ORIE 4520 PRELIM, FALL 2015

Rules:

- You have 1hour and 15 minutes – please turn in your papers by 2:40pm.
- You are allowed one sheet of formulas. No books, cell phones, PDAs, laptops, or any other such aids.
- Write down your final answers clearly in the space below the question. You can use the back side of the sheets for scrap work, but be sure to label your work clearly.

Name:

NetID:

OFFICIAL USE ONLY	
1	/15
2	/20
3	/15
TOTAL	/50

(1) **(Isolated nodes in random graphs)**
Given a set of n nodes $V = [n]$, the $G(n,p)$ random-graph model constructs a graph by independently connecting each pair of nodes (i,j) with probability p. Let N_0 be the number of *isolated nodes* (i.e., not connected to any other node) in the resulting graph. We now show the following *threshold phenomena*: as n grows, N_0 either grows with n or goes to 0, depending on if $p < \ln n/n$ or $> \ln n/n$.

- (Part a, 5 points)
  Prove that:
  $$\mathbf{E}[N_0] = n(1-p)^{n-1}$$

  Solution: Let X_v be the indicator that node v is isolated – then $\mathbf{E}[X_v] = (1-p)^{n-1}$. Moreover, $B_0 = \sum_{v \in V} X_v$. Thus, by linearity of expectations, we have:
  $$\mathbf{E}[N_0] = \sum_{v \in V} \mathbf{E}[X_v] = n(1-p)^{n-1}$$

- (Part b, 5 points)
  For the remaining parts, you should use the approximation $(1-p)^{n-1} \approx e^{-np}$. Now suppose we choose $p = (1-c)\ln n/n$, for some $c \in (0,1)$ – show that, $\mathbf{E}[N_0] = n^c$.
  Moreover, for this choice of p, it can be shown that $Var(N_0) \leq 2\mathbf{E}[N_0]$ (this follows from a straightforward calculation; however you should assume it). Now show that:
  $$\mathbf{P}[N_0 = 0] \leq \frac{2}{n^c}$$
  Hint: Chebyshev's Inequality - $\mathbf{P}\left[|X - \mathbf{E}[X]| \geq t\right] \leq Var(X)/t^2$

  Solution: From before, we have:
  $$\mathbf{E}[N_0] = n(1-p)^{n-1} \approx ne^{-np}$$
  Substituting $p = (1-c)\ln n/n$, we get $\mathbf{E}[N_0] = ne^{-(1-c)\ln n} = n^c$.
  Moreover, we have that $\mathbf{P}[N_0 \leq 0] = \mathbf{P}[N_0 - \mathbf{E}[N_0] \leq -\mathbf{E}[N_0]] \leq \mathbf{P}[|N_0 - \mathbf{E}[N_0]| \leq \mathbf{E}[N_0]]$.
  Now using Chebyshev's inequality, we get:
  $$\mathbf{P}[N_0 = 0] \leq \frac{Var(N_0)}{\mathbf{E}[N_0]^2} \leq \frac{2}{\mathbf{E}[N_0]} \leq \frac{2}{n^c}$$

- (Part c, 5 points) On the other hand, suppose we choose $p = (1+c)\ln n/n$, for some $c > 0$. Show that:
$$\mathbf{P}[N_0 \neq 0] \leq \frac{1}{n^c}$$
Thus, for this choice of p, the number of isolated nodes is 0 with high probability.
Hint: Let A_i be the event that node i is isolated. How can you write the event $\{N_0 \neq 0\}$ in terms of the events A_i?

Solution: First, note that $\mathbf{P}[A_i] = (1-p)^{n-1}$. Next, observe that the number of isolated nodes is non-0 if at least one node i is isolated – this means that $\{N_0 \neq 0\} = \{\cup_i A_i\}$. Now by the union bound, we have:
$$\mathbf{P}[N_0 \neq 0] = \mathbf{P}[\cup_i A_i] \leq \sum_i \mathbf{P}[A_i] = n(1-p)^{n-1} \approx ne^{-np}$$
Substituting $p = (1+c)\ln n/n$, we get $\mathbf{P}[N_0 \neq 0] = ne^{-(1+c)\ln n} = n^{-c}$.

Alternate method: We can also get this result by the first-moment method. Note that $\mathbf{P}[N_0 \neq 0] = \mathbf{P}[N_0 \geq 1]$ – now by Markov's inequality, we have:
$$\mathbf{P}[N_0 \neq 0] \leq \mathbf{E}[N_0] = n(1-p)^{n-1}$$
Contrast this to the use of the second-moment method in part b.

(2) **(Windowed stream sampling)**
We want to process a data stream $\{x_1, x_2, \ldots, x_m\}$ to store a random sample over any lookback-window – at any time (i.e., any m) and for any $w \leq m$, we want to be able to a output uniform random sample from the last w elements. For example, after 100 items have arrived, if we are given $w = 26$, we need to return a uniform random element in $\{x_{75}, x_{76}, \ldots, x_{100}\}$.

- (Part a, 5 points)
Suppose we store each item in the stream as a tuple $\langle t, x_t, \sigma_t \rangle$, where σ_t are i.i.d Uniform$[0,1]$ random variables, one for each time $t \in [m]$. Now for any given lookback-window length w, let $T(w) = \arg\min_{\{\text{last } w \text{ elements}\}} \{\sigma_t\}$, i.e., the index of the element with the minimum σ_t amongst the last w elements. Argue that $T(w)$ is a uniform sample, i.e.:
$$\mathbf{P}\Big[T(w) = t\Big] = \frac{1}{w} \qquad \forall\, t \in \{m-w+1, m-w+2, \ldots, m\}$$
Hint: You can argue this without using anything specific about the Uniform$[0,1]$ *distribution...*

Solution: By symmetry, every element t in a window of size w are equally likely to have the minimum σ_t. Thus $\mathbf{P}[T(w) = t] = 1/w$.

- (Part b, 5 points)
The above result shows that for any given w, we can return $x_{T(w)}$ as a uniform random sample from the last w elements. However, storing all tuples $\langle t, x_t, \sigma_t \rangle$ will take too much space.
We now see how to improve this: Suppose for two distinct times t_1 and t_2 such that $t_1 < t_2 \leq m$, we are told that $\sigma_{t_1} > \sigma_{t_2}$. Can we delete one of the tuples $\langle t_1, x_{t_1}, \sigma_{t_1} \rangle$ or $\langle t_2, x_{t_2}, \sigma_{t_2} \rangle$, and still return a uniform random sample for every w?

Solution: It is clear that $\langle t_1, x_{t_1}, \sigma_{t_1} \rangle$ is not required if we are given $w \leq m - t_1$. However, if $w > m - t_1$, then we will still never return x_{t_1}, as the fact that $\sigma_{t_1} > \sigma_{t_2}$ means that σ_{t_1} is not the element with minimum σ_t in the lookback-window. Thus, we can always delete the tuple $\langle t_1, x_{t_1}, \sigma_{t_1} \rangle$.

- (Part c, 10 points)
  Suppose we store a subset of the tuples $C = \{\langle t, x_t, \sigma_t\rangle\}$, inserting new elements as follows:
  - INSERT(x_{m+1}): Generate $\sigma_{m+1} \sim$ Uniform$[0,1]$, and add $\langle m+1, x_{m+1}, \sigma_{m+1}\rangle$ to C. Delete all tuples $\langle t, x_t, \sigma_t\rangle$ in C where $\sigma_t \geq \sigma_{m+1}$.

  Let C_m denote the set of tuples stored after m elements have arrived in the stream. Argue that:
  $$\mathbf{E}[|C_m|] = H_m$$
  where $H_m = \sum_{i=1}^{m} 1/i$ is the harmonic sum (which we know from before is $\Theta(\log m)$).
  Hint: Define Y_t to be the indicator that the t^{th} item is stored in C_m. Clearly $Y_m = 1$, as we always store the latest element. What is the probability that $Y_{m-1} = 1$? Now, given the last $k+1$ elements $\{x_{m-k}, x_{m-k+1}, \ldots, x_m\}$, what is the probability that $Y_{m-k} = 1$?

Solution: Let Y_t to be the indicator that the t^{th} item is stored in C_m. $Y_m = 1$ as we always store the latest element – for Y_{m-1} to be 1, we need $\sigma_{m-1} < \sigma_m$, which happens with probability $1/2$. More generally, we have that $Y_t = 1$ if and only if it σ_t is the smallest in $\{\sigma_t, \sigma_{t+1}, \ldots, \sigma_m\}$, which is true with probability $1/(m-t+1)$. Finally, we have that $|C_m| = \sum_{t=1}^{m} Y_t$, and hence by linearity of expectation, we have:
$$\mathbf{E}[|C_m|] = \sum_{t=1}^{m} \mathbf{E}[Y_t] = \sum_{t=1}^{m} \frac{1}{m-t+1} = H_m$$

(3) **(Maximum and minimum loaded bin)**

$24\,n\ln n$ balls are thrown in n bins uniformly at random. Let B_i = load of bin i.

- (Part a, 5 points)
  What is $\mathbf{E}[B_i]$?

  Solution: Clearly $B_i = Bin(24n\ln n, 1/n)$, and hence:
  $$\mathbf{E}[B_i] = \frac{24n\ln n}{n} = 24\ln n$$

- (Part b, 10 points)
  Let $B_{\max}$ and $B_{\min}$ denote the maximum and minimum loaded bin respectively. Show that:
  $$\mathbf{P}\left[B_{\max} - B_{\min} \geq \mathbf{E}[B_i]\right] \leq \frac{2}{n}$$
  Hint: For an individual bin, recall we have the following Chernoff bound – for any $\epsilon < 1$:
  $$\mathbf{P}\left[|B_i - \mathbf{E}[B_i]| \geq \epsilon\mathbf{E}[B_i]\right] \leq 2\exp\left(\frac{-\epsilon^2\mathbf{E}[B_i]}{3}\right)$$

Solution: First, for an individual bin B_i, the above Chernoff bound (with $\epsilon = 1/2$) gives us:
$$\mathbf{P}\left[|B_i - \mathbf{E}[B_i]| \geq \frac{1}{2}\mathbf{E}[B_i]\right] \leq 2\exp\left(\frac{-\mathbf{E}[B_i]}{12}\right) = \frac{2}{n^2}$$
Further, by the union bound, we can extend the concentration to *all bins*, to get:
$$\mathbf{P}\left[\cup_i \left\{|B_i - \mathbf{E}[B_i]| \geq \frac{1}{2}\mathbf{E}[B_i]\right\}\right] \leq n\frac{2}{n^2} = \frac{2}{n}$$
Finally, observe that as long as all the B_i are in the range $[\mathbf{E}[B_i]/2, 3\mathbf{E}[B_i]/2]$ (as is true in the complement of the event considered in the above concentration guarantee), then we also have that $B_{\max}, B_{\min} \in [\mathbf{E}[B_i]/2, 3\mathbf{E}[B_i]/2]$ and hence $B_{\max} - B_{\min} < \mathbf{E}[B_i]$. Thus, we get:
$$\mathbf{P}\left[B_{\max} - B_{\min} \geq \mathbf{E}[B_i]\right] \leq \frac{2}{n}$$