CS 422 : Data Mining A20402686

Homework 4

I-I. Leskovec - ch 3

exercise 3.1.1.
Compute Jacaard similarities of each pair of following 3 sets:-

$A = \{1,2,3,4\}$ $B = \{2,3,5,7\}$ $C = \{2,4,6\}$

→ Jacaard similarity

$$S(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

i) $S(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{2}{6} = \frac{1}{3}$

ii) $S(B,C) = \frac{|B \cap C|}{|B \cup C|} = \frac{1}{6}$

iii) $S(A,C) = \frac{|A \cap C|}{|A \cup C|} = \frac{2}{5}$

Exercise 3.2.1
→ First 10- 3-~~s~~ shingles in first sentence:-
The, he_, e_m, _mo, mos, ost, st_, ~~te~~t_e,
_eF, eff

Shingles for words:-
" The most effective", "most effective way",
"effective way to", "way to represent", "to represent document", represent documents as", "documents as sets", "as sets for", "sets for purpose", "for purpose of",

Exercise 3.3.3

Element	S_1	S_2	S_3	S_4	$2x+1$ mod 6	$3x+2$ mod 6	$5x+2$ mod 6
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5.	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

Signature Matrix

	S_1	S_2	S_3	S_4
$h_1(0)$		1		1.
$h_2(0)$		2		2
$h_3(0)$		2		2
$h_1(1)$		1		1
$h_2(1)$		2		2
$h_3(1)$		1		2
$h_1(2)$	5	1		1,
$h_2(2)$	2	2		2
$h_3(2)$	0	1		0
$h_1(3)$	5	1	1	1
$h_2(3)$	2	2	5	2
$h_3(3)$	0	1	5	0
$h_1(4)$	5	1	1	1
$h_2(4)$	2	2	2	2
$h_3(4)$	0	1	4	0
$h_1(5)$	5	1	1	1
$h_2(5)$	2	2	2	2
$h_3(5)$	0	1	4	0

Final Minhash signature matrix:-

5	1	1	1
2	2	2	2
0	1	4	0

b) $h_3(x) = 5x+2 \mod 6$ is true permutation.

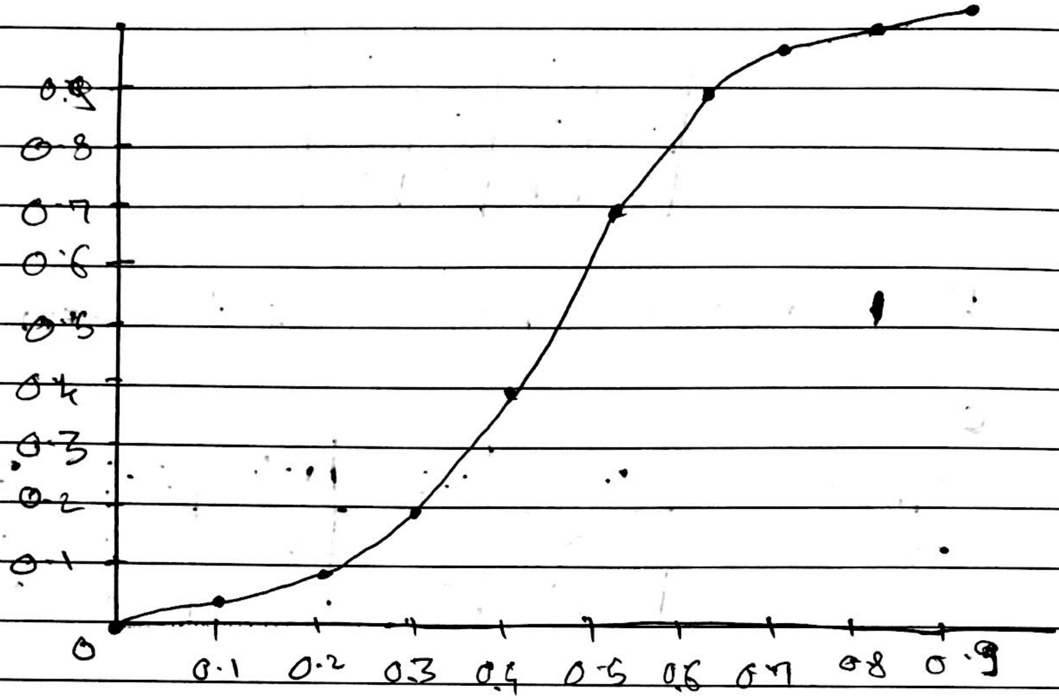
c)

Similarities	1-2	1-3	1-4	2-3	2-4	3-4
True	0	0	0.25	0	0.25	0.25
Estimated	0.33	0.33	0.67	0.67	0.67	0.67

Exercise 3.4.1

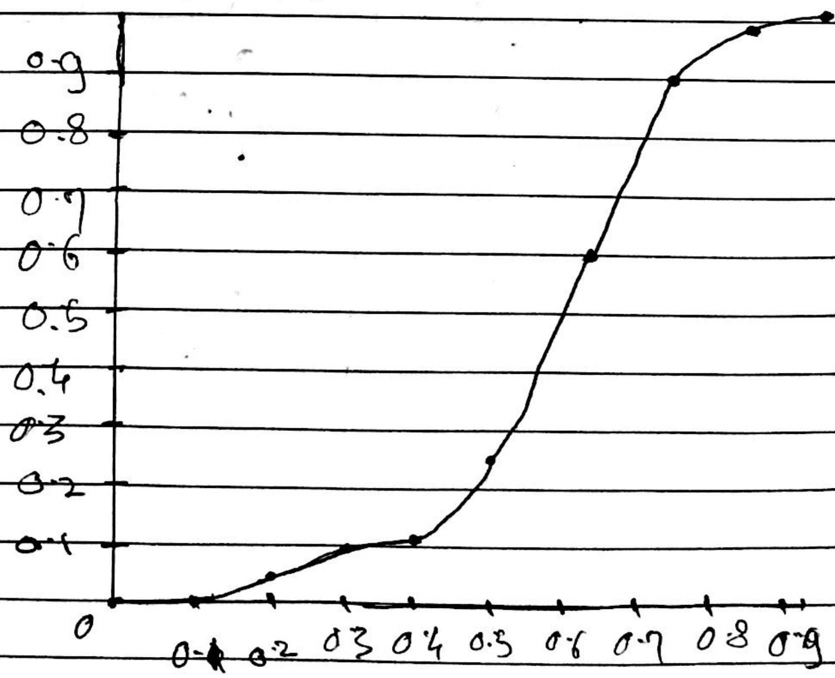
a) $r=3$, $b=10$

s	$1-(1-s^r)^b$
0.1	0.01
0.2	0.07
0.3	0.23
0.4	0.40
0.5	0.73
0.6	0.91
0.7	0.98
0.8	0.9992
0.9	0.9999



b) $r = 6$ & $b = 20$

s	$1-(1-s^r)^b$
0.1	0
0.2	0.001
0.3	0.04
0.4	0.078
0.5	0.27
0.6	0.615
0.7	0.918
0.8	0.997
0.9	0.999



Exercise 3.5.4

Jacard distance $d(x,y) = 1 - SIM(x,y)$

a) $A = \{1,2,3,4\}$ $B = \{2,3,4,5\}$

$$SIM(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{3}{5}$$

$$d(A,B) = 1 - \frac{3}{5} = 2/5$$

b) $A = \{1,2,3\}$ $B = \{4,5,6\}$

$$SIM(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{0}{6} = 0$$

$$d(A,B) = 1 - 0 = 1$$

Exercise 3.5.5.

Cosines of angle:-

a) $(3,-1,2)$ & $(-2,3,1)$

$x = [3,-1,2]$ $y = [-2,3,1]$

$$x \cdot y = 3 \times (-2) + (-1) \times 3 + 2 \times 1$$
$$= -7$$

L_2 norm of $x = \sqrt{3^2 + (-1)^2 + 2^2} = \sqrt{14}$

L_2 norm of $y = \sqrt{(-2)^2 + (3)^2 + 1^2} = \sqrt{14}$

Cosine of angle betn x & y $= \frac{-7}{\sqrt{14}\sqrt{14}} = \frac{-7}{14} = \frac{-1}{2}$

The angle whose cosine is $\frac{-1}{2}$ is 120^0

b) $x=[1,2,3]$ $y=[2,4,6]$

$x.y = 2+8+18 = 28$

L_2 norm of $x = \sqrt{1^2+2^2+3^2} = \sqrt{14}$

L_2 norm of $y = \sqrt{2^2+4^2+6^2} = \sqrt{56} = 2\sqrt{14}$

cosine of angle betn x & y $= \frac{28}{2\times14} = \frac{28}{28} = 1$

The angle whose cosine 1 is $0°$

c) $x=[5,0,-4]$ $y=[-1,-6,2]$

$x.y = -5+0-8 = -13$

L_2 norm of $x = \sqrt{25+16} = \sqrt{41}$

L_2 norm of $y = \sqrt{1+36+4} = \sqrt{41}$

cosine of angle betn x & y $= \frac{-13}{41}$

The angle whose cosine is $-\frac{13}{41}$ is 108.48

d) $x=[0,1,1,0,1,1]$ $y=[0,0,1,0,0,0]$

$x.y = 0+0+1+0+0+0 = 1$

L_2 norm of $x = \sqrt{4} = 2$

L_2 norm of $y = \sqrt{1} = 1$

cosine of angle betn x & y $= 1/2$

angle whose cosine is $1/2$ is $60°$

1.2 Leskovec, ch. 9.
Exercise 9.2.1

Feature	A	B	C
Processor Speed	3.06	2.68	2.92
Disk size	500	320	640
main memory size	6	4	6

$A = [3.06, 500, 6]$
$B = [2.68, 320, 4]$
$C = [2.92, 640, 6]$

$$\cos(A, B) = \frac{A \cdot B}{|A|\,|B|}$$

$$= \frac{(3.06 \times 2.68) + (500 \times 320)(\alpha^2) + (6 \times 4)(\beta)^2}{(\sqrt{3.06^2 + (500\alpha)^2 + (6\beta)^2}\,(\sqrt{2.68^2 + (320\alpha)^2 + (4\beta)^2}}$$

$$= \frac{8.2008 + 160000\alpha^2 + 24\beta^2}{\sqrt{9.3636 + 250000\alpha^2 + 36\beta^2}\,\sqrt{7.1824 + 102400\alpha^2 + 16\beta^2}}$$

$$\cos(B, C) = \frac{B \cdot C}{|B|\,|C|}$$

$$= \frac{(2.68 \times 2.92) + (320 \times 640)(\alpha^2) + (4 \times 6)(\beta^2)}{\sqrt{(2.68)^2 + (320\alpha)^2 + (4\beta)^2}\,\sqrt{292^2 + (640\alpha)^2 + (6\beta)^2}}$$

$$= \frac{7.8256 + 204800\alpha^2 + 24\beta^2}{\sqrt{7.1824 + 102400\alpha^2 + 16\beta^2}\,\sqrt{8.5264 + 409600\alpha^2 + 36\beta^2}}$$

$$\cos(A,C) = \frac{A \cdot C}{|A|\,|C|}$$

$$= \frac{(3.06 \times 2.92) + (500 \times 640)(\alpha^2) + (6 \times 6)(\beta)^2}{\sqrt{(3.06)^2 + (500\alpha)^2 + (6\beta)^2}\ \sqrt{2.92^2 + (640\alpha)^2 + (6\beta)^2}}$$

$$= \frac{8.9352 + 320000\alpha^2 + 36\beta^2}{\sqrt{9.3636 + 250000\alpha^2 + 36\beta^2}\ \sqrt{8.5264 + 409600\alpha^2 + 36\beta^2}}$$

b) $\alpha = \beta = 1$

$\cos(A,B) = 0.999973$

Angle betn vector = 0.13

$\cos(B,C) = 0.999987$

Angle betn vectors = 0.09

$\cos(A,C) = 0.999953$

Angle betn vectors = 0.17

c) $\alpha = 0.01$ & $\beta = 0.5$

$\cos(A,B) = 0.99088$

Angle betn vectors = 7.74

$\cos(B,C) = 0.9691779$

Angle betn vector = 14.26

$\cos(A,C) = 0.9915547$

Angle betn vector = 7.45

d) Average value of disk size $= \frac{500+320+640}{3}$

$= 486.66$

$\alpha = \frac{1}{486.66} = 0.00205$

Average value of main memory $= \frac{6+4+6}{3}$

$= 5.33$

$\beta = \frac{1}{5.33} = 0.1875$

$\cos(A,B) = 0.9944$

Angle betn vectors $= 6.06$

$\cos(B,C) = 0.9822$

Angle betn vectors $= 10.82$

$\cos(A,C) = 0.9956$

Angle betn vector $= 5.37$

Exercise 9.2.3

a) Mean $= \frac{4+2+5}{3} = \frac{11}{3}$

$A = 4 - \frac{11}{3} = \frac{1}{3}$ $\quad B = 2 - \frac{11}{3} = \frac{-5}{3}$

$C = 5 - \frac{11}{3} = \frac{4}{3}$

b) User Profile

Processor speed $= 3.06 \times (\frac{1}{3}) + 2.68 \times (-\frac{5}{3}) + 2.92 (\frac{4}{3})$

$= 0.4467$

Disk size $= 500 \times \frac{1}{3} + 320 \times (-\frac{5}{3}) + 640 \times (\frac{4}{3})$

$= 486.66$

main memory size $= 6 \times \frac{1}{3} + 4 \times (-\frac{5}{3}) + 6 (\frac{4}{3})$

$= 3.33$

Exercise 9.3.1

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

a) Jacard distance $[d(A,B)] = 1 - SIM(A,B)$

$d(A,B) = 1 - 4/8 = 1/2$

$d(B,C) = 1 - 4/8 = 1/2$

$d(A,C) = 1 - 4/8 = 1/2$

b) $\cos(A,B) = \frac{15+15+1+3}{\sqrt{80}\sqrt{40}} = 0.6010$

$\cos(B,C) = \frac{4+9+8+5}{\sqrt{40}\sqrt{64}} = 0.513$

$\cos(A,C) = \frac{8+15+15+6}{\sqrt{80}\sqrt{64}} = 0.615$

c)

	a	b	c	d	e	f	g	h
A	1	1	0	1	0	0	1	0
B	0	1	1	1	0	0	0	0
C	0	0	0	1	0	1	1	1

$d(A,B) = 1 - 2/5 = 3/5$

$d(B,C) = 1 - 1/6 = 5/6$

$d(A,C) = 1 - 2/6 = 2/3$

d) $\cos(A,B) = \dfrac{2}{\sqrt{4}\sqrt{3}} = 0.577$

$\cos(B,C) = \dfrac{1}{\sqrt{3}\sqrt{4}} = 0.288$

$\cos(A,C) = \dfrac{2}{\sqrt{4}\sqrt{4}} = \dfrac{2}{4} = 0.5$

e) $Avg(A) = 20/6 = 10/3$

$Avg(B) = 14/6 = 7/3$

$Avg(C) = 18/6 = 3$

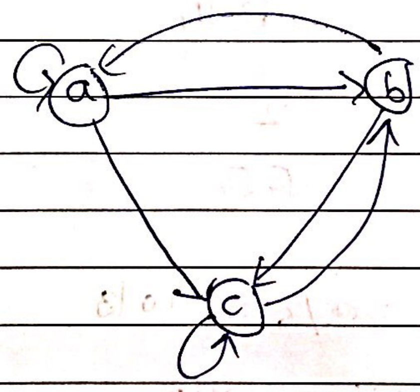
	a	b	c	d	e	f	g	h
A	2/3	5/3		5/3	-7/3		-1/3	-4/3
B		2/3	5/3	2/3	-4/3	-1/3	-4/3	
C	-1		-2	0		1	2	0

f) $\cos(A,B) = \dfrac{10/9 + 10/9 + 28/9 + 4/9}{\sqrt{\frac{120}{9}} \times \sqrt{\frac{65}{9}}} = 0.5743$

$$\cos(B,C) = \frac{-\frac{10}{3} - \frac{1}{3} - \frac{8}{3}}{\frac{1}{3}\sqrt{66} \times \sqrt{10}} = -0.7395$$

$$\cos(A,C) = \frac{\left(-\frac{2}{3}\right) + \left(-\frac{2}{3}\right)}{\frac{\sqrt{120}}{3} \times \sqrt{10}} = -0.1154$$

1.3 Leskovec ch. 5
Exercise 5.1.1



Flow eqn. –

$$r_a = \frac{r_a}{3} + \frac{r_b}{2}$$

$$r_b = \frac{r_a}{3} + \frac{r_c}{2}$$

$$r_c = \frac{r_a}{3} + \frac{r_b}{2} + \frac{r_c}{2}$$

$$M = \begin{array}{c|ccc} & a & b & c \\ \hline a & 1/3 & 1/2 & 0 \\ b & 1/3 & 0 & 1/2 \\ c & 1/3 & 1/2 & 1/2 \end{array}$$

$$r = M \cdot r$$

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1/3 & 1/2 & 0 \\ 1/3 & 0 & 1/2 \\ 1/3 & 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

$$A = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\Rightarrow r = \begin{bmatrix} 1/3 & 1/2 & 0 \\ 1/3 & 0 & 1/2 \\ 1/3 & 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$= \begin{bmatrix} 1/3 & 1/2 & 0 \\ 1/3 & 0 & 1/2 \\ 1/3 & 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 0.2778 \\ 0.2778 \\ 0.444 \end{bmatrix}$$

$r.\text{Norm} = 0.0158$

After 9-10 iterations,

$$r = \begin{bmatrix} 0.2307 \\ 0.307 \\ 0.4615 \end{bmatrix} \approx \begin{bmatrix} 3/13 \\ 4/13 \\ 6/13 \end{bmatrix}$$

$$r_{norm} = 0.000000002001$$

Exercise 5.1.2

$$\beta = 0.8$$

$$r = \beta M + (1-\beta) \left[\frac{1}{N} \right]$$

$$\gamma = \begin{bmatrix} 4/15 & 2/3 & 0 \\ 4/15 & 0 & 2/5 \\ 4/15 & 2/5 & 2/5 \end{bmatrix} + \begin{bmatrix} 1/15 & 1/15 & 1/15 \\ 1/15 & 1/15 & 1/15 \\ 1/15 & 1/15 & 1/15 \end{bmatrix}$$

$$A = \begin{bmatrix} 1/3 & 7/15 & 1/15 \\ 1/3 & 7/15 & 7/15 \\ 1/3 & 7/15 & 7/15 \end{bmatrix}$$

$$\gamma = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

By power method.

$$\gamma = A \cdot \gamma = \begin{bmatrix} 1/3 & 7/15 & 1/15 \\ 1/3 & 1/15 & 7/15 \\ 1/3 & 7/15 & 7/15 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$= \begin{bmatrix} 0.2879 \\ 0.2889 \\ 0.4222 \end{bmatrix}$$

$$= \begin{bmatrix} 1/3 & 7/15 & 1/15 \\ 1/3 & 1/15 & 7/15 \\ 1/3 & 7/15 & 7/15 \end{bmatrix} \begin{bmatrix} 0.2879 \\ 0.2889 \\ 0.4222 \end{bmatrix} = \begin{bmatrix} 0.2583 \\ 0.3126 \\ 0.4282 \end{bmatrix}$$

After 7-8 iterations,

$$\gamma_{norm} = 0.0000000000002977$$

$$\gamma = \begin{bmatrix} 0.2595 \\ 0.3086 \\ 0.4321 \end{bmatrix}$$

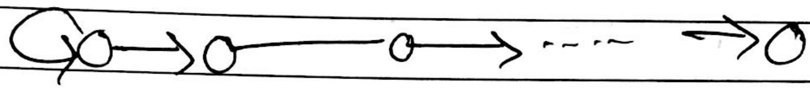
Page rank for,

A → 7/27

B → 25/81

C → 35/81

Exercise 5.1.6

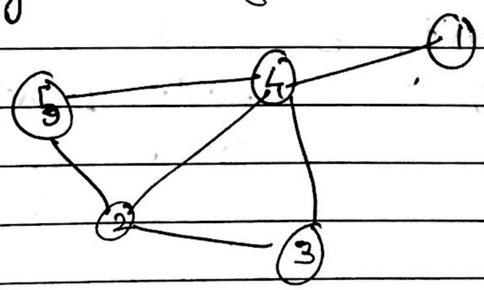


If we remove dead end one by one, only first node with self loop will remain.

Page rank of node would be 1. Page rank of all remaining nodes will be 1/2.

1.4 Centrality Measures:-

a)



(i) Normalized degree centrality for each node:

degree centrality $C_D(v)$ = No. of links a node has (deg(v))

Normalized degree centrality $C_D^*(v)$ $= \frac{1}{n-1} C_D(v)$

$n = 5$

Node 1 :-

$$C_D^*(v) = \frac{1}{4}\times 1 = 1/4$$

Node 2 :-

$$C_D^*(v) = \frac{1}{4}\times 3 = 3/4$$

Node 3 :-

$$C_D^*(v) = \frac{1}{4}\times 2 = 1/2$$

Node 4 :-

$$C_D^*(v) = \frac{1}{4}\times 4 = 1$$

Node 5 :-

$$C_D^*(v) = \frac{1}{4}\times 2 = 1/2$$

b) Normalized closeness centrality of each node :-

Closeness Centrality $C_c(v)$ = sum of length of shortest path betn a node & all other nodes in the graph $\frac{1}{\sum_j d(v,j)}$

Normalized closeness Centrality $C_c^*(v) = (n-1)\ C_c(v)$

Node 1 → $C_c^*(v) = 4\times\frac{1}{1+2+2+2} = \frac{4}{7}$

node 2 → $C_c^*(v) = 4\times\frac{1}{1+1+1+2} = \frac{4}{5}$

Node 3 → $C_c^*(v) = 4\times\frac{1}{1+1+2+2} = 2/3$

Node 4 :-

$$C_c^*(v) = 4 \times \frac{1}{1+1+1+1} = 1$$

Node 5 :

$$C_c^*(v) = 4 \times \frac{1}{1+1+2+2} = 2/3$$

c) Normalized betweenness centralized of each node

Normalized betweeness Centrality $C_B^*(v) = \dfrac{C_B(v)}{2\binom{n-1}{2}}$

Node 1 :-

$$C_B^* = 0$$

Node 2 :-

$3 \to 4 = 0/1$
$3 \to 5 = 1/1$
$3 \to 1 = 0/1$

(shortest path through 2 / Total shortest path)

$$C_B^*(v) = \frac{1 \times 2}{2 \times (4c_2)} = \frac{2}{12} = \frac{1}{6}$$

Node 3 :-

$$C_B^*(v) = 0$$

Node 4:-

$1 \to 2 \quad 1/1$

$1 \to 3 \quad 1/1$

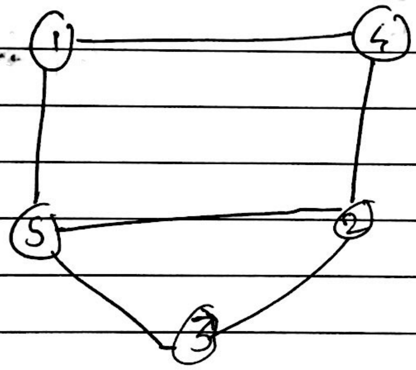
$1 \to 5 \quad 1/1$

$$C_B^*(v) = \frac{3 \times 2}{2 \times ({}^4C_2)} = \frac{1}{2}$$

Node 5:-

$$C_B^*(v) = 0$$

(b)



a) Normalized degree centrality of each node:

$n = 5$

Node 1:-

$$C_D^*(v) = \frac{1}{4} \times 2 = 1/2$$

Node 2:-

$$C_D^*(v) = \frac{1}{4} \times 3 = 3/4$$

Node 3:- $C_D^*(v) = \frac{1}{4} \times 2 = 1/2$

Node 4:- $C_D^*(v) = \frac{1}{4} \times 2 = 1/2$

Node 5:-

$$C_D^*(v) = \frac{1}{4} \times 3 = 3/4$$

b) Normalized closeness centrality of each node:-

Node 1:- $C_c^*(v) = 4 \times \frac{1}{1+1+2+2} = 2/3$

Node 2:- $C_c^*(v) = 4 \times \frac{1}{1+1+1+2} = 4/5$

Node 3:- $C_c^*(v) = 4 \times \frac{1}{1+1+2+2} = \frac{4}{6} = 2/3$

Node 4:-

$$C_c^*(v) = 4 \times \frac{1}{1+1+2+2} = 2/3$$

Node 5:- $C_c^*(v) = 4 \times \frac{1}{1+1+1+2} = \frac{4}{5}$

(c) Normalized betweeness centrality of each node.

Node 1:-

$$C_B^*(v) = \frac{1/2 \times 2}{2 \times (^4C_2)} = 1/12$$

$5 \to 4 = 1/2$ Total $= 1/2$

Node 2: $3 \to 4 = 1/1$
$5 \to 4 = 1/2$ Total $= 1 + \frac{1}{2} = 3/2$

$$C_B^*(v) = \frac{2 \times \frac{3}{2}}{2 \times ({}^4C_2)} = 1/4$$

Node 3:- $C_B^*(v) = 0$

Node 4:-

$1 \to 2 = 1/2$ Total $= 1/2$

$$C_B^*(v) = \frac{2 \times 1/2}{2 \times ({}^4C_2)} = 1/12$$

Node 5 :-
$1 \to 2 = 1/2$, Total $= \frac{1}{2} + 1 = 3/2$
$1 \to 3 = 1/1$

$$C_B^*(v) = \frac{\frac{3}{2} \times 2}{2 \times ({}^4C_2)} = 1/4$$