Problem 1: (The Flajolet-Martin Counter)

In class (and in the prelim!), we looked at an idealized algorithm for finding the number of distinct elements in a stream, where we sampled uniform random variables for each item, and then stored their minimum value. One way to implement this in practice is via the Flajolet-Martin counter:

Suppose we have a stream $(X_1, X_2, \ldots, X_m)$ of m items, where each item X_i corresponds to an element in $[n]$. Assume n is a power of 2, and $k = \log_2 n$. Let h be a hash function that maps each of the elements in $[n]$ to k bits – in particular, let us denote $h(x) = (b_1(x), b_2(x), \ldots, b_k(x))$ for each $x \in [n]$), and assume that each bit k independently satisfies $\mathbf{P}[b_k(x) = 0] = \mathbf{P}[b_k(x) = 1] = 1/2$ for every pair $x \in [n]$. For every $x \in [n]$, let $r(x)$ be the number of *trailing* 0*'s* in $h(x)$ – so for example, for $n = 16$ (i.e., $k = 4$), $h(x) = 0100$ means $r(x) = 2$, $h(x) = 1000$ means $r(x) = 3$, and so on). Finally, let $R = \max_i\{r(X_i)\}$ – i.e., the maximum number of trailing 0's in the hashes of the items in the stream.

Part (a)

For any element $x \in [n]$, let $Y_j(x)$ be the indicator that $r(x) = j$. Argue that $\mathbf{E}[Y_j(x)] = 1/2^{j+1}$.

Solution:

$$\mathbf{E}[Y_j(x)] = \mathbf{P}[r(x) = j] = \left(\frac{1}{2}\right)^j \frac{1}{2} = \frac{1}{2^{j+1}}.$$

Part (b)

Let F_0 be the number of distinct elements in the stream, and define N_j to be the number of elements in the stream for which $r(x) > j$. Show that:

$$\mathbf{E}[N_j] = \frac{F_0}{2^{j+1}} \quad , \quad Var(N_j) = \frac{F_0}{2^{j+1}}\left(1 - \frac{1}{2^{j+1}}\right) \leq \mathbf{E}[N_j]$$

Solution: Let $\mathbb{1}_{\{A\}}$ denote the indicator for any event A. First note that

$$N_j = \sum_{i=1}^{F_0} \mathbb{1}_{\{r(x)>j\}} = \sum_{i=1}^{F_0} \mathbb{1}_{\{r(x)\geq j+1\}}.$$

Therefore, by linearity of expectation,

$$\mathbf{E}[N_j] = \mathbf{E}\left[\sum_{i=1}^{F_0} \mathbb{1}_{\{r(x)\geq j+1\}}\right] = \sum_{i=1}^{F_0} \mathbf{P}\left[r(x) \geq j+1\right] = \frac{F_0}{2^{j+1}}.$$

Moreover, since each element x is hashed independently, hence $\mathbb{1}_{\{r(x)\geq j+1\}}$ are independent. Also, since $\mathbb{1}_{\{r(x)\geq j+1\}} \sim Bernoulli(1/2^{j+1})$, this implies that $Var\left(\mathbb{1}_{\{r(x)\geq j+1\}}\right) = 2^{-(j+1)}\left(1 - 2^{-(j+1)}\right)$.

Thus we have:

$$\begin{aligned} Var(N_j) = Var\left(\sum_{i=1}^{F_0} \mathbb{1}_{\{r(x)\geq j+1\}}\right) &= \sum_{i=1}^{F_0} Var(\mathbb{1}_{\{r(x)\geq j+1\}}) \\ &= F_0 Var(\mathbb{1}_{\{r(x)\geq j+1\}}) = \frac{F_0}{2^{j+1}}\left(1-\frac{1}{2^{j+1}}\right). \end{aligned}$$

Note that this implies $Var(N_j) \leq \mathbf{E}[N_j]$

Part (c)

Suppose we use 2^R as an estimator for F_0. Argue that for any j, $\mathbf{P}[R \geq j] = \mathbf{P}[N_{j-1} > 0]$. Next, assuming that F_0 is a power of 2, show that:

$$\mathbf{P}[R < \log_2(F_0) - c] \leq \frac{1}{2^c}$$

Hint: Part (c) had a mistake in the homework. The original version was:
Suppose we use 2^R as an estimator for F_0. Argue that for any j, $\mathbf{P}[R \geq j] \geq \mathbf{P}[N_j > 0]$. Next, assuming that F_0 is a power of 2, show that:

$$\mathbf{P}[R < \log_2(F_0) - c] \leq \frac{1}{2^c}$$

Although the first statement is correct, it gives a weaker bound of $2/2^c$.

Solution: First, using Chebyshev, we get:

$$\mathbf{P}[N_j = 0] \leq \mathbf{P}\left[|N_j - \mathbf{E}[N_j]| > E[N_j]\right] \leq \frac{Var(N_j)}{(\mathbf{E}[N_j])^2} \leq \frac{\mathbf{E}[N_j]}{(\mathbf{E}[N_j])^2} = \frac{1}{\mathbf{E}[N_j]} = \frac{2^{j+1}}{F_0}.$$

Note that both $R \geq j$ and $N_{j-1} \geq 1$ if and only if there is at least one element x for which $r(x) > j$ – thus, $\mathbf{P}[R < j] = \mathbf{P}[N_{j-1} = 0]$. Now from the previous inequality, we have:

$$\begin{aligned} \mathbf{P}[R < \log_2(F_0) - c] &= \mathbf{P}[N_{\log_2(F_0)-c-1} = 0] \\ &\leq \frac{2^{\log_2(F_0)-c}}{F_0} = \frac{1}{2^c}. \end{aligned}$$

Part (d)

On the other hand, argue that $\mathbf{P}[R \geq j] \leq \mathbf{E}[N_{j-1}]$, and hence show that:

$$\mathbf{P}[R \geq \log_2(F_0) + c] \leq \frac{1}{2^c}$$

Hint: Part (d) had a mistake in the homework. The original version was:
On the other hand, argue that $\mathbf{P}[R \geq j] \leq \sum_{j' \geq j} \mathbf{E}[N_{j'}]$, and hence...

Solution: Note that both $R \geq j$ and $N_{j-1} \geq 1$ if and only if there is at least one element x for which $r(x) > j$. Thus from Markov's inequality, we have:

$$\mathbf{P}[R \geq j] = \mathbf{P}[N_{j-1} \geq 1] \leq \mathbf{E}[N_{j-1}] = \frac{F_0}{2^j}$$

Thus we have $\mathbf{P}[R \geq \log_2(F_0) + c] \leq \frac{F_0}{F_0 2^c} = \frac{1}{2^c}$.

Problem 2: (An Alternate All-Pair Distance Sketch)

In class we saw an All-Pairs Distance Sketch (ADS) by Das-Sarma et al., which for each node v stored a sketch $S(v)$ with distances to $O(\log n)$ other nodes, and then given any pair (u, v), used the to sketches to compute a shortest-path estimate within a multiplicative 'stretch' of $O(\log n)$. We'll now see an alternate ADS proposed by Cohen et al.

We are given an undirected weighted graph $G(V, E)$, where each edge $(u, v) \in E$ has some weight $w_{u,v} \geq 0$ corresponding to its length. The shortest path distance $d^*(u, v)$ between any nodes u and v is the minimum sum of weights over all paths from u to v. For convenience, assume that the weights are such that each pair of nodes has a unique shortest-path distance. Thus for any given node v, we can uniquely sort all nodes in V in increasing order of distance, to get $V_v = \{v, w_1, w_2, \ldots\}$, where $d^*(v, w_i) < d^*(v, w_{i+1})\}$. Moreover, for each node $v \in V$, we generate an i.i.d random variable $R(v)$ which is $Uniform[0, 1]$ distributed.

Part (a)

To generate the ADS $S(v)$, we first initialize $S(v) = (v, 0, R(v))$; then sequentially pick nodes w_i from the shortest-path ordering $V_v = \{v, w_1, w_2, \ldots\}$, and add $(w_i, d^*(v, w_i), R(w_i))$ to $S(v)$ if $R(w_i)$ is smaller than $R(u)$ for all $u \in S(v)$. What is the expected number of nodes in $S(v)$?
Hint: Recall the random sampling from stream algorithm in the prelim.

Solution: Given node v with shortest-path ordering $V_v = \{v, w_1, w_2, \ldots\}$, let X_i be the indicator that node w_i is in the ADS $S(v)$. Note that for w_i to be added, we need $R(w_i) < R(v)$ and $R(w_i) < R(w_j)$ for all $j < i$ – thus $\mathbf{E}[X_i] = \frac{1}{i+1}$. Now, by linearity of expectation, we have:

$$\mathbf{E}[|S(v)|] = 1 + \sum_{i=1}^{n-1} \mathbf{E}[X_i] = H_n = \Theta(\log n)$$

Part (b)

Given two nodes u and v, to estimate their distance, we first find all nodes w in $S(u) \cap S(v)$, and then compute $\widehat{d}(u, v) = \min_{w \in S(u) \cap S(v)} d^*(u, w) + d^*(w, v)$. We now argue that $\widehat{d}(u, v) \leq \log_2 n d^*(u, v)$ with constant probability (which we can then amplify by taking independent ADS).

Argue first that for any pair of nodes u and v, we have that $|S(u) \cap S(v)| > 0$.

Solution: Consider the node $v^* = \arg\min_{v \in V}\{R(v)\}$; as long as the graph is connected (i.e., $d(u,v) < \infty$ for all pairs (u,v)), then it is easy to see that $v^* \in S(v)$ for all nodes $v \in V$. Thus, $|S(u) \cap S(v)|$ is at least 1.

Part (c)

As we did in class, for any node v and distance d, define $B_d(v) = \{u \in V | d^*(v,u) \leq d\}$, i.e., the set of nodes within a distance of d from v. Now for any given pair u, v, let $d = d^*(u,v)$. Argue that:

$$B_d(v) \subseteq B_{2d}(u) \subseteq B_{3d}(v) \subseteq \ldots,$$

and in general, for any $k \in \{1, 2, \ldots, B_{kd}(v) \subseteq B_{(k+1)d}(u)$

Solution: Consider any node $w \in B_d(v)$ – then $d^*(v,w) \leq d$. Now since the graph is undirected, therefore one can go from u to w by first taking the shortest path from u to v and then v to w – thus $d^*(u,w) \leq d^*(u,v) + d^*(v,w) \leq 2d$, and hence $w \in B_{2d}(u)$. The remaining inclusions can be proved in a similar manner.

Part (d)

For convenience, let $m = \log_2 n$ be an even integer. Now for every pair (u,v), given the sequence of nested sets $B_d(v) \subseteq B_{2d}(u) \subseteq B_{3d}(v) \subseteq \ldots \subseteq B_{(m-1)d}(v) \subseteq B_{md}(u)$, argue (by contradiction) that there exists at least one pair of consecutive sets (i.e., $B_{(2k-1)d}(v) \subseteq B_{(2k)d}(u)$ or $B_{2kd}(u) \subseteq B_{(2k+1)d}(v)$) such that that the larger set's cardinality is less than twice that of the smaller set.

Solution: Given the sequence of nested sets $B_d(v) \subseteq B_{2d}(u) \subseteq B_{3d}(v) \subseteq \ldots \subseteq B_{(m-1)d}(v) \subseteq B_{md}(u)$, suppose instead that there are no pair of consecutive sets (i.e., $B_{(2k-1)d}(v) \subseteq B_{(2k)d}(u)$ or $B_{2kd}(u) \subseteq B_{(2k+1)d}(v)$) such that that the larger set's cardinality is less than twice that of the smaller set. Then we have $B_{2d}(u) > 2B_d(v) \geq 2$, $B_{3d}(v) > 2B_{2d}(u) \geq 4$, and so on till we get $\subseteq B_{md}(u) > 2^m = n$ – this however is a contradiction.

Part (e)

Finally, for a given pair u, v with shortest distance d, and any $k \in \{1, 2, \ldots, m\}$, we want to find the probability that a node in $B_{kd}(v)$ is present in both ADS sketches $S(u), S(v)$. Argue that:

$$\mathbf{P}[\exists w \in S(u) \cap S(v)\, s.t.\, w \in B_{kd}(v)] \geq \frac{|B_{kd}(v)|}{|B_{(k+1)d}(u)|}$$

This put together with the previous parts shows that with probability at least 1/2, the ADS sketch returns an estimate $\widehat{d}(u,v) \leq md^*(u,v)$.
Note: The question in the HW was framed incorrectly – this is the correct statement.

Solution: Recall that $B_{kd}(v) \subseteq B_{(k+1)d}(u)$. For any set $B \subseteq V$ let $v^*(B) = \arg\min_{v\in B}\{R(v)\}$. Then we claim that $v^*(B_{(k+1)d}(u))$ is contained in $S(u)$ – to see this, note that if this were not the case, then there is some node w with $R(w) < R(v^*(B_{(k+1)d}(u)))$ which is closer to u than $v^*(B_{(k+1)d}(u))$. By definition, this would imply that $w \in B_{(k+1)d}(u)$, which contradicts the fact that $v^*(B_{(k+1)d}(u))$ had the minimum value of $R(v)$ for nodes in $B_{(k+1)d}(u)$. Note however that if $v^*(B_{(k+1)d}(u)) \in B_{kd}(v)$, then $v^*(B_{(k+1)d}(u)) = v^*(B_{(k)d}(v))$ (as the sets are nested), and hence it is also in $S(v)$. Thus, we have that:

$$\begin{aligned}\mathbf{P}[\exists w \in S(u)\cap S(v)\, s.t.\, w \in B_{kd}(v)] &\geq \mathbf{P}[v^*(B_{(k+1)d}(u)) \in B_{kd}(v)]\\ &= \frac{|B_{kd}(v)|}{|B_{(k+1)d}(u)|},\end{aligned}$$

where the last inequality follows from the same argument as in MinHash – that taking a uniform random permutation of nodes is equivalent to we picking node uniformly at random, without replacement.

Problem 3: (The Galton-Watson Branching Process)

In class we saw the Galton-Watson branching process:

$$X_{n+1} = \sum_{i=1}^{X_n} Y_{n,i},$$

where X_n is the number of descendants at the n^{th} level of the GW tree, and $Y_{n,i} \sim Y$ is i.i.d number of descendants of each node. Let us denote $\mathbf{P}[Y = k] = p_k, \mathbf{E}[Y] = \mu, Var(Y) = \sigma^2$ and $G(s) = \mathbf{E}[s^Y]$. We now try to derive some quantities of interest, in order to understand the GW process.

Part (a)

Prove that $\mathbf{E}[X_{n+1}] = \mu\mathbf{E}[X_n]$, and hence argue that $\mathbf{E}[X_n] = \mu^n$. Using this, show that if $\mu < 1$, then $\mathbf{P}[X_n > 0] = o(1)$ (i.e., it goes to 0 as $n \to \infty$).

Solution: From definition, we have $\mathbf{E}[X_{n+1}] = \mathbf{E}[\sum_{i=1}^{X_n} Y_{n,i}]$ – now we can write:

$$\mathbf{E}[X_{n+1}] = \sum_{k=0}^{\infty} \mathbf{E}[\sum_{i=1}^{k} Y_{n,i}|X_n = k|]\mathbf{P}[X_n = k] = \sum_{k=0}^{\infty} k\mu\mathbf{P}[X_n = k] = \mu\mathbf{E}[X_n]$$

Starting with $\mathbf{E}[X_0] = 1$, we can iterate to get $\mathbf{E}[X_n] = \mu^n$.

Next, assume $\mu < 1$; since X_n takes integer values, hence from Markov's inequality, we have:

$$\mathbf{P}[X_n > 0] = \mathbf{P}[X_n \geq 1] \leq \mathbf{E}[X_n] = \mu^n \to 0 \quad \text{as} \quad n \to \infty.$$

Part (b)

Assuming $\mu \neq 0$, prove that:

$$Var(X_n) = \sigma^2 \mu^{n-1} \left(\frac{1-\mu^n}{1-\mu} \right)$$

Solution: As in the previous part, we have:

$$\begin{aligned}
\mathbf{E}[X_n^2] &= \sum_{k=0}^{\infty} \mathbf{E}\left[(\sum_{i=1}^{k} Y_i)^2 \right] \mathbf{P}[X_{n-1} = k] \\
&= \sum_{k=0}^{\infty} \left(k\mathbf{E}\left[Y_i^2\right] + k(k-1)\mathbf{E}[Y_i]^2 \right) \mathbf{P}[X_{n-1} = k] \\
&= (\sigma^2 + \mu^2)\mathbf{E}[X_{n-1}] + \mu^2 \mathbf{E}[X_{n-1}^2 - X_{n-1}] \\
&= (\sigma^2 + \mu^2)\mu^{n-1} + \mu^2(\mathbf{E}[X_{n-1}^2] - \mu^{n-1}) \\
&= \sigma^2 \mu^{n-1} + \mu^2 \mathbf{E}[X_{n-1}^2]
\end{aligned}$$

Also, we have that $Var(X_n) = \mathbf{E}[X_n^2] + \mathbf{E}[X_n]^2$. Substituting from above, we get:

$$\begin{aligned}
Var(X_n) &= \sigma^2 \mu^{n-1} + \mu^2 \mathbf{E}[X_{n-1}^2] + \mu^{2n} \\
&= \sigma^2 \mu^{n-1} + \mu^2 (Var(X_{n-1}) - \mu^{2n-2}) + \mu^{2n} \\
&= \sigma^2 \mu^{n-1} + \mu^2 (Var(X_{n-1}))
\end{aligned}$$

Now note that $Var(X_0) = 0$. Iterating, we get:

$$Var[X_n] = \sigma^2 \left(\mu^{n-1} + \mu^n + \ldots + \mu^{2n-2} \right) = \sigma^2 \mu^{n-1} \left(\frac{1-\mu^n}{1-\mu} \right)$$

Part (c)

Recall we defined the pgf $G_n(s) = \mathbf{E}[s^{X_n}]$, and also defined T to be the time to extinction, (i.e. $T = \inf_{n>0} \{X_n = 0\}$), $\gamma_n = \mathbf{P}[X_n = 0]$ to be the probability of extinction, and $\gamma = \mathbf{P}[\lim_{n\to\infty} X_n = 0]$ as the probability of ultimate extinction. We now see how we can compute T, γ_n and γ using the pgf.

Suppose $Y \sim Bernoulli(p)$, i.e., $p_1 = p, p_0 = 1-p$. – show that:

$$G_n(s) = (1-p) \left(\sum_{i=0}^{n-1} p^i \right) + p^n s$$

Using the above result, show that T is distributed as $Geometric(1-p)$, i.e., $\mathbf{P}[T = k] = (1-p)p^{k-1}$.

Solution: We can check this by induction. The above formula gives $G_0(s) = s$, which is true since $X_0 = 1$. Also, note that $G(s) = \mathbf{E}[s^Y] = ps + (1-p)$. Now assume the formula is true for $n-1$ – since $G_n(s) = G(G_{n-1(s)})$, we get:

$$\begin{aligned} G_n(s) &= pG_{n-1}(s) + 1 - p \\ &= p\left((1-p)\left(\sum_{i=0}^{n-2} p^i\right) + p^{n-1}s\right) + 1 - p \\ &= (1-p)\left(1 + \sum_{i=1}^{n-1} p^i\right) + +p^n s \\ &= (1-p)\left(\sum_{i=0}^{n-1} p^i\right) + p^n s \end{aligned}$$

Next, recall that in class we showed $\mathbf{P}[T = k] = G_k(0) - G_{k-1}(0)$. Substituting, we get $\mathbf{P}[T = k] = (1-p)p^{k-1}$, and therefore, $T \sim \mathit{Geometric}(1-p)$.

Part (d)

Suppose instead $Y \sim \mathit{Geometric}(1-p)$; find γ, the probability of ultimate extinction.

Solution: If $Y \sim \mathit{Geometric}(1-p)$, then

$$G(s) = \mathbf{E}[s^Y] = \sum_{k=1}^{\infty} s^k p^{k-1}(1-p) = (1-p)s\sum_{k=0}^{\infty}(sp)^k = \frac{(1-p)s}{1-ps}.$$

From class, we know that the probability of ultimate extinction γ satisfies

$$\gamma = \frac{(1-p)\gamma}{1-p\gamma}.$$

Solving, we get $\gamma = \min\{0,1\} = 0$. Note that this makes sense, as we have at least one descendent for every node, and hence the process never terminates.

Problem 4: (The DFS view of the GW Tree and Random Graph)

In class, we also discussed an alternate way of studying the Galton-Watson process, by expanding it using a *depth-first search*. We now develop this further, and use it to study the $G(n,p)$ graph in the subcritical regime.

Part (a)

Let A_n be the set of *active nodes* (i.e., nodes which have been uncovered, but whose descendants have not yet been uncovered) after n steps. We start with $A_0 = 1$, and have: $A_n = A_{n-1} - 1 + Y_n$, where Y_n is the number of offspring of the active node uncovered in the n^{th} step. Let T_{DFS} denote the time when the DFS procedure stops, i.e., the first time when there are no more active nodes to explore. Argue that:

$$T_{DFS} = 1 + \sum_{i=1}^{T_{DFS}} Y_i$$

Note: There was a typo in the recurrence equation in the HW.

Solution: Using the recurrence equation, we have that for all $t \leq T_{DFS}$,

$$A_t = A_0 + \sum_{i=1}^{t} Y_i - t.$$

Moreover, by definition $A_{T_{DFS}} = 0$. Hence, we have that:

$$T_{DFS} = 1 + \sum_{i=1}^{T_{DFS}} Y_i.$$

Part (b)

Let $\mathcal{X}$ be the total number of nodes in a Galton-Watson tree at the moment it stops growing. What is the relation between $\mathcal{X}$ and T_{DFS}? Also argue that:

$$\mathbf{P}[\mathcal{X} > k] \leq e^{-kh(Y)}$$

where $h(Y) = \sup_{\theta \geq 0} \left[\theta - \ln \mathbf{E}[e^{\theta Y}]\right]$.

Solution: In DFS, we explore the nodes one by one – hence $\mathcal{X} = T_{DFS}$, and $\mathcal{X} = 1 + \sum_{i=1}^{\mathcal{X}} Y_i$. Now from the DFS procedure we have:

$$\begin{aligned}\mathbf{P}[\mathcal{X} > k] &= \mathbf{P}[A_1 > 0, \cdots, A_k > 0] \leq \mathbf{P}[A_k > 0] \\ &= \mathbf{P}\left[A_0 + \sum_{i=1}^{k} Y_i - k > 0\right] = \mathbf{P}\left[\sum_{i=1}^{k} Y_i \geq k\right] \\ &= \mathbf{P}\left[e^{\theta_0 \sum_{i=1}^{k} Y_i} \geq e^{\theta_0 k}\right],\end{aligned}$$

for any $\theta > 0$. Now, using Markov's inequality, and since Y_i are independent, we get:

$$\mathbf{P}[\mathcal{X} > k] \leq \inf_{\theta \geq 0} \frac{\mathbf{E}\left[e^{\theta Y}\right]^k}{e^{\theta k}} = e^{-k \sup_{\theta \geq 0}(\theta - \ln \mathbf{E}[e^{\theta Y}])} = e^{-kh(Y)}.$$

Part (c)

Next, we use this DFS technique to study the $G(n,p)$ graph. Starting from some node v, we perform a DFS as in the GW tree – again we have A_k denoting the number of active nodes after k explorations, with $A_0 = 1$. Also note that after k explorations, the number of nodes not yet discovered by the DFS is $U_k = n - k - A_k$. Argue that $U_k \sim \mathit{Binomial}(n-1,(1-p)^k)$.

Solution: After k steps of the DFS, we have explored k nodes, and uncovered A_k active nodes – the number of nodes not yet discovered is thus $U_k = n - k - A_k$. Now note that for a node u to be undiscovered after k explorations, we need all the potential edges from the first k explored nodes to u to be missing – this happens with probability $(1-p)^k$. Now, since we started with 1 discovered and $n-1$ undiscovered nodes, thus $U_k \sim \mathit{Binomial}(n-1,(1-p)^k)$.

Part (d)

Using the above result, we have that $n-1-U_k = A_k + k - 1$ is distributed as $\mathit{Binomial}(n-1, 1-(1-p)^k)$. Also note that a necessary condition for the connected component $C(v)$ around v to be bigger than k is that the active set after k explorations is non-empty. Thus we have: $\mathbf{P}[|C(v)| > k] \leq \mathbf{P}[A_k > 0]$. Let $np = \lambda$ – now show that:

$$\mathbf{P}[|C(v)| > k] \leq e^{-kh(\lambda)},$$

where $h(\lambda) = \sup_{\theta \geq 0}\left[\theta + \lambda(1-e^{\theta})\right]$
Hint: Use the following 'stochastic dominance' relation (try to convince yourself why this is true):

$$\mathbf{P}[Bin(n-1, 1-(1-p)^k) > a] \leq \mathbf{P}[Bin(n, pk) > a]\,, \forall\, a$$

.

Solution: First, note that $n = k + A_k + U_k$, since all nodes are either explored, active or undiscovered. Moreover, if $X \sim \mathit{Binomial}(m,q)$, then $m - X \sim \mathit{Binomial}(m, 1-q)$ – now since $U_k \sim \mathit{Binomial}(n-1,(1-p)^k)$, therefore $n-1-U_k = A_k + k - 1$ is distributed as $\mathit{Binomial}(n-1, 1-(1-p)^k)$.

Moreover, for $C(v) > k$, a necessary condition is that the active set A_k after k explorations is non-empty (this is not sufficient, as the active set could have become empty before the k^{th} exploration) – thus $\mathbf{P}[|C(v)| > k] \leq \mathbf{P}[A_k > 0]$. Now from the previous result, we have:

$$\begin{aligned}
\mathbf{P}[|C(v)| > k] &\leq \mathbf{P}[A_k > 0] = \mathbf{P}[A_k + k - 1 > k] \\
&= \mathbf{P}[\mathit{Binomial}(n-1, 1-(1-p)^k) > k] \\
&\leq \mathbf{P}[\mathit{Binomial}(n, pk) > k] \\
&\leq \inf_{\theta \geq 0} e^{-\theta k}\mathbf{E}[e^{\theta Bin(n,pk)}] = \inf_{\theta \geq 0}(1 - pk(1-e^{\theta}))^n e^{-\theta k} \\
&\leq \inf_{\theta \geq 0} e^{-npk(1-e^{\theta}) - \theta k} = e^{-k \sup_{\theta \geq 0}(\lambda(1-e^{\theta}) + \theta)} = e^{-kh(\lambda)}
\end{aligned}$$

Part (e)

If $\lambda < 1$, show that $h(\lambda) > 0$, and using this, argue that:

$$\mathbf{P}\left[\bigcup_{v\in V}\left\{|C(v)| > \frac{2\log n}{h(\lambda)}\right\}\right] \leq \frac{1}{n}$$

Note: This question was incomplete in the HW.

Solution: First, note that for $\theta > 0$ we have $e^{-\theta} > 1-\theta$. Now we have

$$\begin{aligned}\theta + \lambda(1-e^{\theta}) &= e^{\theta}(e^{-\theta}(\theta+\lambda)-\lambda)\\ &\geq e^{\theta}((1-\theta)(\theta+\lambda)-\lambda)\\ &= \theta e^{\theta}(1-\theta-\lambda)\end{aligned}$$

Now if $\lambda < 1$, then we can always choose $\theta \in (0, 1-\lambda)$ to ensure that $\theta + \lambda(1-e^{\theta}) > 0$. Hence $h(\lambda) = \sup_{\theta\geq 0}(\lambda(1-e^{\theta})+\theta) > 0$.

Now, using the union bound, we have:

$$\begin{aligned}\mathbf{P}\left[\bigcup_{v\in V}\left\{|C(v)| > \frac{2\log n}{h(\lambda)}\right\}\right] &\leq \sum_{v\in V}\mathbf{P}\left[|C(v)| > \frac{2\log n}{h(\lambda)}\right]\\ &\leq ne^{-\frac{2\log n}{h(\lambda)}h(\lambda)} \leq \frac{1}{n}\end{aligned}$$

Problem 5: (The Reed-Frost Epidemic Model)

In class we'll study the SIS epidemic model, where a node alternates between being infected and being susceptible. An alternate model with node recovery is called the SIR model. Here, as in the basic SI model, nodes are initial susceptible (S) to the infection, and get infected (I) due to interactions with infected neighbors – in particular, every pair of nodes meet according to an independent Poisson process with rate λ, and if an infected node meets a susceptible node, then the latter gets infected. In addition, we now model node recovery by assuming that an infected node i remains infected for a time T_i, after which she becomes resistant (R) to the infection – at this point she does not affect any other node in future.

Part (a)

Let $(N_S(t), N_I(t), N_R(t))$ be the number of susceptible, infected and resistant nodes at any time t, and $T_i \sim \mathit{Exponential}(1)$. Argue that $(N_S(t), N_I(t), N_R(t))$ forms a Markov chain, and write the state-transitions for different states (n_s, n_i, n_r). What are the absorbing states (if any) of this Markov chain?

Solution: Since we are considering the SIS infection on the complete graph, hence by symmetry, it is clear that $(N_S(t), N_I(t), N_R(t))$ forms a Markov chain. In particular, the state transitions obey:

$$(n_s, n_i, n_r) \rightarrow (n_s, n_i - 1, n_r + 1) \text{ at rate } n_i$$
$$(n_s, n_i, n_r) \rightarrow (n_s - 1, n_i + 1, n_r) \text{ at rate } \lambda n_s n_i$$

Now if $n_i = 0$, then it is clear that there are no state transitions – thus, all states of the form $(k, 0, n-k)$, $k \in [n]$ are absorbing states of the chain.

Part (b)

Suppose the T_i's are all constant (in particular, let $T_i = 1 \,\forall\, i$), and suppose the infection starts at a single node v. Argue now that the number of nodes that are eventually resistant to the SIR epidemic is the same as the connected component around v in a $G(n, p)$ graph. What value of λ is required to ensure that a constant fraction of all nodes experience the infection?

Solution: Consider any node v which gets infected at time t – now for any neighbor w of v which is susceptible at time t, it gets infected by v if and only if v meets w before recovering. This happens with probability $1 - e^{-\lambda}$ (i.e., it is equivalent to checking if an $Exponential(\lambda)$ r.v. is < 1). Note also that this holds independently for all. More generally, let X_{wv} be the indicator that v meets a neighbor w before recovering – then X_{wv} is i.i.d $Bernoulli(1 - e^{-\lambda})$. Suppose we instead construct a $G(n, 1 - e^{-\lambda})$ graph over the nodes V. Now if we start an infection at a node v, then the edges out of v can be thought to represent all nodes visited while infected – hence these nodes get infected. However, for each of these nodes, their outgoing edges again represent nodes visited while being infected – any of these which were not infected before are now infected.

Thus, the SIR infection (with recovery time = 1) is identical to doing a BFS of the neighborhood of v – clearly the set of nodes infected correspond to the connected component around v. We know that in order to have a giant component, we need $np > 1$ – thus, we need $n(1 - e^{-\lambda}) > 1$, i.e., $\lambda > \ln n - \ln(n-1)$.

Problem 6: (Time to Absorption in the Gambler's Ruin)

In class you saw the gambler's ruin problem, where a pair of gamblers A (starting with \$$a$) and B (starting with \$$b$; here a, b are positive integers), play a series of independent games, each of which results in A winning a dollar from B with probability $1/2$, else losing a dollar to B. Let T be the time when the game ends (i.e., when either player loses all their money). Using first-transition analysis, show that $\mathbf{E}[T] = a \cdot b$.
Hint: Define T_i be the time to absorption if A starts with \$$i$ and B with \$$(a + b - i)$; we want to find $\mathbf{E}[T] = \mathbf{E}[T_a]$. Now write a recurrence relation between the $\mathbf{E}[T_i]$. What is $\mathbf{E}[T_0]$ and $\mathbf{E}[T_{a+b}]$?

Solution: Let T_i be the time to absorption if A starts with \$$i$ and B with \$$(a+b-i)$. Then, for $0 < i < a+b$ we can get the following recurrence:

$$\mathbf{E}[T_i] = \frac{1}{2}(1+\mathbf{E}[T_{i-1}]) + \frac{1}{2}(1+\mathbf{E}[T_{i+1}]).$$

Also, note that $\mathbf{E}[T_0] = 0$ and $\mathbf{E}[T_{a+b}] = 0$. Therefore, from recurrence equations, for $0 \leq i \leq a+b$, we will get

$$\mathbf{E}[T_i] = i(a+b-i).$$

And hence, $\mathbf{E}[T] = \mathbf{E}[T_a] = ab$.