A1Q1

1. (i) $f(S_1, S_2) = 1 - JS(S_1, S_2)$

$$= 1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

$$= 1 - \frac{|S_2 \cap S_1|}{|S_2 \cup S_1|}$$

$$= 1 - JS(S_2, S_1)$$

$$= f(S_2, S_1)$$

Since $|S_2 \cap S_1| \leq |S_1 \cup S_2|$

so $JS(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \leq 1$ $\quad f(S_1, S_2) = 1 - JS(S_1, S_2) \geq 0$

so $f(S_1, S_2) = f(S_2, S_1) \geq 0$

(ii) To prove the sufficiency:

when $S_1 = S_2$ then $|S_1 \cap S_2| = |S_1 \cup S_2| = 1$

$JS(S_1, S_2) = 1$

so $f(S_1, S_2) = 1 - JS(S_1, S_2) = 0$

To prove the necessity:

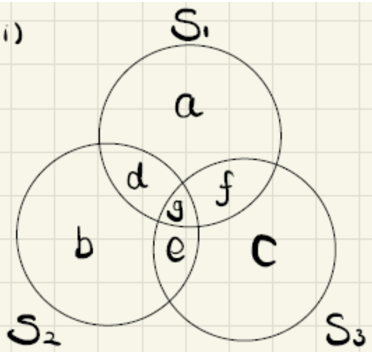
when $f(S_1, S_2) = 1 - JS(S_1, S_2) = 0$

then $JS(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} = 1$

thus $|S_1 \cap S_2| = |S_1 \cup S_2|$

so $S_1 = S_2$

(iii)



To prove $f(S_1, S_3) \le f(S_1, S_2) + f(S_2, S_3)$

is the same as proving $1 - JS(S_1, S_3) \le 2 - JS(S_1, S_2) - JS(S_2, S_3)$

$$JS(S_1, S_3) \ge JS(S_1, S_2) + JS(S_2, S_3) - 1$$

Assume that $JS(S_1, S_3) \le JS(S_1, S_2) + JS(S_2, S_3) - 1$

$$\frac{|S_1 \cap S_3|}{|S_1 \cup S_3|} \le \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} + \frac{|S_2 \cap S_3|}{|S_2 \cup S_3|} - 1$$

$$\frac{f+g}{a+c+d+e+f+g} \le \frac{d+g}{a+b+d+e+f+g} + \frac{g+e}{b+c+d+e+f+g} - 1$$

Since $\frac{f+g}{a+b+c+d+e+f+g} \le \frac{f+g}{a+c+d+e+f+g}$

$$\left.\begin{aligned} \frac{d+g}{a+b+d+e+f+g} &\le \frac{d+g+c}{a+b+c+d+e+f+g} \\ \frac{g+e}{b+c+d+e+f+g} &\le \frac{e+g+a}{a+b+c+d+e+f+g} \end{aligned}\right\}$$ because for any $0 \le x < y$, $0 \le z$ $\quad \frac{x}{y} \le \frac{x+z}{y+z}$

so $$\frac{f+g}{a+b+c+d+e+f+g} \le \frac{c+d+g}{a+b+c+d+e+f+g} + \frac{a+e+g}{a+b+c+d+e+f+g} - 1$$

$$f+g \le c+d+g+a+e+g-a-b-c-d-e-f-g$$

$$f+g \le g-b-f$$

$$f \le -b-f \le 0$$

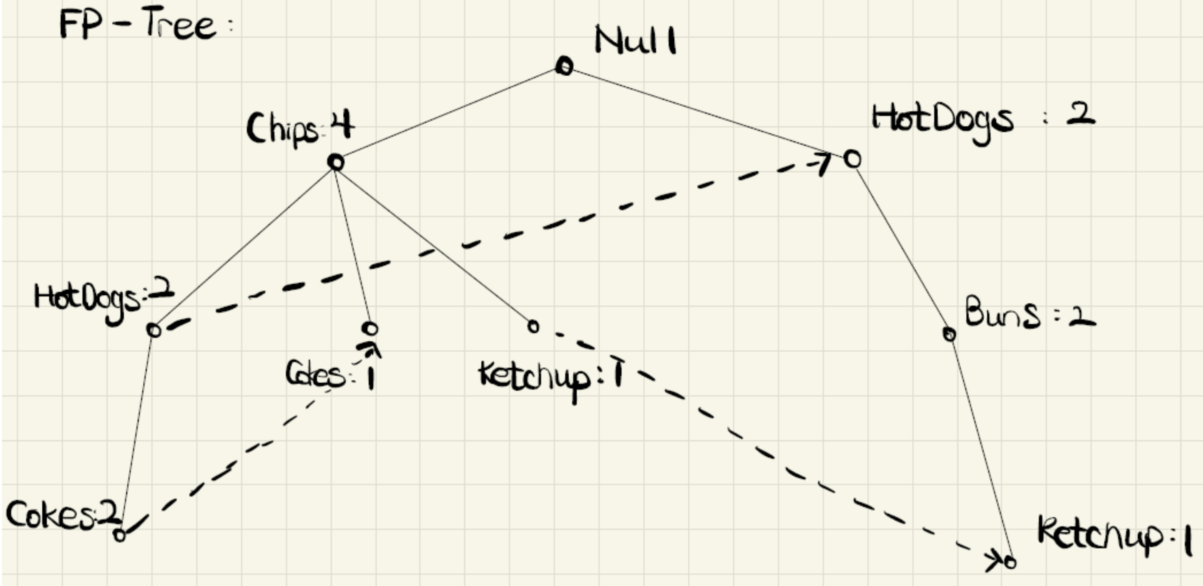
by contradiction, $f(S_1, S_3) \le f(S_1, S_2) + f(S_2, S_3)$ for any S_1, S_2, S_3

A1Q2

Chips	4
HotDogs	4
Coke	3
Buns	2
Ketchup	2

Transaction ID	Items
1	HotDogs Buns Ketchup
2	HotDogs Buns
3	Chips HotDogs Coke
4	Chips Coke
5	Chips Ketchup
6	Chips HotDogs Coke

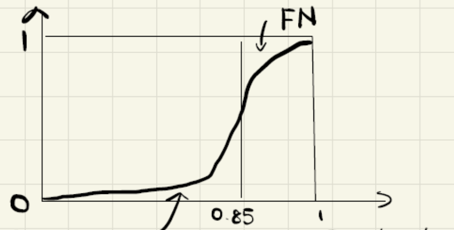
FP-Tree:



A1Q3

$k = 160 = b \times r$

$f(s) = 1 - (1 - s^r)^b$



minimize $\displaystyle\int_0^{0.85} 1 - (1 - s^r)^{\frac{160}{r}}\, ds$ $\longrightarrow$ False Positive

maximize $\displaystyle\int_{0.85}^{1} 1 - (1 - s^r)^b\, ds$

$\Rightarrow$ minimize $\displaystyle\int_{0.85}^{1} (1 - s^r)^b\, ds$ $\longrightarrow$ False Negative

After testing all possible value of r and b, I find that letting $r = 20$, $b = 8$ could minimize the sum of False Positive rate and False Negative rate.

A2Q1(1)

$\rho(S) \geqslant \lambda$ let $\bar{S}$ be $V \setminus S$

$$\frac{\sum_{(u,v)\in E(S)} w_{uv}}{|S|} \geqslant \lambda$$

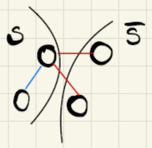
$$\sum_{(u,v)\in E(S)} w_{uv} \geqslant \lambda |S|$$

$$\frac{1}{2}\Big(\sum_{(u,v)\in E,\, u\in S} w_{uv} - \sum_{(u,v)\in E(\bar{S},S)} w_{uv} \Big) \geqslant \lambda |S|$$

$\uparrow$

here, we count the same edge twice.

i.e.



In $\sum_{(u,v):u\in S} w_{uv}$

we count the blue edge twice

we count the red edge once

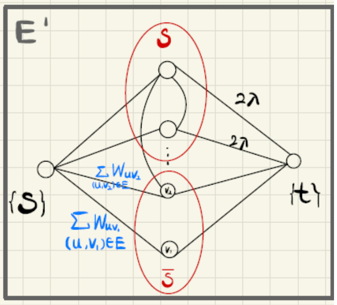
$$\sum_{(u,v)\in E,\, u\in S} w_{uv} + \sum_{(u,v)\in E,\, u\in \bar{S}} w_{uv} - \sum_{(u,v)\in E,\, u\in \bar{S}} w_{uv} - \sum_{(u,v)\in E(S,\bar{S})} w_{uv} \geqslant 2\lambda |S|$$

$$\sum_{(u,v)\in E,\, u\in V} w_{uv} \geqslant 2\lambda|S| + \sum_{(u,v)\in E,\, u\in \bar{S}} w_{uv} + \sum_{(u,v)\in E(S,\bar{S})} w_{uv}$$

Transform $G = \langle V, E\rangle$ to $G' = \langle V\cup\{s,t\}, E'\rangle$, $E \subseteq E'$

$(s,u)\in E'$, $(u,t)\in E'$ for all $u\in V$

let $$w_{uv} = \begin{cases} \sum_{(u,v)\in E,\, v=v} w_{uv} & \text{if } u=s,\ v\in V \\ 2\lambda & \text{if } u\in V,\ v=t \\ w_{uv} & \text{if } u\in V,\ v\in V \end{cases}$$



(s-t)-cut in G':

$WC(\{s\}\cup S,\ (V-S)\cup\{t\})$

$= WC(\{s\}\cup S,\ \bar{S}\cup\{t\})$

$$\Rightarrow \sum_{(u,v)\in E',\, u\in\{s\}\cup S,\, v\in \bar{S}\cup\{t\}} w_{uv} \leqslant \gamma$$

$$\sum_{(u,v)\in E',\, u\in S,\, v=\{t\}} w_{uv} + \sum_{(u,v)\in E',\, u\in\bar{S},\, v=\{s\}} w_{uv} + \sum_{(u,v)\in E',\, u\in S,\, v\in\bar{S}} w_{uv} \leqslant \gamma$$

$$2\lambda|S| + \sum_{(u,v)\in E,\, u\in\bar{S}} w_{uv} + \sum_{(u,v)\in E(S,\bar{S})} w_{uv} \leqslant \gamma$$

---

so by setting $\gamma = \sum_{(u,v)\in E,\, u\in V} w_{uv}$

we prove that these 2 problems are equivalent.

A2Q1(2)

original edge number: $|E|$ construct a new graph $G' = <V, E'>$

For any 2 nodes in original graph, we connect them together

For any edge in G': $W_{uv} = \begin{cases} 1 & \text{if } (u,v) \in E \\ -|E|-1 & \text{if } (u,v) \notin E \end{cases}$

Proof: In this case,

① The density of any clique in the original graph is

$\frac{C(C-1)}{2C} = \frac{C^2-C}{2C} = \frac{C-1}{2}$ which is monotanuously increasing

Therefore, the larger the clique is, the larger its new density is.

② For any subgraph S which is not a clique in the original graph, its density in the new graph G' is definitely smaller than 0.

because we add edges with weight $-|E|-1$ to it.

$\rho_{(S)} = \frac{\sum_{(u,v)\in E(S)} W_{uv}}{|S|}$ E(S) is the edge set of S in the new graph

Obviously, there must be at least a new edge with weight $\leq -|E|-1$

let the new edge be e

so $\rho_{(S)} \leq \frac{\sum_{(u,v)\in \{E(S)-e\}} W_{uv} - |E| - 1}{|S|} \leq \frac{|E| - |E| - 1}{|S|} < 0$

③ For a single node, its density is 0.

Therefore, the densest subgraph in the new graph is also the clique with maximum clique size in the original graph.

A2Q4(2)

Algorithm:

1. Start from k=1
2. Remove all nodes with degree no greater than k and adjust degrees of neighbor nodes of removed nodes
3. Set core number of each removed node as k
4. k=k+1
5. If the graph still has nodes, go to 2
6. If the graph is empty, terminate

Using the Linked-List data structure described in lecture slides, we can always maintain all nodes sorted based on their degrees. Every time when we need to reduce the degree of one node by 1, the time cost is only O(1). Thus, using the Linked-List Data structure, the total time complexity is O(m+n), where m is the number of edges and n is the number of nodes.

A3Q1

1. To prove that Randomized Response is $\ln 3$-differentially private

we need to prove $\exp(-\ln 3) \le \frac{Pr(\vec{y}_1 = \vec{y} | D_1)}{Pr(\vec{y}_2 = \vec{y} | D_2)} \le \exp(\ln 3)$

thus $\frac{1}{3} \le \frac{Pr(\vec{y}_1 = \vec{y} | D_1)}{Pr(\vec{y}_2 = \vec{y} | D_2)} \le 3$

Since D_1 and D_2 are neighbouring database, $D_1 = D \cup \{x_1\}$, $D_2 = D \cup \{x_2\}$, $x_1 \neq x_2$.

let $D_1 = \{a_1, a_2, a_3 \dots a_{n-1}, x_1\}$
$D_2 = \{a_1, a_2, a_3 \dots a_{n-1}, x_2\}$

$$\frac{P(\vec{y}_1 = \vec{y} | D_1)}{P(\vec{y}_2 = \vec{y} | D_2)} = \frac{P(y_{11} = y_1, y_{12} = y_2 \dots y_{1n} = y_n | a_1, a_2, \dots a_{n-1}, x_1)}{P(y_{21} = y_1, y_{22} = y_2 \dots y_{2n} = y_n | a_1, a_2, \dots a_{n-1}, x_2)}$$

$$= \frac{P(y_{11} = y_1 | a_1)}{P(y_{21} = y_1 | a_1)} \times \frac{P(y_{12} = y_2 | a_2)}{P(y_{22} = y_2 | a_2)} \times \dots \times \frac{P(y_{1(n-1)} = y_{n-1} | a_{n-1})}{P(y_{2(n-1)} = y_{n-1} | a_{n-1})} \times \frac{P(y_{1n} = y_n | x_1)}{P(y_{2n} = y_n | x_2)}$$

$$= \frac{P(y_{1n} = y_n | x_1)}{P(y_{2n} = y_n | x_2)}$$

since y_{1n} and y_{2n} are random response variable

we could write the probability of responses as

	Yes	No
$= y_n$	$\frac{3}{4}$	$\frac{1}{4}$
$\neq y_n$	$\frac{1}{4}$	$\frac{3}{4}$

$$\exp(-\ln 3) = \frac{1}{3} = \frac{\frac{1}{4}}{\frac{3}{4}} \le \frac{P(y_{1n} = y_n | x_1)}{P(y_{2n} = y_n | x_2)} \le \frac{\frac{3}{4}}{\frac{1}{4}} = 3 = \exp(\ln 3)$$

therefore, $\frac{1}{3} \le \frac{Pr(\vec{y}_1 = \vec{y} | D_1)}{Pr(\vec{y}_2 = \vec{y} | D_2)} \le 3$

randomize response is $\ln$-3 differentially private

A3Q2

$$\begin{bmatrix} \frac{\sqrt{3}+3}{6} & \frac{3-\sqrt{3}}{6} \\ \frac{3-\sqrt{3}}{6} & \frac{3+\sqrt{3}}{6} \\ \frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} \end{bmatrix} \begin{bmatrix} \sqrt{3+\sqrt{3}} & 0 \\ 0 & \sqrt{3-\sqrt{3}} \end{bmatrix} \begin{bmatrix} \frac{\sqrt{3}+1}{2\sqrt{\sqrt{3}+3}} & \frac{\sqrt{3}+1}{2\sqrt{\sqrt{3}+3}} & 0 & \frac{1}{\sqrt{\sqrt{3}+3}} \\ \frac{1-\sqrt{3}}{2\sqrt{3-\sqrt{3}}} & \frac{1-\sqrt{3}}{2\sqrt{3-\sqrt{3}}} & 0 & \frac{1}{\sqrt{3-\sqrt{3}}} \end{bmatrix}$$