MIDTERM EXAMINATION
Networked Life
CIS 112
Spring 2008
March 6, 2008
Prof. Michael Kearns

This is a closed-book exam. You should have no material on your desk other than the exam itself and a pencil or pen.

Name: _____

Penn ID: _____

Problem 1: _____/20

Problem 2: _____/15

Problem 3: _____/15

Problem 4: _____/10

Problem 5: _____/15

Problem 6: _____/10

Problem 7: _____/15

TOTAL: _____/100

Problem 1 (20 points) Indicate whether each of the follow statements is true or false.

(a) Recall the “contagion” model of economic exchange discussed in class: Given an undirected connected graph, each vertex begins with an equal amount of currency. At each step, every vertex divides its current wealth equally among its neighbors. As this process is repeatedly infinitely, wealth will be distributed uniformly across the vertices.
FALSE

(b) The frequency of English words appearing in the New York Times over the past 20 years is well-approximated by a power law distribution.
TRUE

(c) "Having a diameter greater than 6" is a monotone property of a graph.
FALSE

(d) The Alpha Model with a large value for the parameter alpha corresponds to Watt's “Solaria” world.
TRUE

(e) In Kleinberg's model of navigation in social networks there is a parameter r governing the distribution of “long-distance” connections. Rapid search is not possible in this model when r = 2 because the long-distance connections will actually not travel very far.
FALSE

(f) The maximum number of edges that an undirected graph with N vertices can have is N(N+1)/2.
FALSE

(g) The PageRank of a web page can be interpreted as the probability that a certain kind of web surfer will visit that page.
TRUE

(h) All monotone graph properties exhibit tipping behavior in the Erdos-Renyi model.
TRUE

(i) “Organic” search results refer to those ranked according to objective criteria that cannot be easily manipulated.
FALSE

(j) In matters of spatial distribution, it is relatively straightforward to infer individual preferences from collective behavior.

FALSE

Problem 2 (15 points)

(a) Explain in words the primary differences between a normal or Poisson distribution, and a power law distribution.

Correct comments: Normal/Poisson sharply peaked around mean, has exponential decay away from mean, unlikely to draw values far from mean, etc; power law has long tails, slow decay away from mean, likely to draw values many times larger than the mean, etc.

(b) Describe the standard way of testing whether a set of sampled data points are better fit by a normal or Poisson distribution, or by a power law distribution. Feel free to illustrate your description with diagrams.

Plot y-axis equal to log of the value of the quantity in question (e.g. degree), x-axis equal to log of the number of observations, frequency, probability, etc., with that value. Then power laws will appear (nearly) linear with negative slope equal to the power, while Normal/Poisson will have high curvature away from a line. Plots showing these two cases would be appropriate.

(c) Name a few naturally occurring data sources that empirically seem to obey power law distributions.

Many examples from class: degrees in a social network, North American city sizes, distances traveled by dollar bills, file sizes on a computer, etc.

Problem 3 (15 points) Briefly but precisely describe the main definitions and ideas behind Kleinberg's "Hubs and Authorities" algorithm and the PageRank algorithm. Describe conditions or examples for which you think the two algorithms would disagree on the importance of a page.

Here either pseudo-code for both algorithms as was given in class, or the English descriptions that led to them are satisfactory. E.g., A good authority is a page which is pointed to by a lot of good hubs, and a good hub is a page that points to a lot of good authorities. A page with high PageRank is one which is linked from pages with high PageRank. Both types of answers are equally acceptable.

A page that has many outlinks but no inlinks would be a good candidate for a page that would have high hub weight but low PageRank. Similarly, a page linked from a few such hubs but no other pages would have high authority weight but low PageRank. Other differences can arise due to the fact that in PageRank, a page "p" gives 1/N(p) of its rank to each of the pages it links to, whereas in Hubs & Authorities, a hub p gives h(p) to each of the pages it links to. In dividing by N(p), PageRank discounts transmitted rank from

pages that have many links, whereas Hubs and Authorities does not. This could certainly cause discrepancies.

Problem 4 (10 points) Consider any of the network formation models we discussed in which both "local" and "long-distance" edges are present. Describe such a model as precisely as you can, and discuss what aspects of the real world such a model is intended to capture. Discuss which of the following properties networks generated by the model will and will not have: heavy-tailed degree distribution, small diameter, and high clustering coefficient.

We expect you to describe a model like Kleinberg's, where you start off with a grid (or a line or a cycle) and then add either uniformly random long-distance edges, or long-distance edges added according to a power-law distribution; also fine if instead of adding long-distance edges we instead "rewire" grid edges with fixed probability (as in the early epidemic demo).

For these models, the properties entailed would be small diameter and high clustering, but not heavy-tailed degrees.

5 points for clearly describing the model; 2 point for describing real world aspects the model captures; 3 points for listing the correct properties. Points will be deducted for confusing/conflating models; if you list multiple models, points will be deducted for those that do not meet the criteria specified in the problem.

Unacceptable answers here include the alpha model and Erdoes-Renyi model. However, in case you list one of these unacceptable models but give a reasonable argument that there will indeed be a mixture of distances about these models, a maximum of 5 points can be given. (A reasonable argument must clearly indicate that because in both models, there is a significant probability p with which both long distance connections and local connections are equally likely to be added.)

Models like preferential attachment are simply wrong and 0 point is given. (In the preferential attachment model, many nodes have degree only one!)

Problem 5 (15 points) Consider the classic paper "An Experimental Study of the Small World Problem" by Travers and Milgram. Briefly present and defend any three criticisms of the methodology or findings of this paper.

The following critiques are among the legitimate answers:

1. Small overall sample size --- very few initial parties even contacted by T&M

2. Even smaller completion rate --- only 64 chains completed

3. Bias introduced by failing to account for chains that never complete

4. Conflation of existence of short paths and their discovery by subjects from only local info, which is not pointed out until Kleinberg

5. Lack of analysis of how people decided who to forward the letter to, an issue analyzed in the Dodds et al study

6. Sample biased towards families of higher income.

7. Lack of incentives for subjects to actually try his/her best to find the shortest path.

8. The conclusion that connectors played a major role is probability an artifact of the small sample size of the experiment.

You need to list 3 critiques, 5 points for each correct critique.

Problem 6 (10 points) Consider a search engine which offers both "organic" and "sponsored" search results in response to user queries, and suppose the sponsored search results are listed in the order of the bid-per-click made by advertisers on the search term.

(a) Describe ways in which a person or organization might try to "game" the organic search results --- that is, manipulate the ranking algorithm used by the search engine in order to make their page appear higher in the organic results. You are free to draw on the class discussion of the common elements in organic search employed by search engines such as Google and Yahoo!.

For either (a) or (b), you receive 5 points if you list at least two correct methods. You receive 2.5 points if you list only one correct method. And beyond that, 1 point is deducted for each wrong answer.

For part (a), possible answers include:

1. Gaming PageRank by arranging to have lots of pages point to yours, possibly by creating those pages yourself

2. Adding hidden terms via tags in the html

3. Having terms on the page that don't have anything to do with its content but are just popular search terms

4. Typo squatting --- having your page contain common typos so that your page will appear in response to searches on them (this applied to (b) below as well)

5. Adding a lot of links from one's page to other important but irrelevant web sites

6. Hire a search engine optimization company (SEO) for you.

7. Trading links with some reputable sites

(b) Describe ways in which a person or organization might try to "game" the sponsored search results --- that is, to cause their page to appear higher in the sponsored results, to cause their advertising costs to be lower, or to cause the advertising costs of their competitors to be higher.

Possible answers include:

1. Bid jamming --- making your bid as close to the one above you without exceeding it

2. Clicking on competitors ads in order to cost them advertising budget

3. Typo-squatting, or put false description to lure customers
4. Bid search queries totally unrelated to their business in order to increase exposure.

5. Hire a search engine marketing (SEM) company for you.

Problem 7 (15 points) This problem considers the assigned paper "Graph Structure in the Web" by Broder et al.

(a) Briefly but precisely name and describe the five regions of the web identified in the paper.

Strongly Connected Component (SCC): group of web pages such that for any two pages A and B in the SCC, there exists a directed path of hyperlinks from A to B.

IN: pages for which there exists a directed path of hyperlinks leading to the SCC but not vice versa.

OUT: pages that can be found by following a directed path of links from the SCC but not vice versa.

TENDRILS: pages which can not reach the SCC or be reached from the SCC via directed hyperlink paths. These are pages for which a directed path leads to OUT, or for which a directed path leads from in IN.

DISCONNECTED: pages for which there are no paths to or from any of the pages in the weakly connected component (WCC) = SCC + IN + OUT + TENDRILS.

(b) Which of the five regions does a page belong to the very moment it is created, and why?

The newly created page presumably has no hyperlinks pointing to it since it is brand new. Therefore it cannot be in SCC or OUT. However, any of the other three regions are

possible depending on whether it has no outbound hyperlinks (DISCONNECTED), outbound hyperlinks pointing to IN or the SCC (IN), or outbound hyperlinks point to OUT (TENDRILS).

(c) In what ways is the creator or author of a web page in control of which of the five regions their page falls? In what ways are they not in control?

A web page author controls what pages her page links to but not what pages link to her page. If no pages in the SCC link to her page, the author can ensure her page belongs to IN by linking to pages in the SCC or to pages in IN. If pages in the SCC link to her page, the author can ensure her page belongs to OUT by not linking to pages in SCC or IN.

Similarly, if her page links to a page in IN or SCC, she cannot control whether her page is linked to. Thus she cannot control whether her page ends up in IN or SCC. And if her page does not link to pages in IN or SCC, she cannot control whether her page lands in DISCONNECTED, OUT or even TENDRILS in the case that a page in IN links to her page.