M375T/M396C: Topics in Complex Networks

Problem Set 1

Due date: Tuesday, May 07

Instructions: Starred problems (marked with an *) are additional problems for those taking the course for graduate credit (M396C) that are not necessary for those in M375T. Do not be intimidated by the length of the problem descriptions—they are purposefully made to provide a thorough explanation and to guide you through the analysis of several results we have discussed in lecture. Finally, please feel free to collaborate with others on problems, but your submitted writeup must be your own.

1 Branching processes

1. *Geometric branching process:* Consider a Galton-Watson branching process with geometric offspring distribution with parameter p, i.e.,

$$\mathbb{P}(\xi = k) = p^k(1-p).$$

(A) Compute the probability of extinction occuring in generation n (using generating functions).

(B) Give a general expression for the probability of extinction.

*(C) Derive an expression for the generating function of

$$W = \lim_{n\to\infty} X_n / \left(\mathbb{E}\left(\xi\right)\right)^n,$$

where X_n is the number of individuals in generation n.

2. *Chernoff bound:* Consider a coin with probability p of landing heads and probability $1-p$ of landing tails. Suppose the coin is flipped n times, and let $X_i \sim$ Bernoulli(p) be the indicator random variable that the i^{th} flip lands heads.

(A) Show that the rate function $h(a)$ corresponding to a Bernoulli(p) random variable is

$$h(a) = a\log\left(\frac{a}{p}\right) + (1-a)\log\left(\frac{1-a}{1-p}\right).$$

(B) Use a Chernoff bound to determine a value for n so that the probability that more than half of the coin flips come out heads is less than 0.001.

2 Random graphs and phase transitions

1. *Subgraphs of random graphs:* Fix the probability of any given link forming in an Erdos-Renyi network to be p with $0 < p < 1$. Fix some arbitrary network g on k nodes. Now, consider a sequence of random networks indexed by the number of nodes n, as $n \to \infty$. Show that the probability that a copy of the k-node network g is a subnetwork of the random network on the n nodes goes to 1 as n goes to infinity.

[Hint: Partition the n nodes into as many separate groups of k nodes as possible (with some leftover nodes) and consider the subnetworks that form on each of these groups. Using independence of link formation, show that the probability that none of these match the desired network goes to 0 as n grows.]

2. *Cliques in random graphs:* A k-clique in a graph is a subset of k vertices such that any two of them are directly connected by an edge (i.e., there are $\binom{k}{2}$ edges in a k-clique). Cliques are important objects in many computing problems and they also have simple combinatorial properties that allows one to analyze them well. This exercise guides you towards understanding how they may appear in a random graph.

In this exercise, we always work with a sequence of Erdos-Renyi random graphs $G(n, p)$. We seek a threshold for the event that G_n contains a k-clique. That is, we will find a function $t(n)$ such that:

- Property (i): $\lim_{n\to\infty} \mathbb{P}\left(G_n \text{ contains a } k\text{-clique}\right) = 0$ if $p(n) = o(t(n))$
- Property (ii): $\lim_{n\to\infty} \mathbb{P}\left(G_n \text{ contains a } k\text{-clique}\right) = 1$ if $p(n) = \omega(t(n))$.

(A) What can you say about cliques of size 1 and 2? Can you define a threshold function for them? Prove that your answers satisfy the definition of a threshold.

We now consider the case $k = 4$. It is not as easy to directly characterize the probability of having at least one 4-clique, and we will need to use the second-moment method to prove the existence of any threshold.

Let N_n be the number of 4-cliques in the graph $G(n, p)$. Since the edges of $G(n, p)$ appear randomly, N_n is an $\mathbb{N}$-valued random variable. To be precise, there are $L_n = \binom{n}{4}$ number of choices of 4 elements in the n veritices of $G(n, p)$, and any of these can potentially form a 4-clique provided that the associated edges happen to be present. Denote by $C_1, \ldots, C_{L_n}$ all of these possible choices of 4 elements and for a subset C_l, let X_l be the indicator random variable that C_l is a clique. That is,

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

Note that $N_n = \sum_{l=1,\ldots,L_n} X_l$ as a sum of random variables.

(B) Are the variables $\{X_l\}_{l=1,\ldots,L_n}$ mutually independent? Are they identically distributed?

(C) What is $\mathbb{E}[X_l]$ for a given l? What is $\mathbb{E}[N_n]$?

(D) Use this average analysis to determine a threshold function $t(n)$ for the existence of a 4-clique.

(E) Using Markov's inequality, prove that $t(n)$ satisfies property (i).

(F) For this value of the threshold, if we assume that $p(n) = \omega(t(n))$ then what can be said about the expectation of N_n as $n \to \infty$? Why is this not sufficienty fo conclude that property (ii) is satisfied?

Now recall the following property of the variance of the sum of $\{0, 1\}$-valued random variables:

$$Var\left[N_n\right] \leq \mathbb{E}\left[N_n\right] + \sum_{l \neq m} Cov\left(X_l, X_m\right),$$

where $Cov(X, Y) = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]$ with μ_X and μ_Y the mean of X and Y, respectively. Note that $Cov(X, Y) \leq \mathbb{E}[XY]$ if X and Y are nonnegative random variables.

(G) Show that when C_l and C_m are either disjoint or share a single vertex, then X_l and X_m are independent. What can you conclude about $Cov\left(X_l, X_m\right)$?

(H) Assuming that C_l and C_m share exactly two vertices, give a bound on $Cov\left(X_l, X_m\right)$. What if these subsets share exactly three vertices?

(I) Use the second-moment method to conclude that the threshold $t(n)$ satisfies property (ii).

(J) We now consider the most general case. Use an average analysis to propose a reasonable candidate for the threshold function $t(n)$ of the existence of a k-clique in $G(n, p)$. Prove that this threshold satisfies property (i) using Markov's inequality. No proof of property (ii) is required.

*(K) Imagine we wish to apply the same method to determine when *chordless cycles* appear in a random graph. A cycle is a sequence of edges in the graph forming a path starting and ending in the same node, and is chordless if there are no edges between non-neighboring nodes (i.e., the cycle cannot be made shorter). Will a similar argument hold? Why or why not?

*3. *Clustering in the configuration model:* Consider a graph g with n nodes generated according to the configuration model with a particular degree distribution $P(d)$. Show that the overall clustering coefficient is given by

$$C(g) = \frac{\langle d \rangle}{n} \left(\frac{\langle d^2 \rangle - \langle d \rangle}{\langle d \rangle^2} \right)^2,$$

where $\langle d \rangle$ is the expected degree under distribution $P(d)$, i.e., $\langle d \rangle = \sum_{d=1}^{\infty} dP(d)$ and similarly $\langle d^2 \rangle = \sum_{d=1}^{\infty} d^2 P(d)$.

3 Small-world model and routing

1. *Clustering in the small-world model:* Consider the Watts-Strogatz small-world model, given by a ring lattice in which each node is connected to all nodes within distance k. Let p be the rewiring probability.

(A) Show that when $p = 0$, the overall clustering coefficient of this graph is given by

$$C = \frac{3k-3}{4k-2}.$$

*(B) Show that when $p > 0$, the clustering coefficient is $C = \frac{3k-3}{4k-2}(1-p)^3$.

2. *Norm and number of neighbors on lattices in dimension $k \geq 1$:* This exercise establishes an important step to ansewr the question in following exercises. It does not directly related to the proof seen in lecture but it deals with a fundamental property of lattices that is well worth learning.

A norm on the vector space $\mathbb{R}^k$ is any function $\|\cdot\| : \mathbb{R}^k \to [0, \infty)$ such that for all scalars $a \in \mathbb{R}$ and vectors $x, y \in \mathbb{R}^k$,

$$\begin{aligned} &\|ax\| = |a| \, \|x\| \\ &\|x + y\| \leq \|x\| + \|y\| \\ &\|x\| = 0 \iff x = 0. \end{aligned}$$

It is classical that all norms on finite-dimensional spaces are equivalent—that is, for any two norms $\|\cdot\|_A$ and $\|\cdot\|_B$ there exists $c_1, c_2 > 0$ such that for all $x \in \mathbb{R}^k$,

$$c_1 \|x\|_A \leq \|x\|_B \leq c_2 \|x\|_A .$$

The following functions are all norms:

- l_p-norm, $1 \leq p < \infty$:

  $$\|x\|_p = \left(\sum_{i=1}^{k} |x_i|^p \right)^{1/p} .$$

  In particular, $p = 1$ and $p = 2$ give the l_1- and l_2-norm, respectively:

  $$\|x\|_1 = \sum_{i=1}^{k} |x_i| , \qquad \|x\|_2 = \sqrt{\sum_{i=1}^{k} |x_i|^2}.$$

- l_∞-norm:

  $$\|x\|_\infty = \max \{ |x_1| , \ldots , |x_k| \} .$$

(A) Show that in a lattice $\mathbb{Z}^k$ of dimension $k \geq 1$ and for any norm $\|\cdot\|$, the following holds: there exists $c_1 > 0$ and $c_2 > 0$ such that for all $u \in \mathbb{Z}^k$, we have

$$c_1 j^k \leq \left| \left\{ v \in \mathbb{Z}^k : \|u - v\| \leq j \right\} \right| \leq c_2 j^k$$

for every $j > 0$.

[Hint: The key to answer this easily is to first show it in a well-chosen norm, and then to use the equivalence of norms on finite-dimensional spaces.]

(B) Similarly, show that there exists $c_1 > 0$ and $c_2 > 0$ such that for all $u \in \mathbb{Z}^k$,

$$c_1 j^{k-1} \leq \left|\left\{v \in \mathbb{Z}^k : \|u - v\| = j\right\}\right| \leq c_2 j^{k-1}$$

for every $j > 0$.

(C) We have essentially proven that in a lattice or grid, the number of points at distance j grows polynomially in j (for any distance such as, for example, the number of edges to traverse in the grid). Does the same hold for any graph (where the distance is again given by the number of edges on a path)?

3. *1-D analogue of Kleinberg small-world model:* Nodes are ordered on a line and each node u has a single shortcut, which links it with a node v with probability $\mathbb{P}(u \rightsquigarrow v)$ proportional to $\|u - v\|^{-\alpha}$. Redo the heuristic given the lecture notes (i.e., no need to give a proof) to determine the navigation times for different values of the clustering exponent α. What is the critical value of α that allows greedy routing to work efficiently?

*4. *Analysis of Kleinberg small-world model in $\mathbb{Z}^k$, $k \geq 1$:* A good way to obtain a rigorous proof for the behavior of greedy routing in the Kleinberg model is to study it for different graphs. This exercise is a complement to the rigorous proof of the case $k = 1$ given in the notes, which we will complete step by step. Here, the probability $\mathbb{P}(u \rightsquigarrow v)$ of a shortcut from u to v is proportional to $\|u - v\|^{-r}$ for some constant $r > 0$ (note that we are now using r instead of α to denote this exponent!). You may wish to start this exercise immediately after reading the proof in the lecture notes for the three separate cases $r < 1$, $r = 1$, and $r > 1$ when $k = 1$.

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

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

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

(C) 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, show that there exists $\delta > 0$ and a constant $c_1 > 0$ such that

$$\mathbb{P}(u \rightsquigarrow v) \leq c_1 N^{-\delta}.$$

(D) Let us denote by I_l the set of nodes at distance at most l from the target t:

$$I_l = \{u \in V : \|u - t\| \leq l\}.$$

Which one of the following is an upper bound on the probability that at least one of the first n shortcuts met by the walk drawn using greedy routing connects to a node within I_l? (i) $2c_1 nl/N^\delta$, (ii) $2c_1 nl^k/N^\delta$, (iii) none of the above.

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

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

(G) 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. This will conclude the proof.

*5. *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.

(A) Assuming $r > k$, can you prove that Kleinberg's model extends naturally without modification to an infinite lattice? Why is that impossible when $r \leq 1$?

(B) Assuming $r > k$, 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 $CD\eta$ 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.

(C) For any $\varepsilon > 0$, let $f(x) = 1/\log^{\varepsilon}(x)$. What is $\lim_{x\to\infty} f(x)$? What is f'?

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

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

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

4 Power laws and preferential attachment

1. *Analysis of the copying model:* Through the analysis of the Yule process, we have seen in class the consequence of reinforcement. Reinforcement here denotes the fact that a difference between two entities (e.g. the size of two genera, the number of links received by two webpages) is itself biasing the dynamics so that the difference continues to increase. As a consequence, even starting from a small initial set of equivalent entities, minor difference created by randomness could further lead to major differences. In the case of the Yule process, it provided

a simple model explaining the imbalance of species among genus which is characterized by a power law. In this exercise, we conduct a very similar analysis to model edges created in a graph. The main result is to show that a very simple copying strategy leads to big imbalances, characterized by a power law degree distribution of node in-degrees.

The copying model is given as follows. We start at time $t = 1$ from a directed graph containing $N(1)$ nodes such that each of these nodes has exactly one outgoing edge. We introduce at each time step $t = 2, 3, 4, \ldots$ a new node $v(t)$ with a single outgoing edge $e(t)$ that is initially connected to another node chosen uniformly at random (which we denote by $u(t)$). We assume the following evolution:

- With probability p, the process stops there and the new edge connects $v(t)$ to $u(t)$
- Otherwise (i.e., with probability $1 - p$), $v(t)$ examines the edge that is starting at $u(t)$ and decides to *copy* this edge. This means that the edge from $v(t)$ to $u(t)$ is rewired to one that goes from $v(t)$ to the destination of the edge starting in $u(t)$.

We would like to determine the evolution of node degrees. Since the graph is directed, all nodes have both an out-degree and an in-degree. The out-degree of all nodes in the graph is always equal to 1. The interesting problem is to analyze the evolution of the in-degree of nodes in the graph as t becomes large. Denote by $X_i(t)$ the number of nodes in the graph with an in-degree equal to $i \geq 0$.

(A) To begin, assume that at time $t = 1$ there is a single node and a single edge (i.e., this edge is a self-loop from this single node to itself). Prove that no other self-loops will be created later.

(B) How many nodes (denoted by $N(t)$) and edges (denoted by $E(t)$) are there in the graph as a function of t?

(C) Assuming that $X_0(t)$ (i.e., the number of nodes with no incoming edge) is known, what the possible values of $X_0(t+1)$ and what are the probabilities that each of these values occur?

(D) Derive from the previous question the evolution equation giving the expected value $\mathbb{E}\left[X_0(t+1)\right]$ as a function of $\mathbb{E}\left[X_0(t)\right]$. As seen in lecture, this can be done using conditional expectation.

(E) For the next few parts, we assume $p < 1$. Given a constant c_0, define the sequence

$$\Delta_0(t) = \mathbb{E}\left[X_0(t)\right] - c_0 t.$$

Show that there exists a value of c_0 such that for all $t \geq 1$, $|\Delta_0(t)| \leq |\Delta_0(1)|$. What is the value of c_0? (i) $c_0 = 1/(1-p)$, (ii) $c_0 = 1/(2-p)$, (iii) $c_0 = 1/(1+p)$

(F) Deduce that the following hypothesis is true for $i = 0$: For all $\varepsilon > 0$, there exists an $A > 0$ such that $|\Delta_i(t)| \leq At^{\varepsilon}$.

(G) For a sequence of constants $c_0, c_1, \ldots$, define

$$\Delta_i(t) = \mathbb{E}\left[X_i(t)\right] - c_i t.$$

Show that for any $i > 0$, if the sequence satisfies

$$c_i = c_{i-1}\left(1 - \frac{2-p}{(1+p)+i(1-p)}\right)$$

then we have that

$$\Delta_i(t+1) = \Delta_i(t)\left(1 - \frac{p+i(1-p)}{N(t)}\right) + \Delta_{i-1}(t)\left(\frac{p+(i-1)(1-p)}{N(t)}\right).$$

(H) Assume that the hypothesis of part (F) holds for any $i \geq 0$. What does this tell us about the evolution of degrees in this system? Using that $p < 1$, show that for $i > 0$ we have

$$c_i = c_{i-1}\left(1 - \frac{\beta}{i} + \varepsilon(i)\right) \text{ where } |\varepsilon(i)| \leq \frac{A}{i^2} \text{ and } \beta = \frac{2-p}{1-p}.$$

(I) As shown in lecture, if we neglect the error term $\varepsilon(i)$ we have that c_i is approximately following a power law with coefficient β. For which values of p does the power law become the most imbalanced? How does this compare to your intuition about the dynamics of copying?

(J) Assuming now $p = 1$, how could you characterize the decrease of c_i as i gets large? Relate this behavior to the dynamics of the copying model.

*(K) Complete the proof by showing that that the hypothesis of part (F) is in fact true for all $i \geq 0$.