HW3 solutions

Q1

(a)

Let us denote the sketch of A by $\langle a_1, a_2, \ldots, a_k \rangle$ and sketch of B be $\langle b_1, b_2, \ldots, b_k \rangle$. If there are r positions $i_1, i_2, \ldots, i_r$ such that $a_{i_j} = b_{i_j}, \forall j = 1, 2, \ldots, r$, then our estimate s(A,B) = r/k

We know that

$$Prob(a_i = b_i) = s(A, B)$$

Define an indicator random variable X_i which is 1 if $a_i = b_i$ and 0 otherwise. Note that X_is are all independent. Define $X = \sum_{i=1}^{k} X_i$
and $Y = X/k$

The Y is the estimate that we return.

$$E[X] = \sum_{i=1}^{k} E[X_i] = ks(A, B)$$
$$E[Y] = s(A, B)$$

Now we apply chernoff bound on Y.

$$\begin{aligned} Prob(|Y - E[Y]| > \epsilon E[Y]) &= Prob(k|Y - E[Y]| > \epsilon k E[Y]) = Prob(|X - E[X]| > \epsilon E[X]) \\ &\leq 2e^{-\frac{E[X]\epsilon^2}{3}} \leq 2e^{-\frac{k\epsilon^3}{3}} \end{aligned}$$

Set $k \geq \frac{3}{\epsilon^3} \log \frac{2}{\delta}$. We get

$$Prob(|Y - E[Y]| \geq \epsilon E[Y]) \leq \delta$$

(b)

From the LSH conditions we have :

$$Pr_{h \sim H}[h(D_i) = h(D_j) | d(T_i, T_j) \leq R] \geq 1 - R = p_1$$

$$Pr_{h\sim H}[h(D_i) = h(D_j)|d(T_i, T_j) \geq tR] \leq 1 - tR = p_2$$

Given, $sim(T_q, T_i) >= 0.8 \Rightarrow d(T_q, T_i) \leq 0.2 \Rightarrow p_1 = 0.8$

We are satisfied with a document that has similarity atleast $c \times 0.8 = 1 - d(T_i, T_q) = 1 - tR \Rightarrow p_2 = c \times 0.8$.

We have success probability given by the expression $1 - (1 - p_1^K)^L$. Given success probability is $1 - \frac{1}{e^2}$.
Therefore,$1 - (1 - p_1^K)^L \geq 1 - \frac{1}{e^2}$.

Setting $L \geq \frac{2}{p_1^K}$ gives us the required success probability because
$(1 - p_1^K)^{2/p_1^K} = ((1 - p_1^K)^{1/p_1^K}))^2 = 1/e^2$ (we have used $(1 - \frac{1}{x})^x \approx \frac{1}{e}$)

Query time is given by $O(KL + NLp_2^K)$ and to minimize it we set $Np_2^K = 1$
$\Rightarrow p_2 = (\frac{1}{N})^{1/K}$
$\Rightarrow \frac{p_2}{p_1} = \frac{(\frac{1}{N})^{1/K}}{(\frac{2}{L})^{1/K}}$
$\Rightarrow c \leq (\frac{L}{2N})^{1/K}$

Setting $K = \log N, L = \sqrt{N}$ gives us query time $O(KL) = O(\sqrt{N} \log N)$ and pre-processing time $O(NLK) = O(N^{3/2} \log N)$ which we want.

Q2

Suppose an item s has frequency $\geq \frac{m}{k+1} + 1$. Each copy of s either increments its own counter or decrement the counter of k items from the list. If possible assume s is not present at the end.

In each iteration there are two possiblities (a) either s is in the k elements (b) s is not in the k elements

If s is in the k elements with some count. Consider the situation when count of s is decremented. It has to be the case because s is not in the final k elements. Hence the total number of decrements is k+1 (1 is the new element and k elements in the list whose count is decremented)

If s is not in the k elements, the counter of all the k elements is reduced by 1 when s appears in the sequence. This leads to k+1 decrements.

Hence whenever s occurs it is either added (or increment of its own) or it causes decrements of all. For all these cases we found that there are k deletions corresponding to each occurrence of s because it is not in the final list. Hence the total deletions is atleast $(k + 1) * (\frac{m}{k+1} + 1) > m$ which is a contradiction as there cannot be more than m elements.

Q3

Let us define the following indicator random variables which will be useful.
$X_u = 1$ *if* $u \in V_p$. Otherwise, $X_u = 0$.

(a)

Storage Requirement is given by $S = |V_p| + \Sigma_{u \in V_p} |N(u)|$.
We can rewrite this in terms of our random variables as
$S = \Sigma_{u \in V} X_u + \Sigma_{u \in V} X_u |N(u)|$

$$\begin{aligned} E[S] &= E[\Sigma_{u \in V} X_u + \Sigma_{u \in V} X_u |N(u)|] \\ &= \Sigma_u E[X_u] + \Sigma_u E[X_u |N(u)|] \text{ (using linearity of expectation)} \\ &= \Sigma_u p + \Sigma_u p|N(u)| \\ E[S] &= pN + 2pM \end{aligned}$$

The last equality holds because $\Sigma_u |N(u)| =$ sum of degrees of nodes in $G = 2M$

(b)

We want to bound $E[S]$ in terms of M. We do that by bounding N in terms of M. Given graph G is connected, so it must have at least $N - 1$ edges. As it is not a multi-graph, the maximum number of edges is $\binom{N}{2}$.

$$N - 1 \leq M \leq \binom{N}{2} \leq \frac{N^2}{2}$$

$\Rightarrow N \leq M + 1$ and $N \geq \sqrt{2M}$

Therefore,

$$p\sqrt{2M} + 2pM \leq E[S] \leq 3pM + p$$

Substituting $p = \frac{60}{\sqrt{M}}$ gives us :

$$60\sqrt{2} + 120\sqrt{M} \leq E[S] \leq 180\sqrt{M} + \frac{60}{\sqrt{M}}$$

$$E[S] = \Theta(\sqrt{M})$$

(c)

$$D_p = \frac{1}{|V|} \Sigma_{u \in V_p} deg_{G_p}(u)$$

Observe that $deg_{G_p}(u) = |\{v | (u, v) \in E_p\}| = |N(u)| = deg_G(u)$

Using random variables we defined earlier, we can rewrite D_p as :

$$D_p = \frac{1}{|V|} \Sigma_{u \in V} X_u deg_G(u)$$

$$\hat{D} = \frac{D_p}{p}$$

(i)

$$E[\hat{D}] = \frac{1}{p}E[D_p]$$

$$\begin{aligned}
E[D_p] &= \frac{1}{|V|}\Sigma_{u\in V}E[X_u deg_G(u)] \\
&= \frac{1}{N}\Sigma_{u\in V}E[X_u]deg_G(u) \\
&= \frac{1}{N}\Sigma_{u\in V}p(deg_G(u)) \\
&= \frac{p}{N}\Sigma_{u\in V}deg_G(u) \\
&= \frac{2pM}{N}
\end{aligned}$$

$$E[\hat{D}] = \frac{2pM}{pN} = \frac{2M}{N}$$

(ii)

Observe that random variables X_u defined earlier are mutually independent. Therefore we can use linearity of variance.

$$Var[\hat{D}] = \frac{1}{p^2}Var[D_p] = \frac{1}{p^2N^2}Var[\Sigma_u X_u deg_G(u)]$$

$$Var[\hat{D}] = \frac{1}{p^2N^2}\Sigma_u deg_G(u)^2 Var[X_u]$$

We know variance of a Bernoulli random variable is $p(1-p)$.

$$Var[\hat{D}] = \frac{p(1-p)}{p^2N^2}\Sigma_u deg_G(u)^2 = \frac{1-p}{pN^2}\Sigma_{u\in V}deg_G(u)^2$$

(iii)

$$D = \frac{1}{|V|}\Sigma_u deg_G(u) = \frac{2M}{N}$$

Therefore, $E[\hat{D}] = D$.

Since D_p is a **weighted** sum of Bernoulli random variables, we cannot directly apply the Chernoff Bound that we have learnt in the class.

We can apply Chebyshev's inequality here to get :

$$Pr[|D - \hat{D}| > \frac{D}{2}] = Pr[|\hat{D} - E[\hat{D}]| > \frac{D}{2}] \leq \frac{Var[\hat{D}]}{(\frac{D}{2})^2}$$

We use the following identity which can be shown easily :
$\Sigma_{u \in V} deg_G(u)^2 = \Sigma_{(u,v) \in E}\{deg_G(u) + deg_G(v)\}$.

Given $max_u deg_G(u) \leq \sqrt{M}$.
So, $\Sigma_u deg_G(u)^2 = \Sigma_{(u,v) \in E}\{deg_G(u) + deg_G(v)\} \leq 2\sqrt{M}M = 2M^{3/2}$

$$Var[\hat{D}] = \frac{1-p}{pN^2}\Sigma_{u \in V} deg_G(u)^2 \leq \frac{1-p}{pN^2}2M^{3/2} = \frac{2M^{3/2}(1-p)}{pN^2} \leq \frac{2M^{3/2}}{pN^2}$$

The last inequality holds because $0 < p \leq 1$.
Therefore,

$$Pr[|D - \hat{D}| > \frac{D}{2}] \leq \frac{2M^{3/2}/pN^2}{(M/N)^2} = \frac{2}{M^{1/2}p} = \frac{1}{30}$$

(d)

Compute the number of distinct pairs of vertices (u, v) that are reachable via paths of length atmost 2 in G_p where $u \in V_p$ and $v \in V_p$. Let that count be T_p. Our estimate $\hat{T} = \frac{1}{p^2}T_p$.

We have to show that $E[\hat{T}] = T$.

We can write T_p as follows : $T_p = \Sigma_{u,v \in V_p} P_{uv}$.

Here, we define $P_{uv} = 1$ if there is a path of length atmost 2 between u and v in G, otherwise $P_{uv} = 0$

Using random variables we defined earlier, we can rewrite T_p as :

$$T_p = \Sigma_{u,v \in V} P_{uv} X_u X_v$$

The above equation is valid because we include P_{uv} in our estimate only when $X_u = X_v = 1$ which means $u, v \in V_p$.

$$E[T_p] = \Sigma_{u,v \in V} E[P_{uv} X_u X_v] = \Sigma_{u,v \in V} P_{uv} E[X_u X_v] = \Sigma_{u,v \in V} P_{uv} E[X_u] E[X_v]$$

The last equality is true because X_u, X_v are independent.

$$E[T_p] = p^2 \Sigma_{u,v \in V} P_{uv}$$

$$E[\hat{T}] = E[\frac{1}{p^2}T_p] = \frac{1}{p^2}E[T_p] = \frac{p^2}{p^2}\Sigma_{u,v \in V} P_{uv} = \Sigma_{u,v \in V} P_{uv} = T$$

This is because $T = \Sigma_{u,v \in V} P_{uv}$ i.e, number of all distinct pairs of vertices which are atmost 2 hops away from each other.