ norm, defined as: $\|x\|_{\mathbb{L}^\infty} = \max_{i=1,2,\ldots,k} |x_i|$.

1. ($\curvearrowright$) Show that in a lattice $\mathbb{Z}^k$ of dimension $k \geq 1$ and, for any norm $||.||$, the following holds:
   $\exists\ \alpha > 0$ and $\beta > 0$ such that, $\forall u \in \mathbb{Z}^k$, we have $\forall j > 0$: $\alpha j^k \leq \#\left\{\ v \in \mathbb{Z}^k\ \mid \|u - v\| \leq j\ \right\} \leq \beta j^k$.
   (hint: the key to answer this question quickly is to prove it first for a well chosen norm).
2. ($\curvearrowright$) Show similarly that we have:
   $\exists\ \alpha > 0$ and $\beta > 0$ such that, for any u and j: $\alpha j^{k-1} \leq \#\{v \in V | \|u - v\| = j\} \leq \beta j^{k-1}$.
3. ($\nrightarrow$) We have essentially proved that, in a lattice or a grid, the number of points at distance j grows polynomially in j (for any distance such as, *e.g.*, the number of edges to traverse in the grid). Does the same hold for any graph (where the distance again is given by the number of edges on a path)?

Exercise 3: Analysis of small world property in dimension $k \geq 1$ A good way to master the proof on augmented lattice is to manipulate it for different graphs. This exercise is a complement to the detailed proof of the case $k = 1$, found at the end of this document and in the scribing note, it will allow you to do that step by step. You may if you wish even start the exercise immediately after reading the proof in the three separate case $r < 1$, $r > 1$ and $r = 1$.

1. ($\curvearrowright$) Deduce from the previous exercise that for a finite lattice of dimension k (with length $L - 1$, containing $N = L^k$ nodes) there exist $\alpha > 0$, $\beta > 0$ independent of N such that:

$$\alpha \sum_{j=1}^{\lfloor L/2 \rfloor} \frac{1}{j^{r-(k-1)}} \leq \sum_{v \neq u} \frac{1}{\|u - v\|^r} \leq \beta \sum_{j=1}^{L} \frac{1}{j^{r-(k-1)}}\ .$$

2. ($\rightarrowtail$)From this inequality can you briefly justify why the value $r = k$ is critical for dimension k?

3. ($\curvearrowright$)Assuming $r < k$ show that, wherever u and v are located on the lattice the probability that u is connected to v by a shortcut becomes polynomially small as N grows. In other words, there exists $\delta > 0$ and a constant $c_1 > 0$ such that: $\mathbb{P}\left[u \rightsquigarrow v\right] \leq \frac{c_1}{N^\delta}$.

4. ($\rightarrowtail$) Let us denote by I_l the set of nodes at distance at most l from the target:

$$I_l = \{\ u \in V\ \mid \|u - t\| \leq l\ \}\ ,$$

Which one of the following is an upper bound on the probability that at least one the n first shortcuts met by the walk drawn using greedy routing connects to a node within I_l?

(*i*) $\frac{2c_1 nl}{N^\delta}$ (*ii*) $\frac{2nl^k c_1}{N^\delta}$ (*iii*) none of the above

5. ($\rightarrowtail$) Conclude that greedy routing needs in expectation at least a constant multiplied by $N^{\frac{k-r}{k(k+1)}}$ steps to succeed.

6. ($\curvearrowright$) We now assume $r > k$. Prove that the probability that u shortcuts has length greater than m is less than $\frac{c_3}{m^{r-k}}$. Conclude that greedy routing needs in expectation at least a constant times N^η steps for $\eta > 0$.

7. ($\rightarrowtail$) Assuming $r = k$ what can you deduce on the normalizing constant? Prove that the probability for a node in phase j to be connected to a node in phase $j' < j$ does not depend on j and becomes small slowly with N. Conclude.

Exercise 3: Extension of the small world result to an infinite lattice One of the limitation of the above proof is to deal frequently with normalizing constant and finite networks. In this exercise we prove that for at least two cases of the one studied above, a formulation using an infinite lattice can be drawn.

In an infinite lattice, one cannot hope to have any bound on the time to connect two arbitrarily far away nodes. On the other hand, one may hope that on an infinite lattice that starting from a node at a fixed distance D from the target, greedy routing finds a path whose length grows slowly with D.

1. ($\curvearrowright$)Assuming $r > k$, can you prove that the random biased augmented lattice (which we also called Kleinberg's model) extends naturally without modification to an infinite lattice. Why is that impossible when $r \leq 1$?

2. ($\curvearrowright$)Assuming $r > k$, can you quickly justify why there exists a constant $C > 0$ and $\eta > 0$ such that, in expectation, the path found by greedy routing requires at least CD^η steps when starting from a node at distance D from the target.

   We will now prove that Kleinberg model can be modified so that the proof of the critical case $r = k$ applies to an infinite lattice.

3. ($\rightarrowtail$)For an $\varepsilon > 0$ Let $f : x \mapsto \frac{1}{\ln^\varepsilon(x)}$, what is f' the derivative of f ? what is the limit $\lim_\infty f$?

4. ($\curvearrowright$)Prove that for any $\varepsilon > 0$, one can naturally extend Kleinberg model to an infinite lattice for $r = k$, assuming that the probability to have a shortcut $u \rightsquigarrow v$ is

$$\mathbb{P}\left[u \rightsquigarrow v\right] = \frac{1}{\|u - v\|^k \ln^{1+\varepsilon}(\|u - v\|)}\ .$$

5. ($\curvearrowright$)Prove that greedy routing in the above model uses at most $O(\ln^{2+\varepsilon}(D))$ steps when starting at distance D from the target.

Part B
Concepts at large

Exercise 1: In a future, not so far away ... Let's imagine, for the sake of a thought experiment, a planet in which live the Erdosians. The Erdosians is a strict society that made the choice of canceling all biases in social relationship between its member. Through a pervasive computer interface, occasions of social gathering (attending a course, a movie, a party, etc.) are randomly assigned. As a consequence the set of people that anyone knows at any time is drawn in a uniform random way from the population irrespective of geography and social classes.

To prove the superiority of their social organization, the Erdosians demonstrates to their visitors their "connect" application which allows any member of the society to find its shortest paths to any other. A constant monitoring over a century has shown that this distance remains remarkably small in spite of the population increase. The "connect" application made available for all members have quickly become extremely popular to keep up with friends through a multi-hop referral. One important reason why it is used is that the Erdosians, liberated from any geographical biases, have started to adopt an increasing nomadic life, making them hard to locate.

Unfortunately, a major natural disaster is responsible for a black-out of the "connect" application and its associated data, leaving each Erdosians with only a way to contact their immediate friends, that were initially cached on their cell-phone.

- Using elements from the course, can you comment on the gravity of the situation? Assuming that it is still possible for an Erdosian to call her friend, and ask for referral, could they possibly exchange information to keep up with each other through a short chains of acquaintance?

(NB: I know you will not believe me, but as I was writing this exercise in La Guardia airport one morning, I was approached two minutes later by a teenager and her mum who asked if she could use my laptop to contact her friend after a delayed flight. When I said that I was not online but that they could use my phone for email, her answer was "Oh thank you, but it's going to be too complicated, I have to contact her on Facebook!". I could not help but wonder "Are Erdosians already landed? Why choose La Guardia Terminal D for a first contact?")

Exercise 2: The data-set and the phone company You are working with a group of students on studying the way people use cell phones to connect to each other, and the small world property of the network among them. A connection with a telephone provider allows you to get access to some anonymous data set that would turn out invaluable to answer empirically some important questions.

Unfortunately, after some concerns of public data release, your contact in the company mention that it would not possible to release the full data sets and that members of her team would simply accept to release the information about people who make call between each other very regularly (e.g., every 3 days). They mention to you that these are the most important ties of the network anyway, and that they should be sufficient for your study. Tomorrow you will meet with responsible of the team to present your project and explain how the data will be used. What arguments could you use to convince them that they should provide a more complete data sets?

(This exercise was inspired by D. Easley and J. Kleinberg)