Algorithms for Data Science - 590D HW-3

Krishna Prasad Sankaranarayanan & Vinayak Mathur

Q1 i) Given two sets A and B we hash them to get equivalent sets which contain numbers. Once we have these sets, we define 't' different random permutations $\pi_1, \pi_2 \cdots \pi_t$

Now, we calculate the min wise signature for each of our set. To calculate the signature we simply check which number in the random permutation π_i matches first in our set.

Once we have the minwise signature for each set Jaccard similarity can be estimated by the following

Approximate Jaccard similarity = $\frac{\text{Number of Matching positions}}{\text{Total positions}}$

$= \frac{|S \cap T|}{|S \cup T|}$

ii) Size of Min-Hash computing

Consider that we have 2 sets s_1 & s_2. We would like to estimate Jaccard Similarity.

Define indicator random variable X_i which is 1 if the i^{th} min-hash match.

$$X = \sum_{i=1}^{t} X_i = r$$

∴ For both S_1 & S_2, the i^{th} min-hash will be same.

Since all $S_1 \cup S_2$ elements are equally likely to come first in the permutation, chance that 2 min-hashes are same is exactly $\left|\frac{S_1 \cap S_2}{S_1 \cup S_2}\right|$ which is same as Jaccard similarity of s_1 & s_2

$\Rightarrow$ Pr [$X_i = 1$] = Jaccard similarity (S_1, S_2)

$$\therefore E[X] = \sum_{i=1}^{t} E[X_i] = \sum_{i=1}^{t} Pr[X_i = 1]$$

$$= t \cdot \text{Jaccard Similarity } (S_1, S_2)$$

By the chernoff bound

$$Pr[X \in E[x](1 \pm \varepsilon)] \leq 2e^{\frac{-t \text{JaccardSim}(S_1,S_2)\varepsilon^2}{3}}$$

If we want that bad probability to be less the δ, we get

$$e^{\frac{-t \text{JaccardSimilarity}(S_1,S_2)\varepsilon^2}{3}} \leq \frac{2}{\delta}$$

(or)

$$t \geq \frac{3}{\text{JaccardSimilarity}(S_1,S_2)\varepsilon^2} \ln\frac{2}{\delta}$$

∴ Suppose we want to measure JacCard similarity to an accuracy $\pm\varepsilon$ only when similarity is at least $1/10$

∴ Size of hash =)

$$t \geq \frac{30}{\varepsilon^2} \ln\frac{2}{\delta}$$

ii) Similarity ≥ 0.8 document is selected

$\therefore 1 - 0.8 \leq |0.2| = k$

$p_1 \geq 1 + k = 1 - 0.2 = 0.8$

$p_2 \leq 1 - e^{-k} = 1 - e^{-(0.2)}$

$$\rho = \frac{\log 1/p_1}{\log 1/p_2} \Rightarrow \frac{\log 1/0.8}{\log 1/p_2} \Rightarrow \frac{\log 10/8}{\log 10^2/8^2} \Rightarrow p_2 = \frac{8^2}{10^2}$$

$p = 0.64$

Similarity (atleast) $\geq c * 0.8$

$\Rightarrow$ dist $\leq c * k \leq 0.36$

$\Rightarrow$ Similarity ≥ 0.64 $(c * 0.8)$

$\Rightarrow$ Maximum possible value of c in our scheme

$\Rightarrow c \leq 0.8$

2. Proof: algorithm frequent

for each $s \in \{1, 2, \ldots m\}$, counter on the list is atleast the number of occurrences of s in the stream minus $m/(k+1)$.

Suppose, in particular, if some s does not occur in the list, the counter $= 0$.

Also it says that it occurs fewer than $m/k+1$ times in the stream.

Consider each decrement counter step as eliminating some items.

An item is eliminated if it is currently being read and there are already 'k' symbols different from it and k other items are simultaniously eliminated.

Thus, the elimination of each occurance of an $s \in \{1, 2, \ldots m\}$ is elimination of $k+1$ items.

∴ No more than $n/(k+1)$ occurences of any symbol can be eliminated.

∴ Item x not eliminated.

⇒ It will be on the list at the end

∴ Any item that has frequency > $\frac{m}{k+1}$ times occur in the list.

Q3a The expected storage requirement of the algorithm will depend on the number of nodes selected and the count of their neighbours.

$\mathbb{E}[\text{storage requirement}] =$

$= \mathbb{E}[\quad]$

First let us define an indicator random variable

$X_u = \begin{cases} 1 & \text{if node } u \text{ is sampled} \\ 0 & \text{otherwise} \end{cases}$

$\mathbb{E}[\text{storage requirement}] = \mathbb{E}\left[\sum_{u \in V_p} X_u + \sum_{u \in V_p} X_u |N(u)|\right]$

using linearity of expectation

$= \mathbb{E}\left[\sum_{u \in V_p} X_u\right] + \mathbb{E}\left[\sum_{u \in V_p} X_u \, N(u) \, 2M\right]$

$= \sum \mathbb{E}[X_u] + \sum \mathbb{E}[X_u \, N(u) \, 2M]$

$= \sum p X_u + \sum p X_u \, N(u) \, 2M$

$= NP + 2NP \, N(u) M$

$= N \frac{60}{\sqrt{M}} + 2N \frac{60}{\sqrt{M}} N(u) M$

Q3b: In order to compute the expected storage requirement as a function of M, consider

Lower Bound

For the lower bound, consider a graph which has minimum possible edges

ie $M = N-1$

$\Rightarrow N = M+1$.

Substituting this in our previous result we get

Ans: $(M+1) \cdot \frac{60}{\sqrt{M}} + (M+1) \cdot \frac{60}{\sqrt{M}} \cdot 2M$

Upper Bound

For the upper bound, consider a fully connected graph

ie $M = {}^{N}C_2 = \frac{N!}{2! \times (N-2)!}$

$= \frac{N(N-1)}{2}$

$2M = N^2 - N$

$N^2 - N - 2M = 0$

$N = \frac{1 \pm \sqrt{1+8M}}{2}$ substituting in the previous result

Ans: $\frac{1 \pm \sqrt{1+8M}}{2} \cdot \frac{60}{\sqrt{M}} \left(1 + \frac{2M}{N(N-1)}\right)$

Q3c

i) Define an indicator random variable

$$X_u = \begin{cases} 1 & \text{if node } u \text{ is sampled} \\ 0 & \text{otherwise} \end{cases}$$

$$E[\hat{D}] = E\left[\frac{1}{p} D_p\right] = E\left[\frac{1}{p} \cdot \frac{1}{|V|} \sum_{u \in V_p} \deg G_p(u)\right]$$

$$= E\left[\frac{1}{p} \cdot \frac{1}{|V|} \sum_{u \in V} X_u \deg G(u)\right]$$

$$= \frac{1}{p} \cdot \frac{1}{|V|} \sum_{u \in V} E\left[X_u \deg G(u)\right]$$

$$= \frac{1}{p} \cdot \frac{1}{|V|} \; p \cdot \deg G(u)$$

$$= \frac{1}{|V|} \deg G(u)$$

$$E[\hat{D}] = D$$

QED

ii) $\text{Var}[\hat{D}] = ?$

$$\text{Var}[\hat{D}] = E[\hat{D}^2] - (E[\hat{D}])^2$$

$$= E\left[\left(\frac{1}{p} \cdot \frac{1}{|V|} \sum_{u \in V_p} \deg G_p(u)\right)^2\right] - (E[\hat{D}])^2$$

Let $X_u = 1$ if node u is sampled
0 otherwise

$$\text{Var}[\hat{D}] = \frac{1}{p^2|V|^2} \mathbb{E}\left[\left(\sum_{j=0}^{|V|} X_{u_j} \deg_G(u_j)\right)^2\right] - D^2$$

$$= \frac{1}{p^2|V|^2} \mathbb{E}\left[\sum_{j=0}^{|V|} (X_{u_j} \deg_G(u_j))^2 + \sum_{i \neq j}^{|V|} X_{u_i} \deg_G(u_i) \cdot X_{u_j} \deg_G(u_j)\right]$$

$$\left[\because \left(\sum a_i\right)^2 = \sum_j a_i^2 + \sum_{i \neq j} a_i a_j\right]$$

$$= \frac{1}{p^2|V|^2} \sum_{j=0}^{|V|} \mathbb{E}\left[(X_{u_j} \deg_G(u_j))^2\right] + \sum_{i \neq j}^{|V|} \mathbb{E}\left[X_{u_i} \deg_G(u_i) \cdot X_{u_j} \deg_G(u_j)\right]$$

Since each vertex is sampled independently

$$= \frac{1}{p^2|V|^2} NP \cdot (\deg_G(u_j))^2 \sum_{i=0}^{2|V|} \sum_{\substack{j=0 \\ i \neq j \\ j \neq i}}^{|V|} \mathbb{E}\left[X_{u_i} \deg_G(u_i) \cdot X_{u_j} \cdot \deg_G(u_j)\right]$$

$$= \frac{1}{p^2|V|^2} NP(\deg_G(u_j))^2 \cdot NP \deg_G(u_i) \cdot NP \deg_G(u)$$

$$= \frac{1}{p^2|V|^2} N^3 P^3 (\deg_G(u))^4$$

$$= NP(\deg_G(u))^4 \quad [\because |V| = N]$$

Q3c (iii) $\text{Prob}\left(|D - \hat{D}| \geq \frac{D}{2}\right)$

$= \text{Prob}\left(\left|\,|\hat{D}| - E(\hat{D})\,\right| \geq \frac{E[\hat{D}]}{2}\right)$

$\leq \frac{\text{Var}(\hat{D})}{(E[\hat{D}])^2} \cdot 4$

$\leq \frac{4 \cdot \text{Var}[\hat{D}]}{D^2}$

$\leq \frac{4 \cdot NP \cdot (\deg_4(u))^4}{D^2}$

$\leq \frac{4 \cdot N \cdot P\ |E|^2}{D^2}$

$\leq \frac{4 \cdot NP \cdot |E|^2 \cdot N^2}{|E|} = 4PN^3M$

$\text{Prob}\left(|D - \hat{D}| \geq \frac{D}{2}\right) \leq 4PN^3M$

3 d) Let G be the entire graph which is going to be sampled.

Given the $\hat{T}$ - number of distinct pair of vertices that are reachable via paths of length - 2. from G_p

G_p - sampled graph with probability p.

Given that $E[T] = \bar{T}$ => proving that $\hat{T}$ is no of distinct pairs of vertices that are reachable via length - 2

For every edge $e \in C_G(G, G_p)$ define an indicator random variable

$$X_e = \begin{cases} 1 & \text{if } e \text{ is sampled} \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow X = C_{G_p}(s, \bar{s})$$

$$= \sum_{e \in C_G(s,\bar{s})} X_e$$

$$E[X] = \sum_{e \in C_G(s,\bar{s})} E[X_e] = \sum_{e \in C_G(s,\bar{s})} E[X_e]$$

$$= p \cdot C_G(s, \bar{s})$$

$$E\left[\frac{X}{p}\right] = C_G(S,\bar{S})$$

The cut (sampling) value of the sampled graph is same as the orginal graph

∴ using chernoff bound ⇒

$$P\left[\left|\frac{X}{p} - C_G(S,\bar{S})\right| > \epsilon\, C_G(S,\bar{S})\right]$$

for length between 1 to 2 edges of 1 and atmost 2 hop.

2 hop

$$T = \left| \sum_{u,v \in C_G(S,\bar{S})} c(u,x) , \sum_{u,v \in C_G(S,\bar{S})} c(x,v) \right|$$

$$\therefore P\left[\frac{X}{p} - C_G(S,\bar{S})\right| > \epsilon \cdot C_G(S,\bar{S})]$$

$$\Rightarrow P[|X - p \cdot C_G(S,\bar{S})| > t\, p\, C_G(S,\bar{S})$$

$$= P[|X - E[X]| > t \cdot E[X]$$

$$\leq e^{-\frac{E[X]\epsilon^2}{2}}$$

$$\leq e^{-\frac{pc\epsilon^2}{2}}$$

∴ probability value ⇒ we need to reduce the deviation.

$p \geq \frac{2d \log n}{c\epsilon^2}$ ⇒

$$P\left[\frac{X}{p} \cdot t \cdot (1 \pm \epsilon) \cdot c\right] > \left(1 - \frac{1}{n^d}\right)$$

∴ Hence $E[T'] = T$ [From the above statement]

∴ Hence Justified that sampling of 2 hops in G is equal to sampling in G_p