# Problem Set 2

Daniel Molitor (djm484)

**(1)** Consider a long, straight road, which we model as a line segment of length $n$. We drop a set of $k$ sensors randomly on this road — so each lands in a location selected uniformly and independently from the interval $[0, n]$.

Now, each sensor has a transmitting range of 2, so it can communicate with any other sensor within a distance 2 of it. This means that the random placement of the sensors defines a random $k$-node graph $G$, in which the nodes are the sensors, and we connect two by an edge if they can communicate with each other. We'd like to choose $k$ large enough so that $G$ is connected with high probability, and we can do this by reasoning as follows.

**(a) For an integer $j$ from $1, 2, \dots, n$, let $E_j$ denote the event that no sensor lands in the interval $[j-1, j]$. Give a formula for Pr$(E_j)$ in terms of $n$ and $k$, together with a brief explanation.**

The probability density function (p.d.f.) of the uniform distribution on this interval is $f(x) = \frac{1}{n}$, and thus the probability that it doesn't fall in $[j-1, j]$ is $1 - \frac{1}{n}$, or equivalently, $\frac{n-1}{n}$. Let $X_1, \dots, X_k$ be a random sample from the given distribution. The event $E_j$ is the joint probability that no sensors fall in $[j-1, j]$, so $E_j = \left(\frac{n-1}{n}\right)^k$.

**(b) Argue that if none of the events $E_j$ occurs, then the random graph $G$ defined above is connected.**

If none of the events $E_j$ occurs, then every interval $[j-1, j]$ for $j \in \{1, 2, \dots, n\}$ must contain a sensor. It follows that the farthest apart any two sensors can be is $(j+1)-(j-1) = 2$. So, every sensor must

be connected to at least one sensor in each of its neighboring intervals. It then is obvious that you can pick any two sensors and find a connecting path between their respective intervals, which is equivalent to saying you can find a connecting path between those sensors, since all sensors in a given interval must be connected.

**(c) Show using the Union Bound that if we drop $k = 2n \ln n$ sensors at random, then with high probability the graph $G$ will be connected. (In particular, with probability converging to $1$ as $n \to \infty$.)**

Let $E$ represent the event that there is some isolated node in graph $G$. Given the structure of our graph, this means that there is some interval $[j-2, j]$ for $j \in [2, n]$ with no sensor, denoted $E_j$. From part *a*, the probability that a given sensor falls in $[j-2, j]$ is $\frac{2}{n}$, and so

$$\begin{aligned} \Pr(E_j) &= \left(1 - \frac{2}{n}\right)^k \\ &= (1-p)^k. \end{aligned}$$

Now, let $p = \frac{c}{k}$ for some constant $c$; Then,

$$\Pr(E_j) = \left(\left(1 - \frac{c}{k}\right)^{\frac{k}{c}}\right)^c$$

which is bounded between $4^{-c}$ and $e^{-c}$. Now, $E = \bigcup_j E_j$, so by the Union Bound,

$$\begin{aligned} \Pr(E) &\leq \int_2^n e^{-c} dj \\ &= e^{-c}(n-2). \end{aligned}$$

Since $p = \frac{c}{k}$, then $e^{-c} = e^{-\frac{2k}{n}}$.

$$\begin{aligned}
\Pr(E) &\leq e^{-4\ln(n)}(n-2) \\
&\leq e^{-4\ln(n)}(n) \\
&= n^{-4}(n) \\
&= \frac{1}{n^3}.
\end{aligned}$$

Since $\lim_{n\to\infty} \frac{1}{n^3} = 0$, it follows that $\lim_{n\to\infty} \Pr(E) = 0$ and thus the probability $G$ is connected $\to 1$.

**(2)** 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.

Many of these networks 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.

**(a) Let $Z_j$ be a random variable corresponding to the in-degree of node $v_j$ (i.e., the number of in-coming links to node $v_j$). Describe a way of writing $Z_j$ as a sum of random variables that each take the value $0$ or $1$.**

For $i \in \{j+1, \dots, n\}$, $X_i$ is now a random variable from a Bernoulli distribution with probability $p$ where $1$ indicates that node $v_i$ connected to node $v_j$. The sequence of probabilities $p_{j+1}, p_{j+2}, \dots, p_n$ corresponds to the following sequence: $\frac{1}{j}, \frac{1}{j+1}, \dots, \frac{1}{n-1}$. From this,

$$Z_j = \sum_{i=j+1}^{n} X_i.$$

**(b) Applying linearity of expectation to your expression in (a), give a formula for the expected in-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.**

By linearity of expectation,

$$\begin{aligned} E(Z_j) &= E\left(\sum_{i=j+1}^{n} X_i\right) \\ &= \sum_{i=j+1}^{n} E(X_i) \\ &= \sum_{i=j+1}^{n} p_i \\ &= \frac{1}{j} + \frac{1}{j+1} + ... + \frac{1}{n-1} \\ &= H_{n-1} - H_{j-1}. \end{aligned}$$

**(c) Continuing with this model, here's another way to look at the imbalances in the load on different nodes. Let $X_j$ be a random variable equal to $1$ if node $v_j$ has no incoming links, and equal to $0$ otherwise. What is the expected value of $X_j$? Provide a justification for your answer.**

For $i \in \{j+1, ..., n\}$, $X_i$ is now a random variable from a Bernoulli distribution with probability $p$ where 1 indicates that node $v_i$ did **not** connect to node $v_j$. The sequence of probabilities $p_i = \{p_{j+1}, p_{j+2}, ..., p_n\}$ now corresponds to the following sequence: $\{1 - \frac{1}{j}, 1 - \frac{1}{j+1}, ..., 1 - \frac{1}{n-1}\}$. It follows that $\Pr(X_j)$ is the joint probability of $\{p_i\}$, so $\Pr(X_j) = \Pi_i\{p_i\}$. Since $X_j$ is a random variable from a Bernoulli distribution with probability $p_j = \Pr(X_j)$, by properties of the Bernoulli distribution we know that

$$E(X_j) = p_j = \frac{j-1}{j} * \frac{j}{j+1} * ... * \frac{n-2}{n-1} = \frac{j-1}{n-1}.$$

**(d) Building on your answer to (c), and again using linearity of expectation, give a formula for the expected number of nodes with no incoming 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.**

For $j \in \{1, \ldots, n\}$ let $X_j$ be a random variable as described in part *c*, and let $G_n$ represent the expected number of nodes with no incoming links in a randomly grown network. We can express this value as

$$G_n = E\left(\sum_{j=1}^{n} X_j\right)$$

which by linearity of expectation is equal to

$$\begin{aligned}\sum_{j=1}^{n} E(X_j) &= \sum_{j=1}^{n} \frac{j-1}{n-1} \\ &= \frac{0}{n-1} + \frac{1}{n-1} + \cdots + \frac{n-1}{n-1} \\ &= \frac{\frac{(n-1)n}{2}}{n-1} \\ &= \frac{n}{2}.\end{aligned}$$

**(3)** 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 the given figures. 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 a blank table with just the initial contestants filled in, 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 $j$ 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 $j$.

**(a) Write an expression for $p(j)$ as a function of $j$ and $n$. Provide a justification for your answer.**

At any round $j \in \{2, \ldots, k\}$, let's focus on one random match, $X_j$. To correctly guess one of the participants, we must also have correctly guessed the outcome of every preceding match that this individual participated in. The probability of this event occurring is $0.5^{j-1}$ and the probability it doesn't occur is $1 - 0.5^{j-1}$. Since there are $2^{k-j+1}$ contestants in each round $j$, the probability of incorrectly guessing all the contestants is $(1 - 0.5^{j-1})^{2^{k-j+1}}$. Doing some algebra, we get

$$\begin{aligned} n = 2^k \Rightarrow \\ \log_2 n = k \Rightarrow \\ 2^{k-j+1} &= 2^{\log_2(n)-j+1} \\ &= (2\log_2 n)(2^{-j})(2) \\ &= n2^{1-j}. \end{aligned}$$

From this it follows that

$$p(j) = 1 - (1 - 2^{1-j})^{n2^{1-j}}.$$

**(b) Is there a constant value $\alpha^*$ 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 $\alpha^*$ and justify why it is critical, or argue why there is no critical value of $\alpha^*$.**

Yes, $\alpha^* = 0.5$ is this critical value. When we set $j = \alpha \log_2 n$, we find that

$$p(j) = 1 - (1 - \frac{2}{n^\alpha})^{2n^{1-\alpha}}.$$

Intuitively, $\alpha^* = 0.5$ is this critical value, because when $\alpha < 0.5$, the $\frac{2}{n^\alpha}$ term gets dominated by the $2n^{1-\alpha}$ term which drives $(1 - \frac{2}{n^\alpha})$ to 0 as $n \to \infty$. And the reverse is true; when $\alpha > 0.5$ the $\frac{2}{n^\alpha}$ term dominates the $2n^{1-\alpha}$ term which drives $(1 - \frac{2}{n^\alpha})$ to 1.