The Structure of Information Networks	**Homework 2**
CS 6850 Spring 2011	**Due Friday, February 25, 2011**

(1) Consider the following abstract model for a network that implements a kind of associative memory. There is a set of *sensor nodes* and a set of *storage nodes*; for simplicity, we will assume that each of these two sets has the same size, denoted n. There is a bipartite graph G on the sensor and storage nodes: each edge has one end equal to a sensor node and one end equal to a storage node. Now, every time a sensor node records an observation, it writes it to each of the storage nodes to which it has an edge; and similarly, each sensor node is able to read all the observations recorded at any storage node to which it has an edge.

Given this set-up, we would like the graph G to have the following property:

> *($*$) For every two sensor nodes u and v, there is some storage node w such that both u and v have an edge to w.*

If property ($*$) holds, then every sensor node will be able to read all the observations recorded by all the other sensor nodes. On the other hand, if property ($*$) fails to hold, then there are sensor nodes u and v such that neither can read the observations of the other. Thus, ($*$) is a crucial property for maintaining access to information by all sensor nodes in our network-based memory structure.

One way to achieve this property would be to have G be a complete bipartite graph, with an edge from every sensor node to every storage node. However, we can achieve property ($*$) by a structure where no node of G takes part in nearly this many edges, and one way to do this is via a random construction.

Suppose that for some exponent α, we connect each sensor node and each storage node by an edge independently with probability $n^{-\alpha}$. Give a value α^* that is *critical* for property ($*$) in the following sense:

> *If $\alpha > \alpha^*$ then property ($*$) holds with probability converging to 0 as $n \to \infty$; and if $\alpha < \alpha^*$ then property ($*$) holds with probability converging to 1 as $n \to \infty$.*

Give a justification for your answer. *(Hint: you can use the fact that for any constants $c > 0$ and $\varepsilon > 0$, it is the case that $\lim_{n\to\infty} n^c e^{-n^\varepsilon} = 0$.)*

(2) Consider a standard k-round, single-elimination tournament, such as you see in championship competitions for a number of different sports. Specifically, $n = 2^k$ contestants start out at the beginning, and in each round the current contestants play each other in specified pairs, with the winners moving on to the next round. An example with $k = 3$ and $2^3 = 8$ initial contestants is depicted in Figures 1 and 2. One can also picture this as a complete binary tree with the initial contestants at the leaves, and each internal node corresponding to a match-up of two surviving contestants.

Suppose that a group of people get together to bet on the results of the tournament, predicting the outcome of each match. If you're taking part in this group, you get a copy of

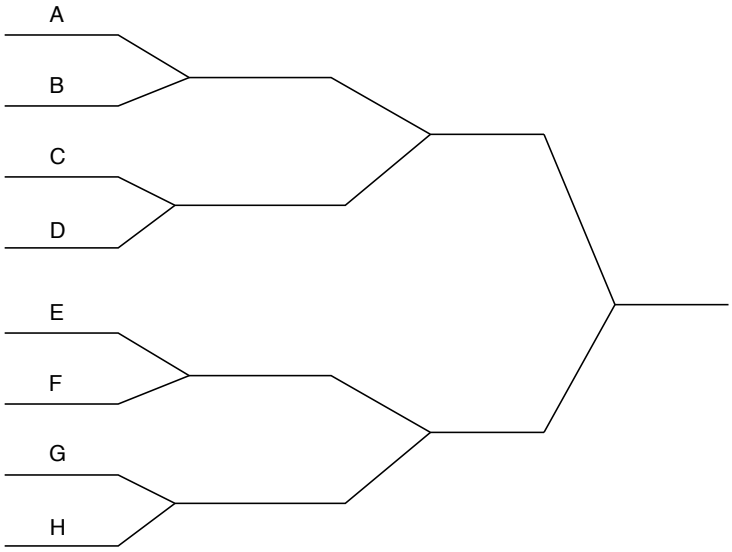

Figure 1: An elimination tournament with just the initial contestants, before any matches have been played.

a blank table with just the initial contestants filled in (as in Figure 1), and you're asked to fill in all the entries for the subsequent rounds. Crucially, all entries must be filled in before any matches are played. The filling-in should be consistent, in that if you write a person's name as a winner in round j, you should also have guessed that they're a winner in the previous round $j-1$.

Let's consider how well we'd expect someone to do at guessing these results, if they had no information about any of the contestants and were just guessing at random. We'll model this lack of information by assuming that after they fill in their table, each match's outcome is determined by an independent fair coin flip. (So each contestant is equally likely to win and advance to the next round, where their fate will be determined by the next coin flip).

Let $p(j)$ denote the probability that at least one of a person's guesses for the entries in round k turn out to be correct. That is, we look at all their guesses for the contestants in round j, and see if any of these contestants actually made it to round j.

Is there a value α^* that is *critical* for this probability in the following sense?

If $\alpha < \alpha^*$ *then*

$$\lim_{n\to\infty} p(\alpha \log n) = 1,$$

while if $\alpha > \alpha^*$ *then*

$$\lim_{n\to\infty} p(\alpha \log n) = 0.$$

(Recall that $k = \log n$; all logarithms here are base 2.) In your answer, either specify such a value of α^* and justify why it is critical, or argue why there is no critical value of α^*.

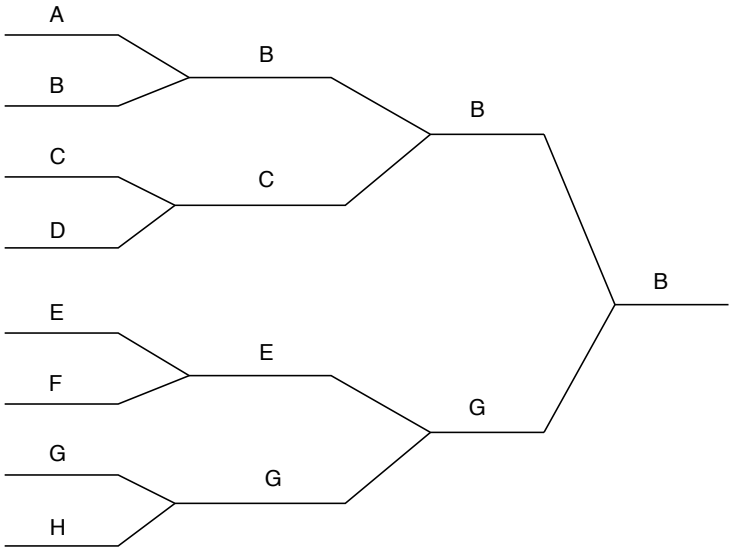

Figure 2: The results of the elimination tournament, after all matches have been played.

(3) A number of peer-to-peer systems on the Internet are based on *overlay networks*: rather than using the physical Internet topology as the network on which to perform computation, these systems run protocols by which nodes choose collections of virtual "neighbors" so as to define a higher-level graph whose structure may bear little or no relation to the underlying physical network. Such an overlay network is then used for sharing data and services, and it can be extremely flexible compared with a physical network, which is hard to modify in real-time to adapt to changing conditions.

Peer-to-peer networks tend to grow through the arrival of new participants, who join by linking into the existing structure, and this growth process has an intrinsic effect on the characteristics of the overall network.

Here's a simple model of network growth, for which we can begin to analyze some structural consequences. The system begins with a single node v_1. Nodes then join one at a time; as each node joins, it executes in a protocol whereby it forms a directed link to a single other node chosen uniformly at random from those already in the system. More concretely, if the system already contains nodes $v_1, v_2, \ldots, v_{k-1}$ and node v_k wishes to join, it randomly selects one of $v_1, v_2, \ldots, v_{k-1}$ and links to this node.

Suppose we run this process until we have a system consisting of nodes $v_1, v_2, \ldots, v_n$; the random process described above will produce a directed network in which each node other than v_1 has exactly one out-going edge. On the other hand, a node may have multiple in-coming links, or none at all. The in-coming links to a node v_j reflect all the other nodes whose access into the system is via v_j; so if v_j has many in-coming links, this can place a large load on it. To keep the system load-balanced, then, we'd like all nodes to have a roughly comparable number of in-coming links, but that's unlikely to happen here, since

nodes that join earlier in the process are likely to have more in-coming links than nodes that join later. Let's try to quantify this imbalance as follows.

Give a formula for the expected degree of node v_j, as a function of n and j. In particular, express your formula as the difference of two harmonic numbers, $H_a - H_b$ for choices of a and b that you should find (in terms of the parameters n and j in the model). Recall that the harmonic numbers are defined as

$$H_k = \sum_{j=1}^{k} \frac{1}{j}.$$

Provide a justification for your answer.

(4) Continuing with the model from the previous question, here's another way to look at the imbalances in the load on different nodes. Give a formula for the expected number of nodes with no in-coming links in a network grown randomly according to this model. Try to write your formula if possible in "closed form" — that is, written without any lengthy summations or $\sum$ notation. Provide a justification for your answer.