CS224W: Social and Information Network Analysis Autumn 2013

Problem Set 1

Due 9:30am October 10, 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 Fighting Reticulovirus Avarum [20 points – Christie, Ashley, Justin]

One of the main applications of network analysis is the study of how viruses spread. By modeling hosts as nodes and host A infecting host B as a directed edge from A to B, we have a very natural and useful way to represent the spread of a virus as a network. To understand how a virus might spread, one first needs to understand the underlying network on which it is spreading.

In this problem, we will apply the network analysis techniques that you've learned in class to gain insight into how an email virus might propagate. We concentrate on email viruses because we have very accurate information on the underlying networks through which they spread. The data we'll use is an anonymous email network (http://www.stanford.edu/class/cs224w/data/email_network.txt).

An evil TA has written a nasty email virus called Reticulovirus avarum (R. avarum) that can propagate through this network. Your job is to analyze the worst-case scenario to obtain an upper bound on how bad the spread could be: Our working assumption will be that once R. avarum infects a host, it always infects all of the host's contacts if they are not already infected (probability of infection given contact is 100%).

(a) [**2 points**] If you know the first node to contract the virus, which network analysis concept captures the set of nodes that will eventually also be infected? In other words, given an initially infected node v, how would you define the set of nodes, in terms of v, that will be eventually infected?

(b) [**5 points**] In this part, we'll compute a simplified version of the bow-tie structure of the email network. How big (as a percentage of the nodes in the graph) are: (1) the largest SCC, (2) the in-component of the largest SCC, (3) the out-component of the largest SCC, and (4) the disconnected components (the components that aren't connected to the "bow-tie" components you just computed). Don't worry about the tendrils or tubes. (Hint: see http://snap.stanford.edu/class/cs224w-readings/broder00bowtie.pdf.)

(c) [**5 points**] Assume R. avarum is only introduced once into the email network. Assuming that the initially affected node is chosen uniformly at random, what is the probability that R. avarum becomes a large-scale epidemic? Assume "large-scale" means at least 30% of the population gets infected. Explain your reasoning. (Note: use *only* the numbers you computed in part (b). No additional computation is needed to answer this question. Ignore tendrils, tubes and the disconnected components.)

(d) The evil TA modified his virus to spread through Twitter instead. Consider an extremely simplified representation of the Twitter network which has 100 million nodes. The SCC has 40 million nodes, the out-component has 30 million nodes, and the in-component has 20 million nodes. The disconnected components of the Twitter network consist of 10 million nodes. There are no tendrils or tubes. Again, assume a node always infects all of its contacts[1].

(i) [**2 points**] Suppose R. avarum infects a node in the SCC. What is the resulting outbreak size? Explain.

(ii) [**3 points**] The NSA wants to know what the worst-case outbreak size of R. avarum is. Knowing only the component sizes of the Twitter network, you argue that you can only answer their question if you can make some assumptions about the network's structure.

Suppose R. avarum can now infect any node in the network. What's the worst possible outbreak size? Give a description of a network structure satisfying the component sizes from above (SCC, in-component, out-component, and disconnected components), and a starting node for the outbreak that would lead to this epidemic size. You can use a diagram in your explanation if that helps.

(iii) [**3 points**] Happy with your work from (ii), the NSA has given you permission to modify the edges of the Twitter network to minimize the spread of R. Avarum.

[1] Or followers, in Twitter-speak.

They now want you to reduce the size of the worst-case outbreak by at least 10 million nodes, but are only letting you remove one edge. You again argue that you need to make assumptions about the network's structure to answer this question.

Give a description of a network structure satisfying the component sizes from above, where removing a single edge reduces the size of the worst-case outbreak of R. Avarum by at least 10 million nodes. Identify an edge in this network, that when removed, achieves this reduction in worst-case outbreak size, and state the reduction in outbreak size in this case. This network structure does not need to be the same as the one you described in (ii). You can use a diagram in your explanation if that helps. Again, no computation is needed to answer this question; thoughtfully reasoned answers will receive full credit.

What to submit

(a) State the network analysis concept.

(b) The size (as a percentage of total nodes) of the largest SCC, in-component, out-component, and disconnected components. Print out your code and also submit it online.

(c) The probability that R. avarum becomes a large-scale epidemic and your reasoning.

(d) (i) The resulting outbreak size and explanation. (ii) A description of a possible network structure, the node at which the outbreak begins, and the worst-case outbreak size. (iii) A description of a possible network structure, a possible edge that can be removed, and the reduction in the worst-case outbreak size.

2 Network Characteristics [35 points – Yoni, Ashwin]

One of the goals of network analysis is to find mathematical models that characterize real-world networks and that can then be used to generate new networks with similar properties. In this problem, we will explore two famous models—Erdős-Rényi and Small World—and compare them to real-world data from an academic collaboration network. Note that in this problem all networks are *undirected*.

- $G(n, m)$ *Random Network:* To construct this undirected network, use $n = 5242$ nodes and pick $m = 14496$ edges at random.
- *Small-World Network:* You can generate this undirected graph as follows: Begin with $n = 5242$ nodes arranged as a ring, i.e., imagine the nodes form a circle and each node is connected to its two direct neighbors (e.g., node 399 is connected to nodes 398 and 400), giving us 5242 edges. Next, connect each node to the neighbors of its neighbors (e.g., node 399 is also connected to nodes 397 and 401). This gives us another 5242

edges. Finally, randomly select 4012 pairs of nodes not yet connected and add an edge between them. In total, this will make $m = 5242 \cdot 2 + 4012 = 14496$ edges.

- *Real-World Collaboration Network:* You can download this *undirected* network from http://snap.stanford.edu/data/ca-GrQc.txt.gz. Note that some edges may appear twice in the data, once for each direction. You should consider edges as undirected and store at most one edge between any pair of nodes, for a total of 5,242 nodes and 14,496 edges. (Don't worry if you get 14,490 edges, as there are 6 self-loops in the graph).

(a) Degree Distribution [10 points]

Generate or read in the above three networks. Plot and compare the log-log degree distribution[2] of all three networks. Superimpose the plots of the three networks. Briefly describe the shape of each distribution and explain the differences between the networks.

(b) Excess Degree Distribution [15 points]

An important concept in network analysis is the *excess degree distribution*, denoted as q_k, for $k \geq 0$. Intuitively, q_k gives the probability that a randomly chosen edge goes to a node of degree $k+1$. Excess degree can be calculated as follows:

$$q_k = \frac{q'_k}{\sum_i q'_i}, \qquad q'_k = \sum_{i \in V} \sum_{(i,j) \in E} I_{[k_j = k+1]},$$

where $I_{\text{condition}} = 1$ when condition is true and 0 otherwise. V denotes the set of nodes, E the set of edges and k_j the number of neighbors of node j (equivalently, the degree of node j). Additionally, the *expected excess degree* is defined as $\sum_{k \geq 0} k \cdot q_k$. The degree distribution of the network is denoted by $\{p_k | k \geq 0\}$, i.e., p_k is the proportion of nodes having degree exactly k. The *expected degree* can consequently be computed as $\sum_{k \geq 0} k \cdot p_k$.

Your Tasks:

- **(i) [10 points]** Plot and compare the log-log excess degree distributions of all three networks. Report the expected degree and the expected excess degree (the formulæ are given above) for each network. Also, focus on the tail of the distributions (large degrees) and briefly explain how and why these plots differ from the degree distribution plots (especially for the collaboration network) that you obtained in part (a).
- **(ii) [5 points]** Can you calculate the excess degree distribution $\{q_k\}$ using the degree distribution $\{p_k\}$? Find this formula.

[2] Generate a plot with the horizontal axis representing node degrees and the vertical axis representing the proportion of nodes with each degree. By log-log we mean that both the horizontal and vertical axis must be in logarithmic scale.

(c) Clustering Coefficient [10 points]

Recall that the local clustering coefficient for a node v_i was defined in class as

$$C_i = \frac{2|e_i|}{k_i \cdot (k_i - 1)},$$

where k_i is the degree of v_i and e_i is the number of edges between the neighbors of v_i. Find the *average clustering coefficient* for all the three networks, which is defined as

$$C = \frac{1}{|V|} \sum_{i \in V} C_i.$$

(if a node has 0 or 1 neighbor, we ignore it). Compare the average clustering coefficient of the three network types and explain why it is in the order you see. For this question, write your own implementation to compute the clustering coefficient, instead of using a built-in library function.

What to submit:

(a) Log-log degree distribution plot for all three networks (on the same graph). Brief description of shape of each distribution and explanation of differences. Print out your code, and also submit it online.

(b) (i) Log-log excess degree distribution plot for all three networks (on the same graph). Expected degree and expected excess degree for each network. Comparison against plots from part (a). Print out your code, and also submit it online.

(ii) Formula for calculating $\{q_k\}$ using $\{p_k\}$.

(c) Clustering coefficient for each network. Brief comparison of the values and reasoning for the order. Print out your code, and also submit it online.

3 Decentralized Search [45 points – Zhemin, Peter, Bell]

In class, we saw a decentralized search algorithm based on geography that seeks a path between two nodes on a grid by successively taking the edge towards the node closest to the destination.

Here we will examine an example of a decentralized search algorithm on a network whose nodes reside in the leaves of a tree. The tree may, for instance, be interpreted as representing the hierarchical organization of a university where one is more likely to have friends inside the same department, a bit less likely in the same school, and the least likely across schools.

Let us organize students at Stanford into a tree hierarchy, where the root is Stanford University and the second level contains the different schools (engineering, humanities, etc.). The third level represents the departments and the final level (i.e., the leaves) are the Stanford students. Tom, a student from the computer science department, wants to hang out with Mary, who is in sociology. If Tom does not know Mary, he could ask a friend in the sociology department to introduce them. If Tom does not know anybody in the sociology department, he may seek a friend in the Stanford humanities school instead. In general, he will try to find a friend who is "close" to Mary in the tree.

There are three parts in this problem. The first two parts explore an effective decentralized search algorithm on the hierarchical model in a specific setting. The third part involves simulation experiments on the model under a more general setting. There are many subproblems, but do not panic. A good understanding of the math involved in the lattice model we covered in class will make this problem easy and shorten each subproblem to just a few line proofs.

(a) Basic Tree Properties [15 points]

This part covers some basic facts of the setting of the hierarchical model and defines the notion of tree based "distance" between the nodes.

Consider a complete, perfectly balanced b-ary tree (each non-leaf node has b children and $b \geq 2$) T, and a network whose nodes are the leaves of T (the red nodes in the picture). Let the number of network nodes (equivalently, the number of leaves in the tree) be N and let $h(T)$ denote the height of T.

One important thing to keep in mind: In this problem, the network nodes are only the leaf nodes in the tree. Other nodes in the tree are virtual and only there to determine the edge creation probabilities between the nodes of the network.

(i) [5 points] Write $h(T)$ in terms of N.

(ii) [5 points] Next we want to define the notion of "tree distance." The intuition we want to capture is that students that share the same department are closer than for example students sharing schools. For instance, in the tree in Figure 1 nodes u and t are "closer" than nodes u and s. We formalize the notion of "distance" as follows:

Given two network nodes (leaf nodes) v and w, let $L(v, w)$ denote the subtree of T rooted at the lowest common ancestor of v and w, and $h(v, w)$ denote its height (that is, $h(L(v, w))$). In Figure 1, $L(u, t)$ is the tree in the circle and $h(u, t) = 2$. Note that we can think of $h(v, w)$ as a "distance" between nodes v and w.

For a given node v, what is the maximum possible value of $h(v, w)$?

(iii) [5 points] Given a value d and a network node v, show that there are $b^d - b^{d-1}$ nodes satisfying $h(v, w) = d$.

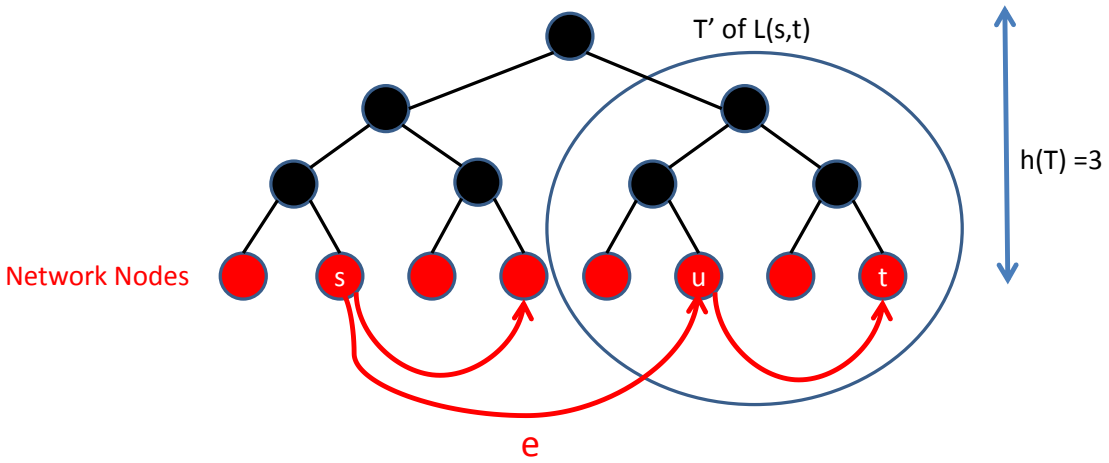

Figure 1: Illustration for Question 3 of the hierarchical graph. Black nodes and edges are used to illustrate the hierarchy and structure, but are not part of our network. Red nodes (leaf nodes) and red edges are the ones in our network. The lowest common ancestor of s and t is the root of the tree. The decentralized search proceeds as follows. Denote the starting node by s and the destination by t. At each step, the algorithm looks at the neighbors of the current node s and moves to the one "closest" to t, that is, the algorithm moves to the node with the lowest common ancestor with t. In this graph, from s we move to u.

(b) Network Path Properties [20 points]

This part helps you design a decentralized search algorithm in the network.

We will generate a random network on the leaf nodes in a way that models the observation that a node is more likely to know "close" nodes than "distant" nodes according to our university organizational hierarchy captured by the tree T. For a node v, we define a probability distribution of node v creating an edge to any other node w:

$$p_v(w) = \frac{1}{Z} b^{-h(v,w)}$$

where $Z = \sum_{w \neq v} b^{-h(v,w)}$ is a normalizing constant. By symmetry, all nodes v have the same normalizing constant.

Next, we set some parameter k and ensure that every node v has exactly k outgoing edges. We do this with the following procedure. For each node v, we repeatedly sample a random node w according to p_v and create edge (v, w) in the network. We continue this until v has exactly k neighbors.Equivalently, after we add an edge from v to w, we can set $p_v(w)$ to 0 and renormalize with a new Z to ensure that $\sum_w p(w) = 1$. This results in a k-regular directed network.

(iv) [5 points] Show that $Z \leq \log_b N$. (Hint: use the results in (ii) and (iii).)

(v) [5 points] For two leaf nodes v and t, let T' be the subtree of $L(v, t)$ satisfying:

- T' is of height $h(v,t)-1$,
- T' contains t,
- T' does not contain v.

For instance, in Fig. 1, T' of $L(s,t)$ is the tree in the circle.

Let us consider node v and an edge e from v to a random node u sampled from p_v. We say that e points to T' if u is a leaf node of T'. Show that the probability of e pointing to T' is no less than $\frac{1}{b\log_b N}$.

(vi) [5 points] Let the out-degree k for each node be $c\cdot(\log_b N)^2$, where c and b are constants. Show that when N grows very large, the probability of v not having any edge pointing to T' is asymptotically no more than $N^{-\theta}$, where θ is a positive constant which you need to compute.

(Hints: Use the result in (v) and recall that each of the k edges is independently created. Also, use $\lim_{x\to\infty}(1-\frac{1}{x})^x=\frac{1}{e}$.)

Argue why the above result indicates that for any node v, we can, with high probability, find an edge to a (leaf) node u satisfying $h(u,t)<h(v,t)$.

(vii) [5 points] Show that starting from any (leaf) node s, within $O(\log_b N)$ steps, we can reach any (leaf) node t. You do not need to prove it in a strict probabilistic argument. You can just assume that for any (leaf) node v, you can always get to a (leaf) node u satisfying $h(u,t)<h(v,t)$ and argue why you can reach t in $O(\log_b N)$ steps.

(c) Simulation [10 points]

In (i) to (vii), we have set the theory to find an efficient decentralized search algorithm, assuming that for each edge of v, the probability of it going to w is proportional to $b^{-h(v,w)}$. Now we experimentally investigate a more general case where the edge probability is proportional to $b^{-\alpha h(v,w)}$. Here $\alpha>0$ is a parameter in our experiments.

In the experiments below, we consider a network with the setting $h(T)=10$, $b=2$, $k=5$, and a given α. That is, the network consists of all the leaves in a binary tree of height 10; the out degree of each node is 5. Given α, we create edges according to the distribution described above.

(viii) [7 points] Create random networks for $\alpha=0.1,0.2,\ldots,10$. For each of these networks, sample 1,000 random (s,t) pairs ($s\neq t$). Then do a decentralized search starting from s as follows. Assuming that we are currently at (leaf) node s, we pick its neighbor u (also a leaf node) with smallest $h(u,t)$ (break ties arbitrarily). If $u=t$, the search succeeds. If $h(s,t)>h(u,t)$, we set s to u and repeat. If $h(s,t)\leq h(u,t)$, the search fails.

For each α, pick 1,000 pairs of nodes and compute the average path length for the searches that succeeded. Then draw a plot of the average search time (number of steps it takes to

reach t) as a function of α. Also, plot the search success probability as a function of α.

(ix) [**3 points**] Briefly comment on the plots and explain the shape of the curve.

What to submit

(i) Write the expression and a short explanation.

(ii) Write the expression and a short explanation.

(iii) Write a short proof.

(iv) Write a short proof.

(v) Write a short proof.

(vi) Write a short proof, give an expression for θ and brief argument.

(vii) Write a short proof.

(viii) Print out your code and also submit it online. Print out both plots.

(ix) Write a brief comment.