CS224W: Social and Information Network Analysis | Fall 2013

Problem Set 2

Due 9:30am October 24, 2013

General Instructions

These questions require thought, but do not require long answers. Please be as concise as possible. Please fill the cover sheet (http://cs224w.stanford.edu/cover.pdf) and submit it as a front page with your answers. We will subtract 2 points for failing to include the sheet. Include the names of your collaborators (see the course website for the collaboration or honor code policy). You are allowed to take maximum of 1 late period (see course website for a definition of a late period).

Regular (non-SCPD) students should submit hard copies of the answers either in class or in the submission cabinet (see course website for location). You should also include source code and any other files you used in the paper submission.

SCPD students should submit their answers through SCPD. The submission must include all the answers, the source code and the usual SCPD routing form (http://scpd.stanford.edu/generalInformation/pdf/SCPD_HomeworkRouteForm.pdf).

Additionally, all students (Regular and SCPD) should upload all code to http://snap.stanford.edu/submit/.

Questions

1 Musings of the Chief Social Engineer of the World [20 points-Ashley, Peter]

1.1 [5 points]

Suppose every person in the world has a large number of friends. Each friendship has a strength associated with it, such that each node's friends can be ranked by friendship strength. While each node has many friends in the real world, the social networking site, Friendarchy, everyone is using allows each user to maintain only the 10 friendships that he/she considers the strongest. For example, Sue may consider Martha her strongest friendship but Martha may not reciprocate the feeling.

While working at Friendarchy, you discover that an administrative error has given you the power to dictate the structure of Friendarchy as well as the ability to reassign everyone's friendships. Because you are egocentric, you decide to maximize the number of people *you* *personally* can reach in m hops (e.g., by doing a breadth-first search of depth m starting from your node) by modifying the network. However, you are still constrained by every human having exactly 10 links.

What is the optimal network structure for your mischievous intent? How many people can you reach in m hops with this structure?

1.2 [3 points]

Administration discovered their mistake and is working quickly to revoke your network modification privileges. The company has successfully restored the Friendarchy network to its original state (before you tampered with it). Before your privileges disappear, you decide to perform one more experiment on the original Friendarchy network. Instead of allowing everyone to maintain only their 10 strongest friendship links, now everyone is only allowed to maintain their friendship links numbered $i + 1$ through $i + 10$, for various values of $i \in \{0, 10, 20, 30, \dots\}$ (the numbering of friendships is in the order of friendship strength).

How do you expect the clustering and diameter of the network to behave as you increase the value of i?

(a) Clustering increases, diameter increases

(b) Clustering increases, diameter decreases

(c) Clustering decreases, diameter increases

(d) Clustering decreases, diameter decreases

Explain your choice.

1.3 [2 points]

(i) What is the expansion of the complete, balanced tree with $n = 2^{h+1} - 1$ nodes? Note: h is the height of the tree.

(ii) What is the expansion of the complete graph on n nodes?

1.4 [10 points]

In lecture, we showed that a random 3-regular graph has diameter $O(\log(n))$ with high probability. In this problem we will show that if $p = 2\frac{\ln(n)}{n}$, then with high probability $G_{n,p}$ contains a random 3-regular subgraph. Thus $G_{n,p}$ has diameter $O(\log(n))$ with high probability.

(i) Fix one of the nodes v. As we generate $G_{n,p}$, what is the probability that the degree of v is less than 3?

(ii) Show that the probability that there exists a node in $G_{n,p}$ with degree less than 3 is at most

$$n(1 + np + n^2p^2)(1 - p)^{n-3}.$$

Hint: use part 1 and a union bound.

(iii) If $p = 2\frac{\ln(n)}{n}$, show that the probability that $G_{n,p}$ contains a node having degree less than 3 goes to 0 as n goes to infinity. From this, conclude that with high probability $G_{n,p}$ has a diameter $O(\log(n))$. *(Hint: use part (ii) and the fact that* $1 - x < e^{-x}$ *for* $x > 0$*).*

What to submit

1.1 Optimal network structure and number of people you can reach.

1.2 Answer selection and your reasoning.

1.3 Expansion of two graphs.

1.4 (i) Probability that the degree is less than 3.

(ii) Short proof.

(iii) Short proof.

2 Signed Networks over Time [30 points - Justin, Christie]

2.1 [8 points (2 points each sub question)]

Consider the following simple model for constructing random signed networks, which we'll call the G^+ model. Start with a complete graph on n nodes. For each edge e mark its sign as positive with probability p (and thus negative with probability $1 - p$). All edges are undirected.

Let G_B denote the event that a graph G is balanced. In this question, we'll show that $P(G_B) \to 0$ as $n \to \infty$ for graphs generated according to the G^+ model. Assume that $p = 1/2$.

(i) Let T be a maximum set of disjoint-edge triangles in G. A "disjoint-edge" set of triangles is one in which every edge is in exactly one triangle. Give a simple lower bound for $|T|$, the number of triangles that don't share any edges in G (the bound should be an increasing function of n).

(ii) For any triangle in G, what is the probability that it is balanced?

(iii) Using the simple lower bound from part (i), give an upper bound on the probability that *all* of the triangles in T are balanced. Show that this probability approaches 0 as $n \to \infty$.

(iv) Explain why the last part implies that $P(G_B) \to 0$ as $n \to \infty$.

2.2 [5 points]

If balanced signed networks don't show up by chance, as we showed in the previous part of the question, how do they arise? One class of mechanisms that researchers have proposed and studied are *dynamic processes*, in which signed networks can evolve over time to become more balanced. The following describes a very simple example of such a dynamic process:

(I) Pick a triad at random.

(II) If it's balanced, do nothing.

(III) Otherwise, choose one of the edges uniformly at random and flip its sign (so that the triad becomes balanced).

Consider the following claim: in this process, the number of balanced triads can never decrease. Is this true? If so, give a proof, otherwise give a counterexample.

2.3 [8 points]

Now let's run simulations of this process on small networks.

(I) Create a complete network on 10 nodes.

(II) For each edge, choose a sign $(+, -)$ at random ($p = 1/2$).

(III) Run the dynamic process described in the previous part for 1,000,000 iterations.

Repeat this process 100 times (so that you run the dynamic process described in the previous part for 100 different graphs generated from the G^+ model). What fraction of the networks end up balanced? [Hint: To check whether a network is balanced or not, remember that a graph is balanced if and only if it's possible to separate the nodes into two "factions" such that every pair of nodes in the same faction is linked by a positive edge and every pair of nodes in different factions is linked by a negative edge. To speed things up, you can stop the process once the network is balanced, because once the network becomes balanced it will never change.]

2.4 [4 points]

Another way signed networks can evolve over time is if new nodes join the network and create new signed edges to nodes already in the network. Consider the network shown in Figure 1. Is it possible to add a node D such that it forms signed edges with all existing nodes (A, B, and C), but isn't itself part of any unbalanced triangles? Justify your answer.

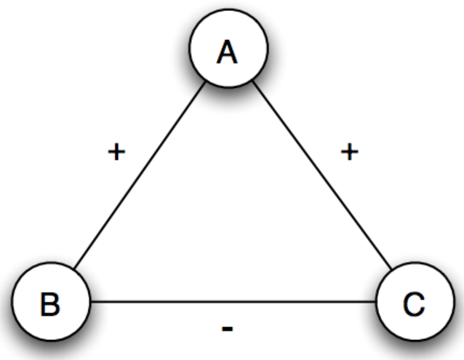

Figure 1: An unbalanced network with one triangle.

2.5 [5 points]

Using your answer to the previous part, consider the following question. Take any complete signed network, on any number of nodes, that is unbalanced. When (if ever) is it possible for a new node X able to join the network and form edges to all existing nodes in such a way that it does not become involved in any unbalanced triangles? If it's possible, give an example. If it's not, give an argument explaining why.

What to submit

2.1) (i) A simple lower bound for $|T|$. (ii) The probability that any triangle in G is balanced. (iii) An upper bound on the probability that all triangles in T are balanced, and show that the probability approaches 0 as $n \to \infty$. (iv) Explain why the initial claim holds.

2.2) Whether the number of balanced triads never decreases. A proof (if true) or a counterexample (if false).

2.3) Your code (both printed and submitted online). The fraction of networks that end up balanced.

2.4) Whether it is possible to add a node D, and a justification.

2.5) Whether it is possible for a new node X to join. An example (if possible) or argument (if impossible).

3 Decision Based Cascades: A Local Election [25 points – Yoni, Zhemin]

It's election season and two candidates, Candidate A and Candidate B, are in a hotly contested city council race in sunny New Suburb Town. You are a strategic advisor for Candidate A in charge of election forecasting and voter acquisition tactics.

Based on careful modeling, you've created two possible versions of the social graph of voters. Each graph has 10,000 nodes, each denoted by an integer id between 0 and 9999. The edge list for Graph 1 can be found at: `http://www.stanford.edu/class/cs224w/homeworks/hw2/g1.edgelist` and the edge list for Graph 2 can be found at: `http://www.stanford.edu/class/cs224w/homeworks/hw2/g2.edgelist`. Both graphs are **undirected**.

Given the hyper-partisan political climate of New Suburb Town, most voters have already made up their minds. 40% know they will vote for A, 40% know they will vote for B, and the remaining 20% are undecided. Each node's support is determined by the last digit of their id. If the last digit is 0-3, the node supports A. If the last digit is 4-7, the node supports B. And if the last digit is 8 or 9, the node is undecided.

The undecided voters will go through a 10-day **decision period** where they choose a candidate based on the majority of their friends. The decision period works as follows:

1. The graphs are initialized with every voter's initial state.
2. In each iteration, every undecided voter is assigned a vote. If the majority of their friends support A, they now support A. If the majority of their friends support B, they now support B. "Majority" for A means that more of their friends support A than the number of their friends supporting B, and vice versa for B.
3. If they have an equal number of friends supporting A and B, we assign A or B in alternating fashion, starting with A. This alternation happens at a global level for the whole network, across all rounds, so we break the first tie with A, the second with B, and so on (not on a per node basis). Keep a single global variable that keeps track of whether the current alternating vote is A or B, and initialize it to A in the first round. Then as you iterate over nodes in order of increasing ID, whenever you assign a vote using this alternating variable, change its value afterwards.
4. When updating the votes, use the values from the current iteration. So, for example, when updating the votes for node 10, you should use the votes for nodes 0-9 from the current iteration, and nodes 11 and onwards from the previous iteration.
5. The process described above happens 10 times.
6. On the 11th day, it's election day, and the votes are counted.

3.1 Basic setup and forecasting [5 points]

Read in the two graphs using software of your choice. Assign initial vote configurations to the network. Then, simulate the 10 day voting process. Which candidate wins in Graph 1, and by how many votes? Which candidate wins in Graph 2, and by how many votes?

3.2 TV Advertising [8 points]

Luckily you've amassed a substantial war chest of 9,000 dollars. You decide to spend this money by showing ads on the local news. Unfortunately only 100 New Suburb Townians watch the local news, those with ids 3000-3099, but your ads are extremely persuasive, so anyone who sees the ad is immediately swayed to vote for candidate A. For each 1,000 dollars you spend, you reach 10 additional voters, starting with 3000-3009. This advertising happens before the decision period.

Now, you will simulate the effect of advertising spending on the two possible social graphs. First, you'll read in the two graphs again, and assign the initial configurations as before. But now, before the decision process, you will run your k dollars worth of ads, then you will go through the decision process of counting votes.

For each of the two social graphs, please plot k (the amount you spend) on the x-axis (for values 1000, 2000, ..., 9000) and the number of votes you win by on the y-axis (this is a negative number if you lose). What's the minimum amount you can spend to win the election in each of the two social graphs?

3.3 Wining and Dining the High Rollers [8 points]

TV advertising is only one way to spend your campaign war chest, however. You have another idea to have a very classy $1000 per plate event for the high rollers of New Suburb Town (the people with the highest degree in the social graph). You invite high rollers in order of how many people they know, and everyone that comes to your dinner is instantly persuaded to vote for candidate A. This event will happen before the decision period. When there are degree ties, the rollers with lowest node ID get chosen first.

Now, you will simulate the effect of the high roller dinner on the two graphs. First, read in the graphs and assign the initial configuration as before. Now, before the decision process, you will spend k dollars on the fancy dinner, then you will go through the decision process of counting votes.

For each of the two social graphs, please plot k (the amount you spend) on the x-axis (for values 1000, 2000, ..., 9000) and the number of votes you win by on the y-axis (this is a negative number if you lose). What's the minimum amount you can spend to win the election in each of the two social graphs?

3.4 Analysis [4 points]

Briefly (1-2 sentences) explain some property of the 2 graphs that can explain the results of the last 2 questions. Also, explain how you'll spend your 9,000 dollars and why (there's no specific answer we're looking for here, just come up with a reasonable strategy and explain it in 1-2 sentences).

What to submit

3.1) Which candidate wins and by how many votes. Your code (both printed and submitted online).

3.2) The two plots and minimum amount you can spend to win the election in each of the two social graphs. Your code (both printed and submitted online).

3.3) The two plots and minimum amount you can spend to win the election in each of the two social graphs. Your code (both printed and submitted online).

3.4) The explanations.

4 Complex Contagions [25 points – Ashwin, Bell]

In this question we will explore one way in which network topology can affect sensitivity to cascading behavior. Specifically, we will see how the standard picture of graph connectivity is somewhat misleading in cases where the contagion requires that an individual be exposed to several other adopters before she herself adopts. Contagions of this sort are called *complex*. In contrast to *simple* contagions, which only require a single exposure in order to activate an individual, the dynamics of complex contagions require that a dense set of activated nodes persist throughout the cascade. In highly clustered, real-world networks, if the activated set becomes too widely distributed, this will diminish the probability that an unactivated node is sufficiently connected to this set. For example, this sort of model provides one hypothesis for why certain cultural behaviors tend to spread geographically (rather than over the internet, say) due to the presence of continuously adjacent neighborhoods.

We formalize the notion of a contagion's *complexity* as the number of neighbors, a, who must already be infected in order for a new node to be infected. This is similar to the game-theoretic model of cascades which we explored in class except that here we are interested in the absolute number of activated neighbors rather than the proportion.

To examine this phenomenon we will use a regular, highly clustered ring lattice in which each node is connected to its k nearest neighbors. You can assume that k is even to simplify the proofs. Also, assume that $N \gg k$, where N is the number of nodes in the ring.

Due to the requirement of multiple exposures in the spread of complex contagions it is useful to think about *neighborhoods* of nodes infecting one another rather than single nodes infecting one another. We define a *neighborhood* to just be a focal node along with its adjacent nodes (or equivalently, a node's ego network). Consider the scenario depicted in Figure 2 which shows the neighborhoods of i and l, which we will denote as I and L respectively. The black and gray-black nodes correspond to I and the gray and gray-black nodes correspond to L.

We define a bridge from one neighborhood to another as the set of edges which connect the intersection of those two neighborhoods to the nodes of the second neighborhood which are not in the first. Formally for neighborhoods, A and B the bridge from A to B is the set of edges which connect the intersection of the two sets with the difference of the second set minus the first, or

$$R_{AB} = \{(u, v) | u \in (A \cap B), v \in (B - A)\}$$

The *width* of this bridge is just number of edges in that set, $W_{AB} = |R_{AB}|$. The notion of the width of a bridge will be useful for analyzing the spread of complex contagions.

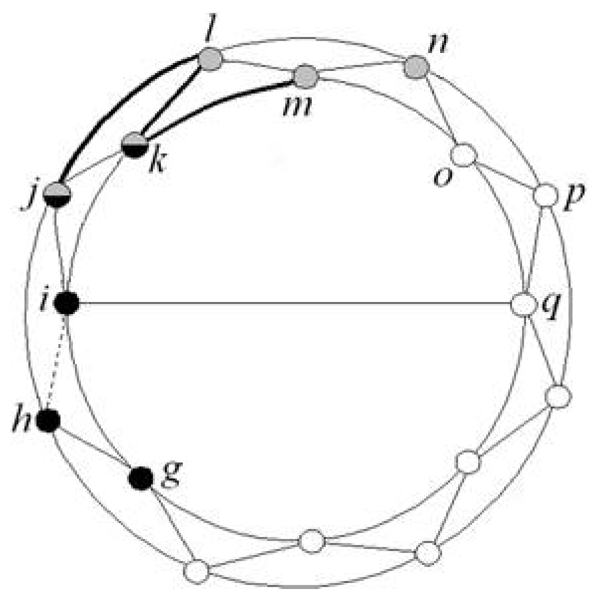

Figure 2: Diagram of a neighborhood infection event on a ring graph in which each nodes is connected to its 4 nearest neighbors.

4.1 [5 points]

Find an expression in terms of k (the number of neighbors of each node) for the maximum overlap between two neighborhoods, such that the focal node in each neighborhood is not a part of the other neighborhood (For example, neighborhoods I and K do not satisfy this).

4.2 [10 points]

Consider the two neighborhoods I and L in the 4-regular ring network depicted in Figure 2. The width of the connecting bridge, W_{IL}, is 3, which we can see from the emboldened edges. For a regular ring network in which each node is connected to its k nearest neighbors, find an expression for the maximal width of a bridge, called W_{max}, between any two neighborhoods in terms of k.

4.3 [5 points]

Now, consider what we will call the *critical width*, W_C, defined as the minimum number of "nonredundant" ties required for a contagion to propagate to an unactivated neighborhood. Ties are nonredundant so long as there are no more than a bridge ties to a single member of $B - A$, where a is the number of neighbors who must already be infected in order for a new node to be infected. For example, in the given figure, suppose $a = 2$, *critical width* is 3. If there were three bridge ties to any grey node, then one of these ties would be redundant. Find an expression for the critical width on a k-regular ring network as a function of a. (*Hint: Think of this part independently of the value of* k)

4.4 [5 points]

What relation should hold between W_C and W_{max} so that the contagion is unable to spread through the network. Using this relation, derive the maximum value of a so that the contagion can spread in a k-regular ring graph?

What to submit

4.1) Write the expression in terms of k and a short explanation.

4.2) Write the expression for W_{max} and a short explanation.

4.3) Write the expression for W_C and a short explanation.

4.4) Write expressions for relation between W_C and W_{max}, and the maximum value of a.