<u>Definition</u>

Let D and D' be transactional (0-1) datasets. Define a relation $\equiv$ by

$D \equiv D' \iff D$ and D' are equivalent up to the order of transactions.

<u>Claim:</u>

Suppose we swapped edges $(i,j),(k,l) \in E(G)$ s.t $(i,l),(k,j) \notin E(G)$.

A different dataset $D' \not\equiv D$ is generated iff

$\exists c \in C \setminus \{j,l\}$ s.t. $D(i,c) \neq D(k,c)$.

set of column indices

<u>Proof:</u> (WTS: $D \not\equiv D' \iff \exists c \in C \setminus \{j,l\}$ s.t. $D(i,c) \neq D(k,c)$)

($\Longrightarrow$) (WTS: $\forall c \in C \setminus \{j,l\}$, $D(i,c) = D(k,c) \Rightarrow D \equiv D'$)

If we swapped eges $(i,j),(k,l) \in E(G)$ s.t. $(i,l),(k,j) \notin E(G)$ and

$\forall c \in C \setminus \{j,l\}$, $D(i,c) = D(k,c)$, then transactions $D(i) = D'(k)$ and

$D(k) = D'(i)$. Since only transactions i,k changed, we have that $D \equiv D'$.

($\Longleftarrow$) (WTS: $\exists c \in C \setminus \{j,l\}$ s.t. $D(i,c) \neq D(k,c) \Rightarrow D \not\equiv D'$)

If we swapped eges $(i,j),(k,l) \in E(G)$ s.t. $(i,l),(k,j) \notin E(G)$ and

$\exists c \in C \setminus \{j,l\}$ s.t. $D(i,c) \neq D(k,c)$, then transactions $D(i) \neq D'(k)$ and

$D(k) \neq D'(i)$. Since only transactions i,k changed, we have that $D \not\equiv D'$. $\square$

$$Q_{x,y} = \frac{\#\text{ of swappable pairs from } x \text{ to } y}{\#\text{ of swappable pairs}}$$

States:

S_1:

a	b	c
c	d	
d		
d		

S_2:

b	c	d
a	c	
d		
d		

S_3:

b	c	d
c	d	
a		
d		

S_4:

a	c	d
b	c	
d		
d		

S_5:

a	c	d
c	d	
b		
d		

S_6:

a	b	d
c	d	
c		
d		

S_7:

b	c	d
a	d	
c		
d		

S_8:

a	c	d
b	d	
c		
d		

Swappable pairs:

S_1: $(1,5) \to S_2$, $(1,6) \to S_3$, $(1,7) \to S_3$, $(2,5) \to S_4$, $(2,6) \to S_5$, $(2,7) \to S_5$, $(3,6) \to S_6$, $(3,7) \to S_6$

S_2: $(1,4) \to S_4$, $(3,4) \to S_1$, $(4,6) \to S_3$, $(4,7) \to S_3$, $(5,6) \to S_7$, $(5,7) \to S_7$

S_3: $(1,6) \to S_5$, $(2,6) \to S_6$, $(3,6) \to S_1$, $(4,6) \to S_7$, $(5,6) \to S_2$

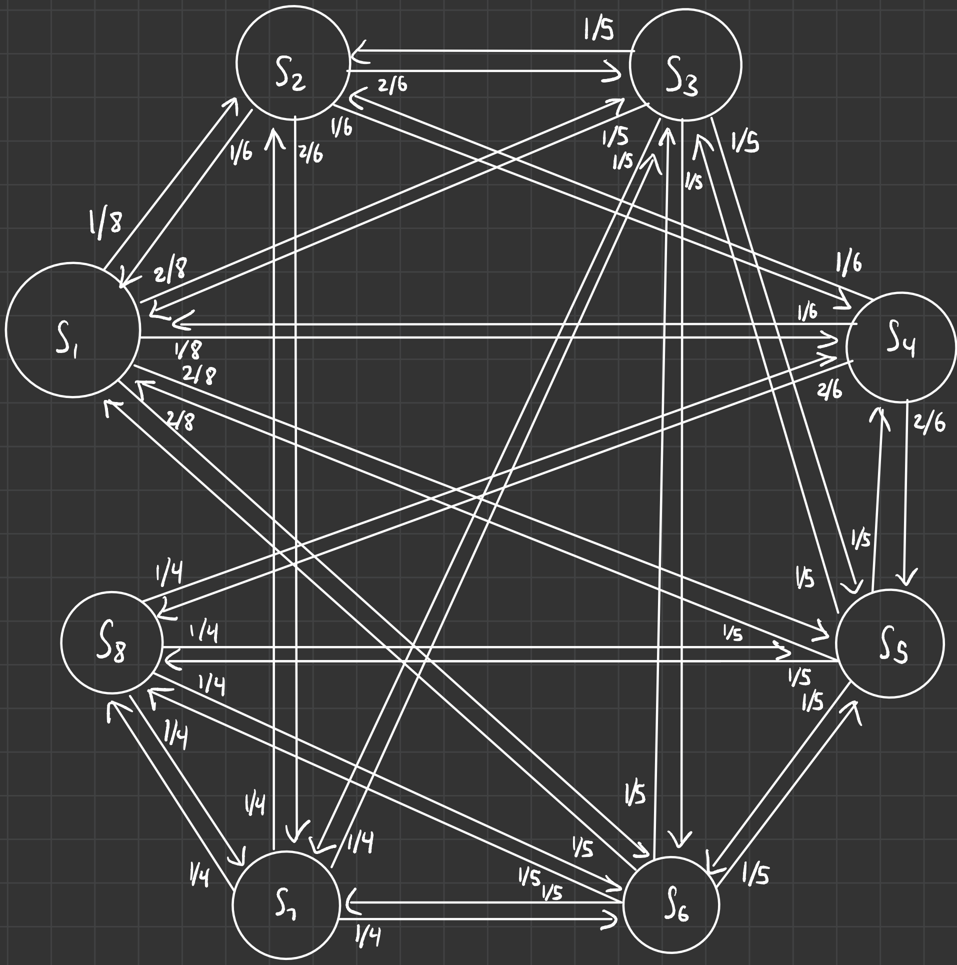
S_4: $(1,4) \to S_2$, $(3,4) \to S_1$, $(4,6) \to S_5$, $(4,7) \to S_5$, $(5,6) \to S_8$, $(5,7) \to S_8$

S_5: $(1,6) \to S_3$, $(2,6) \to S_6$, $(3,6) \to S_1$, $(4,6) \to S_8$, $(5,6) \to S_4$

S_6: $(1,4) \to S_7$, $(1,6) \to S_3$, $(2,4) \to S_8$, $(2,6) \to S_5$, $(3,6) \to S_1$

S_7: $(1,4) \to S_8$, $(2,4) \to S_6$, $(4,6) \to S_3$, $(5,6) \to S_2$

S_8: $(1,4) \to S_7$, $(2,4) \to S_6$, $(4,6) \to S_5$, $(5,6) \to S_4$



$$\begin{bmatrix} \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \end{bmatrix}
\begin{array}{c|cccccccc}
 & S_1 & S_2 & S_3 & S_4 & S_5 & S_6 & S_7 & S_8 \\
\hline
S_1 & 0 & \frac{1}{8} & \frac{2}{8} & \frac{1}{8} & \frac{2}{8} & \frac{2}{8} & 0 & 0 \\
S_2 & \frac{1}{6} & 0 & \frac{2}{6} & \frac{1}{6} & 0 & 0 & \frac{2}{6} & 0 \\
S_3 & \frac{1}{5} & \frac{1}{5} & 0 & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 \\
S_4 & \frac{1}{6} & \frac{1}{6} & 0 & 0 & \frac{2}{6} & 0 & 0 & \frac{2}{6} \\
S_5 & \frac{1}{5} & 0 & \frac{1}{5} & \frac{1}{5} & 0 & \frac{1}{5} & 0 & \frac{1}{5} \\
S_6 & \frac{1}{5} & 0 & \frac{1}{5} & 0 & \frac{1}{5} & 0 & \frac{1}{5} & \frac{1}{5} \\
S_7 & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} \\
S_8 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0
\end{array}
= \begin{bmatrix} \frac{7}{60} & \frac{89}{960} & \frac{37}{240} & \frac{89}{960} & \frac{37}{240} & \frac{23}{160} & \frac{59}{480} & \frac{59}{480} \end{bmatrix}$$

Not uniform!

M-H

$$P_{x,y} = Q_{x,y} \min(Q_{y,x}/Q_{x,y}, 1) = \min(Q_{x,y}, Q_{y,x}) \quad \text{if } x \neq y \wedge y \in N(x)$$

$P_{s_1,s_2} = \min(\frac{1}{8}, \frac{1}{6}) = \frac{1}{8}$ $P_{s_1,s_3} = \min(\frac{2}{8}, \frac{1}{5}) = \frac{1}{5}$ $P_{s_1,s_4} = \min(\frac{1}{8}, \frac{1}{6}) = \frac{1}{8}$

$P_{s_1,s_5} = \min(\frac{2}{8}, \frac{1}{5}) = \frac{1}{5}$ $P_{s_1,s_6} = \min(\frac{2}{8}, \frac{1}{5}) = \frac{1}{5}$ $P_{s_1,s_1} = 1 - 2(\frac{1}{8}) - 3(\frac{1}{5}) = \frac{3}{20}$

$P_{s_2,s_1} = \min(\frac{1}{6}, \frac{1}{8}) = \frac{1}{8}$ $P_{s_2,s_3} = \min(\frac{2}{6}, \frac{1}{5}) = \frac{1}{5}$ $P_{s_2,s_4} = \min(\frac{1}{6}, \frac{1}{6}) = \frac{1}{6}$

$P_{s_2,s_7} = \min(\frac{2}{6}, \frac{1}{4}) = \frac{1}{4}$ $P_{s_2,s_2} = 1 - \frac{1}{8} - \frac{1}{5} - \frac{1}{6} - \frac{1}{4} = \frac{31}{120}$

$P_{s_3,s_1} = \min(\frac{1}{5}, \frac{2}{8}) = \frac{1}{5}$ $P_{s_3,s_2} = \min(\frac{1}{5}, \frac{2}{6}) = \frac{1}{5}$ $P_{s_3,s_5} = \min(\frac{1}{5}, \frac{1}{5}) = \frac{1}{5}$

$P_{s_3,s_6} = \min(\frac{1}{5}, \frac{1}{5}) = \frac{1}{5}$ $P_{s_3,s_7} = \min(\frac{1}{5}, \frac{1}{4}) = \frac{1}{5}$ $P_{s_3,s_3} = 1 - 5(\frac{1}{5}) = 0$

$P_{s_4,s_1} = \min(\frac{1}{6}, \frac{1}{8}) = \frac{1}{8}$ $P_{s_4,s_2} = \min(\frac{1}{6}, \frac{1}{6}) = \frac{1}{6}$ $P_{s_4,s_5} = \min(\frac{2}{6}, \frac{1}{5}) = \frac{1}{5}$

$P_{s_4,s_8} = \min(\frac{2}{6}, \frac{1}{4}) = \frac{1}{4}$ $P_{s_4,s_4} = 1 - \frac{1}{8} - \frac{1}{6} - \frac{1}{5} - \frac{1}{4} = \frac{31}{120}$

$P_{s_5,s_1} = \min(\frac{1}{5}, \frac{2}{8}) = \frac{1}{5}$ $P_{s_5,s_3} = \min(\frac{1}{5}, \frac{1}{5}) = \frac{1}{5}$ $P_{s_5,s_4} = \min(\frac{1}{5}, \frac{2}{6}) = \frac{1}{5}$

$P_{s_5,s_6} = \min(\frac{1}{5}, \frac{1}{5}) = \frac{1}{5}$ $P_{s_5,s_8} = \min(\frac{1}{5}, \frac{1}{4}) = \frac{1}{5}$ $P_{s_5,s_5} = 1 - 5(\frac{1}{5}) = 0$

$P_{s_6,s_1} = \min(\frac{1}{5}, \frac{2}{8}) = \frac{1}{5}$ $P_{s_6,s_3} = \min(\frac{1}{5}, \frac{1}{5}) = \frac{1}{5}$ $P_{s_6,s_5} = \min(\frac{1}{5}, \frac{1}{5}) = \frac{1}{5}$

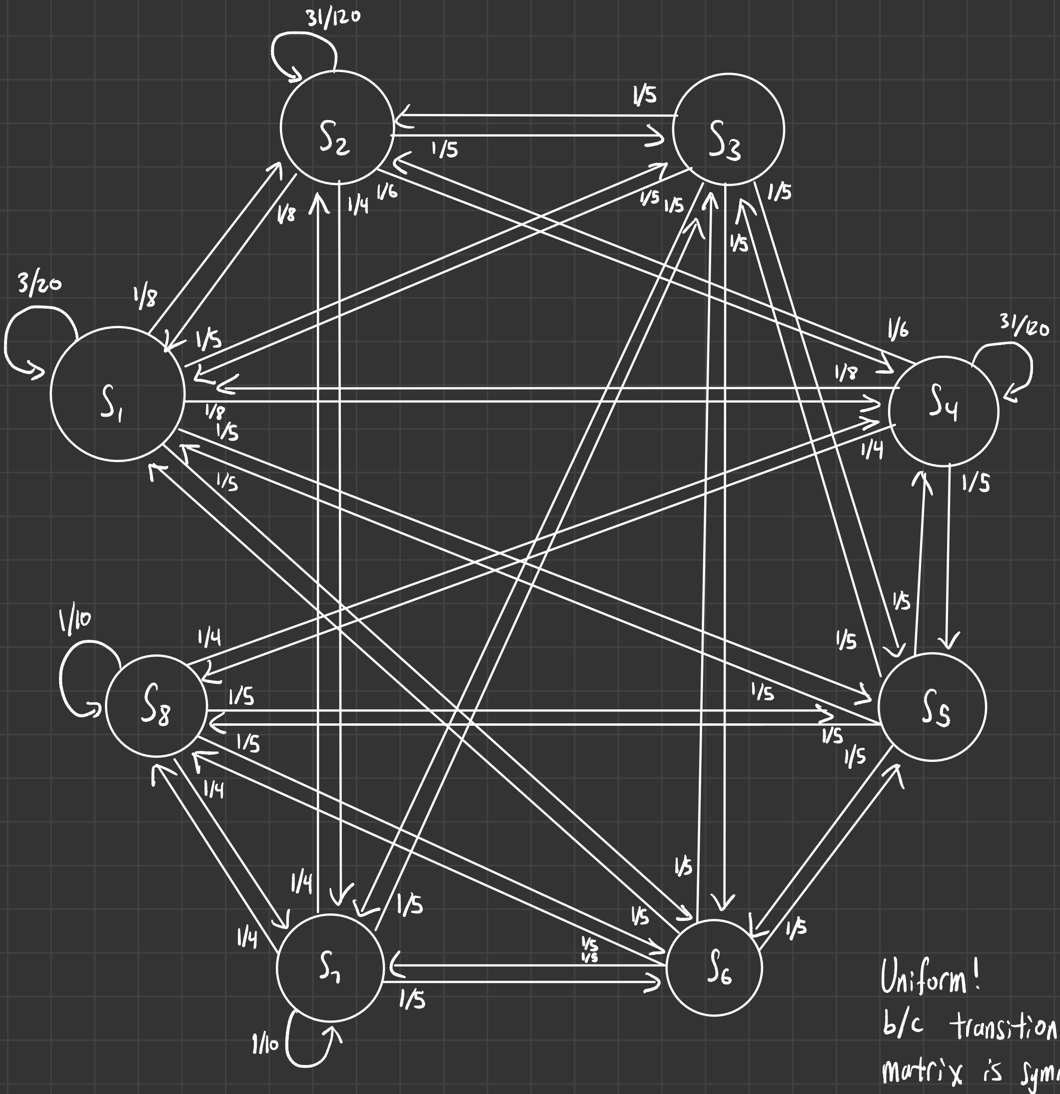
$P_{s_6,s_7} = \min(\frac{1}{5}, \frac{1}{4}) = \frac{1}{5}$ $P_{s_6,s_8} = \min(\frac{1}{5}, \frac{1}{4}) = \frac{1}{5}$ $P_{s_6,s_6} = 1 - 5(\frac{1}{5}) = 0$

$P_{s_7,s_2} = \min\left(\frac{1}{4}, \frac{2}{6}\right) = \frac{1}{4}$ $\quad P_{s_7,s_3} = \min\left(\frac{1}{4}, \frac{1}{5}\right) = \frac{1}{5}$ $\quad P_{s_7,s_6} = \min\left(\frac{1}{4}, \frac{1}{5}\right) = \frac{1}{5}$

$P_{s_7,s_8} = \min\left(\frac{1}{4}, \frac{1}{4}\right) = \frac{1}{4}$ $\quad P_{s_7,s_7} = 1 - 2\left(\frac{1}{4}\right) - 2\left(\frac{1}{5}\right) = \frac{1}{10}$

$P_{s_8,s_4} = \min\left(\frac{1}{4}, \frac{2}{6}\right) = \frac{1}{4}$ $\quad P_{s_8,s_5} = \min\left(\frac{1}{4}, \frac{1}{5}\right) = \frac{1}{5}$ $\quad P_{s_8,s_6} = \min\left(\frac{1}{4}, \frac{1}{5}\right) = \frac{1}{5}$

$P_{s_8,s_7} = \min\left(\frac{1}{4}, \frac{1}{4}\right) = \frac{1}{4}$ $\quad P_{s_8,s_8} = 1 - 2\left(\frac{1}{4}\right) - 2\left(\frac{1}{5}\right) = \frac{1}{10}$



Uniform!
b/c transition matrix is symmetric

Calculating $Q_{x,y}$:

$Q_{x,y} = \dfrac{\sigma_{x,y}}{\tau_x}$, where $\sigma_{x,y}$ is the # of swappable pairs that go from state x to state y and $\tau_x = \sum_y \sigma_{x,y}$ is the total # of swappable pairs for state x.

number of possible swaps that Gionis et al. use

$$\tau_G = d(G) - H(G),$$

$$H(G) = \sum_{\substack{i,k \in U \\ i<k}} \mathbb{1}[(r_i = r_k) \wedge (h_{i,k} = 2)],$$

// # swaps that lead to the same dataset up to the order of transactions

where $h_{i,k} = |\{v \in V : D(i,v) \neq D(k,v)\}|$

<u>Claim</u>: The number of pairs of edges $(i,j), (k,\ell) \in E(G)$ s.t. $(i,\ell), (k,j) \notin E(G)$ that lead to the same dataset up to the order of transactions is $H(G) = \sum_{\substack{i,k \in U \\ i<k}} \mathbb{1}[(r_i = r_k) \wedge (h_{i,k} = 2)]$,

U is the set of row indices

where r_i is the number of 1's in row i and

V is the set of column vertices

$h_{i,k} = |\{v \in V : D(i,v) \neq D(k,v)\}|$

Claim (rephrased):

Let $W = \{((i,j),(k,\ell)) : i<k \wedge (i,j),(k,\ell) \in E(G) \wedge (i,\ell),(k,j) \notin E(G)\}$.

Let $A = \{((i,j),(k,\ell)) \in W : \forall c \in C \setminus \{j,\ell\},\ D(i,c) = D(k,c)\}$.

A is the set of pairs of edges that lead to the same dataset

Let $H = \{(i,k) \in U : i<k \wedge r_i = r_k \wedge h_{i,k} = 2\}$.

Then, $|A| = |H|$.

Proof:

Suppose $f : A \to H$ is defined by $f((i,j),(k,\ell)) = (i,k)$, where $((i,j),(k,\ell)) \in A$ and $(i,k) \in H$.

Let $a_1, a_2 \in A$ s.t. $a_1 = ((i_1,j_1),(k_1,\ell_1))$ and $a_2 = ((i_2,j_2),(k_2,\ell_2))$.

$f(a_1) = f(a_2) \implies f((i_1,j_1),(k_1,\ell_1)) = f((i_2,j_2),(k_2,\ell_2))$

$\implies (i_1,k_1) = (i_2,k_2)$

$\implies i_1 = i_2$ and $k_1 = k_2$

$\implies j_1 = j_2$ and $\ell_1 = \ell_2$ because there are only 2 columns where the values are different between rows $i_1 = i_2$ and $k_1 = k_2$

$\implies ((i_1,j_1),(k_1,\ell_1)) = ((i_2,j_2),(k_2,\ell_2))$

$\implies a_1 = a_2$

Thus, f is injective.

Let $h \in H$ s.t. $h = (i,k)$. Since $h_{i,k} = 2$, there are exactly 2 columns where values in row i and k differ. Let's call these 2 columns j and l.

Thus, $D(i,j) \neq D(k,j)$ and $D(i,l) \neq D(k,l)$.

Further, since $r_i = r_k$, $D(i,j) \neq D(i,l)$ and $D(k,j) \neq D(k,l)$.

So we either have $D(i,j)=1$, $D(i,l)=0$, $D(k,l)=1$, $D(k,j)=0$ or

$$D(i,j)=0,\ D(i,l)=1,\ D(k,l)=0,\ D(k,j)=1.$$

Hence, either $((i,j),(k,l)) \in W$ or $((i,l),(k,j)) \in W$.

W.L.O.G., suppose $((i,j),(k,l)) \in W$.

Then, if $a \in A$ s.t. $a = ((i,j),(k,l))$,

$f(a) = f((i,j),(k,l)) = (i,k) = h$.

Therefore, f is surjective.

Since $f: A \to H$ is bijective, $|A| = |H|$. $\square$

Claim:

Suppose we swapped edges $(i,j),(k,\ell)\in E(G)$ s.t. $(i,\ell),(k,j)\notin E(G)$ and $\exists c\in C\setminus\{j,\ell\}$ s.t. $D(i,c)\neq D(k,c)$ to get a new dataset $D'\neq D$. Let $\sigma_{D,D'}$ be the number of swappable pairs of edges that go from D to D'. $\sigma_{D,D'}=(\#\text{ rows }=i)\cdot(\#\text{ row }=k)$.

Claim (rephrased):

Suppose we swapped edges $(i,j),(k,\ell)\in E(G)$ s.t. $(i,\ell),(k,j)\notin E(G)$ and $\exists c\in C\setminus\{j,\ell\}$ s.t. $D(i,c)\neq D(k,c)$ to get a new dataset $D'\neq D$.

Let $S=\{((w,x),(y,z)) : w<y \wedge \text{Swapping } (w,x) \text{ and } (y,z) \text{ goes from } D \text{ to } D'\}$.

Let $B=\{(w,y)\in U : w<y \wedge D(w)=D(i) \wedge D(y)=D(k)\}$.

Then, $|S|=|B|$. // Note: $|B|=(\#\text{ rows }=i)\cdot(\#\text{ rows }=k)$

Proof:

Suppose $f:S\to B$ is defined by $f((w,x),(y,z))=(w,y)$, where $((w,x),(y,z))\in S$ and $(w,y)\in B$.

Let $s_1,s_2\in S$ s.t. $s_1=((w_1,x_1),(y_1,z_1))$ and $s_2=((w_2,x_2),(y_2,z_2))$.

$f(s_1)=f(s_2)\implies f((w_1,x_1),(y_1,z_1))=f((w_2,x_2),(y_2,z_2))$

$$\Rightarrow (w_1, y_1) = (w_2, y_2)$$

$$\Rightarrow w_1 = w_2 \text{ and } y_1 = y_2$$

$$\Rightarrow x_1 = x_2 \text{ and } z_1 = z_2$$ b/c there can only be 2 columns when swapping across rows $w_1 = w_2$ and $y_1 = y_2$ that lead to D'

$$\Rightarrow ((w_1, x_1), (y_1, z_1)) = ((w_2, x_2), (y_2, z_2))$$

$$\Rightarrow s_1 = s_2$$

Thus, f is injective.

Let $b \in B$ s.t. $b = (w, y)$.

Hence, $D(w) = D(i)$ and $D(y) = D(k)$.

So swapping (w, j) and (y, l) also lead to D'.

Then, if $s \in S$ s.t. $s = ((w, j), (y, l))$,

$f(s) = f((w, j), (y, l)) = (w, y) = b.$

Therefore, f is surjective.

Since $f: S \to B$ is bijective, $|S| = |B|$. $\square$

Calculating ΔH:

Recall: $H(G) = \sum_{\substack{i,k \in U \\ i<k}} \mathbb{1}[(r_i = r_k) \wedge (h_{i,k} = 2)]$, $h_{i,k} = |\{v \in V : D(i,v) \neq D(k,v)\}|$

Let MM be a hash map where the key value pairs are defined as

row margin $\longrightarrow$ list of row indices that have row margins equal to the key.

Let HM be a hash map where the key value pairs are defined as

$(i,k) \rightarrow h_{i,k}$, where $i,k \in U$.

Suppose we swapped edges (i,j) and (k,ℓ).

$D(i,j) = 1 \rightarrow D'(i,j) = 0 \qquad D(k,\ell) = 1 \rightarrow D'(k,\ell) = 0$

$D(i,\ell) = 0 \rightarrow D'(i,\ell) = 1 \qquad D(k,j) = 0 \rightarrow D'(k,j) = 1$

Only rows i,k and columns j,ℓ change.

So $\forall u \in MM[r_i]$, $h_{i,u}$ and $h_{k,u}$ might have changed.

```
get ΔH():
    ΔH ← 0
    for u in MM[r_i]:
        for w ∈ {i, k}:
            if (u = w) ∨ (u = i ∧ w = k) ∨ (u = k ∧ w = i):
                continue
            h_{u,w} ← HM[(u,w)]    // (u,w) is in increasing order
            for c ∈ {j, l}:
                if D(u,c) = D(w,c):
                    HM[(u,w)] ← HM[(u,w)] - 1
                else:
                    HM[(u,w)] ← HM[(u,w)] + 1
            if HM[(u,w)] = 2 ∧ h_{u,w} ≠ 2:
                ΔH ← ΔH + 1
            else if HM[(u,w)] ≠ 2 ∧ h_{u,w} = 2:
                ΔH ← ΔH - 1
    return ΔH

// O(m), where m is the number of rows
```

Calculating $\sigma_{D,D'}$:

Recall: $\sigma_{D,D'} = (\#\text{rows} = i) \cdot (\#\text{rows} = k)$, assuming we swapped edges (i,j) and (k,ℓ).

Let RM be a hash map where the key value pairs are defined as

row vector $\rightarrow$ number of rows equal to the key.

Then, $\sigma_{D,D'} = RM[D(i)] \cdot RM[D(k)]$

After swapping (i,j) and (k,ℓ),

for $u \in \{i, k\}$:

$RM[D(u)] \leftarrow RM[D(u)] - 1$

$RM[D'(u)] \leftarrow RM[D'(u)] + 1$