

בשיעור שעבר...

FDR הוא אחוז שגיאות מסוג 1 ביחס לכמות הגילויים. מאפשר לבחור את סף הגילוי תוך בקרה על

אחוז הגילויים השגויים.

$$FDR = \frac{False\ Positives}{True\ Positives + False\ Positives}$$

מרחק אוקלידי, קורלציית פירסון, קורלציית ספירמן

מטריצת מרחקים (distance matrix)

0	73.2775	136.2525	262.5041
73.2775	0	156.1022	289.6258
136.2525	156.1022	0	143.6521
262.5041	289.6258	143.6521	0

מרחק המינג (Hamming)

מרחק המינג (Hamming) הוא המרחק בין שני רצפים. חישוב מרחק המינג נעשה ע"י ספירת מספר הפעמים בהם

יש הבדל בין שני רצפים במיקום מסויים

String	Hamming Distance
1 0 0 0 0 1 0 0 1 1 0 0	2
1 0 1 0 0 1 0 0 0 1 0 0	
P oint	1
P aint	

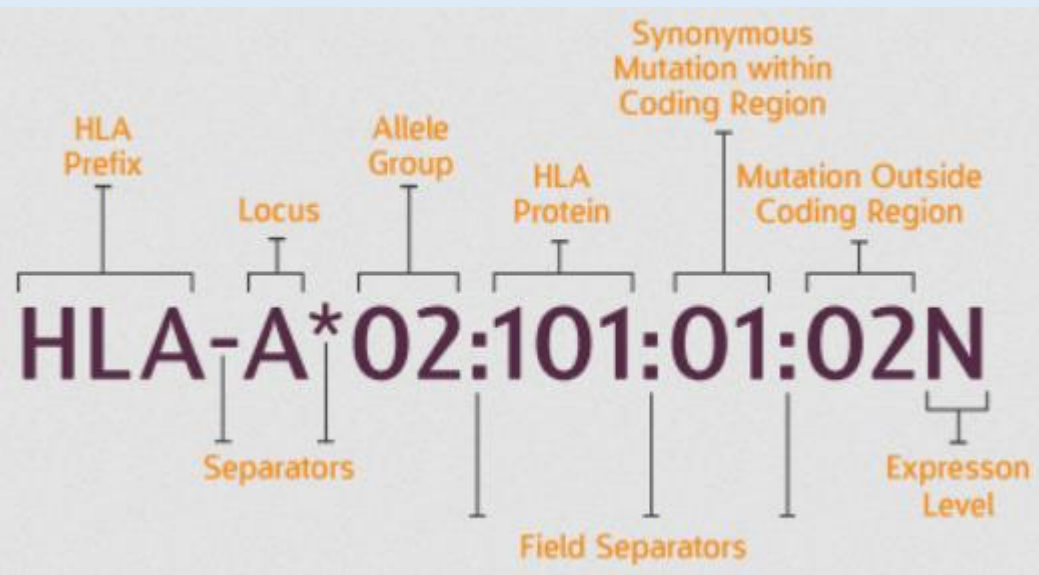
אפשר להשתמש במרחק המינג על מנת לחשב את המרחקים בין רצפי נוקלאוטידים, את המרחק בין אנשים

במולקולות HLA (עבור תרומת מח עצם), רצפי חומצות אמינו וכו'.

מרחק Hamming

על רוב התאים שבגופנו נמצאות מולקולות ממשפחת ה-HLA Human Leukocyte Antigen או HLA. מערכת החיסון משתמשת במולקולות אלו על מנת לזהות אילו תאים שייכים לגוף ואילו הם זרים.

על מנת לתרום מח עצם, דם טבורי או איברים, יש צורך להתאים את המולקולות האלו בין התורם לנתרם במידה רבה ככל האפשר על מנת שלא תהיה דחייה של השתל. לתהליך הזה קוראים HLA Typing או HLA Matching.



הבעיה היא, שבניגוד לתרומת דם למשל, כאן יש צורך להתאים בין מספר גדול של מולקולות HLA. בדרך כלל רצוי שיהיו בין 8-10 מולקולות תואמות על מנת לבצע את ההשתלה.

מרחק Hamming

בקובץ `hladb.RData` ישנם נתונים עבור 19 מולקולות HLA עבור 90 אנשים – כאשר הראשון מביניהם הוא הנתרם ('Recipient').
אנו צריכים למצוא מי מבין 89 האנשים האחרים יכול לתרום לו בעזרת חישוב מרחק Hamming בינו לבין שאר האנשים.



מרחק המינג

```
5 load("hladb.RData")
6
7 samples = hladb$Sample
8 hamming = data.frame()
9 sample1 = hladb[which(hladb$Sample == "Recipient"),-1]
10
11 for(i in 2:length(samples)) {
12   sample2 = hladb[which(hladb$Sample == samples[i]),-1]
13
14   hammingDistance = sum(sample1 != sample2) # This is how we ham
15
16   hamming = rbind(hamming, data.frame(samples[i], hammingDistance))
17 }
18 colnames(hamming) = c("Sample2", "Hamming")
19
20 # We need at 10 matches so we say it's a match
21 # But Hamming measures DISTANCE and not SIMILARITY
22 # So let's calculate the maximum hamming distance for it to be a match:
23 # We have 19 measurements, out of which we need 10 to match - meaning we can have
24 # a maximum distance of 19-10 = 9
25
26 hamming$Sample2[hamming$Hamming <= 9]
```

שורה	מה עושים בשורה?
5	טוענים את מטריצת הנתונים בעזרת פונקציית load לתוך אובייקט שקוראים לו Data
7	וקטור samples יכיל את שמות כל האנשים בניסוי
8	מייצרים data.frame ריק
9	מבודדים את ערכי ה-HLA השונים לנתרם (Recipient), ומורידים את העמודה הראשונה בה נמצא שמו
11	יצירת לולאת for שנותנת למשתנה i ערך עולה בין 2 (השורה הראשונה היא הנתרם) למספר האנשים בכל איטרציה (כלומר עבור כל אדם שהוא לא הנתרם)
12	מבודדים את ערכי ה-HLA השונים לאדם בשורה i, ומורידים את העמודה הראשונה בה נמצא שמו
14	מחשבים מרחק המינג בין הנתרם לאדם בשורה i
16	מכניסים את הערכים המתקבלים ל-data.frame שיצרנו בשורה 8 עם שתי עמודות – אחת שמכילה את שם התורם והשנייה את מרחק המינג
18	נותנים שמות לעמודות של ה-data.frame
26	בודקים אילו אנשים בעלי מרחק המינג קטן מ-9 (הסבר בשורות 20-24)

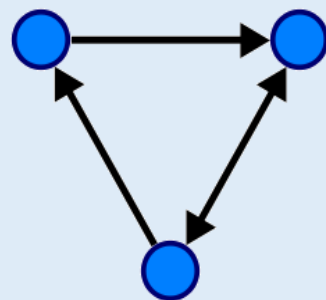


רשתות

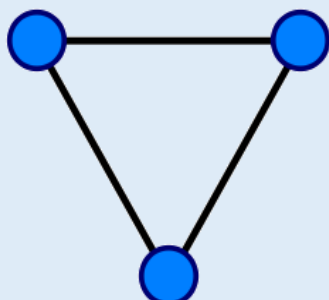
כאשר אנחנו עושים ניתוח נתונים רב-מימדיים למידע ביולוגי, נצפה שיהיו קשרים רבים בין שניים או יותר מהמשתנים.

כדי להבין את הקשרים האלו נרצה לבנות רשתות (networks) של קשרים, כשהמרחק בין שתי דוגמאות/מימדים הוא הבסיס

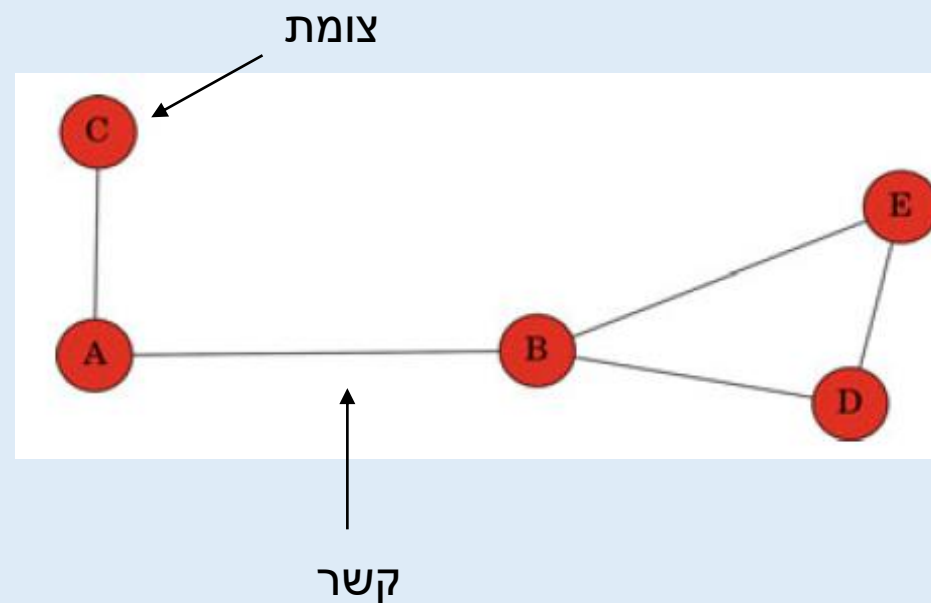
לקביעה יש קשר בין שתי הדוגמאות/מימדים הללו. את הרשתות שנמצא נוכל לייצג באמצעות גרפים.



directed



undirected



רשתות

כדי להבין איך כל צומת קשורה לצומת אחרת, נשתמש במטריצת המרחקים עליה למדנו בשיעור הקודם.

המרחק במטריצה מצביע לנו על הקשר בין כל שני צמתים.

0	73.2775	136.2525	262.5041
73.2775	0	156.1022	289.6258
136.2525	156.1022	0	143.6521
262.5041	289.6258	143.6521	0

מטריצת סמיכות (Adjacency matrix) – מטריצת המרחקים לאחר שקבענו סף הנחשב בעינינו ל"קשר" – ערכים גדולים או

שווים (בערך מוחלט) לסף, יקבלו 1, וכל השאר – 0.

איך בונים רשת?

שאלה 1:

נתונה מטריצת דמיון המחושבת לפי קורלציית פירסון בין גנים.

שרטט את רשת הגנים המתקבלת מסף 0.9 (בערך מוחלט). מה המשמעות של הרשת?

$$\begin{pmatrix} 1 & 0.98 & -0.9 & 0.2 \\ 0.98 & 1 & -0.5 & 0.95 \\ -0.9 & -0.5 & 1 & 0.8 \\ 0.2 & 0.95 & 0.8 & 1 \end{pmatrix}$$



תשובה:

שימו לב שמעניין אותנו **אם** קיימת קורלציה. ולא מעניין אותנו כמה חזקה הקורלציה.

$$\left(\right)$$



איך בונים רשת?

שאלה 1:

נתונה מטריצת דמיון המחושבת לפי קורלציית פירסון בין גנים.

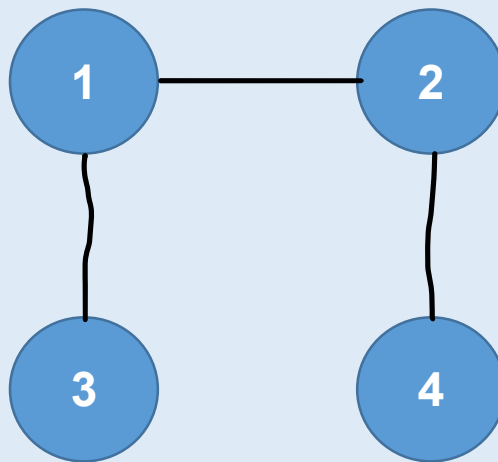
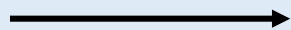
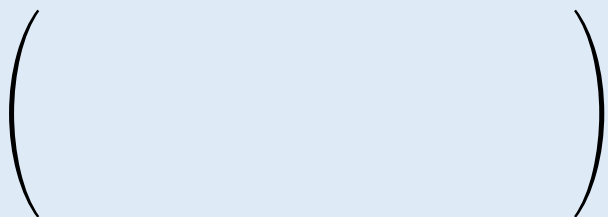
שרטט את רשת הגנים המתקבלת מסף 0.9 (בערך מוחלט). מה המשמעות של הרשת?

$$\begin{pmatrix} 1 & 0.98 & -0.9 & 0.2 \\ 0.98 & 1 & -0.5 & 0.95 \\ -0.9 & -0.5 & 1 & 0.8 \\ 0.2 & 0.95 & 0.8 & 1 \end{pmatrix}$$



תשובה:

שימו לב שמעניין אותנו **אם** קיימת קורלציה. ולא מעניין אותנו כמה חזקה הקורלציה.



הגדרות נוספות

1. משקל (weight)

מרחק/ערך כלשהו שמצביע על עוצמת הקשר בין שני צמתים

2. מסלולים (paths)

דרך המחברת בין שתי נקודות

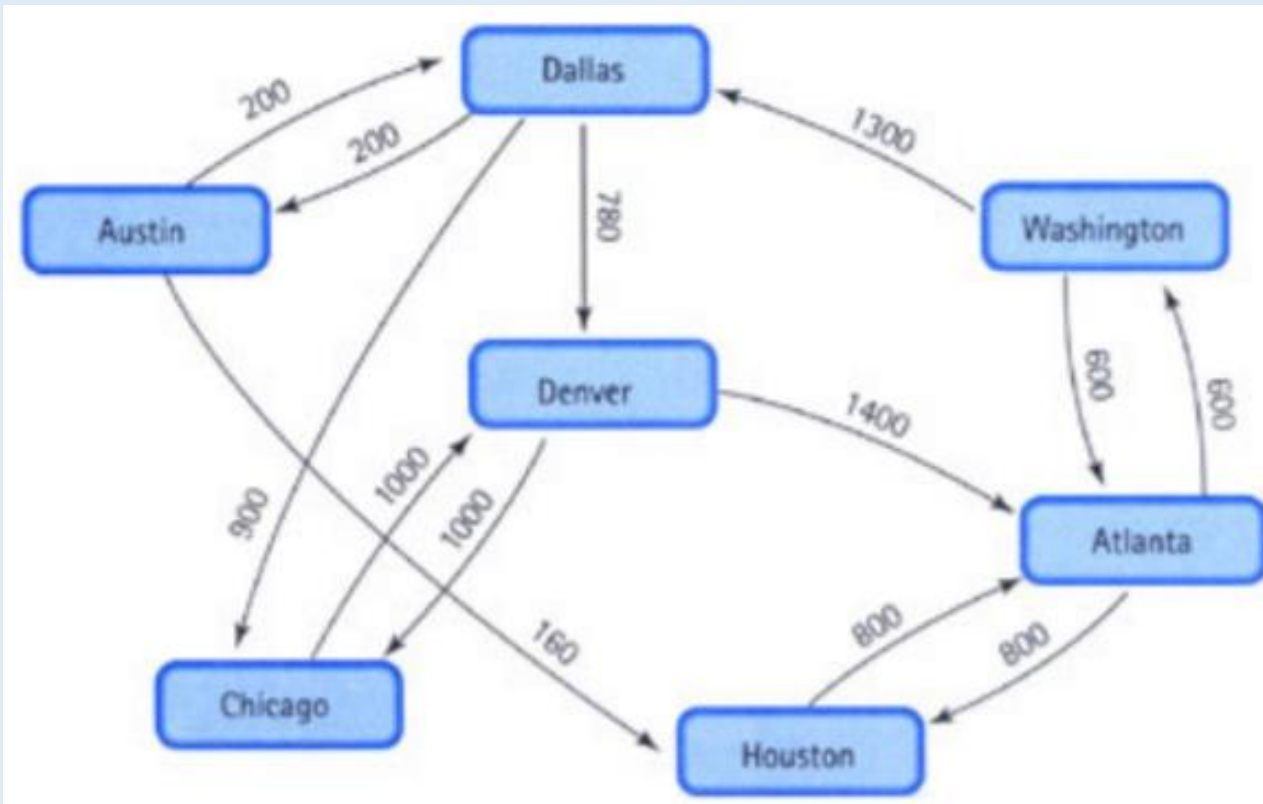
3. מרחק/מסלול גיאודזי

המסלול הקצר ביותר בין שני צמתים

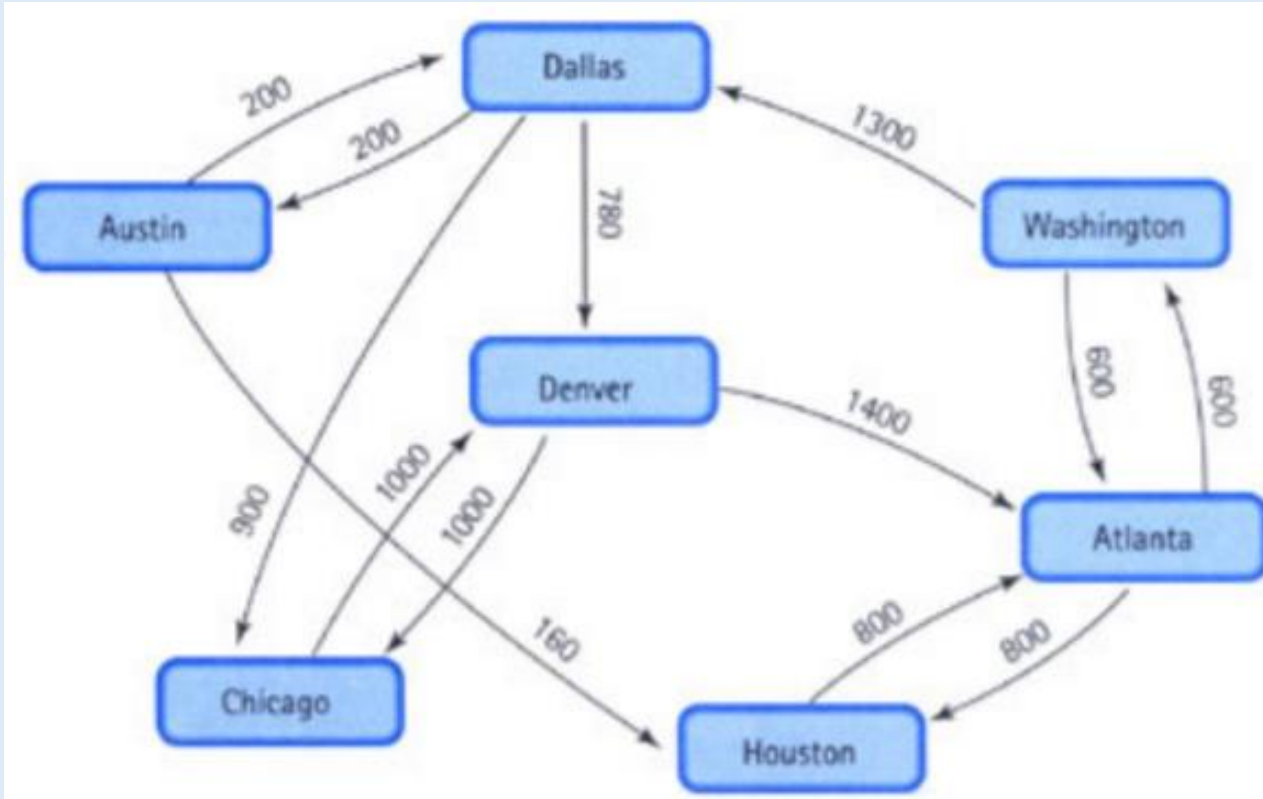
4. מסלול ממוצע

ממוצע המרחק הגיאודזי בין כל הצמתים ברשת

למעט (singletons)



הגדרות נוספות



5. דרגה (degree)

מספר הקשרים שיש לצומת

1. Indegree – מספר הקשתות הנכנסות לצומת

2. Outdegree – מספר הקשתות היוצאות מהצומת

6. Hub

צומת בעל קישוריות גבוהה

7. קליקה (Clique)

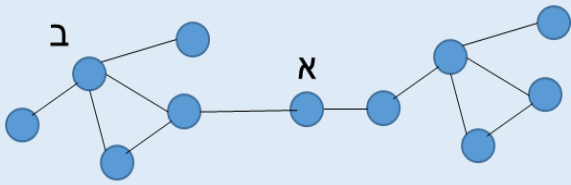
אזור בו כל הצמתים מקושרים

8. Singleton

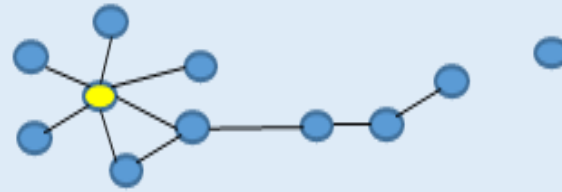
צומת ללא קישוריות



הגדרות נוספות



Betweenness Centrality – מספר הפעמים שצומת מופיע במסלול הגיאודזי ביותר בין שני צמתים.



Centrality .9

מדד לחשיבות הצומת (node).

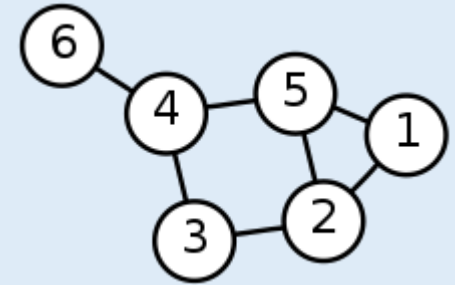
10. קוטר הגרף (diameter)

המסלול הכי קצר בין שתי הנקודות הכי רחוקות.

הגדרה נוספת: המסלול הכי ארוך מכל המסלולים הגיאודזים

דוגמא

From	To	Shortest Distance
1	2	1
1	3	2
1	4	3
1	5	1
1	6	3
2	3	1
2	4	2
2	5	1
2	6	3
3	4	1
3	5	2
3	6	2
4	5	1
4	6	1
5	6	2



- מה המרחקים האפשריים בין כל שתי נקודות
- מהו המרחק הקצר ביותר בין כל שתי נקודות?

שאלה 2

נתונות שתי מטריצות סמיכות:

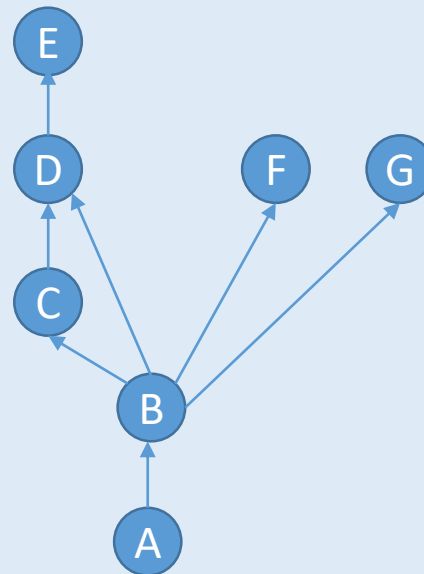
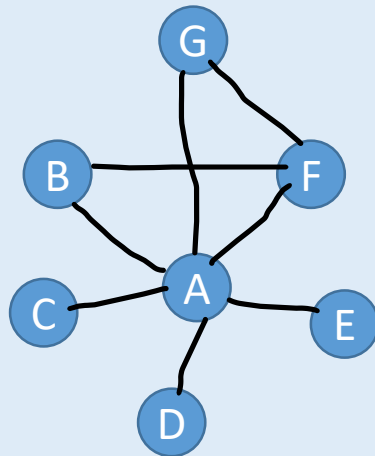
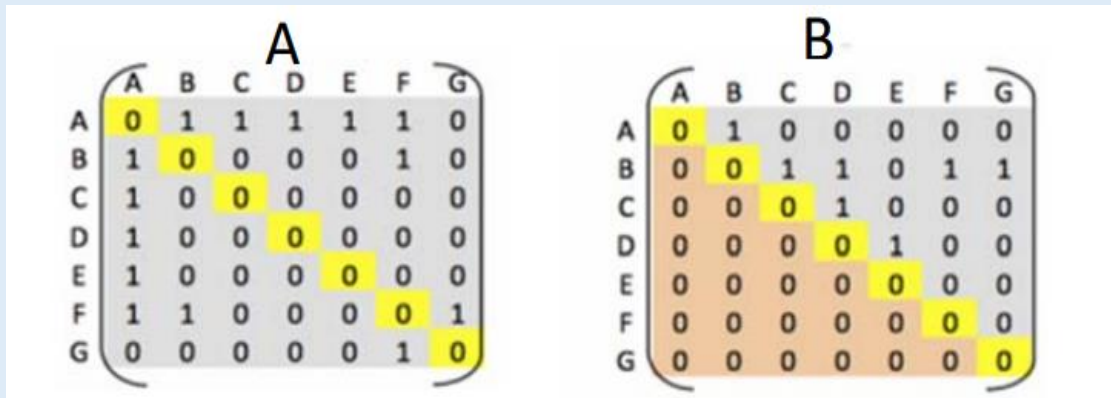
	A	B	C	D	E	F	G
A	0	1	1	1	1	1	0
B	1	0	0	0	0	1	0
C	1	0	0	0	0	0	0
D	1	0	0	0	0	0	0
E	1	0	0	0	0	0	0
F	1	1	0	0	0	0	1
G	0	0	0	0	0	1	0

	A	B	C	D	E	F	G
A	0	1	0	0	0	0	0
B	0	0	1	1	0	1	1
C	0	0	0	1	0	0	0
D	0	0	0	0	1	0	0
E	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0

1. שרטט את הרשתות המוגדרות מכל מטריצה.
2. זהה בכל רשת את הצומת בעל ה- קישוריות (degree) הגדולה ביותר.
3. מה הקוטר של כל רשת?
4. מה ההשלכות של להוציא את צומת B מכל אחד משתי הרשתות?
5. מה המשמעות של להוציא את צומת A מכל אחד משתי הרשתות?

שאלה 2

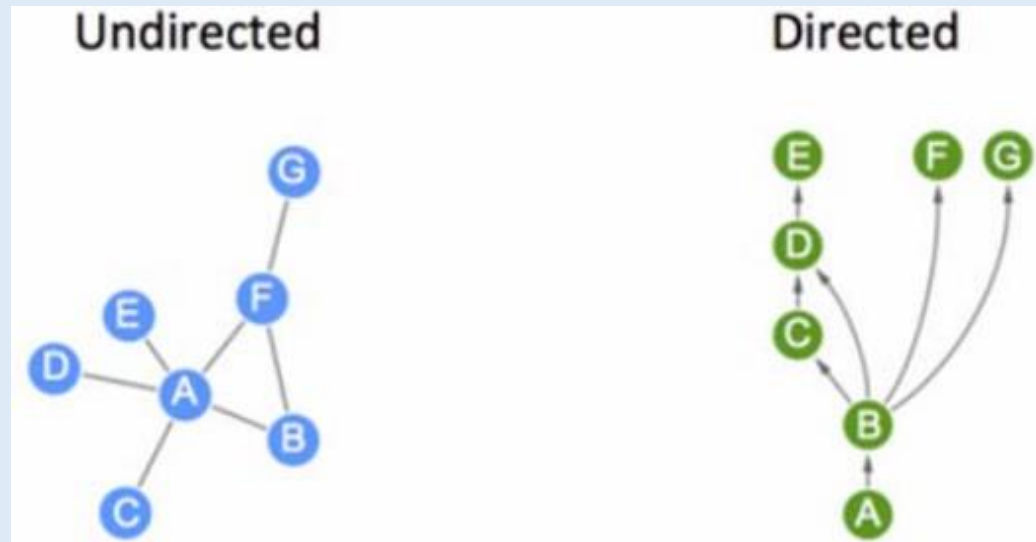
שרטט את הרשתות המוגדרות מכל מטריצה.



שאלה 2

זוהי בכל רשת את הצומת בעל ה- קישוריות (degree) הגדולה ביותר.

Degree:

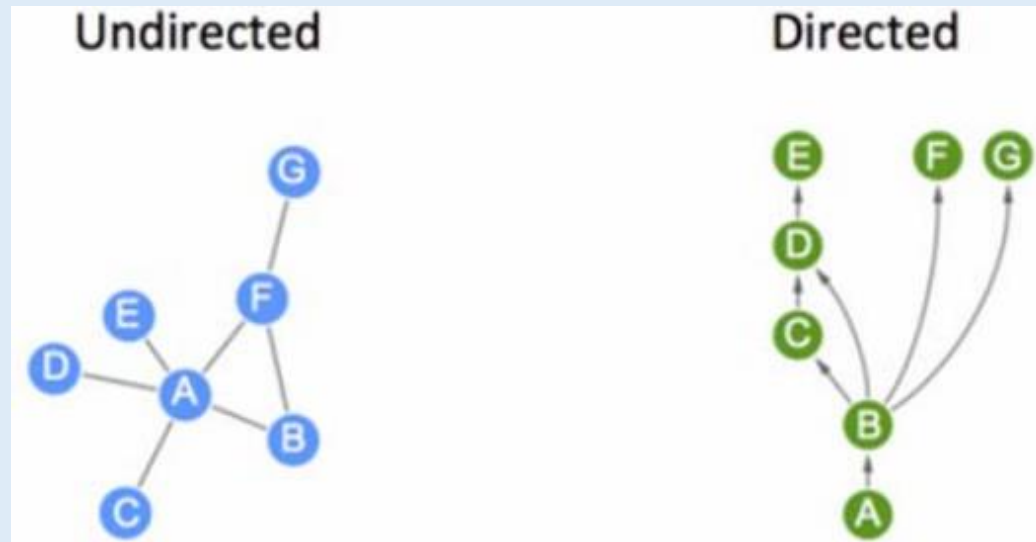


Indegree:

Outdegree:

שאלה 2

זוהי בכל רשת את הצומת בעל ה- קישוריות (degree) הגדולה ביותר.



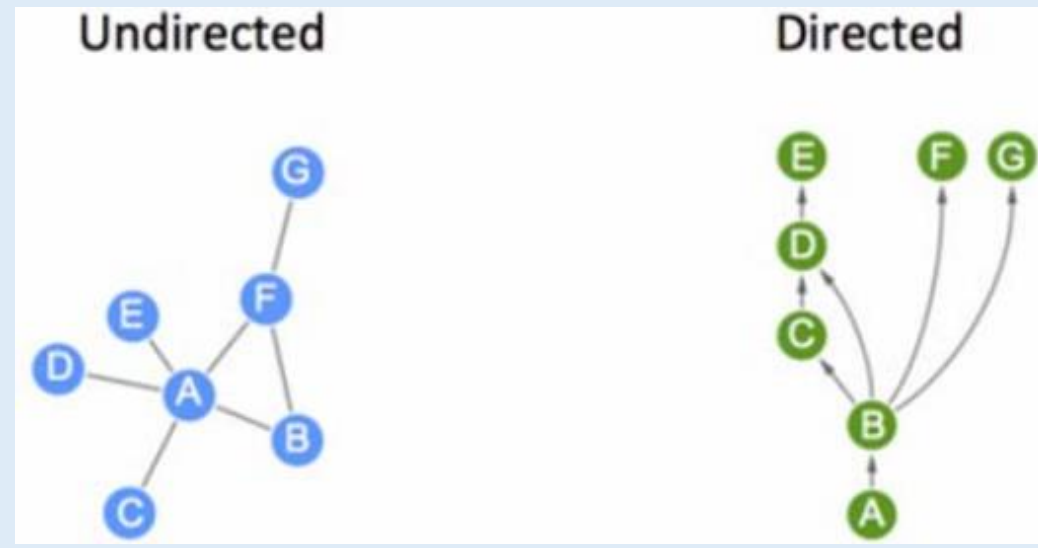
Degree: A

Indegree: D

Outdegree: B

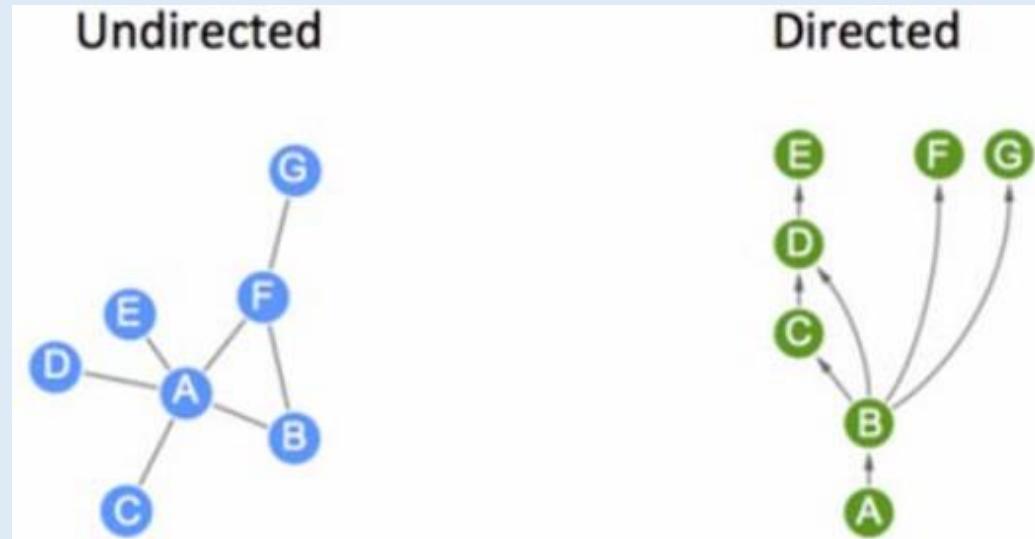
שאלה 2

מה הקוטר של כל רשת?



שאלה 2

מה הקוטר של כל רשת?



3

3

שאלה 2

מה ההשלכות של להוציא את צומת B מכל אחד משתי הרשתות?



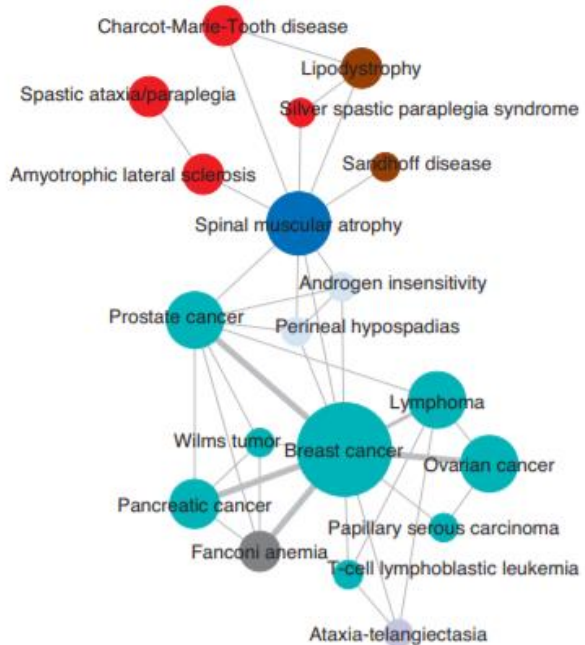
שאלה 2

מה ההשלכות של להוציא את צומת A מכל אחד משתי הרשתות?

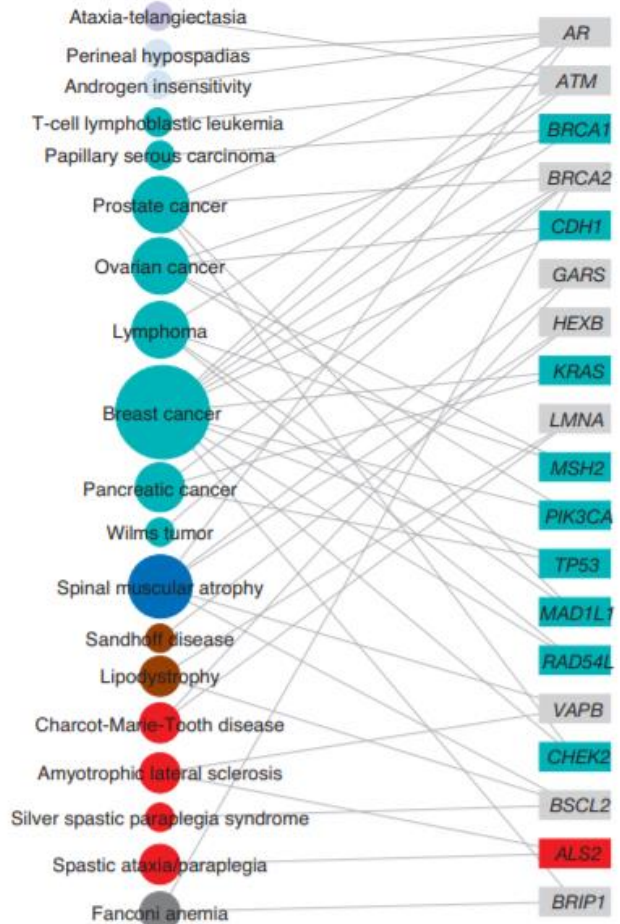


שאלה 3

Human Disease Network (HDN)



disease phenotype disease genome

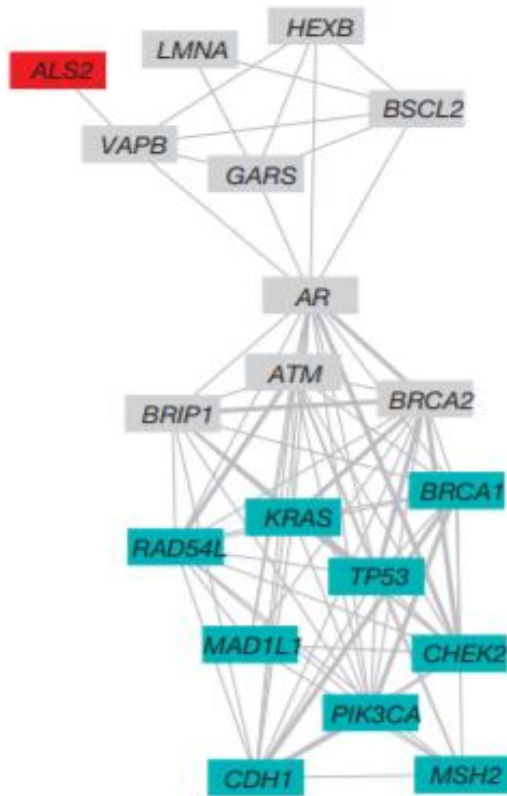


The human disease network (Goh et al. PNAS 2007) ומתארות את הקשר בין הרשתות הבאות נלקחו מהמאמר The human disease network (Goh et al. PNAS 2007) ומתארות את הקשר בין מחלות לבין גנים שמוטציה בהם נמצאה כקשורה למחלה. גודל כל צומת מתארת את מספר הגנים שנמצאו קשורים למחלה והצבע מצביע על סוג המחלה.

האם ניתן להציג את המידע הזה בצורת רשת אחרת המבוססת גנים? אם כן תאר מה תהיה משמעות הצמתים ומה תהיה משמעות הקשתות?

שאלה 3

*Disease Gene Network
(DGN)*

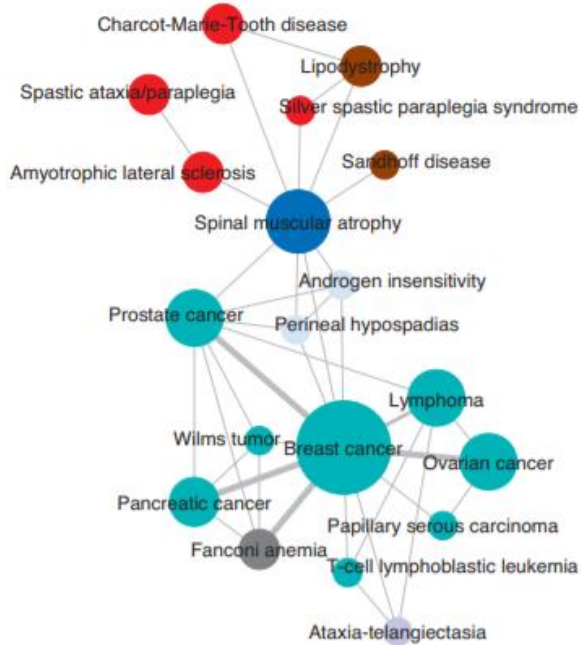


האם ניתן להציג את המידע הזה בצורת רשת אחרת המבוססת גנים? אם כן תאר מה תהיה משמעות הצמתים ומה תהיה משמעות הקשתות?

כאן הצמתים הם הגנים והקשתות מחברות את הגנים אם הם מתבטאים באותה המחלה

שאלה 3

Human Disease Network (HDN)



disease phenome

Ataxia-telangiectasia
Perineal hypospadias
Androgen insensitivity
T-cell lymphoblastic leukemia
Papillary serous carcinoma
Prostate cancer
Ovarian cancer
Lymphoma
Breast cancer
Pancreatic cancer
Wilms tumor
Spinal muscular atrophy
Sandhoff disease
Lipodystrophy
Charcot-Marie-Tooth disease
Amyotrophic lateral sclerosis
Silver spastic paraplegia syndrome
Spastic ataxia/paraplegia
Fanconi anemia

disease genome

AR
ATM
BRCA1
BRCA2
CDH1
GARS
HEXB
KRAS
LMNA
MSH2
PIK3CA
TP53
MAD1L1
RAD54L
VAPB
CHEK2
BSCL2
ALS2
BRIP1

מה מייחד את מחלת Spinal Muscular Atrophy?

ניתן לראות בגרף השמאלי שהמחלה היא hub בעלת betweenness centrality גבוהה המשתפת גנים הקשורים למחלות סרטן כמו גם למחלות אחרות.

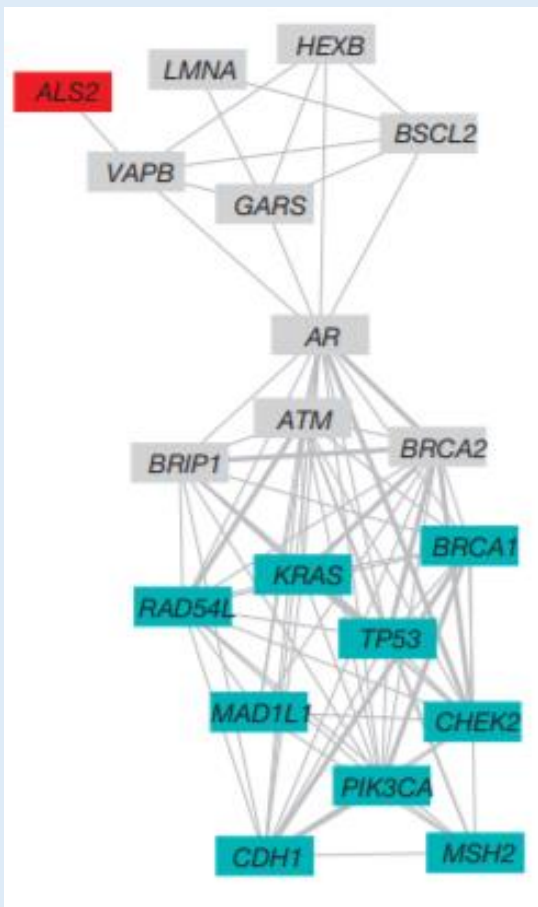
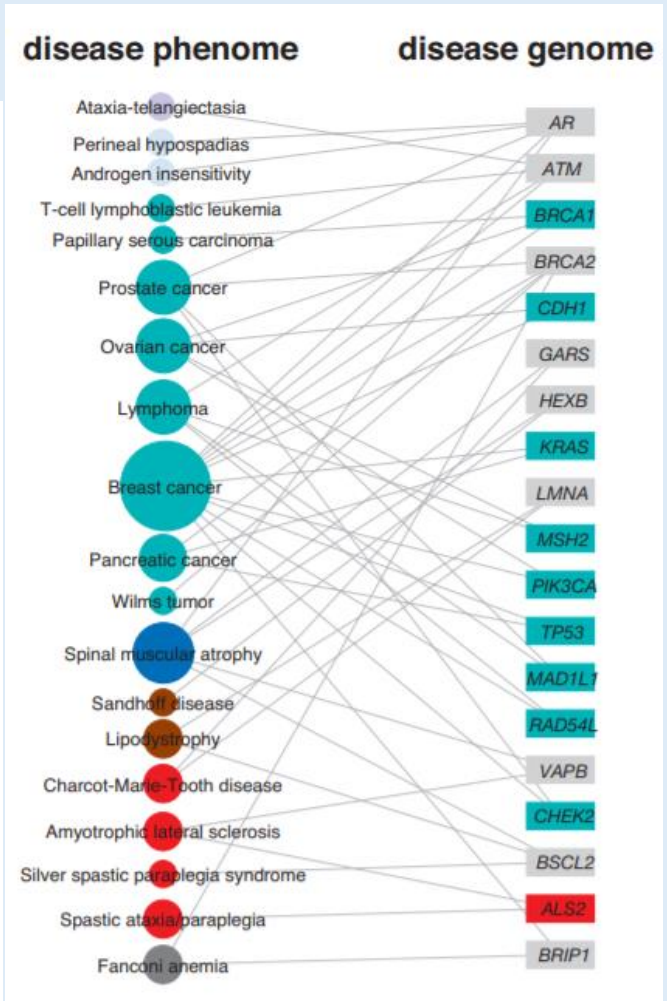
שאלה 3

ביכולתנו לרצף גנים על מנת למצוא מוטציות. האם יש גן אחד אותו נוכל לרצף כדי לזהות סרטן?
אם כן – מיהו הגן?

ניתן לראות כי אין גן אחד האחראי על כל סוגי הסרטן.

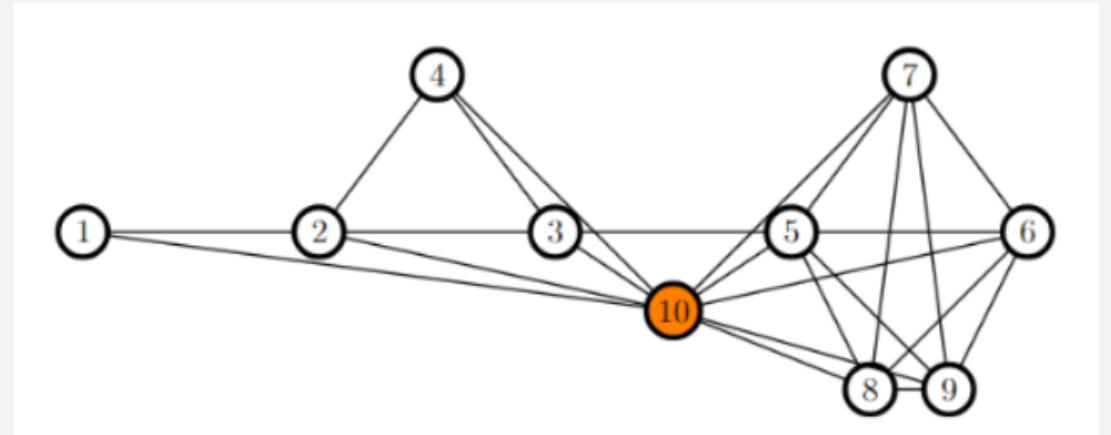
אם נחפש מספר גנים נראה שלמשל הסרטן T-cell Lymphoblastic
Leukemia מקושר רק לגן ATM. אך הגן ATM מקושר למחלה שהיא
לא סרטנית.

מסיבה זו, אין גם סט מינימלי שיכול להבדיל לנו בין סוגי הסרטן.



שאלות מטלת בית

נתונה הרשת הבאה:

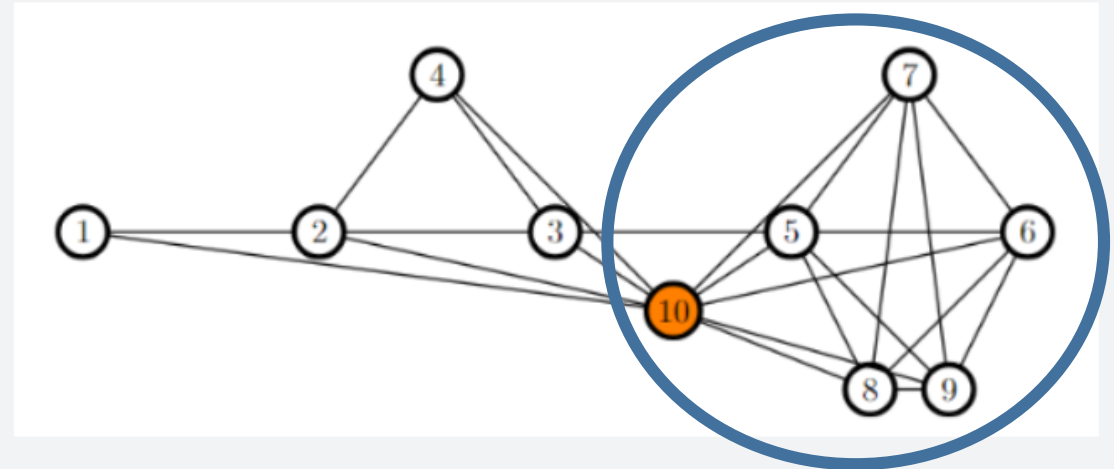


כמה צמתים יש לקליקה (clique) הגדולה ביותר?

תשובה:

שאלות מטלת בית

נתונה הרשת הבאה:



כמה צמתים יש לקליקה (clique) הגדולה ביותר?

תשובה:

שאלות מטלת בית

כאשר אנו בודקים רשתות של נוירונים, מה נכון להגיד על מדד betweenness centrality גבוה?

יש לבחור תשובה אחת:

- מצביע על נוירונים אשר בעיקר שולחים סיגנלים
- מצביע על נוירונים שלא מקבלים סיגנלים כלל
- מצביע על נוירונים שנמצאים במרכז הולכת הסיגנלים
- מצביע על נוירונים אשר בעיקר מקבלים סיגנלים

שאלות מטלת בית

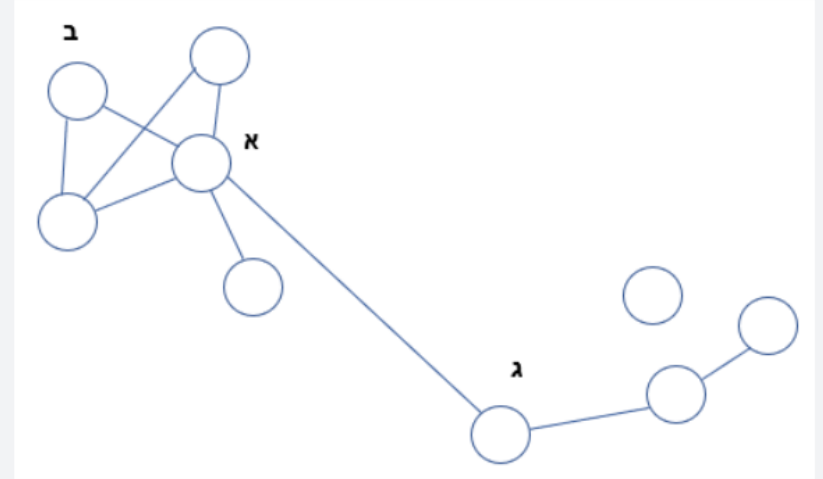
כאשר אנו בודקים רשתות של נויירונים, מה נכון להגיד על מדד betweenness centrality גבוה?

יש לבחור תשובה אחת:

- מצביע על נויירונים אשר בעיקר שולחים סיגנלים
- מצביע על נויירונים שלא מקבלים סיגנלים כלל
- מצביע על נויירונים שנמצאים במרכז הולכת הסיגנלים
- מצביע על נויירונים אשר בעיקר מקבלים סיגנלים

שאלות מטלת בית

נתונה הרשת הבאה:



מה נכון להגיד?

יש לבחור תשובה אחת:

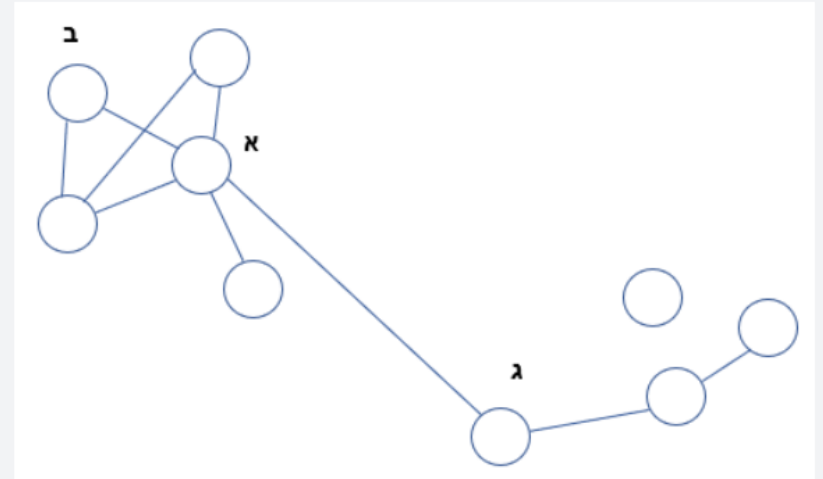
- לצומת ג' יש degree גבוה יותר מלצומת ב'
- לצומת ג' יש centrality גבוה יותר מזה של צומת ב'
- בצומת א' ה-indegree גבוה יותר מה-outdegree
- צומת א' היא לא hub

כמה קליקות (cliques) יש ברשת?

תשובה:

שאלות מטלת בית

נתונה הרשת הבאה:



מה נכון להגיד?

יש לבחור תשובה אחת:

- לצומת ג' יש degree גבוה יותר מלצומת ב'
- לצומת ג' יש centrality גבוה יותר מזה של צומת ב'
- בצומת א' ה-indegree גבוה יותר מה-outdegree
- צומת א' היא לא hub

כמה קליקות (cliques) יש ברשת?

2

תשובה: