Sid Banerjee (sbanerjee@cornell.edu)

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 $\mathbf{E}[Y_j(x)] = 1/2^{j+1}$.

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:

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

Hint: Write N_j in terms of $Y_j(X_i)$.

Part (c)

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:

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

Hint: Use Chebyshev to bound $\mathbf{P}[N_j = 0]$.

Part (d)

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

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

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.

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 $|S(u) \cap S(v)| > 0$.

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

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.

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:

$$\mathbf{P}[w \in S(u) \cap S(v) | w \in B_{kd}(v)] = \frac{|B_{kd}(v)|}{|B_{(k+1)d}(u)|}$$

This put together with the previous parts shows that with probability at least $1/2$, the ADS sketch returns an estimate $\widehat{d}(u, v) \leq md^*(u, v)$.
Hint: Recall the MinHash sketch.

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 $n \to \infty$).

Part (b)

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

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

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:

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

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

Part (d)

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

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: $A_{n+1} = A_n + Y_n$, 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:

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

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

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 $U_k \sim Binomial(n-1, (1-p)^k)$.

Part (d)

Using the above result, we have that $n - 1 - U_k = A_k + k - 1$ is distributed as $Binomial(n-1, 1-(1-p)^k)$. Also note that a sufficient 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):

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

.

Part (e)

If $\lambda < 1$, then it is easy to show that $h(\lambda) < 0$

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 $T_i \sim Exponential(1)$. 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_t). What are the absorbing states (if any) of this Markov 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?

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 $\mathbf{E}[T_{a+b}]$?